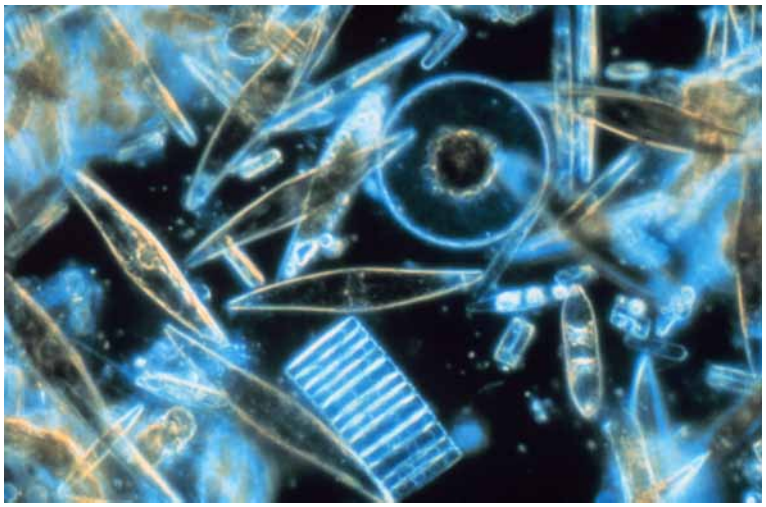
Méthodes statistiques appliquées à la télédétection de la chlorophylle a du lagon de Nouvelle-Calédonie
Rapport de stage M2 ISN

Professeur encadrant : François RÉCHER
Encadrant : Cécile DUPOUY
Stage de : Guillaume WATTELEZ

17 juin 2011

Projets Valhysat et Valhybio



Université Lille1 Sciences et Technologies

IRD Institut de recherche pour le développement

Remerciements

Je tiens à remercier :

– Cécile DUPOUY pour son encadrement, ses conseils et sa proposition de stage que je trouve fort intéressant ;

– Morgan MANGEAS pour ses bonnes idées, ses coups de pouce et sa disponibilité malgré un emploi du temps très chargé, ainsi que Marc DESPINOY pour leur accueil dans l'Unité ESPACE-DEV ;

– Jérôme LEFÈVRE pour son aide et ses explications concernant les données satellites (notamment l'extraction de ces données) ;

– Christophe BIERNACKI et François RÉCHER pour leur disponibilité concernant toutes les questions techniques à propos du stage, et pour le temps passé lors d'essais pour les oraux ;

– les stagiaires et l'ensemble du personnel du centre IRD de Nouméa pour leur accueil et la bonne ambiance ;

– toutes les personnes m'ayant aidé et soutenu durant ce stage.

Table des matières

1 Présentations

1.1 Institut de Recherche pour le Développement (IRD)

1.1.1 Généralités

L'IRD est un établissement public français à caractère scientifique et technologique placé sous la double tutelle du ministère de l'enseignement supérieur et de la recherche, et du ministère des affaires étrangères et européennes. Il intervient depuis plus de soixante ans dans les pays du Sud. Ses activités (recherche, expertise, valorisation et formation) ont pour but de contribuer au développement économique, social et culturel de ces pays. Les chercheurs, ingénieurs, techniciens et personnels locaux interviennent dans une cinquantaine de pays et participent à de nombreux programmes nationaux, européens et internationaux.

Le siège de l'IRD est localisé à Marseille. L'institut dispose par ailleurs de trente implantations dont deux en France métropolitaine (Bondy, Montpellier), cinq dans les régions et collectivités d'outre-mer (la Réunion, Guyane, Martinique, Nouvelle-Calédonie - le lieu du stage -, Polynésie Française) et vingt-trois dans des pays situés dans la zone intertropicale, en Afrique, en Méditerranée, en Asie et en Amérique Latine.

1.1.2 Historique

Durant l'existence de l'Institut, les objectifs ont plus ou moins évolué. Voici quelques dates importantes qui marquent des tournants dans l'Histoire de l'Institut.

- **1937 :** Création du Comité consultatif des recherches scientifiques de la France d'outre-mer et du Conseil supérieur de la recherche scientifique pour la coordination de la recherche nationale. Les institutions sont chargées de l'organisation de la "science coloniale" (c'était encore la période coloniale).
- **1943 :** Création de l'Office de la recherche scientifique coloniale (ORSC). L'ORSC est placé sous l'autorité du Secrétaire d'État à la marine et aux colonies. Le Conseil d'Administration est présidé par le directeur du CNRS. L'ORSC a pour mission de constituer un corps de chercheurs, de créer une formation scientifique de haut niveau spécialisée dans le monde tropical et de mettre en place un réseau de centres de recherche dans l'outre-mer français.
- **1944 - 1953 :** L'Office change deux fois d'appellation, tout d'abord " ORSOM " (Office de Recherche Scientifique d'Outre-Mer) puis " ORSTOM " (Office de la Recherche Scientifique et Technique Outre-Mer).
- **1960 :** L'indépendance des pays africains entraîne de profonds changements pour l'Office.
  - Nouvelle tutelle conjointe : le ministère de l'éducation nationale et secrétariat d'état aux relations avec les états de la communauté.
  - Nouvelles missions : entreprendre des recherches fondamentales en vue du développement des pays tropicaux et prémisse de la politique de coopération scientifique et technique avec les pays du tiers-monde.
- **1960 - 1983 :** Les principales raisons d'être de l'ORSTOM sont : la consolidation de l'organisation scientifique, le renforcement des infrastructures en Afrique dans les DOM et le développement d'une coopération avec des pays d'Amérique du Sud

et d'Asie du Sud-Est ainsi qu'avec des pays arabes.
- **1984 :** Réforme en profondeur de l'Office. L'ORSTOM prend le nom d'Institut français de recherche scientifique et technique pour le développement en coopération, tout en conservant son acronyme. L'Institut se retrouve sous double tutelle du ministère de la Recherche et de celui de la Coopération. Il acquière son statut d'établissement public à caractère scientifique et technologique (EPST). Ses missions ont désormais pour but de promouvoir et de réaliser des recherches scientifiques et techniques susceptibles de contribuer de façon durable aux progrès économique, social et culturel des pays en développement.
- **1998 :** L'ORSTOM devient l'Institut de Recherche pour le Développement (IRD). Il résulte une mise en place d'un fonctionnement par unités thématiques. Cinq départements scientifiques constituent l'Institut :
  1. Milieux et environnement
  2. Ressources vivantes
  3. Société et santé
  4. Expertise et valorisation
  5. Soutien et formation des communautés scientifiques du Sud

1.1.3 Quelques succès auprès des populations du Sud

Voici quelques avancées scientifiques, technologiques et humanitaires réalisées en partie avec le soutien de l'IRD :
- **Le Plumpy'nut :** pâte à base d'arachides prête à l'emploi qui a révolutionné la lutte contre la malnutrition. Conçu par un nutritionniste de l'IRD, cet aliment est aujourd'hui largement utilisé sur le terrain humanitaire.
- **L'onchocercose en voie d'éradication en Afrique de l'Ouest :** l'IRD a participé, au côté de l'OMS, à la lutte contre l'onchocercose. Dans de nombreuses régions d'Afrique de l'Ouest, la "cécité des rivières" n'est plus considérée comme un problème de santé publique.
- **Des moustiquaires imprégnées pour lutter contre le paludisme :** promues par l'OMS, les moustiquaires imprégnées d'insecticide constituent l'un des meilleurs moyens de prévention contre le paludisme.
- **Limiter la transmission du VIH mère-enfant :** grâce aux essais cliniques, réalisés par l'IRD et ses partenaires, on sait que la zidovudine (AZT) permet de diminuer le risque de transmission du SIDA de la mère à l'enfant d'un facteur 3 à 10. Associée à la névirapine, elle constitue un traitement efficace et peu coûteux. Ces résultats ont eu une incidence directe sur les politiques de santé et de prévention en Thaïlande dès 2003.
- **Recherche sur les glaciers andins :** dans les Andes, la fonte des glaciers tropicaux s'accélère. L'IRD et ses partenaires ont déployé en Bolivie, au Pérou et en Équateur, un réseau d'observation de ces glaciers, unique dans l'hémisphère sud. Ces observations et les modèles hydrologiques permettent d'évaluer et de prédire les conséquences du changement climatique sur la ressource en eau, essentielle pour l'agriculture, l'alimentation des villes et la production hydroélectrique.
- **Des herbiers :** depuis près de cinquante ans, l'IRD gère deux herbiers de référence au niveau international, en Guyane et en Nouvelle-Calédonie. Aujourd'hui informa-

tisés et régulièrement enrichis, ils constituent des outils d'identification indispensables aux recherches en botanique tropicale et en biodiversité végétale. Ils servent également de support à l'étude des pharmacopées traditionnelles.
- **Les épisodes *El Niño* :** *El Niño* est la plus importante anomalie climatique dans la zone intertropicale. Menées depuis de nombreuses années, les recherches de l'IRD et de ses partenaires permettent d'améliorer la prévisibilité de ce phénomène dont les conséquences sont particulièrement importantes pour les populations (inondations, sécheresses, pêche...).

1.1.4 Unités d'accueil

Le stage se déroule au sein de deux Unités Mixtes de Recherche (UMR) de l'IRD ayant une implantation secondaire à Nouméa.

1.1.4.1 UMR LOPB (1M213) Le LOPB (Laboratoire d'Océanographie Physique et Biogéochimique de Marseille) est l'une des trois UMR de l'Observatoire des Sciences de l'Univers "Centre d'Océanographie de Marseille", créé en 1989. La problématique générale a été centrée sur l'étude des impacts des changements globaux (climatiques et anthropiques) sur la structuration des écosystèmes marins, leur fonctionnement et les flux biogéochimiques associés par une approche couplée physique-biogéochimie.

Structuré en deux équipes, le LOPB a pour vocation de contribuer au progrès des connaissances sur le couplage des processus physiques, chimiques, et biologiques qui contrôlent la production, le devenir, et le transfert de matière dans les écosystèmes marins, en s'appuyant sur des programmes nationaux et internationaux. L'Unité centre ses activités sur l'étude des processus physiques ou biologiques affectant la production et le transport de matière au sein des écosystèmes marins :
- L'équipe 1 : "Océanographie physique, réponse biologique, environnement côtier" a pour objet d'étude principal la circulation des masses d'eau à différentes échelles et les processus d'échange au niveau de leurs interfaces, ainsi que l'implication directe de ces phénomènes - tant advectifs [1] que diffusifs - sur le compartiment biologique et les flux biogéochimiques. Lors de ces dernières années, l'équipe a renforcé son activité dans le domaine de l'environnement côtier, avec un recentrage de son activité autour de la question du rôle des perturbations climatiques et anthropiques sur le fonctionnement des écosystèmes côtiers.
- L'équipe 2 : "Biogéochimie marine et structure fonctionnelle des communautés pélagiques [2]" s'intéresse au cycle des éléments biogènes dans la couche de surface. Les recherches sont plus spécifiquement orientées sur l'influence de la disponibilité relative des nutriments majeurs, N, P et Si sur le contrôle de la composition de la communauté planctonique, de la structure du réseau trophique pélagique et, *in fine*, du flux de matière biogène exportée hors de la couche de surface. Depuis 2008, les efforts ont porté sur les impacts de certains forçages physiques (température, régimes de vents, hydrodynamique à mésoéchelle) ou chimiques (CO_2, apports en nutriments par les fleuves ou l'atmosphère) résultant du changement global et susceptibles de modifier la structuration et le fonctionnement des écosystèmes planc-

1. L'advection est le déplacement horizontal d'une masse d'air avec transfert de ses propriétés.
2. c'est-à-dire vivant loin des côtes.

toniques marins, leur biodiversité, et en conséquence, les cycles biogéochimiques et les flux d'éléments aux interfaces (échanges avec l'atmosphère, export vers les fonds océaniques, échanges eau/sédiment, échanges côte/large).

Le LOPB est une Unité Mixte CNRS-IRD-Université de la Méditéranée (Aix-Marseille II). C'est une des trois unités qui vont fusionner pour former le MIO (Mediteranean Institute of Oceanography) en janvier 2012 et rejoindre également le Laboratoire d'Astronomie de Marseille pour former l'Institut PYTHEAS.

1.1.4.2 UMR ESPACE-DEV L'UMR ESPACE-DEV développe et met en œuvre des méthodologies de spatialisation des connaissances en environnement par télédétection spatiale pour le développement durable des territoires, de l'acquisition des données au processus décisionnel. L'unité propose des méthodologies de spatialisation des dynamiques de l'environnement pour permettre aux sociétés du Sud une adaptation aux changements globaux. Elle est organisée autour de trois types d'activités :

- les activités de recherche
- les activités de services/observations à destination de la communauté scientifique internationale et des pays du Sud. L'UMR développe et exploite un réseau de stations de réception de satellites d'observation de la Terre qui contribue à mettre l'espace au service du développement durable.
- les activités de formation essentiellement dans les universités de Montpellier, de La Réunion et de la Nouvelle-Calédonie. L'UMR conçoit des modules de formation à la carte et des modules d'enseignement à distance.

Enfin, l'UMR ESPACE-DEV est sous tutelle de l'IRD, l'Université de Montpellier (UMII), l'Université des Antilles et de la Guyane (UAG) et de l'Université de la Réunion.

En Nouvelle-Calédonie, l'unité s'implique essentiellement dans les systèmes d'information en environnement et la gestion intégrée des zones côtières dans le cadre d'un partenariat scientifique avec l'Université de Nouvelle-Calédonie.

1.1.5 Projets partenaires

1.1.5.1 Valhybio Le projet Valhybio - VAlidation HYperspectrale d'un modèle BIOgéochimique - (2007 - 2010) est un Programme National de Télédétection Spatiale (PNTS). Il a pour objectif de déterminer l'impact des forçages sur la distribution de la chlorophylle par comparaison entre les simulations du modèle biogéochimique ECO3M (ECOlogical Mechanistic and Modular Modelling) (Baklouti et al., 2006) couplé au modèle hydrodynamique MARS3D (3D Model for Applications at the Regional Scale) [3] développé par IFREMER, adapté au lagon de Nouvelle-Calédonie (Douillet et al., 2001) et la chlorophylle extraite des données satellites de la couleur de l'eau (MODIS). Ceci permettra un suivi spatio-temporel de la chlorophylle de surface dans le sud-ouest du lagon de Nouvelle-Calédonie. Dans le cadre de ce projet, une campagne en mer a été réalisée en 2008 sur le navire Alis de l'IRD afin de collecter des mesures de terrain ; depuis, un suivi mensuel a également lieu depuis 2008 avec le navire Coris.

3. Voir les sites http ://www.com.univ-mrs.fr/LOPB/spip.php ?rubrique45 et http ://www.previmer.org/comment_ca_marche/modele_mars3d_manche_gascogne.

1.1.5.2 Valhysat Le projet Valhysat - VALidation HYperspectrale d'une chaîne de traitement/fusion de données SATellitales à haute et moyenne résolution - est le complément du projet Valhybio. Il a pour objectif de comparer les données *in situ* issues du projet Valhybio ainsi que les données antérieures de Camélia (CAractérisation et Modélisation des Échanges dans les Lagons sous Influences terrigènes et Anthropiques) avec des données de la couleur de l'eau entre 1997 et 2010. Valhysat est soutenu financièrement par SPIRALES (un appel d'offre interne de l'IRD et de la DSI : Direction Scientifique de l'Informatique) [4].

1.2 Le sujet : la télédétection de la chlorophylle a (chla)

1.2.1 Qu'est-ce que la chlorophylle a ?

La **chlorophylle** est le principal pigment assimilateur [5] des végétaux photosynthétiques. Il existe plusieurs formes de chlorophylle, différenciables selon leur structure chimique :

- la chlorophylle a (chla) est le pigment photosynthétique le plus commun du règne végétal ; il est présent chez tous les végétaux aquatiques et terrestres.
- la chlorophylle b se trouve chez les végétaux supérieurs et les algues vertes à des teneurs moindres.
- les chlorophylles c (C1 et C2) chez les algues brunes.
- la chlorophylle d chez les algues bleues.
- la chlorophylle f.

On s'intéresse particulièrement à la chlorophylle a (chla).

1.2.2 Intérêts pour la concentration de chlorophylle a

La mesure de la concentration de chla dans l'eau est utilisée comme indicateur de la quantité de plancton végétal (phytoplancton).

Remarque : La concentration de chlorophylle dans l'eau est souvent exprimée en μg chla.L^{-1}.

D'où l'intérêt de s'intéresser à la détection de chlorophylle. En effet, le plancton végétal est le point de départ de toute l'activité biologique de la mer, à la base de toutes les chaînes alimentaires aquatiques. L'énergie solaire est utilisée pour fabriquer de la matière organique. Les algues microscopiques qui constituent le phytoplancton produisent de grandes quantités d'oxygène, nécessaire à la vie dans l'eau. Elles participent également à l'oxygénation de la planète par des échanges gazeux.
L'état du plancton végétal est également un indicateur d'un équilibre ou déséquilibre biologique. Dans certaines conditions (apports élevés de nutriments : souvent des matières organiques, nitrates ou phosphates), un excès de plancton conduit à une situation dite d'eutrophisation, voire de dystrophisation [6]. La concentration de chlorophylle dans une

4. Voir le site http ://www.ird.fr/informatique-scientifique/projets/valhysat/.

5. Composé chimique permettant la transformation de l'énergie solaire en énergie chimique chez les organismes effectuant la photosynthèse.

6. Un mileu aquatique eutrophe est un milieu riche en éléments nutritifs. L'eutrophisation est la modification et la dégradation du milieu aquatique liées à un apport excessif de substances nutritives (provenant généralement de l'agriculture, de la pollution automobile, etc). La dystrophisation est l'état extrême de l'eutrophisation, qui se traduit par la mort des organismes végétaux et animaux supérieurs.

zone devient alors un indicateur du niveau de pollution.

1.2.3 La télédétection "couleur de la mer" dans le visible

Par définition, la télédétection est la mesure ou l'acquisition d'informations sur un objet ou un phénomène, par l'intermédiaire d'un instrument de mesure n'ayant pas de contact avec l'objet étudié. L'instrument de mesure dont on se servira est un satellite. Le capteur passif utilisé mesure l'éclairement solaire rétrodiffusé et absorbé par la mer. La couleur représente la variation spectrale des réflectances dans le visible (de 412 à 600 nm). D'autres canaux de longueurs d'onde sont utilisés pour des corrections atmosphériques. Grâce à des algorithmes connus, les données de longueurs d'onde du satellite sont traitées pour déduire, en chaque point disponible, toutes sortes de quantités physiques et biologiques : la température de surface de l'eau, la turbidité, les concentrations de chlorophylle, etc. Pour connaître la concentration de chla, deux algorithmes différents sont à notre disposition : OC3 (algorithme développé par la NASA) et OC5 (adaptation d'OC3 par IFREMER pour les côtes bretonnes peu profondes). On évoquera un peu plus loin leurs caractéristiques et leurs différences ; mais on peut d'ores et déjà signaler que les calculs de ces algorithmes sont réalisés à partir des rapport des réflectances (notées souvent Rrs) perçues par le satellite, essentiellement les canaux bleus (absorption de la chlorophylle)/verts (rétrodiffusion du phytoplanction), pour faire une estimation de la concentration de chlorophylle.

Avec cette méthode de télédétection par satellite, on peut avoir cependant des problèmes notamment à cause des nuages, de l'éclairement su Soleil etc. Il y a donc ce qu'on appelle des valeurs "drapeaux" (ou masques) dans les données fournies par le satellite qui indiquent telle ou telle anomalie dans la mesure. Il faut également se méfier des résultats pour des eaux proches des côtes ou des estuaires à cause de la présence simultanée de matière organique dissoute, de particules et de pigments chlorophylliens. Des comparaisons ont été réalisées entre les données satellites et les données *in situ* (les données terrains supposées exactes) pour vérifier si les résultats donnés par le satellite correspondaient à la réalité. Pour faire ces comparaisons, plusieurs méthodes également : comparer la valeur *in situ* avec la valeur du pixel de l'image satellite non masqué le plus proche, ou avec la valeur d'une moyenne, pondérée par la distance, des pixels non masqués les plus proches. Il se trouve que des différences notoires ont été relevées [7]. À cela, plusieurs explications : les aérosols, des algues (trichodesmium) à la surface de l'eau, des fonds peu profonds, des couleurs de fonds différentes, etc.

1.2.4 Le but du stage

Ayant conscience de l'utilité de connaître la concentration de chla dans l'océan, le but du stage sera de déterminer, à l'aide des données satellites à disposition, comment la concentration de chla varie dans le temps. On pourra déterminer par exemple s'il y a un certain cycle dans cette concentration, si la période de l'année a une importance, si certains évènements ont un impact sur la concentration. On essaiera de quantifier ces variations.

7. On donnera plus de détails sur la quantification des différences et leur correction dans la partie 2.2.

La zone qui nous occupera est principalement la partie du sud du lagon de Nouvelle-Calédonie. De plus, les eaux très profondes (au-delà de 1500m) seront écartées de l'étude. On effectuera le traitement des données à l'aide du logiciel R.

FIGURE 1 – Zone concernée par le stage

La zone concernée par le stage a été encadrée sur la carte donnée par la figure 1.

1.3 Description des données

On traitera des données sur neuf ans (de 2002 à 2010) dans la zone du sud de la Nouvelle-Calédonie. On utilise des données collectées par le satellite MODIS Aqua. Il balaie l'ensemble de la Terre. Mais pour certains jours, le satellite ne balaie pas (ou pas entièrement) la zone qui nous intéresse. On n'a donc pas accès à tous les jours pour ces neuf années. Il arrive aussi que les nuages soient trop importants sur la zone considérée pour pouvoir faire quelque chose des données disponibles. Cela enlève également des jours dans notre base de données. Finalement, on dispose en tout de 1444 jours différents.

Pour faciliter la récupération des données satellites, on dispose d'un serveur OPeNDAP [8] qui contient les données triées (de notre zone) qu'on utilisera. Ces données sont au format HDF ou NetCDF. Sur le serveur, des données satellites de 2002 à 2010 sont disponibles (en tout 1444 jours différents durant ces années pour une zone située entre les latitudes -24,5° et -19,5° et les longitudes 164,5° et 169°). Les variables sont : les valeurs des différentes réflectances (Rrs) pour les longueurs d'onde de 412, 443, 488, 531, 555, 667, et 678 nm ; la concentration de chla calculée par OC3 (chl_oc3) et par OC5 (chl_oc5) ; valeurs drapeaux (l2_flags) qui indiquent le plus souvent un nuage ou la terre ; un indice de qualité pour la température de surface de l'eau (qual_sst) ; la valeur de la température de surface de l'eau (sst) et un masque composé spécialement pour notre zone à partir

8. Acronyme pour "Open-source Project for a Network Data Access Protocol". C'est une architecture de transport de données. Ces données sont structurées et peuvent être décrites.

de masques SeaDAS (SeaWIFS Data Analysis System) [9] (valhysat_masque). Pour chaque jour disponible, chaque variable disponible peut être représentée par un tableau avec 922 lignes et 1113 colonnes où dans chaque case se trouve la valeur de la variable pour la latitude et la longitude correspondant à la case en question, la longitude variant sur les lignes et la latitude variant sur les colonnes.

On ne manipulera pas 1444 fichiers (un pour chaque jour) de données différents. Sur OPeNDAP, on dispose aussi d'un "fichier agrégé" qui contient toutes les données décrites dans le paragraphe précédent dans un seul fichier. C'est principalement ce fichier qu'on utilisera durant le stage. Ce fichier peut être vu comme un fichier comportant des "tableaux 3D", dans lequel il y a autant de "tableaux 3D" que de variables. Chaque "tableau 3D" contient toutes les valeurs (en temps et en espace) de la variable qui lui correspond. Ce "tableau 3D" est formé de 1444 "étages" (un étage pour chaque jour disponible dans la base de données) qui contiennent chacun un tableau décrit dans le paragraphe précédent (tableau des données rangées par latitude et longitude).

Dans ces données, on peut rencontrer des valeurs négatives. Ces valeurs indiquent que la valeur de la concentration n'a pas été calculée (en effet, une valeur de concentration négative est absurde) pour différentes raisons : présence de nuage, proximité de côte, etc. Également, un masque a été mis au point pour ces données l'année dernière : le masque Valhysat (accessible, comme on le signalait plus haut dans la base de données). Ce masque permet de trier les données et de retirer en plus des valeurs absurdes que le satellite n'aurait pas masquées de lui-même.

Pour palier aux difficultés liées aux fonds, il était suggéré de différencier les points selon leur bathymétrie [10] : les profondeurs inférieures à 20 m, entre 20 et 70 m, et supérieures à 70 m (le large). Pour les deux algorithmes à disposition (OC3 et OC5), on a alors essayé de faire une "adaptation" des résultats du satellite (donnés par les algorithmes en question) pour qu'ils soient plus proches des résultats *in situ*. Cette méthode semblait efficace sur nos données [11].

Comme on dispose de deux algorithmes différents pour calculer la concentration de chla à partir des images satellites, l'idéal est d'utiliser les résultats de l'algorithme le plus adapté à la situation. Or chacun des deux algorithmes a ses défauts. Par exemple, l'algorithme OC3 a tendance à surestimer les concentrations de chla dans les eaux peu profondes du lagon à cause justement du fonds. Lorsqu'avec les images satellites les concentrations de chla calculées par OC3 sont très fortes dans le lagon, les fonds sont peu profonds. Apparemment, l'algorithme OC5 n'a pas ce défaut de biais par le fonds. Par contre, l'algorithme OC5 met une valeur minimum de concentration de 0,1 $\mu g.L^{-1}$. Or dans l'océan (où les eaux sont oligotrophes), il n'est pas rare d'avoir des concentrations inférieures à cette valeur. L'algorithme OC5 a aussi tendance à minimiser la concentration de chla en général (souvent dans des eaux assez profondes). On préfèrera donc utiliser l'algorithme OC3 dans les eaux plutôt profondes (bathymétrie supérieure à 20 m) ; par contre on utili-

9. Ces masques sont : GLINT, HIZEN, TURB, CLOUD, LAND, BATHY, TOA, STRAY, LIGHT et ICE.

10. La bathymétrie est la profondeur.

11. Il y aura une discussion de cette adaptatation dans la partie 2.2.1.

sera plus volontiers l'algorithme OC5 dans les eaux peu profondes du lagon (bathymétrie inférieure à 20 m).

2 Une première zone restreinte : le lagon et l'océan autour de Nouméa

Compte tenu du nombre et de la taille des données, on se restreint pour commencer à la zone du lagon autour de Nouméa, c'est-à-dire aux latitudes entre -22,8° et -22,2° et longitudes entre 166,1° et 166,7° [12]. Dans cette zone, se trouvent cinq stations sur lesquelles des mesures sont effectuées tous les mois. C'est avec ces mesures que des comparaisons ont pu être faites avec les valeurs des concentrations données par le satellite. Cette zone correspond également à la première emprise du modèle décrit en 1.1.5.1 (Faure et al., 2010, Special Marine Pollution Bulletin).

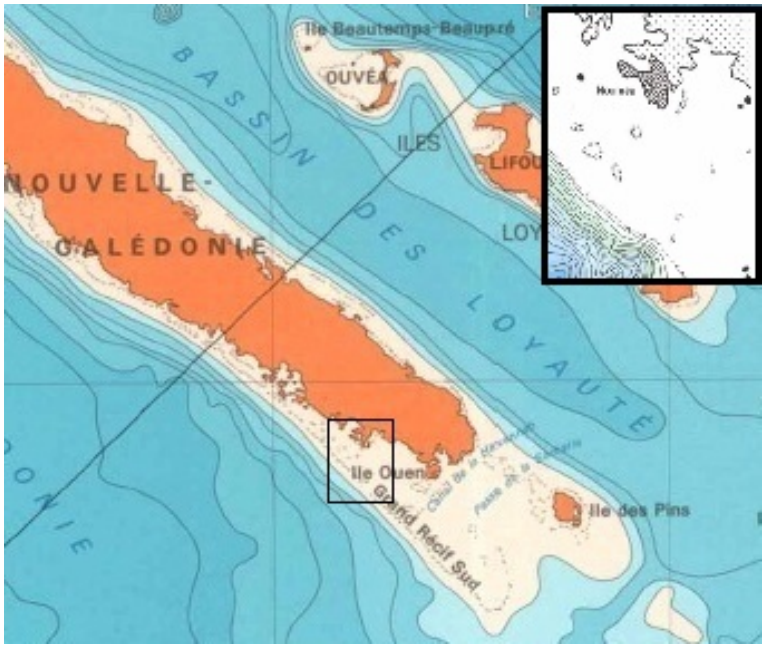


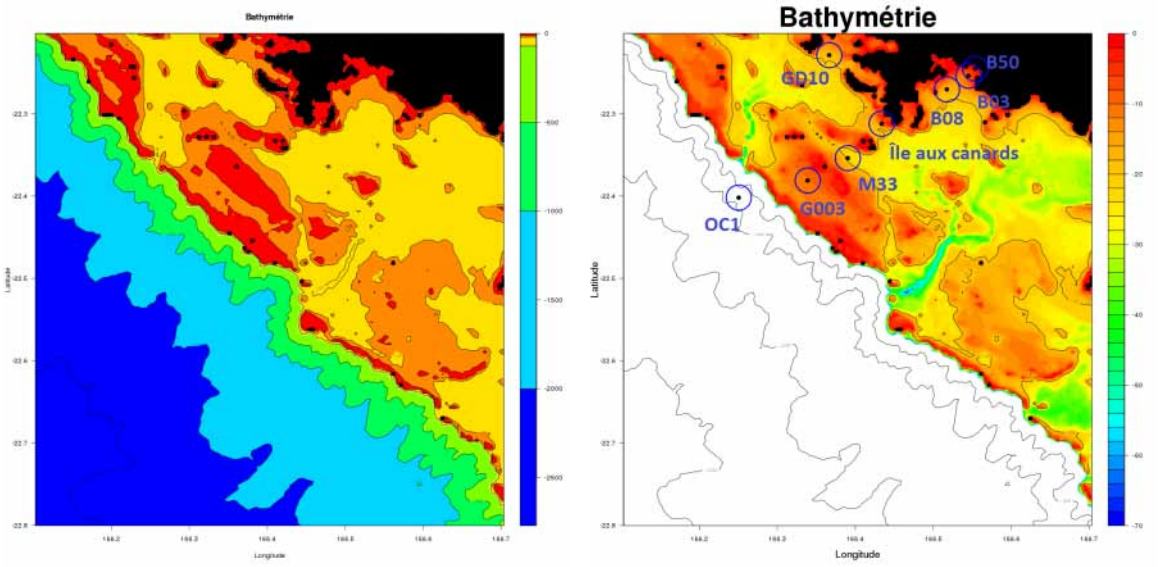
FIGURE 2 – Première zone concernée par le stage : autour de Nouméa

2.1 Première approche des données

2.1.1 Représentation du fond

On commence par faire une représentation de la profondeur de notre zone. On verra que ceci peut être très utile, surtout pour voir où les variables qui nous intéressent prennent les valeurs les plus importantes et où elles varient le plus. En effet, un des objectifs de Valhybio était d'estimer l'effet du fond sur les réflectances. Comme le lagon est clair (oligotrophe) la plupart du temps, on suspecte la réflectance du fond d'être une composante importante du signal perçu par le satellite.

12. Ainsi, on conserve les 1444 jours différents mais on a restreint la taille des tableaux spatiaux avec 124 lignes et 134 colonnes.



(a) Bathymétrie d'ensemble (b) Bathymétrie particulière au lagon

FIGURE 3 – Cartes de la bathymétrie dans la zone de Nouméa

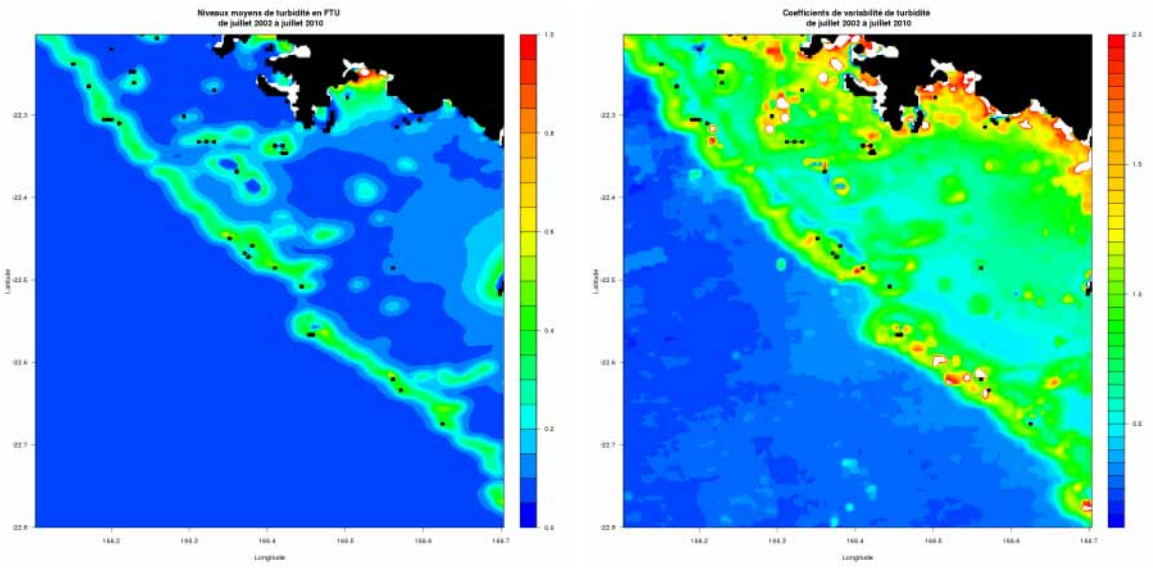
Les terres sont représentées en noir. Sur la carte 3(a), les couleurs chaudes sont les couleurs pour le lagon et les couleurs froides pour l'océan. Dans le lagon, le rouge est pour les profondeurs de 0 à 10m, le orange de 10 à 20m et le jaune de 20 à 70m. Les stations évoquées tout au long du rapport sont indiquées par des points avec leurs noms respectifs. Le récif corallien marque la limite entre l'océan et le lagon. Sur la carte 3(b), seule la profondeur du lagon a été représentée (on a toutefois représenté les courbes de niveau pour le fond de l'océan). On remarque deux "rivières sous-marines" débouchant dans l'océan : ce sont des lits fossiles de rivières qui coulaient avant que le lagon ne s'enfonce. La plus au nord débouche environ au point (166,25° ; -22,35°) et la deuxième au point (166,43° ; -22,50°).

2.1.2 Statistiques de base

Dans un premier temps, des moyennes, des écart-types et des coefficients de variation ont été calculés pour voir quelles valeurs nos différentes variables (principalement la turbidité, la température de surface et la concentration de chla) pouvaient prendre en général, quelles étaient leurs variabilités, et où ces valeurs et variabilités étaient les plus fortes. On a pu aussi avoir des renseignements sur les valeurs "habituelles" grâce aux valeurs mesurées sur le terrain. Par exemple, la concentration de chla varie souvent entre 0 et 1,5 $\mu g.L^{-1}$. C'est ainsi que sur les cartes représentant la répartition de chlorophylle, on a mis le seuil maximal à 1,5 $\mu g.L^{-1}$ mais dans nos données, on a conservé toutes les valeurs positives disponibles. En effet, il peut arriver que dans les tableaux, se trouvent des valeurs négatives. Ceci est en fait un masque du satellite ou de l'algorithme utilisé qui indique la présence d'un nuage, que le pixel en question est proche d'un îlot (ce qui en général introduit un biais dans l'estimation de la concentration de chla), ou d'autres complications qui empêchent de réaliser l'estimation en bonne due forme, comme cela a déjà été signalé dans le 1.3. Cependant, ces valeurs négatives ne sont pas prises en compte lorsque l'on applique le masque valhysat.

On a donc fait une moyenne sur chaque pixel spatial ; c'est-à-dire qu'on a au maximum pour chaque pixel, 1444 valeurs (une valeur par jour pour les jours où ce pixel n'est pas masqué) et on fait la moyenne sur ces valeurs (c'est une moyenne temporelle). On obtient aini un nouveau tableau (à deux dimensions celui-ci) dans lequel se trouve la valeur moyenne de la variable concernée pour chaque case (une case correspondant à un pixel et ce pixel étant identifié géographiquement par rapport à la place occupée dans le tableau : les longitudes varient en ligne et les latitudes en colonne). Cette opération a été faite pour la turbidité, la température de surface et pour la concentration de chla aussi bien avec l'algorithme OC3 qu'avec l'algorithme OC5, et enfin avec le choix de l'algorithme suivant la bathymétrie. Voici des cartes des moyennes réalisées grâce au logiciel R en partie.

Pour l'ensemble des cartes, les zones noires représentent la terre et les zones blanches sont soit des valeurs masquées (à cause de la terre trop proche en général [13]), soit des valeurs hors échelle (souvent trop fortes, et considérées comme abhérentes). On ne représente ici que les coefficients de variations (et non pas les coefficients de variation et écart-types) parce qu'ils sont plus intéressants dans notre cas que les écart-types.

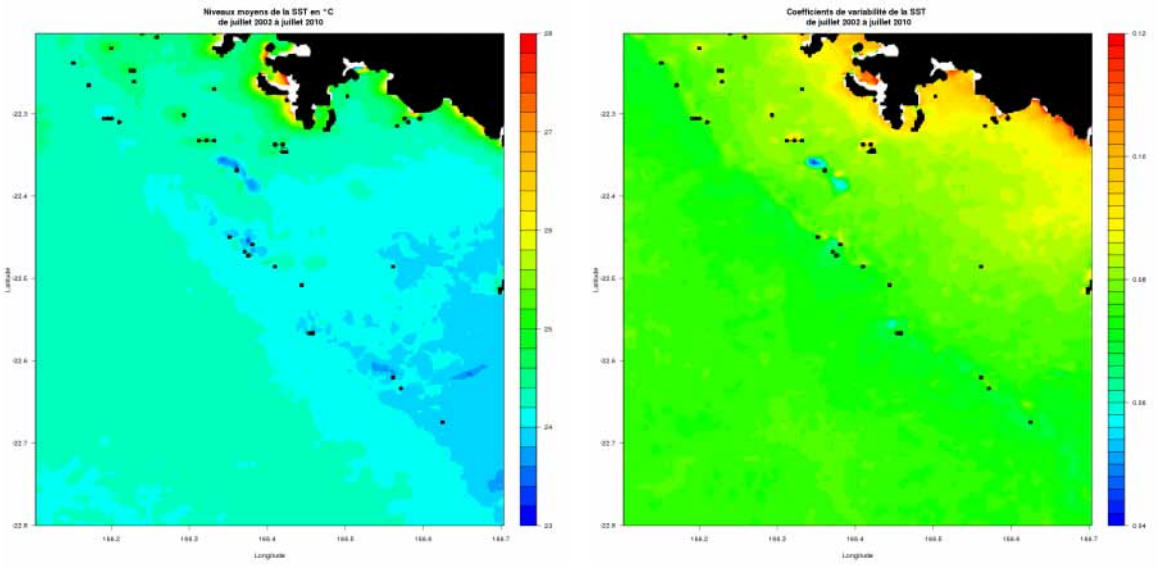


(a) Moyennes de turbidité (b) Coefficients de variation de turbidité

FIGURE 4 – Cartes des valeurs moyennes de turbidité et leurs coefficients de variation dans la zone de Nouméa

La turbidité est donnée en FTU (Formazin Turbidity Unit). Les valeurs de notre base de données ont été calculées à partir d'un algorithme développé par Sylvain OUILLON. On remarque sur ces cartes que la turbidité est la plus forte où la profondeur est la moins forte. Et qu'elle varie aussi beaucoup dans ces zones. D'ailleurs, la carte 4(b) montre bien que dans le lagon, la variation est beaucoup plus forte que dans l'océan, en particulier pour les endroits proches des côtes ou du récif.

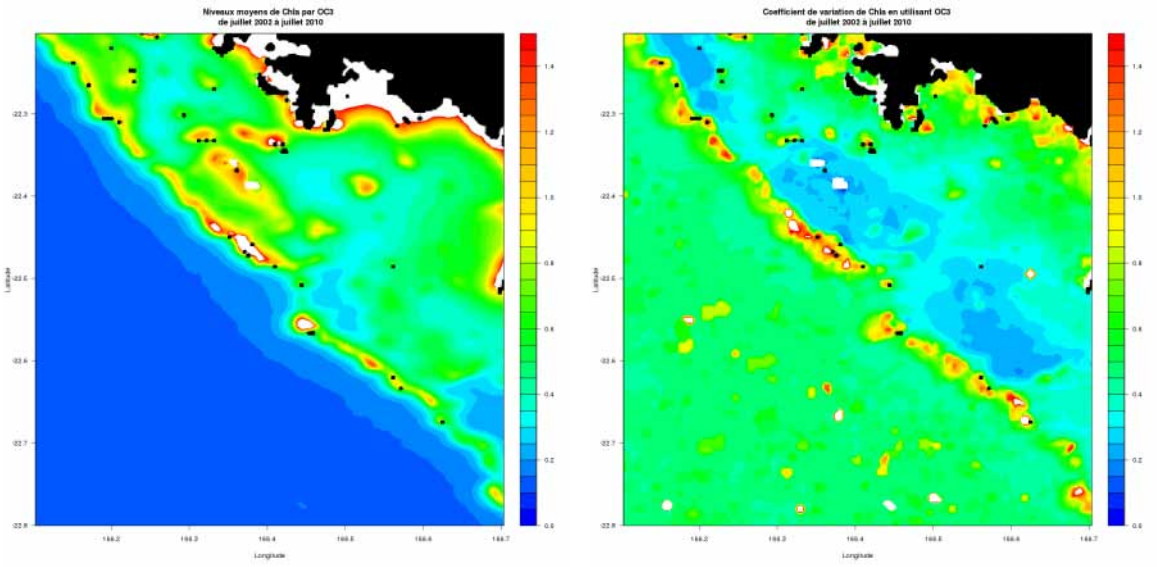
13. Le masque à cause des nuages n'intervient probablement pas ici puisqu'on fait une moyenne sur 1444 jours différents, à moins que le pixel en question masqué pour cause de nuage (ou un autre évènement variable) ait été masqué les 1444 jours de notre base de donnée...



(a) Moyennes de la température de surface (b) Coefficients de variation de la température de surface

FIGURE 5 – Cartes des valeurs moyennes de la température de surface et leurs coefficients de variation dans la zone de Nouméa

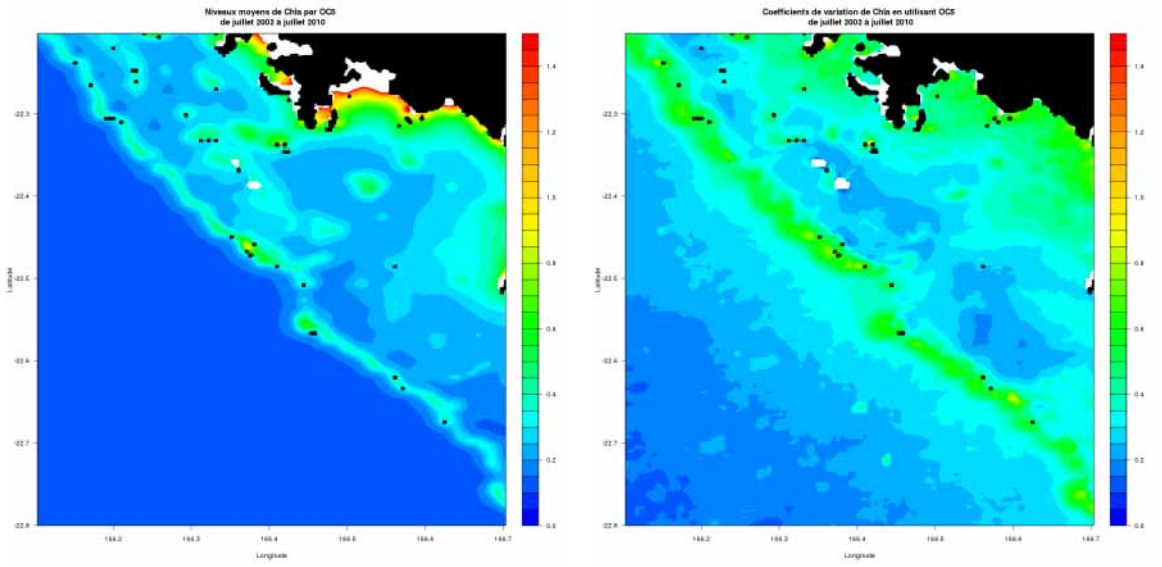
En moyenne, la température de surface, est plus forte près des terres que dans l'océan. Mais là, on ne voit pas vraiment de limite entre le lagon et l'océan et le récif n'est pas visible (contrairement à ce qu'on avait pu voir avec la moyenne de turbidité). La variation est également liée à la proximité avec la terre. Il faut cependant faire attention à l'échelle de la variation parce qu'elle n'est pas très étendue (de 0,04 à 0,12) en comparaison avec celle de la variation de la turbidité (de 0 à 2). Il y a donc des variations mais elles restent faibles pour la température de surface.



(a) Moyennes de la concentration de chla par OC3 (b) Coefficients de variation de la concentration de chla par OC3

FIGURE 6 – Cartes des valeurs moyennes de la concentration de chla calculée à partir de l'algorithme OC3 et leurs coefficients de variation dans la zone de Nouméa

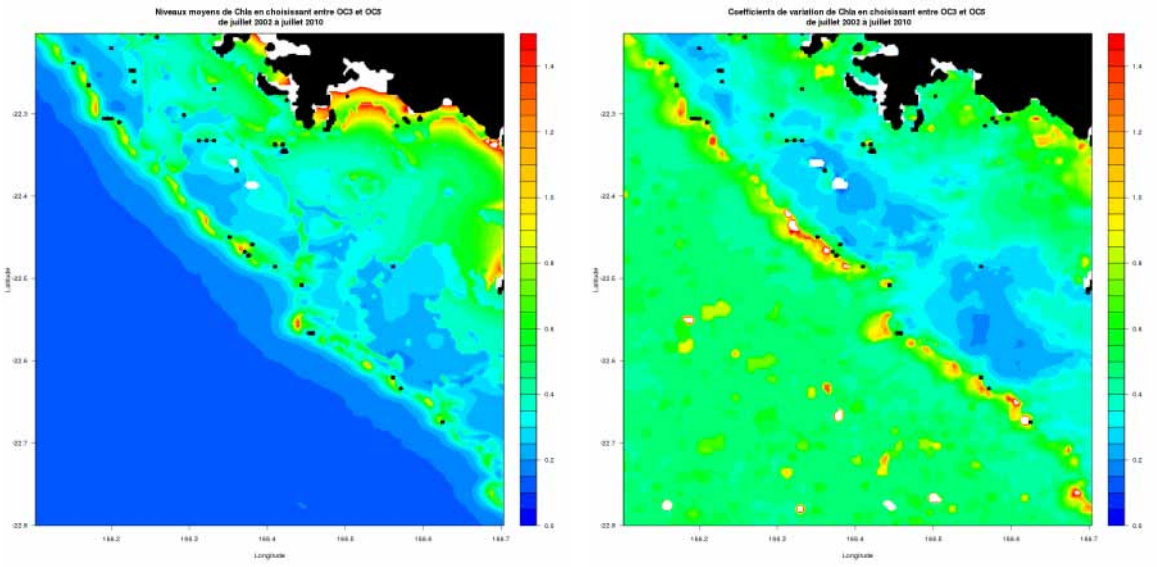
On regarde maintenant l'évolution dans l'espace de la concentration de chla. Là aussi, c'est une carte de moyennes et une carte des coefficients de variation de la concentration de chla estimée grâce à l'algorithme OC3. Pour la moyenne, on voit clairement l'influence de la profondeur. Il y a une nette distinction entre les valeurs dans le lagon et les valeurs dans l'océan. Dans l'océan, les valeurs varient entre 0 et 0,3 $\mu g.L^{-1}$ alors que dans le lagon, elles sont plutôt comprises entre 0,2 et 1,5 $\mu g.L^{-1}$, voire au-delà. En effet, on remarque que près des côtes ou au niveau du récif (c'est-à-dire où la profondeur est faible), certaines valeurs sont hors échelle (au-dessus de 1,5 $\mu g.L^{-1}$), ce qui est visible par le blanc. En revanche, les variations sont assez fortes sur l'ensemble de notre zone (que ce soit le lagon ou l'océan) avec cependant quelques exceptions. On voit une variabilité qui semble plus élevée autour des côtes et du récif mais on remarque également une faible variabilité dans trois zones en particulier. Ces trois zones se situent de part et d'autre de "rivières sous-marines".



(a) Moyennes de la concentration de chla par OC5 (b) Coefficients de variation de la concentration de chla par OC5

FIGURE 7 – Cartes des valeurs moyennes de la concentration de chla calculée à partir de l'algorithme OC5 et leurs coefficients de variation dans la zone de Nouméa

En ce qui concerne l'estimation de concentration de chla par l'algorithme OC5, on fait à peu près le même constat qu'avec OC3 : les concentrations semblent plus fortes dans le lagon par rapport à celles de l'océan et la démarquation du récif est une fois de plus bien visible. Seulement, avec OC5, les valeurs sont généralement plus faibles dans le lagon par rapport aux valeurs données par OC3. Les valeurs données par le satellite sont en quelques sortes "écrasées" pour des bathymétries faibles par rapport à celles d'OC3. Par exemple, sur le récif, on ne voit plus de valeurs hors échelle (au-dessus de 1,5 $\mu g.L^{-1}$) et autour de la côtes, les valeurs hors échelles sont moins étendues par rapport à ce qu'on avait avec OC3. C'est une des raisons qui nous pousse à choisir l'algorithme OC5 dans les zones de faibles profondeurs (OC3 surestimant la concentration pour ces zones). Les coefficients de variation également souvent beaucoup plus faibles avec l'algorithme OC5 qu'avec OC3. On n'a pas l'impression de voir de valeurs de coefficients supérieures à 1,3 (contrairement à OC3). Les variations semblent être les plus fortes autour des côtes et le long du récif corallien. Dans l'océan, le coefficient prend ses valeurs entre 0 et 0,4 ce qui est faible par rapport aux variations qu'on avait avec OC3. Ici aussi, on peut voir dans le lagon les trois zones de faibles variations comme pour OC3.



(a) Moyennes de la concentration de chla en choisissant l'algorithme suivant la bathymétrie

(b) Coefficients de variation de la concentration de chla en choisissant l'algorithme suivant la bathymétrie

FIGURE 8 – Cartes des valeurs moyennes de la concentration de chla calculée en choisissant l'algorithme suivant la bathymétrie et leurs coefficients de variation dans la zone de Nouméa

Comme il était suggéré, on a finalement calculé les moyennes en choisissant l'algorithme selon la profondeur [14]. On rappelle que pour des profondeurs inférieures à 20m, on choisit OC5, sinon, on choisit OC3. Les observations de ces cartes sont donc les mêmes que celles faites pour l'algorithme OC3 pour des profondeurs inférieures à 20m et ce sont les mêmes que pour OC5 dans les profondeurs supérieures à 20m. Il y a toutefois un léger défaut à cette méthode : on remarque une forme de croissant et une autre forme allongée dans les concentrations de chla le long de la côte environ à -22,3° de latitude et entre 166,5° et 166,7° de longitude. L'explication est la suivante : à ces endroits, il y a plus de profondeur ; on choisit donc les résultats d'OC3 (qui est très fort dans cette zone) alors que plus près de la côte, on choisit les résultats d'OC5.

On remarque que quel que soit l'algorithme utilisé (OC3 et OC5), le fond est très visible pour ce qui est de l'estimation de la concentration de chlorophylle : dans les endroits peu profonds (près des côtes, des îlots ou du récif), la concentration est forte alors que dans des endroits plus profonds elle est moins forte. Est-ce que ce phénomène est dû au fait qu'il y a effectivement plus de chlorophylle dans les zones peu profondes ou est-ce que le signal perçu par le satellite est faussé à cause de la vision du fond ?

14. Ce choix est expliqué dans la partie suivante : 2.2.

2.2 Problèmes d'algorithmes et solutions

Le présent stage est en fait continuité d'un précédent stage effectué par Tatiana Savranski qui a, entre autre, utilisé des données de terrain et la base de données dont on dispose, pour pouvoir comparer les deux informations. Lors des comparaisons, la stagiaire a effectué des régressions : les valeurs de la concentration calculées par OC3 et OC5 en fonction des valeurs de concentration mesurées *in situ*. C'est ainsi qu'on a pu s'apercevoir des erreurs commises par les algorithmes utilisés, erreurs déjà mentionnées dans la partie 1.2.3. On avait vu dans cette partie que principalement deux méthodes ont été utilisées pour faire cette comparaison (en prenant la valeur du pixel non masqué le plus proche ou en prenant la valeur de la moyenne pondérée par la distance des pixels aux alentours non masqués) ; en annexe A, sont présentées les régressions pour la méthode de la moyenne pondérée, méthode qui semble la plus adaptée dans notre cas. Les deux algorithmes dont on dispose sont présentés [15] et les graphes sont des graphes à l'échelle logarithmique.

Comme on le disait dans la partie 1.2.3, l'algorithme OC3 semble bien fonctionner pour les eaux du large et les eaux du lagon dont la bathymétrie est supérieure à 20m (et inférieure à 70m). En revanche, il surestime la concentration pour les zones peu profondes (bathymétrie inférieure à 20m). Et pour l'algorithme OC5, les concentration dans les zones du lagon (bathymétrie inférieure à 70m) sont plutôt bien estimées pour de faibles profondeurs ; par contre, dans les eaux du lagon plus profondes la concentation est sous-estimée. D'où une des solutions envisagées : choisir OC3 pour les eaux assez profondes (eaux du large et eaux du lagon avec une bathymétrie supérieure à 20m) et OC5 pour les eaux moins profondes (bathymétrie inférieure à 20m) comme on l'a signalé dans 2.1.2.

Mais comme on le signalait dans ce paragraphe, on observe des discontinuités avec cette méthode (discontinuités qui se voient notamment par la forme de croissant dans la carte moyenne 8(a)) ce qui est contre-intuitif par rapport à l'idée qu'on se fait de la forme de la variation de la concentration de chla dans l'espace. On cherche donc des méthodes pour adapter les algorithmes ou utiliser des algorithmes mieux adaptés à notre situation.

2.2.1 Première idée d'amélioration : utiliser une ACP

Premièrement, comme on l'a appris dans le cours d'Analyse de données au chapitre 4 ([C-BIE : Cours AD]), les premiers axes d'une ACP sont les axes qui expliquent le plus la dispersion des données. Ce sont aussi les axes qui minimisent les distances axe-points. Deuxièmement, comme on veut que les estimations par satellite soient le plus proche possible de la réalité (ici ce sont les données *in situ*) ; géométriquement on veut en fait que les points soient le plus proche possible de la première bissectrice (sur nos graphiques de régression), d'équation : $y = x$.

L'idée est donc de calculer l'équation du premier axe de l'ACP réalisée avec nos données : un individu est une observation et les deux variables sont la mesure *in situ* et l'estimation par un algorithme (ici, ce sera OC3 ou OC5). Une fois l'équation connue, on pourra "redresser" nos observations, de façon à ce que nos données corrigées soient plus proches de la droite $y = x$. Pour vérifier la validité de cette méthode, avant de l'appliquer,

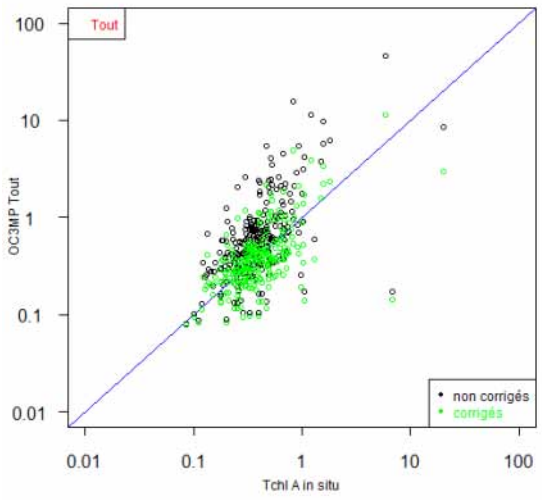
15. Annexe A.1 pour OC3 et annexe A.2 pour OC5.

on a conservé aléatoirement un échantillon de données, représentant 75 % des données, pour faire un test. On verra alors si on améliore vraiment les résultats.

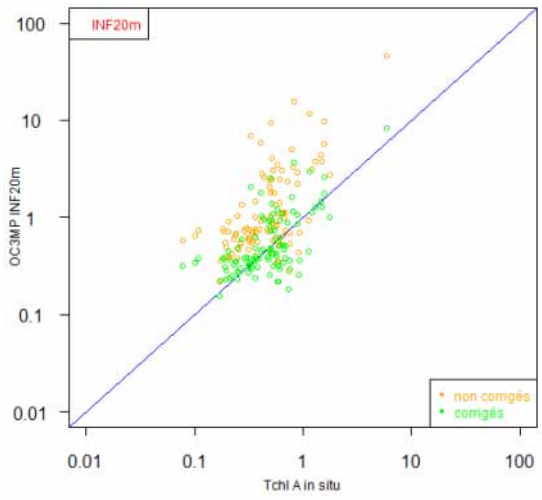
2.2.1.1 Description de la méthode On utilise un fichier Excel réalisé lors du stage de l'année dernière. Dans ce fichier, plusieurs variables sont présentes dont les mesures de concentration de chla à différentes stations, les concentrations estimées par satellite (OC3 et OC5, par le pixel le plus proche et par moyenne pondérée par la distance, avec un possible "écart temporel" de ±5 jours). On dispose en tout de 986 mesures différentes (nos individus pour l'ACP) dont 419 pour une bathymétrie inférieure à 20m, 392 pour une bathymétrie supérieure à 20m (et inférieure à 70m) et 175 pour le large (bathymétrie supérieure à 70m). La méthode a été utilisée d'abord pour tous les points (sans distinction de profondeur), puis en les distinguant selon la profondeur. Voici les étapes du programme pour mettre en œuvre la méthode :

1. "Choisir au hasard" les deux échantillons : 75 % pour apprendre et les 25 % restant pour tester.
2. Réaliser l'ACP sur l'échantillon pour apprendre
3. Grâce aux coefficients COR, déduire la pente a de la droite qui explique le mieux la dispersion (elle est parallèle au premier axe de l'ACP) :
$$a = \frac{COR(2,1)}{COR(1,1)}$$
4. Connaissant les moyennes des deux variables (la variable M représente les mesures *in situ* et la variable S représente les estimations par satellite), calculer la valeur b de l'ordonnée à l'origine :
$$b = \overline{S} - a \times \overline{M}$$
5. On dispose ainsi de l'équation de la droite qui explique le mieux la dispersion : $y = ax + b$
6. Pour redresser les données, il suffit de soustraire b et de diviser par a :
$$M_{corr} = \frac{M - b}{a}$$
où M_{corr} est le nom pour la variable des données corrigées par cette méthode.

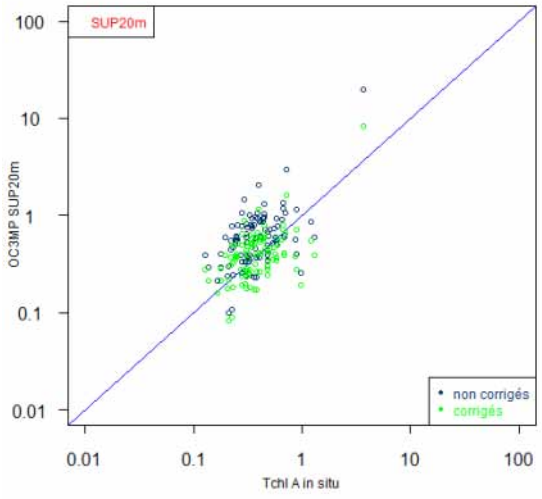
2.2.1.2 Résultats de la méthode Cette méthode a été testée sur un échantillon non utilisé pour avoir les résultats. Voici les nuages de points correspondant pour la méthode appliquée à OC3-calcul par moyenne pondérée :



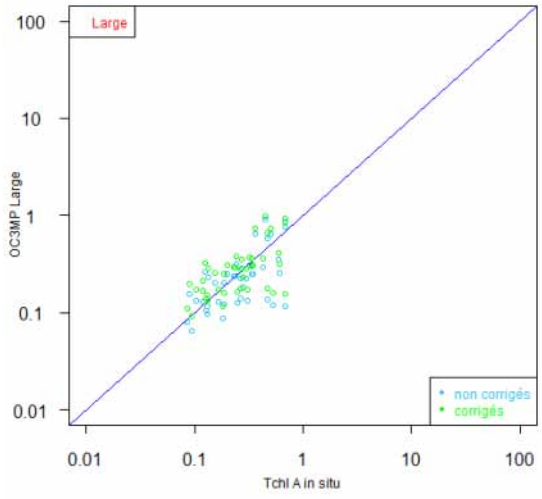
(a) Correction d'OC3 pour tous les points



(b) Correction d'OC3 pour les points dont la bathymétrie est inférieure à 20m



(c) Correction d'OC3 pour les points dont la bathymétrie est supérieure à 20m



(d) Correction d'OC3 pour les points du large

FIGURE 9 – Correction d'OC3 par moyenne pondérée par le premier axe d'ACP

On remarque qu'en général, le nuage de points s'est "rapproché" de la première bissectrice. Comme suggéré par [S-OUI] (qui s'est lui-même appuyé sur d'autres auteurs comme Toole (2000) et Darecki et Stramski (2004)), pour quantifier l'éventuelle amélioration, on utilise les indicateurs suivants : MNB (Mean Normalized Bias) et $RMSE$ (Root Mean Square Error). Voici leurs expressions :

$$MNB = moyenne\left(\frac{y_{algo} - y_{obs}}{y_{obs}}\right).100$$
$$= \frac{1}{n}\sum_{i=1}^{n}\frac{y_{alg_i} - y_{obs_i}}{y_{obs_i}}.100$$

$$RMSE = \sigma\left(\frac{y_{algo} - y_{obs}}{y_{obs}}\right).100$$
$$= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_{alg_i} - y_{obs_i}}{y_{obs_i}} - \overline{\frac{y_{alg} - y_{obs}}{y_{obs}}}\right)^2}.100$$

Avec y_{alg} les valeurs calculées par l'algorithme à tester et y_{obs} les valeurs observées correspondantes.

Le MNB représente en fait le taux d'erreur moyen commis par l'algorithme utilisé. La multiplication par 100 permet de l'exprimer en pourcentage. Donc s'il est petit, l'algorithme estime bien les observations en moyenne. Et grâce à cet indicateur, on peut également comparer deux algorithmes pour voir lequel est le plus adapté à notre cas. Quant au $RMSE$, il représente l'écart-type du taux d'erreur (là aussi, il est exprimé en pourcentage dans notre cas). S'il est petit, le taux d'erreur ne varie pas beaucoup. Pour avoir un "bon algorithme", il est donc nécessaire d'avoir à la fois un faible MNB et un faible $RMSE$[16].

On a représenté les résultats comparatifs dans le tableau suivant (TABLE 1). On compare ainsi les résultats des données corrigées par l'axe de l'ACP et les données OC3 "brutes" calculées par moyenne pondérée (OC3MP).

TABLE 1 – Comparaison entre OC3MP et OC3MP corrigé par le premier axe d'ACP

OC3MP	MNB	MNB corrigé	$RMSE$	$RMSE$ corrigé
Lagon : bathymétrie inférieure à 20m	246,07	38,52	335,53	96,61
Lagon : bathymétrie entre 20m et 70m	73,37	11,47	104,49	57,22
Large : bathymétrie supérieure à 70m	-0,89	21,91	65,74	76,45
Tous les points	124,56	31,80	196,33	88,41

On peut dire qu'on a de bons résultats avec cette méthode. Les MNB et $RMSE$ sont toujours plus faibles lorsqu'on a corrigé l'algorithme, sauf pour les zones du large. Mais les zones qui nous intéressent le plus sont les zones du lagon.

Remarque 1 : Pour voir les comportements de la correction pour les autres algorithmes (OC3 avec le pixel le plus proche noté OC3CL, OC5 avec la moyenne pondérée noté OC5MP et OC5 avec le pixel le plus proche noté OC5CL), voir l'annexe B (TABLE

16. Si le MNB est faible mais le $RMSE$ grand, il se peut qu'il y ait "compensation" entre des valeurs de l'algorithme sous-estimées et des valeurs surestimées.

7).

Remarque 2 : Les chiffres donnés dans les tableaux de comparaison mentionnés sont "relatifs" : ils dépendent du partage échantillon pour apprendre et échantillon gardé pour le test. Mais la méthode a été lancée plusieurs fois avec la plupart du temps les mêmes conclusions. Pour encore mieux appréhender la correction, on aurait pu faire une validation croisée. Cependant, cela n'a pas été jugé utile puisqu'on utilisera dans la suite, plutôt la méthode décrite dans la partie suivante 2.2.2, méthode qui nous semble plus performante.

2.2.2 Deuxième idée d'amélioration : utiliser l'algorithme SVM

Comme on sait que les algorithmes dont on dispose ne sont pas tout à fait adaptés à notre zone d'étude (le lagon de Nouvelle-Calédonie), une autre idée est d'utiliser les données *in situ* dont on dispose avec les données satellites correspondantes pour faire un algorithme entièrement adapté à notre cas. Pour cela, on va utiliser un algorithme appelé SVM (Support Vector Machine). On fait une description de cette méthode dans l'annexe C.

2.2.2.1 Description de la méthode Il existe dans le logiciel R, un package (e1071) contenant une fonction nommée "svm" qui calcule une régression grâce à la méthode du SVM. On a sélectionné les variables à utiliser pour la régression parmi les variables dont on dispose, pour expliquer la concentration de chla *in situ*. On espère ainsi avoir un algorithme qui donne une concentration calculée par les variables du satellite assez proche de ce qu'on peut espérer par rapport aux valeurs *in situ* à disposition. On a choisi les variables qui semblaient les plus pertinentes. Pour cela, on a fait tourner des modèles SVM différents sur les mêmes données que celles utilisées pour la précédente méthode (c'est-à-dire que la valeur mesurée est comparée aux valeurs de la bathymétrie et des réflectances sélectionnées), puis la corrélation entre les valeurs *in situ* et les valeurs fournies par le modèle a été calculée. On a enlevé les variables pour lesquelles la suppression n'entraînait pas de forte baisse de la corrélation. Les variables conservées sont finalement : la bathymétrie et les réflectances 443 nm, 488 nm et 555 nm [17]. Ces résultats sont rassurants puisque pour le calcul de l'algorithme OC3, ce sont justement ces réflectances qui sont utilisées. On espère donc logiquement une amélioration par rapport à cet algorithme puisque la variable bathymétrie est introduite en plus des réflectances utilisées habituellement pour le calcul de concentration de chla. En effet, en introduisant en plus la variable bathymétrie, on espère, en plus, s'affranchir d'un effet de vision des fonds par le satellite. Ainsi, on pourra éventuellement avoir un élément de réponse afin de savoir si les eaux peu profondes sont, comme nous le montre les algorithmes habituels disponibles (OC3 et OC5), plus riches que des eaux plus profondes ou non.

Une fois le modèle établi, il a été appliqué à l'ensemble de nos données [18]. On a ainsi en sortie une concentration de chla calculée à partir d'un modèle prenant en compte des valeurs des concentration mesurées.

17. Le code pour la construction du modèle SVM est en annexe E.2.
18. Le code de l'application du modèle sur les données satellites est mis en annexe E.3.

2.2.2.2 Résultats de la méthode Pour vérifier le modèle, une validation croisée ("leave-one-out cross validation") du modèle a été réalisée avant de l'appliquer à l'ensemble de nos données. Puis, différents indicateurs ont été calculés pour comparer avec les algorithmes déjà disponibles. Voici quelques résultats :

TABLE 2 – Comparaison entre le SVM et les autres algorithmes

Algorithme	Corrélation	Corrélation entre log	Erreur quadratique moyenne	MNB	RMSE
OC3MP	0,42	0,61	3,80	1,21	2,59
OC5MP	0,43	0,63	1,62	-0,02	0,87
Choix entre OC3 et OC5 selon la bathymétrie	0,36	0,57	1,04	0,37	1,09
SVM	0,44	0,70	0,71	0,07	0,47

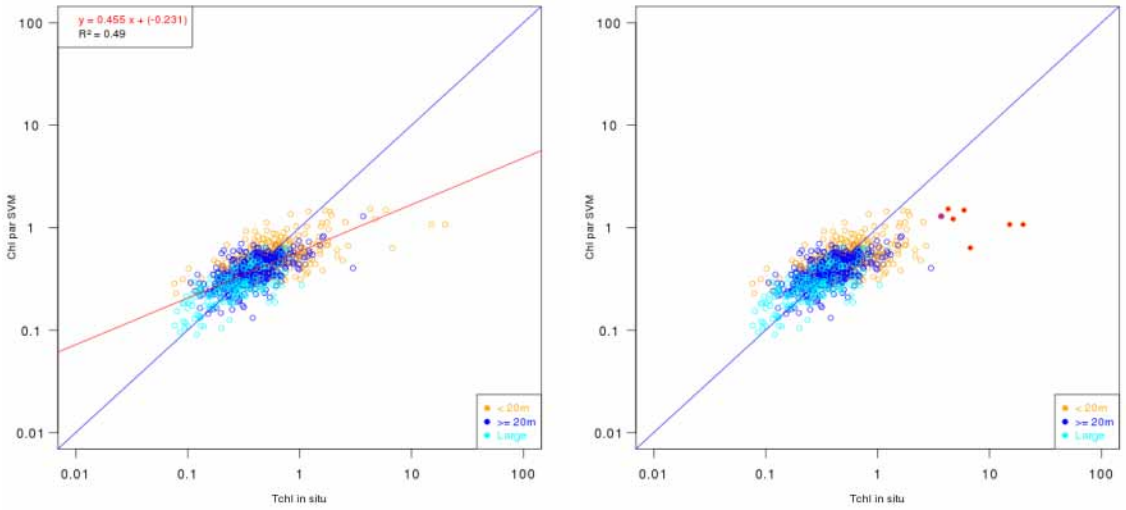
On remarque qu'en général, l'amélioration est intéressante. La corrélation est légèrement améliorée, l'erreur quadratique moyenne est nettement meilleure, le MNB est également toujours meilleur, sauf par rapport à OC5. Mais le fait d'avoir le $RMSE$ et l'erreur quadratique moyenne permet de "tempérer" un peu cette différence. En effet, le $RMSE$ est plus faible pour notre algorithme SVM ce qui indique une variabilité plus faible dans l'erreur (les compensations surestimation - sous-estimation sont moins flagrantes) ; l'erreur quadratique moyenne ne tient pas compte non plus de la surestimation ou de la sous-estimation mais plutôt de la différence entre ce qui est mesuré et le résultat de l'algorithme. Le fait que le MNB soit positif (0,07) indique une légère surestimation de la part de notre algorithme.

On a représenté le nuage de points : concentration de chlorophylle estimée par le SVM (les résultats fournis par la validation croisée) en fonction des résultats *in situ* de la même façon qu'on a pu le faire pour les algorithmes OC3 et OC5.

On remarque que le nuage de points est globalement moins dispersé que pour les autres algorithmes (OC3 ou OC5) et il semble plus proche de la droite d'équation $y = x$ comme le laissaient supposer les indicateurs d'erreur. Contrairement aux autres algorithmes, on voit également qu'il n'y a pas de "zones bathymétriques" étant surestimées ou sous-estimées par rapport aux autres. Néanmoins, certains points se démarquent clairement de l'ensemble du nuage : des points de faible profondeur (4, 5, 10 et 21m) représentés en rouge sur le graphique 10(b) ; ce sont les points pour les stations : GD1, D65, D46, R05, D47, GR12 et D39. Cependant, leur approximation n'est pas aberrante, sauf pour deux ou trois de ces points qui ont une estimation bien inférieure à la valeur mesurée. Ces points "à part" sont soit situés au niveau des baies, soit des mesures prises à des moments exceptionnels (à la suite du cyclone Érica en mars 2003 par exemple).

Finalement, pour la suite, on va conserver les résultats de ce modèle pour les appliquer aux données satellites. On utilisera donc la bathymétrie, les réflectances pour les longueurs d'onde 443, 488 et 555 nm afin de déterminer une valeur de concentration pour les pixels et les jours disponibles.

Remarque : Cette méthode est bien adaptée pour notre zone d'étude mais on ne



(a) Estimation par SVM (en validation croisée) en fonction des mesures *in situ* pour tous les points

(b) Points se démarquant du nuage

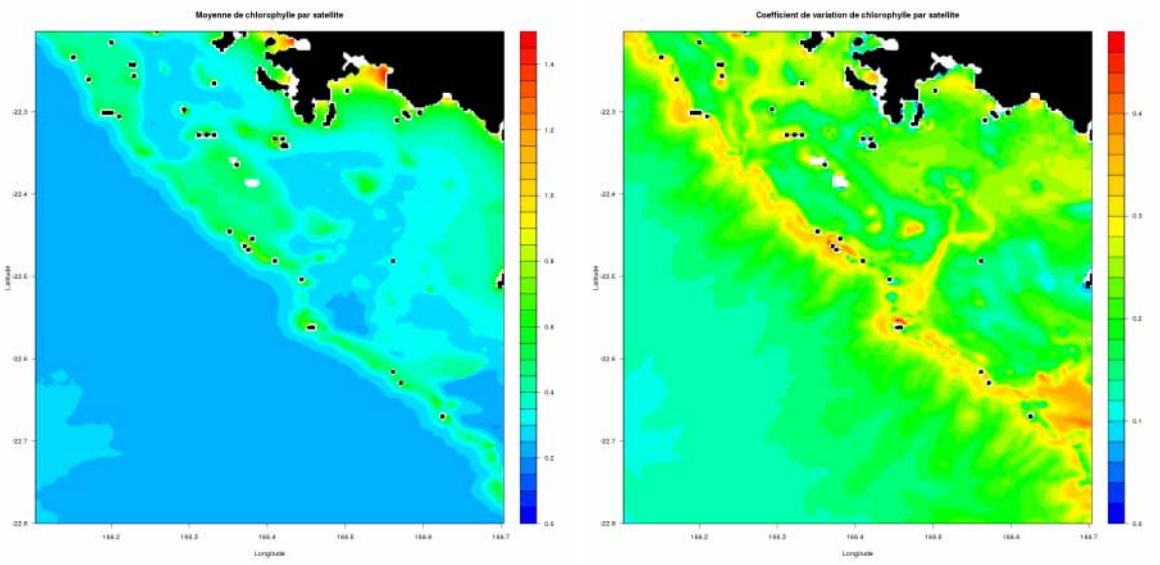
FIGURE 10 – Nuage des résultats du SVM par validation croisée

pourra pas généraliser pour d'autres parties du monde. Si on veut utiliser le SVM pour d'autres endroits, il faudra encore avoir une base de données de mesures *in situ* confrontées à des données satellites pour le même jour (ou à ± 5 jours comme la base utilisée a été construite).

2.3 Traitement des données avec le nouvel algorithme

2.3.1 Visualisation des résultats pour le modèle SVM appliqué aux données satellites

Comme pour les autres algorithmes, on a commencé par représenter la carte des moyennes de concentration de chla et la carte des coefficients de variation pour chaque pixel.



(a) Moyenne de chlorophylle de 2002 à 2010 par le modèle SVM

(b) Coefficient de variation de la chlorophylle de 2002 à 2010 par le modèle SVM

FIGURE 11 – Cartes résultats de l'application du modèle SVM

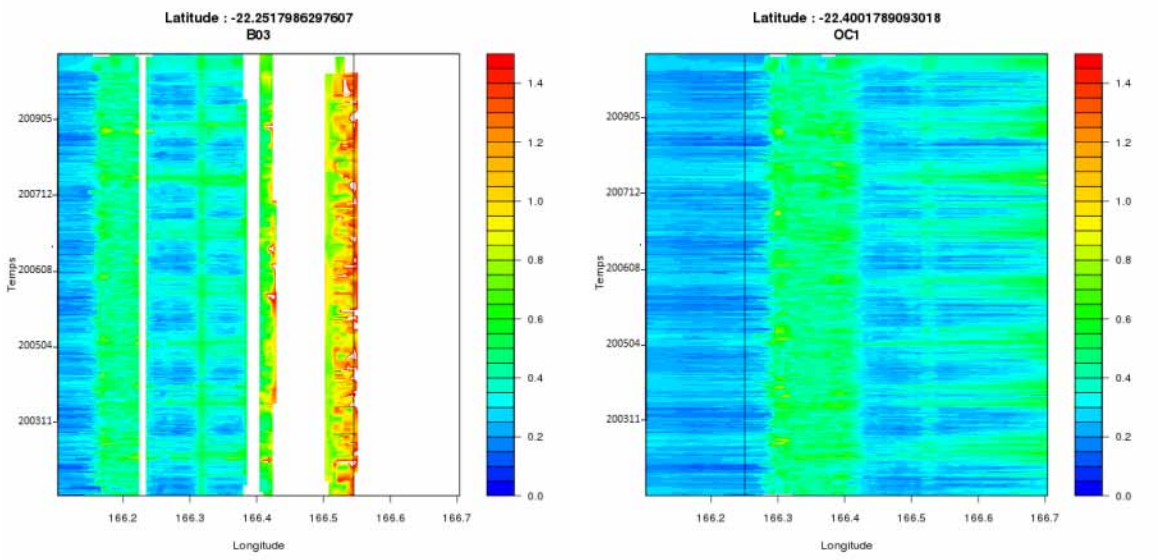
Grâce à la carte moyenne, on peut voir des formes dans la répartition de la concentration de chla. Ces formes sont très semblables à celles de la bathymétrie. On le voit aussi très clairement sur le coefficient de variation. À la question qui était posée dans la partie 2.1.2, savoir : le satellite est-il faussé à cause de la vision du fond, ceci montre qu'il y a sûrement une part de vrai dans le signal. En effet, on rappelle que la bathymétrie était une variable intégrée dans le modèle SVM, donc l'effet de la bathymétrie a été intégré pour l'estimation de concentration de chla [19].

Toutefois, il existe un autre paramètre qui pourrait entrer en jeu : la couleur du fond. Une carte des types (donc de la couleur) des fonds est disponible dans une certaine zone (dont la zone de Nouméa qui nous occupe dans cette partie) ; mais comme la zone sera agrandie dans la suite, on ne disposera pas de toutes les informations. Mais rien n'empêchera, une fois que le type de fond sera connu dans toute la zone, de refaire un modèle SVM avec le paramètre supplémentaire "type de fond" ; et voir si effectivement, l'estimation est améliorée.

Des graphes de variation temporelle (pour une latitude donnée, en abscisses la longitude et en ordonnée le temps) ont été faits également, suite à une interpolation temporelle [20], pour quelques stations où des mesures sont faites régulièrement. Voici deux exemples de ces graphes :

19. Une étude parallèle, plutôt basée sur des mesures *in situ*, est menée afin de voir si effectivement la concentration de chla est significativement plus élevée dans des eaux peu profondes.

20. Les raisons de cette interpolation et le type d'interpolation seront vus dans la partie suivante 2.3.2.



(a) Évolution dans le temps et en longitude de la concentration de chlorophylle pour la latitude de la station B03

(b) Évolution dans le temps et en longitude de la concentration de chlorophylle pour la latitude de la station OC1

FIGURE 12 – Cartes espace-temps résultats de l'application du modèle SVM

Ces deux stations choisies sont des stations "extrêmes" : l'une (OC1) est au bord de la barrière de corail du côté océan et l'autre (B03) est dans une baie très proche de la côte. La ligne noire verticale sur chaque graphe représente la position de la station. Les zones blanches sont en général des terres ou des îlots.

On remarque sur ces graphes que la concentration semble avoir une période aussi bien dans le lagon que dans l'océan, bien que les variations dans l'océan semblent moins marquées que pour le lagon. On a du mal à identifier, à première vue, une période pour les endroits situés au-dessus du récif corallien. Et pour les zones côtières (voir 12(a)), il semble qu'il n'y ait aucune période. Afin d'étudier plus précisément l'évolution dans le temps de la concentration de chla, passons aux séries temporelles correspondantes.

2.3.2 Séries temporelles pour la concentration de chla

Pour chaque pixel de la carte, on dispose d'une série de 1444 points pour une période s'étendant du 2 juillet 2002 au 7 juillet 2010. Afin d'avoir des données journalières, une interpolation temporelle est nécessaire. Notre choix s'est porté sur une "simple interpolation linéaire" avec R.

L'étude de ces séries nous permettra de dégager une tendance et une saisonnalité. Ceci permettra, entre autre, de voir si la période de l'année joue sur les valeurs de concentration. Enfin, l'étude du bruit de la série permettra de voir si un évènement climatique particulier (fortes pluies, cyclone, ...) a influencé la concentration sur un certain laps de temps. Chaque pixel a subi le même traitement pour réaliser la série interpolée.

2.3.2.1 Traitement des séries Dans un premier temps, on a cherché à dégager une tendance pour chaque pixel. Par les représentations de ces séries, on a déduit que la ten-

dance devait être linéaire. Le coefficient linéaire était souvent très faible (de l'ordre de 10^{-6} à 10^{-3} au maximum, et pouvant être positif ou négatif). Les tests de significativité des coefficients ont été vérifiés. Pour la plupart des points, on pouvait accepter l'hypothèse de nullité du coefficient linéaire. On a choisi finalement de prendre la moyenne de la série pour la tendance : l'erreur commise est très minime au regard des ordres des coefficients linéaires par rapport aux ordres des coefficients d'interception (environ 10^{-1}).

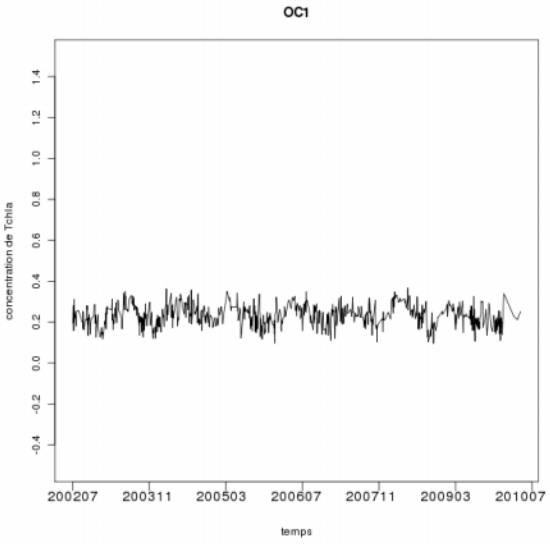
Ensuite, on a calculé une saisonnalité annuelle. Pour cela, on a procédé en plusieurs étapes, on a :

- appliqué une moyenne mobile à notre série afin de la lisser et d'éviter de prendre en compte des pics intempestifs. Même si, après l'application de la moyenne mobile, la série a encore des pics très forts, ils sont moindres par rapport à la série de départ.
- calculé la moyenne pour chaque mois disponible dans nos données, en différenciant chaque année.
- redéterminé, à l'aide d'une interpolation (avec un pas régulier) par splines cubiques, des valeurs "théoriques" (autant que de jours dans la base de données) sur chaque pas.
- calculé une moyenne par jour (du 1^{er} janvier au 31 décembre) à partir de la dernière interpolation.
- centré la dernière série afin d'obtenir une saisonnalité de moyenne nulle.

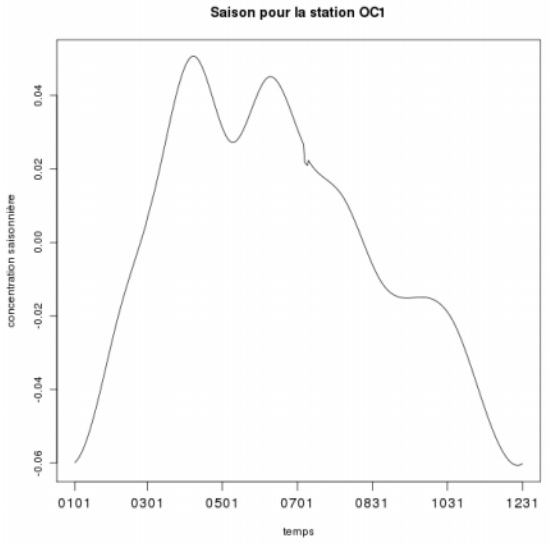
Enfin, le bruit de la série a été déduit en soustrayant la tendance et la saisonnalité de notre série de départ.

2.3.2.2 Exemples de représentation des séries pour deux stations Voici, pour deux des stations visitées, les séries et leur décomposition [21] :

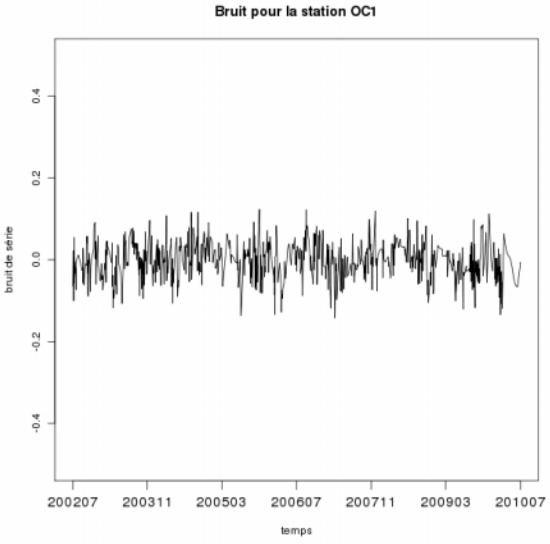
21. En abscisse, les dates sont indiquées. Pour les grahes a, c et d, on a indiqué l'année et le mois (par exemple, 200207 correspond à juillet (07) 2002) et pour le graphe b, on a indiqué le jour et le mois de l'année (par exemple, 0301 correspond au 1^{er} mars).



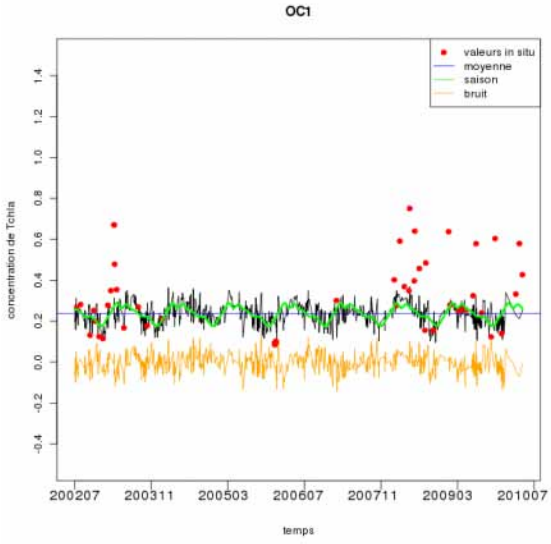
(a) Série pour la station OC1



(b) Saisonnalité extraite de la série pour OC1



(c) Bruit de la série pour la station OC1



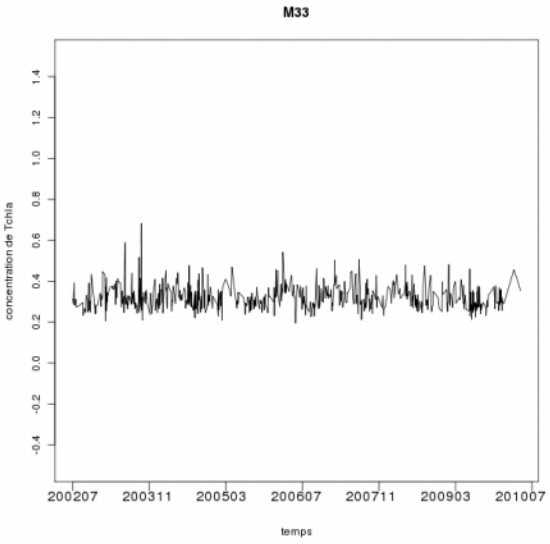
(d) Décomposition de la série pour OC1 avec les valeurs mesurées pour cette station

FIGURE 13 – Série temporelle pour la station OC1 et sa décomposition

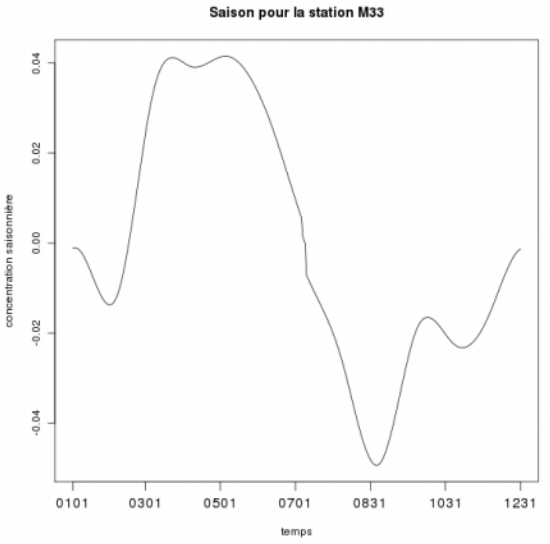
La série varie, grossièrement entre 0,1 et 0,4 $\mu g.L^{-1}$.La saisonnalité indique qu'en général, la concentration de chlorophylle est plus forte lors de la saison fraîche (les valeurs positives de la saison s'étallent de mars à août, soit de la fin de la saison chaude à la fin de la saison fraîche). On l'avait déjà remarqué avec les cartes de variation de concentration dans le temps, dans la partie précédente. Ce phénomène est assez rassurant puisque c'est conforme avec ce qu'observent généralement les scientifiques par leurs mesures. Le fait d'avoir deux extrémas locaux (vers avril et fin juin) est particulier à cette station. D'autres pixels ont ce phénomène également mais ce n'est pas une généralité (comme on pourra le voir dans l'exemple suivant). Le bruit varie entre -0,15 et 0,1 $\mu g.L^{-1}$ (ce qui veut dire que pour un certain jour, par rapport au jour type correspondant, on peut avoir des valeurs plus faibles de 0,15 $\mu g.L^{-1}$ et des valeurs plus fortes de 0,1 $\mu g.L^{-1}$), ce qui

est une différence très forte par rapport aux valeurs "théoriques" de concentration (dans le pire des cas, cela donne environ 150 % de différence). Enfin, le dernier graphe est une sorte de résumé de ce que l'on a pu tirer de la série. On a ajouté la valeur de la tendance à la saisonnalité afin de la faire "coïncider" avec la série (ce qui semble plus visuel). Les points rouges sont les valeurs mesurées pour ces jours. On les a représentés afin de pouvoir comparer les valeurs que donne l'estimation avec des valeurs censées être exactes. On voit que les valeurs faibles sont bien estimées mais les valeurs fortes (surtout vers la fin de la série) ne sont pas reconnues par notre estimation (avec parfois des valeurs mesurées valant plus du double des valeurs estimées).

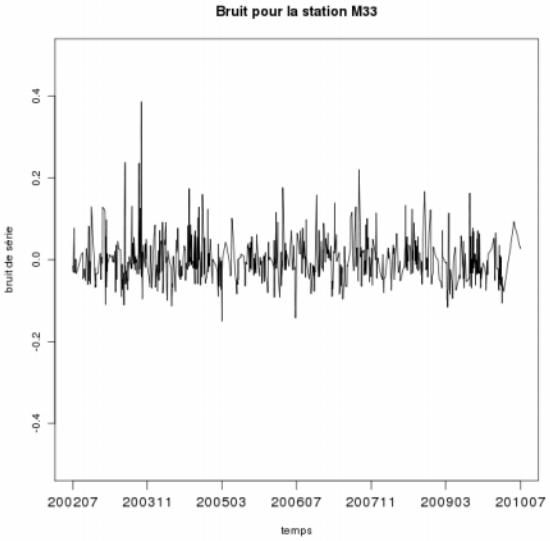
Pour les graphes suivants, on a choisi de représenter la série d'une station dans le lagon : M33. Cette station a une bathymétrie supérieure à 20m (22,9m exactement). On a décidé de ne pas présenter B03, comme dans la partie précédente, parce que le pixel correspondant à cette station a souvent été masqué par le satellite dans nos données. La série extraite et interpolée ne peut être représentative de l'évolution dans le temps de la concentration de chlorophylle pour cette station.



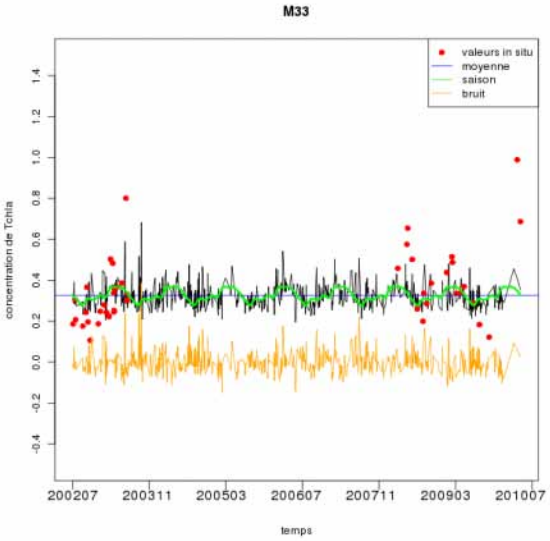
(a) Série pour la station M33



(b) Saisonnalité extraite de la série pour M33



(c) Bruit de la série pour la station M33



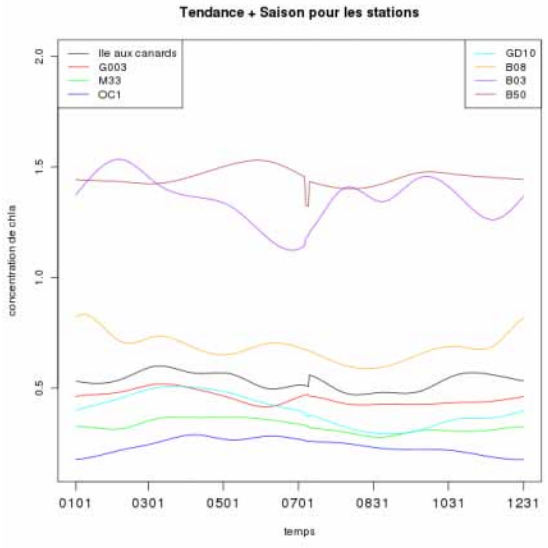
(d) Décomposition de la série pour M33 avec les valeurs mesurées pour cette station

FIGURE 14 – Série temporelle pour la station M33 et sa décomposition

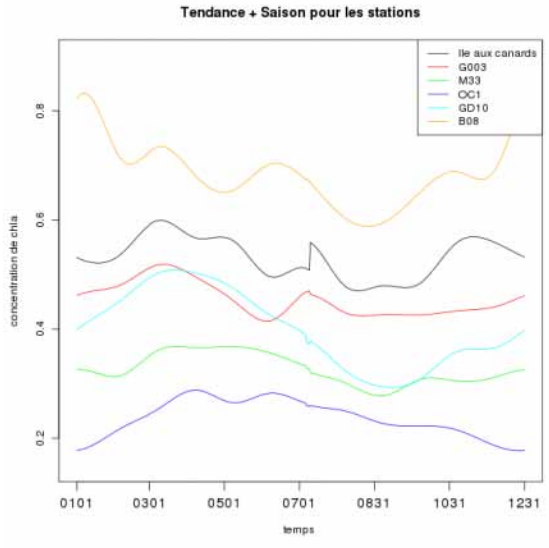
Pour cette série, la concentration varie entre 0,2 et 0,7 $\mu g.L^{-1}$ environ. Ces concentrations sont plus fortes que pour la station OC1 (ceci est conforme à ce qu'on attendait par rapport aux cartes moyennes qu'on a pu tracer précédemment). Là encore la concentration est plus forte lors de la saison fraîche, avec cependant une différence notable par rapport à la station précédente : les valeurs positives sont moins étendues mais augmentent et diminuent plus vite. L'effet des "deux extrémas" est moins marqué aussi. Il y a enfin une forte baisse jusqu'au mois de septembre, mois qui a les concentrations les plus basses, pour les valeurs "théoriques". La série du bruit est également plus perturbée, variant de -0,15 à 0,4 $\mu g.L^{-1}$. Ce sont, là aussi, des fluctuations très fortes. Enfin, sur le dernier graphe, où la tendance est représentée, on voit qu'elle est effectivement plus forte que pour la station OC1, dans l'océan. Il semble que l'on ait un compromis par rapport à nos valeurs *in*

situ entre les valeurs surestimées et les valeurs sous-estimées (encore vers la fin de la série).

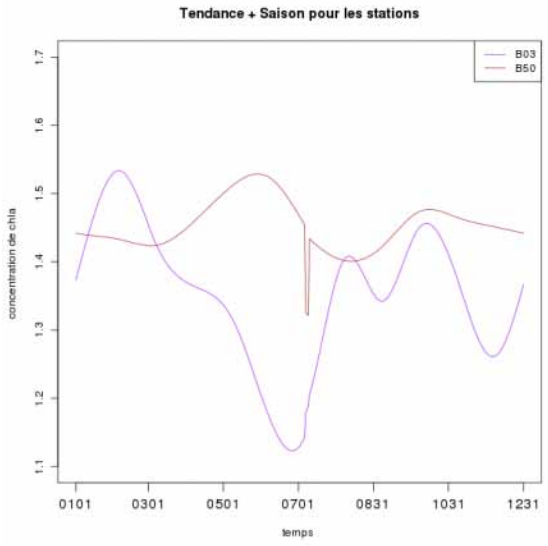
2.3.2.3 Composante saisonnière pour huit stations Afin de mettre en évidence les différences de comportement sur nos huit stations, des graphes avec la saisonnalité (centrée) et la saisonnalité ajoutée à la tendance ont été tracés :



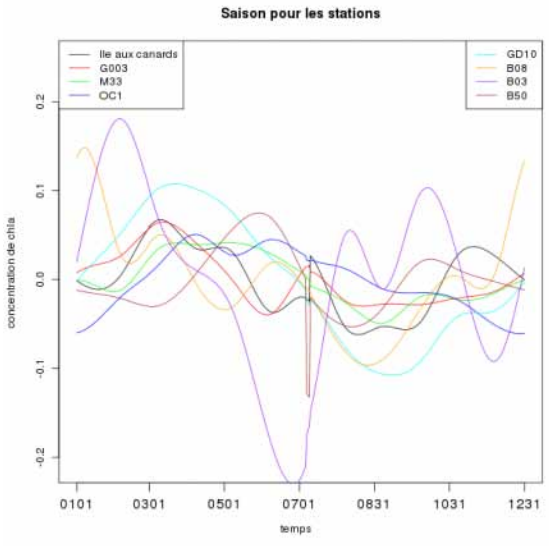
(a) Tendance + Saison



(b) Tendance + Saison pour les stations sauf B03 et B50



(c) Tendance + Saison pour les stations B03 et B50



(d) Saison (centrée) pour les stations

FIGURE 15 – Saison et Tendance + Saison pour les stations visitées régulièrement

Il faut cependant être très prudent envers les stations n'ayant que très peu d'observations dans nos données satellites (surtout pour la station B50 et peut-être un peu pour B03 et Île aux canards). En effet, sur nos huit années d'observations, on ne dispose que d'une vingtaine d'observations pour B50 ; pour B03 et Île aux canards, il y en a plus (plus

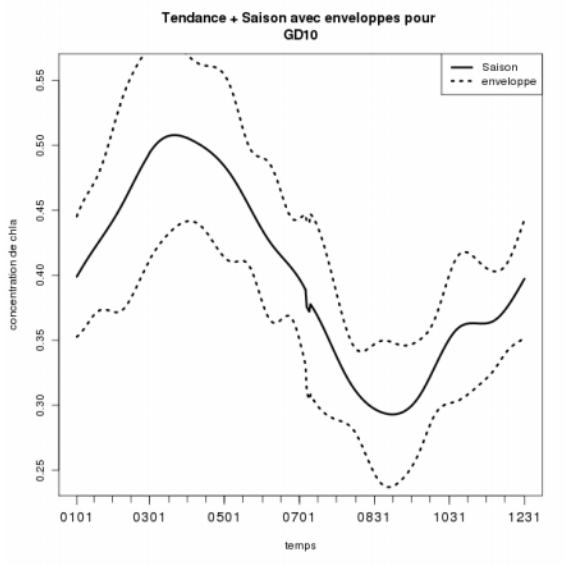
d'une centaine) mais cela reste peu par rapport aux 1444 jours visualisables [22].

Dans les baies (par exemple aux stations B03 et B50), on remarque toutefois que les valeurs de concentration sont beaucoup plus fortes que pour d'autres zones. La tendance saisonnière est un peu plus "chaotique" aussi, mais cela n'est pas étonnant vu le nombre restreint d'observations pour ces endroits. On remarque que les stations M33 et GD10 ont un comportement à peu près similaire : une concentration plus forte au début de l'année et au contraire plus basse à la fin de l'année. On retrouve des similarités entre les stations G003 et Île aux canards : des pics en mars et juillet. Enfin, la station B08 a des maximaux locaux en janvier, mars, juin et octobre, le plus important étant celui de janvier. Le minimum pour cette station est atteint en août, comme pour toutes les stations, à l'exception de G003 et OC1 qui atteignent leur minimum respectivement en juin et fin décembre - début janvier.

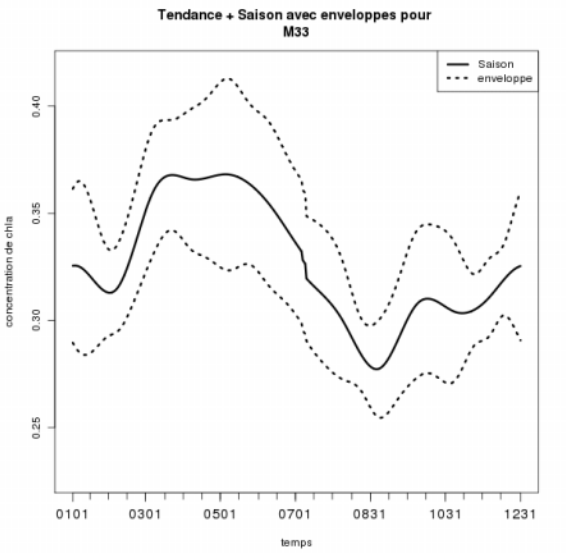
Afin de voir si l'on peut faire confiance aux formes des courbes des composantes saisonnières, on a tracé ces courbes avec l'enveloppe correspondante (qui est un indicateur de la variation pour chaque jour de l'année, la variation interannuelle). L'enveloppe prise correspond à la valeur de la courbe plus ou moins l'écart-type. Voici les graphes pour cinq de nos stations et ensuite pour trois de ces stations en particulier (GD10, M33 et OC1) :

22. On rappelle toutefois que pour la plupart des pixels, on ne dispose pas de 1444 observations (à cause de différentes difficultés liées à la télédétection) mais on dispose souvent d'environ 700 observations exploitables.

(a) Tendance + Saison



(b) Tendance + Saison pour GD10 (0,23 - 0,57)



(c) Tendance + Saison pour M33 (0,20 - 0,45)

(d) Tendance + Saison pour OC1 (0,10 - 0,35)

FIGURE 16 – Tendance + Saison avec les enveloppes pour des stations visitées régulièrement

Sur la figure 16(a), sont représentées les composantes saisonnières ajoutées aux tendances pour cinq de nos stations. La variabilité dans les baies (station B08) est très importante ; il faut donc être très prudent sur ce genre de pixels (près de la côte) pour pouvoir généraliser et affirmer qu'une "composante saisonnière type" est telle que représentée sur les graphes. Par contre, sur d'autres types de pixels, comme les pixels correspondant aux courbes 16(b) à 16(d), les enveloppes sont plus proches de la courbe de composante saisonnière, la généralisation est donc plus sûre. Pour l'océan (OC1), les variations inter-annuelles ont l'air moins fortes que dans le lagon (GD10 et M33). De plus, la composante saisonnière évolue grossièrement entre 0,30 et 0,50 (soit une amplitude de 0,20) pour GD10 et entre 0,18 et 0,28 (soit une amplitude de 0,10). Les variations intra-annuelles (saisonnières) semblent plus marquées dans le lagon que dans l'océan. On peut peut-être

rapprocher cette observation des résultats de ***S. Ouillon et al., C.R. Geoscience 337 (2005)*** ([S-OUI2]), qui faisait remarquer : "Le lagon amplifie à la côte les variations saisonnières et interannuelles de TSM [Température de Surface de la Mer] que connaissent les eaux océaniques". Il existe sûrement une liaison entre la température de surface de la mer et la concentration de chlorophylle.

On va désormais analyser les séries pour voir :

1. si des zones se différencient par rapport à la façon dont évolue la concentration de chlorophylle, évolution visible grâce au comportement de la série temporelle pour les pixels.
2. si des périodes dans le temps se différencient par la façon dont est répartie la chlorophylle dans le lagon et aux alentours.
3. des évènements particuliers.

2.3.3 ACP sur les pixels

Pour chaque pixel, on rappelle qu'on dispose d'une série décrivant la variation de concentration de chlorophylle dans le temps. Pour pouvoir classer les pixels en zones et éventuellement trouver des périodes de temps particuliers, on décide de réaliser une ACP sur les pixels. Les individus seront donc les pixels (classés dans les lignes d'un tableau) et les variables seront les jours (classés dans les colonnes du tableau). Pour faire cette ACP, on utilise la fonction PCA du package R FactoMineR. Des données manquantes (NA) sont présentes dans nos données ; cette fonction remplace les données manquantes par la moyenne de la variable, donc dans notre cas les données manquantes seront remplacées par la moyenne de la journée sur notre zone. On ne peut *a priori* pas choisir la façon de remplacer les données manquantes pour cette fonction.

Comme on a une série pour chaque pixel, notre tableau de données est un tableau à trois dimensions. Il est donc nécessaire de le transformer en un tableau à deux dimensions pour faire cette ACP. Pour cela, on transforme chaque tableau journalier (tableau des concentrations par pixel pour un jour particulier) en un vecteur, en prenant soin de retirer les pixels correspondant à des pixels terrestres (ce qui évite d'allouer de la mémoire pour des lignes de valeurs non renseignées : NA). Ces pixels terrestres sont retirés grâce à nos données de bathymétrie (les bathymétries positives correspondant à la terre).

On commence par s'intéresser aux valeurs propres pour choisir le nombre d'axes à conserver. Voici le graphe de décroissance des dix premières valeurs propres :

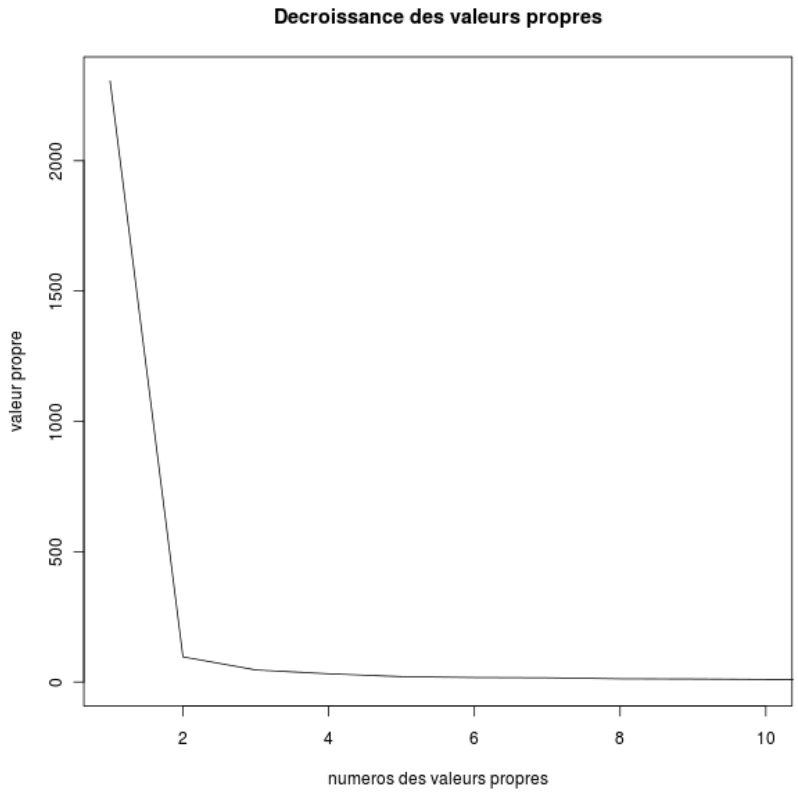
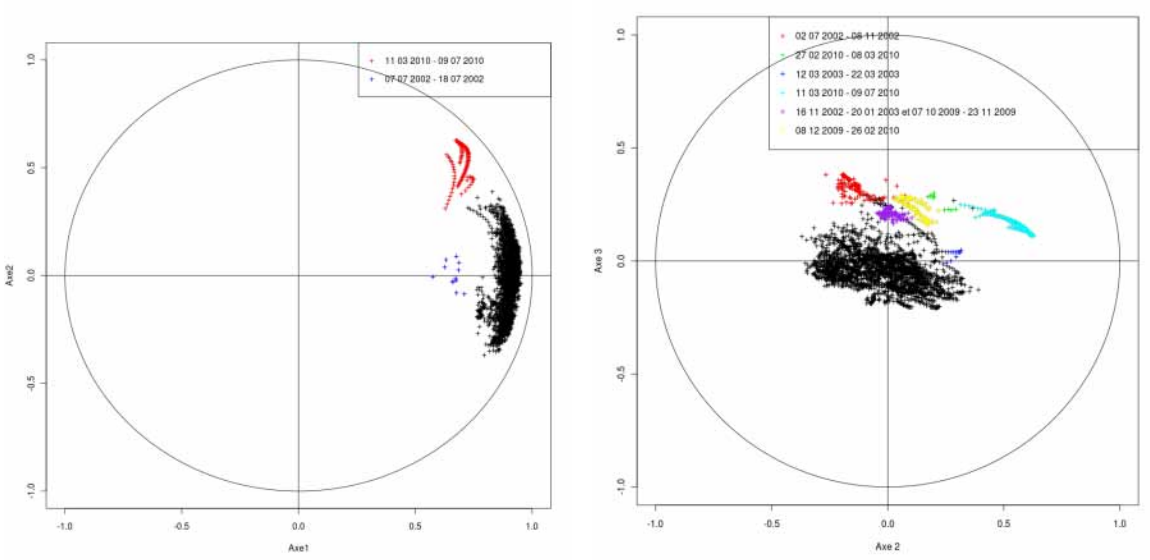


FIGURE 17 – Décroissance des valeurs propres de l'ACP sur les pixels

On voit clairement un décrochement des valeurs propres à partir de la deuxième valeur propre. On a donc envie de conserver deux axes. Les deux premiers axes représentent 82% de l'inertie totale, dont près de 79% pour l'axe 1.

On s'intéresse au cercle des corrélations dans différents plans de l'ACP. Les variables sont représentées (inhabituellement) sous forme de points (et non de vecteurs) par soucis de clarté graphique :



(a) Cercle des corrélations dans le plan (1 ;2) (b) Cercle des corrélations dans le plan (2 ;3)

FIGURE 18 – Cercles des corrélations

On voit que toutes les dates (variables) sont corrélées positivement avec le premier axe de notre ACP. On a un effet taille. Par contre, pour l'axe 2, les dates sont réparties de part et d'autre de l'axe et on croit discerner un groupe de jours plus forts que les autres. Même si l'analyse des valeurs propres suggérait de ne conserver que deux axes, on a essayé d'interpréter les trois premiers axes de l'ACP.

Certains points ont été mis en évidence parce qu'ils semblent se différencier des autres points du nuage dans le cercle des corrélations. Dans le plan (1 ;2), on distingue trois parties du nuage. La plus importante, de couleur noire, possède une forte corrélation positive avec l'axe 1 (avec des coordonnées dans cet axe supérieures à 0,6) et les points sont situés de part et d'autre de l'axe 2, comme on a pu le décrire précédemment. En rouge, ce sont les points correspondant aux jours du 11 mars 2010 au 9 juillet 2010. Ils ont une coordonnée plus importante que les autres jours dans l'axe 2. Cependant, il faut être méfiant. En effet, lors de cette période, on dispose uniquement de six jours (pour environ quatre mois), dans nos données satellites initiales ; ce qui est très peu. Comme les jours manquants ont été complétés par interpolation linéaire, il est logique que tous ces jours soient assez proches les uns des autres. Toutefois, les six jours disponibles sont bien dans cette partie du plan. Enfin, en bleu sont représentés les points correspondant aux jours du 7 juillet 2002 au 18 juillet 2002. Ils semblent être très peu corrélés avec l'axe 2 et ont des coordonnées dans l'axe 1 plutôt faibles, en comparaison avec les autres points. Là encore il faut être prudent puisque le satellite commençait son service lors ces jours ; or on sait qu'il y a toujours un certains temps avant que les capteurs soient bien calibrés.

Dans le plan (2 ;3), on retrouve (en cyan cette fois) les points correspondant aux jours du 3 mars 2010 au 9 juillet 2010. On a essayé de différencier d'autres groupes de points qui semblent se détacher de la grosse majorité des jours. Cette grosse majorité est concentrée autour du centre du cercle, ceci indique une faible corrélation de la majorité de nos données avec ce plan ; alors que les points mis en évidence ont une coordonnée positive avec l'axe 3

plus forte que les autres points. Cependant, en discutant, on n'a pas trouvé d'évènements spéciaux pour ces périodes [23]. Un autre groupe de points a également été mis en avant sans pour autant se différencier grandement des autres jours : les jours du 12 mars au 22 mars 2003. On voulait voir où se situaient dans ce plan les points correspondant aux jours d'un cyclone. En effet, c'est durant cette période que le cyclone Érica est arrivé sur Nouméa. Cependant, comme on le signalait, ils ne se démarquent pas spécialement de l'ensemble des autres jours.

2.3.3.1 Interprétation des axes Passons à l'explication de l'axe 2, l'axe 1 n'étant pas interprété à cause de l'effet taille. Par le nombre de données, l'interprétation est difficile. On a donc essayé de synthétiser au mieux les informations fournies par la procédure d'ACP. Pour cela, on a "compté" le nombre de jours dans chaque mois ayant une bonne contribution pour l'axe, puis le nombre de jours (également dans chaque mois) ayant des coefficients positifs et négatifs pour cet axe. Des tableaux ont été déduits de ces comptages. Voici un exemple pour l'axe 2 :

TABLE 3 – Tableau récapitulatif des contributions des mois pour l'axe 2 : nombre de jours du mois ayant une bonne contribution pour l'axe 2

Années — Mois	01	02	03	04	05	06	07	08	09	10	11	12
2002							0	5	14	17	2	0
2003	0	3	27	25	1	0	6	25	30	11	13	0
2004	0	0	10	0	0	0	17	30	30	31	5	0
2005	0	0	0	15	0	0	1	28	23	6	0	0
2006	0	0	7	6	3	2	8	31	30	31	10	0
2007	0	0	8	20	0	0	2	11	1	9	0	3
2008	0	0	28	30	3	0	1	12	8	13	10	0
2009	0	0	10	5	0	0	0	19	4	0	0	3
2010	0	3	30	30	31	30	9					

23. On traitera plus tard des données de précipitation sur Nouméa et peut-être qu'en mettant en relation ces données avec des dates particulières, on arrivera à détecter un lien qui nous aidera éventuellement pour l'interprétation d'axes.

TABLE 4 – Tableau récapitulatif des coordonnées des mois pour l'axe 2 : nombre de jours du mois ayant une coordonnée négative dans l'axe 2

Années — Mois	01	02	03	04	05	06	07	08	09	10	11	12
2002							20	31	30	31	17	6
2003	5	3	0	0	0	14	31	31	30	31	30	30
2004	8	5	0	23	13	15	22	31	30	31	30	26
2005	0	7	2	0	27	30	30	31	30	30	28	15
2006	1	0	0	1	0	10	31	31	30	31	30	31
2007	23	4	0	0	0	2	19	31	27	9	14	6
2008	5	0	0	0	0	20	31	31	29	31	29	15
2009	0	0	0	10	1	2	15	31	30	18	21	6
2010	0	0	0	0	0	0	0					

TABLE 5 – Tableau récapitulatif des coordonnées des mois pour l'axe 2 : nombre de jours du mois ayant une coordonnée positive dans l'axe 2

Années — Mois	01	02	03	04	05	06	07	08	09	10	11	12
2002							5	0	0	0	13	25
2003	26	25	31	30	31	16	0	0	0	0	0	1
2004	23	24	31	7	18	15	9	0	0	0	0	5
2005	31	21	29	30	4	0	1	0	0	1	2	16
2006	30	28	31	29	31	20	0	0	0	0	0	0
2007	8	24	31	30	31	28	12	0	3	22	16	25
2008	26	29	31	30	31	10	0	0	1	0	1	16
2009	31	28	31	20	30	28	16	0	0	13	9	25
2010	31	28	31	30	31	30	9					

Remarques :

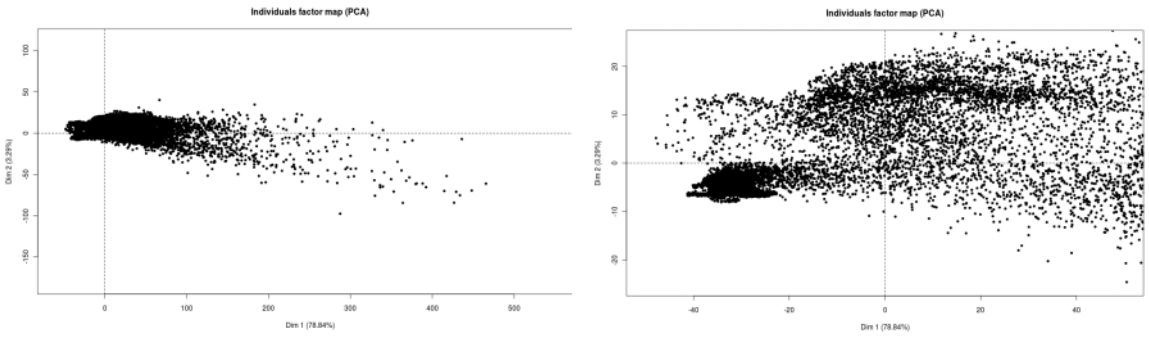
1. Pour le mois de juillet 2002, on dispose des vingt-cinq derniers jours et pour le mois de juillet 2010, on dispose des neuf premiers jours ; d'où lorsque l'on fait la somme du nombre de jours ayant des coordonnées positives avec le nombre de jours ayant des coordonnées négatives, on ne retrouve pas 31 pour ces deux mois.
2. Les cases en jaune sont pour les mois ayant tous leurs jours pris en compte dans le tableau considéré, les cases en orange pour les mois ayant plus de 20 jours pris en compte dans le tableau. Pour les tableaux 4 et 5, les cases en rouge sont pour les mois ayant plus de vingt jours ayant bien contribué à la construction de l'axe 2 et plus de vingt jours pris en compte dans le tableau, et les cases en cyan sont pour les mois ayant plus de vingt jours ayant bien contribué à la construction de l'axe 2 et ayant plus de vingt jours pris en compte dans le tableau complémentaire (tableau des jours positifs pour le tableau des jours négatifs et vice-versa), c'est-à-dire en général moins de dix jours pris en compte dans le tableau même.

À voir ces tableaux, on a envie d'interpréter l'axe 2 comme étant l'axe de la saison. En effet, on remarque que la plupart des mois de janvier à mai sont corrélés positivement avec l'axe 2 et que la plupart des mois de juillet à novembre sont corrélés négativement avec

l'axe 2. Pour les mois "intermédiaires", juin et décembre, on voit que suivant les années, on a tendance à avoir des coefficients tantôt négatifs, tantôt positifs ; et parfois ces mois sont partagés. Mais en regardant de plus près (mêmes si ces jours n'ont pas fortement contribué à la construction de l'axe), on se rend compte que souvent, les jours du début des mois de juin sont corrélés positivement avec l'axe 2 ; et lorsque des jours de juin sont corrélés négativement avec l'axe 2, ils sont en général à la fin du mois. Inversement pour les mois de décembre, lorsque des jours de décembre sont corrélés positivement avec l'axe 2, ils sont souvent à la fin du mois également. On peut expliquer ces différences d'une année à l'autre par des saisons plus ou moins tardives. Enfin, lorsque des "anomalies" (un mois du début d'année qui aurait plus de jours corrélés négativement avec l'axe 2, comme janvier 2007 par exemple) sont repérées, peu ou pas de jours ont contribué à la construction de l'axe 2, ce qui incite à ne pas y faire trop attention, quoiqu'il se peut que la saison soit tardive pour cette année en particulier. Finalement, l'axe 2 représente la saison. On déduit donc que la saison joue effectivement un rôle dans la concentration de chlorophylle, comme on avait déjà pu le remarquer plus haut.

Pour interpréter l'axe 3, un travail similaire a été réalisé, les tableaux sont présentés dans l'annexe D.1. Avec les coordonnées des jours dans l'axe, on déduit que cet axe met en opposition les mois de juillet 2002 à janvier 2003 et octobre 2009 à juillet 2010 (avec une bonne contribution et des coordonnées positives) avec les mois de juin 2004, avril à juin 2005, juin 2006, avril et mai 2007, et mars à mai 2008 (avec une bonne contribution et des coordonnées négatives). Cependant, aucun évènement particulier ou différenciation particulière connus ne caractérisent ces mois.

2.3.3.2 Coordonnées et représentation des pixels Passons maintenant aux résultats de l'ACP concernant les pixels (individus). Voici la représentation du nuage des individus dans le plan (1 ;2).



(a) Le plan (1 ;2) des individus (b) Zoom sur le plan (1 ;2) des individus

FIGURE 19 – Plan (1 ;2) des individus de l'ACP sur les pixels

Les individus sont très dispersés sur l'axe 1 avec des coordonnées pouvant atteindre près de 500. Une forte concentration de pixels est visible dans le rectangle ayant pour sommet haut gauche le point (-50 ;40) et pour sommet bas droit (100 ;-50). On a donc fait un zoom dans ce rectangle pour voir si des différences (non visibles sur la figure 19(a)) pouvaient se voir, d'où la figure 19(b). Et on voit effectivement deux amas de pixels. Le premier est dans les coordonnées négatives de l'axe 2 et a des coordonnées grossièrement

comprises entre -40 et -20 pour l'axe 1. Le deuxième groupe est plus étendu. Les points le composant ont des coordonnées dans l'axe 1 comprises entre -20 et 30, et dans l'axe 2 entre 10 et 20. On s'intéresse donc à savoir quels peuvent être ces points. Pour cela, on représente ces deux groupes sur une carte, une fois que l'emplacement des pixels a été extrait.

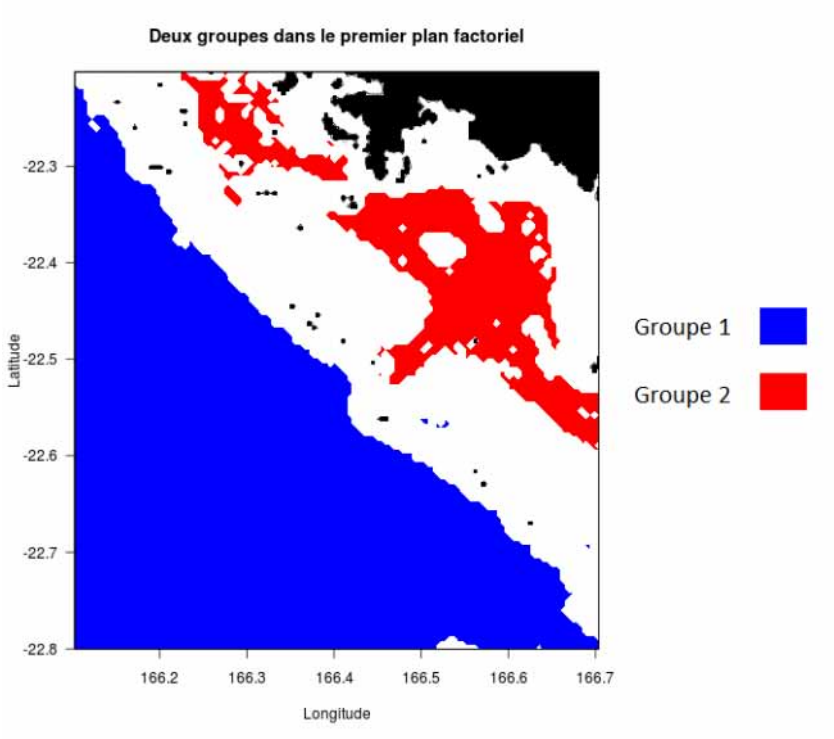
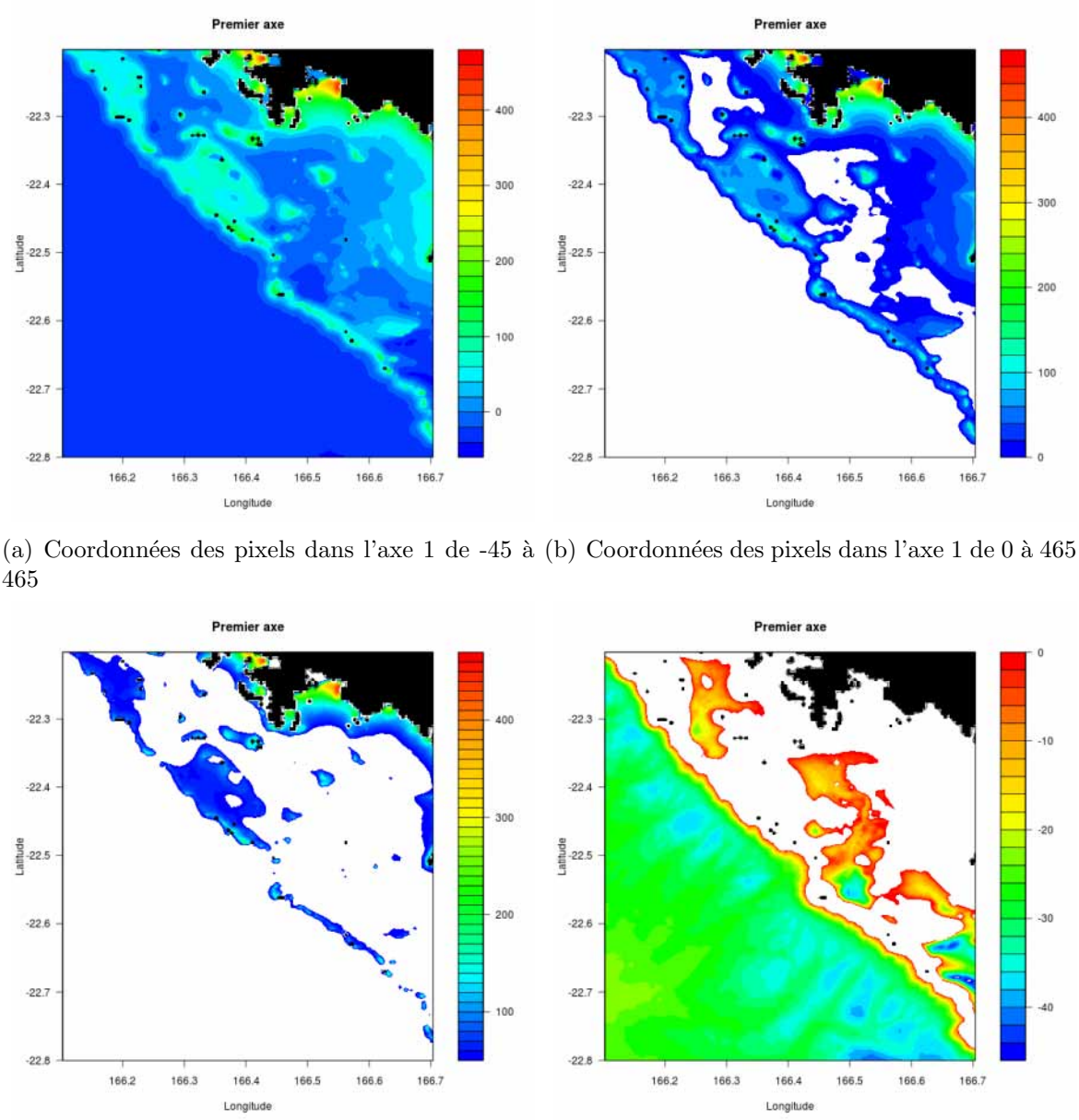


FIGURE 20 – Deux groupes dans le plan (1 ;2) des individus

Les deux groupes sont bien dans des zones distinctes. Le groupe 1 (en bleu) est constitué des pixels situés dans l'océan et le groupe 2 (en rouge) est constitué des pixels du lagon ayant une forte bathymétrie[24]. Pour s'en convaincre, on peut se reporter à la carte 3. On voit clairement que la forme de notre zone rouge est la même que la forme de la zone du lagon avec une bathymétrie plus forte.

Une idée semblable est de représenter une carte avec les pixels en fonction de leurs coordonnées dans un axe. Par exemple, voici cette carte pour l'axe 1. Les cartes présentées sont en fait une même carte avec une échelle de valeurs différente afin d'avoir une vue d'ensemble d'une part, et d'autre part de bien mettre en évidence les zones avec des coordonnées positives , négatives et de fortes coordonnées positives (il n'y a pas de pixels se démarquant particulièrement par une coordonnée négative).

24. Sur cette carte 20, les zones blanches représentent les pixels n'appartenant à aucun de ces deux groupes.



(a) Coordonnées des pixels dans l'axe 1 de -45 à 465

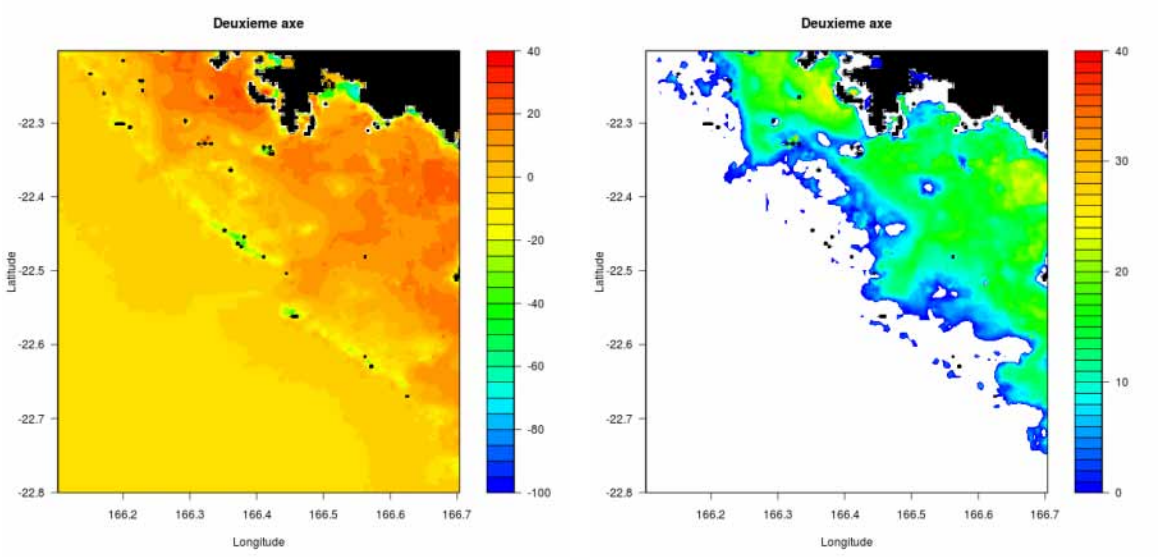
(b) Coordonnées des pixels dans l'axe 1 de 0 à 465

(c) Coordonnées des pixels dans l'axe 1 de 50 à 465

(d) Coordonnées des pixels dans l'axe 1 de -45 à 0

FIGURE 21 – Coordonnées des pixels dans l'axe 1

Grâce à ces cartes, il est visible que les pixels ayant une forte coordonnée dans l'axe 1 (au-delà de 50) sont des pixels avec une faible bathymétrie : proches de la côte, d'îlots ou du récif. Les pixels ayant une coordonnée négative sont des pixels avec une bathymétrie plus forte, et ceux qui en ont une sous -20, sont souvent dans l'océan à quelques exceptions près (exceptions situées dans le lagon avec des bathymétries souvent supérieures à 30 m). Le premier axe de notre ACP est donc corrélé positivement avec les pixels ayant une faible bathymétrie et négativement avec les pixels ayant une forte bathymétrie. On a pu faire des représentations similaires pour l'axe 2.



(a) Coordonnées des pixels dans l'axe 2 de -97 à 40

(b) Coordonnées des pixels dans l'axe 2 de 0 à 40

(c) Coordonnées des pixels dans l'axe 2 de -97 à 0

(d) Coordonnées des pixels dans l'axe 2 de -40 à 0

FIGURE 22 – Coordonnées des pixels dans l'axe 2

Sur cet axe, les coordonnées sont moins fortes que pour l'axe 1 : elles ne dépassent pas, en valeur absolue, 97. La plupart des pixels ayant une coordonnée positive dans cet axe sont des pixels du lagon ; et inversement, la plupart des pixels ayant une coordonnée négative sont des pixels de l'océan. Cependant, il existe des pixels du lagon (voir par exemple aux alentours du point de coordonnées longitude-latitude (166,55 ;-22,25)) ayant une coordonnée dans cet axe très fortement négative. Il est donc difficile de déduire que ces zones sont semblables. Toutefois, en se souvenant de l'interprétation faite de l'axe 2 (corrélé positivement avec les mois du début de l'année et négativement avec les mois de fin d'année), on peut penser que cet axe indique la façon dont la série des pixels évolue dans l'année : pour les pixels étant corrélés positivement avec l'axe, la concentration de chlorophylle est plus forte au début de l'année et plus faible en fin d'année ; et pour les

autres pixels, la variation est différente.

Remarque : Les cartes des contributions des pixels dans la construction des axes ont été mises dans l'annexe D.2.

2.3.3.3 Vérification des interprétations des axes On a voulu enfin faire une vérification des interprétations pour les axes. Pour voir si effectivement la bathymétrie est une variable expliquant le premier axe de notre ACP, on a simplement calculé le coefficient de corrélation de Spearman entre les coordonnées des individus dans le premier axe et les bathymétries de ces individus. Ce coefficient est de 0,70. Une forte liaison existe donc entre le premier axe d'ACP et la bathymétrie.

En ce qui concerne le deuxième axe, on annonçait que les pixels étant fortement corrélés positivement avec cet axe, avaient une composante saisonnière plus forte en début d'année et moins forte en fin d'année. Voici le graphe de la composante saisonnière des cinq pixels ayant les coordonnées les plus fortes dans l'axe 2 :

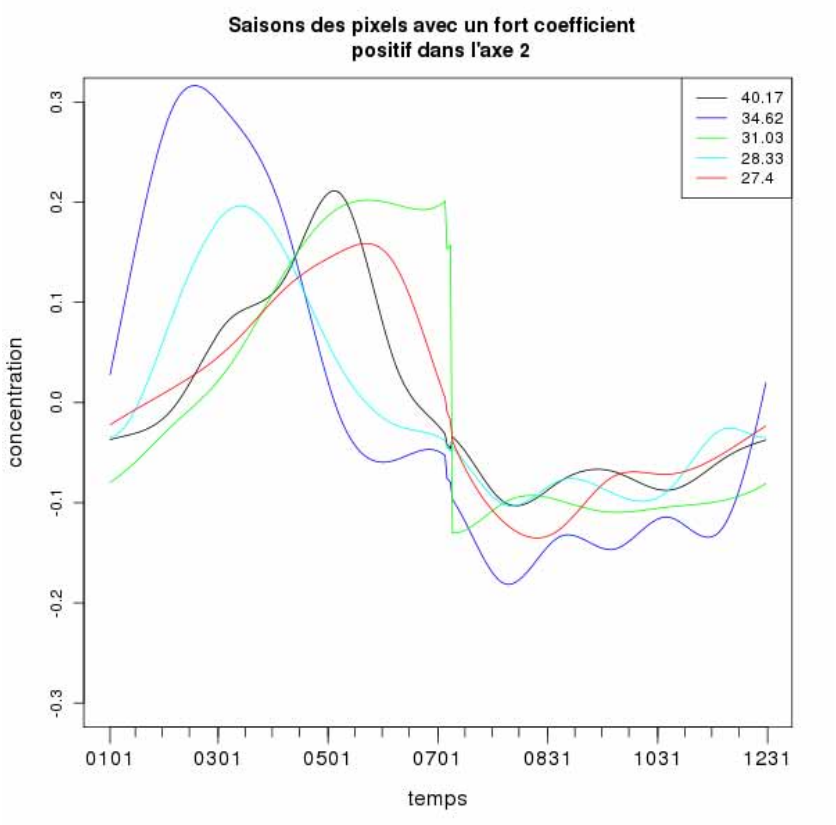
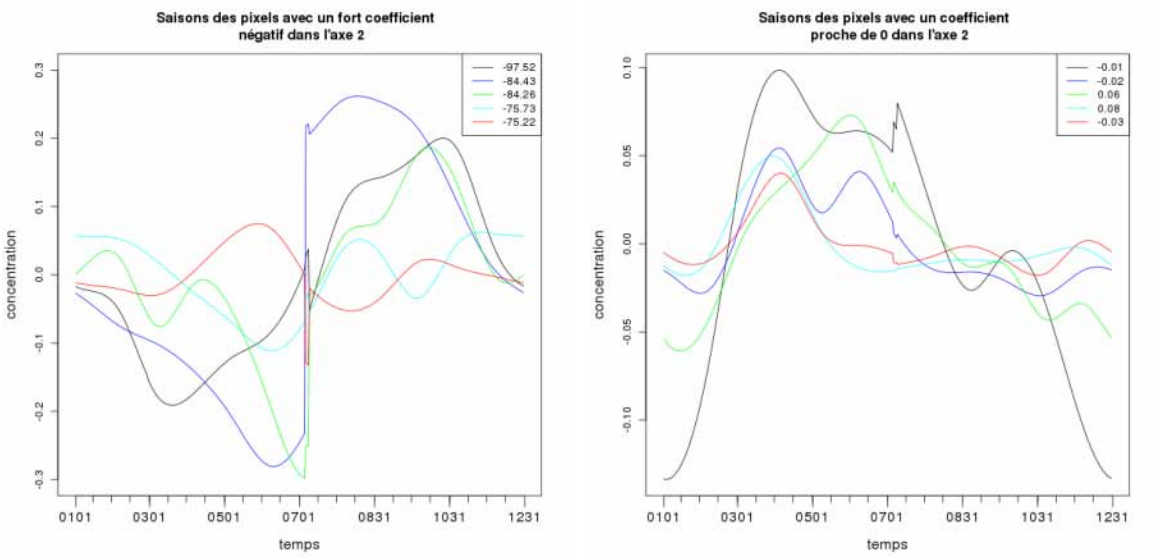


FIGURE 23 – Composantes saisonnières des cinq pixels ayant les coordonnées les plus fortes dans l'axe 2

Dans la légende sont indiquées les coordonnées des pixels dans l'axe 2. Effectivement, ces courbes montrent bien la tendance qu'on décrivait plus haut. Les mêmes représentations ont été faites pour des pixels ayant une forte corrélation négative et avec une corrélation presque nulle avec le deuxième axe :



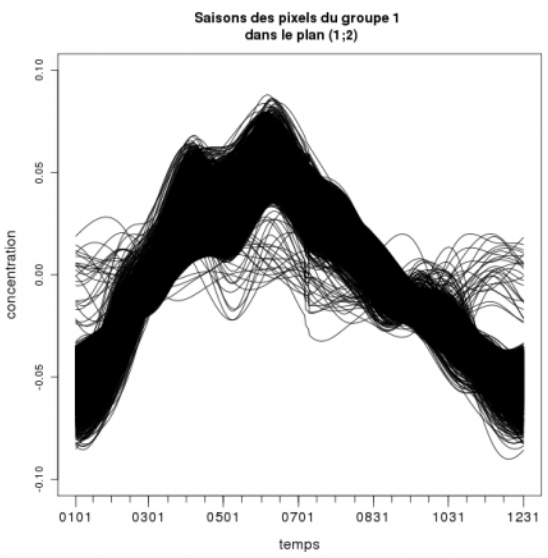
(a) Composantes saisonnières des cinq pixels ayant les coordonnées les plus faibles dans l'axe 2

(b) Composantes saisonnières des cinq pixels ayant des coordonnées proches de 0 dans l'axe 2

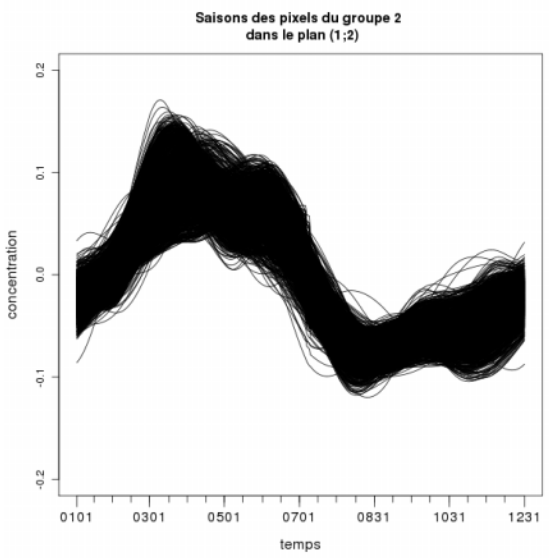
FIGURE 24 – Composantes saisonnières pour des pixels avec des coordonnées négatives ou proches de 0 dans l'axe 2

La forme des courbes pour ces composantes saisonnières sont bien différentes de celles de la figure 23. On remarque même sur la figure 24(a) que certaines de ces courbes ont une forme à l'opposé des courbes visibles pour les pixels avec une coordonnée fortement positive dans l'axe 2. Cependant, il faut une fois de plus être prudent quant à la forme de ces courbes. En effet, comme le montrent les cartes de la figure 22, les points avec une coordonnée très négative sont proches des côtes généralement. Cela ne veut pas dire que l'interprétation de l'axe 2 est fausse, juste que la composante saisonnière pour ces pixels particuliers est peut-être différente, ou tout simplement que sa forme n'est pas généralisable à toutes les années (comme signalé dans la partie 2.3.2.3 avec les figures 16).

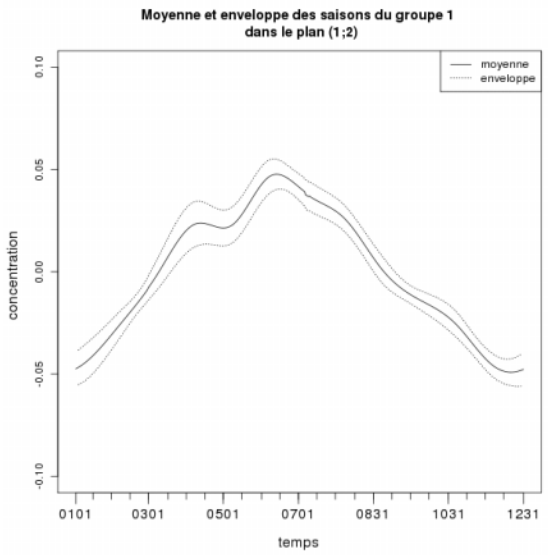
Enfin, toujours en s'intéressant aux composantes saisonnières, on les représente pour l'ensemble des pixels des groupes mis en évidence dans le 2.3.3.2 sur la figure 20 ; puis la moyenne et l'enveloppe (toujours déterminée à partir de l'écart-type) de la composante saisonnière pour tous ces pixels sont calculées par groupe et également représentées.



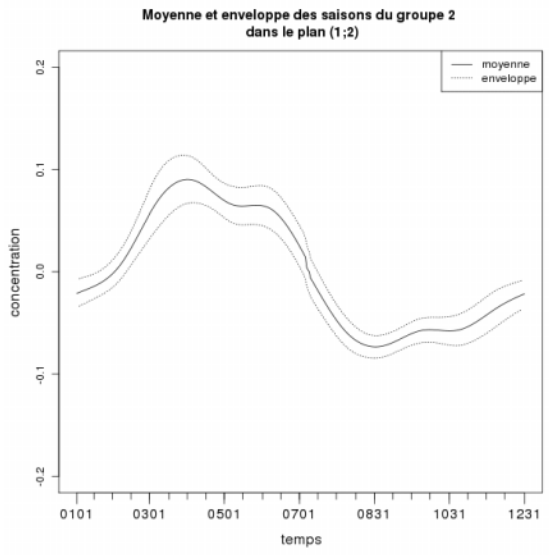
(a) Composantes saisonnières des pixels du groupe 1 du plan (1 ;2) (-0,10 - 0,10)



(b) Composantes saisonnières des pixels du groupe 2 du plan (1 ;2) (-0,2 - 0,2)



(c) Moyenne et enveloppes des composantes saisonnières des pixels du groupe 1 du plan (1 ;2) (-0,1 - 0,1)



(d) Moyenne et enveloppes des composantes saisonnières des pixels du groupe 2 du plan (1 ;2) (-0,2 - 0,2)

FIGURE 25 – Composantes saisonnières pour des pixels des groupes 1 et 2 du plan (1 ;2) de l'ACP

Les composantes saisonnières pour ces pixels sont très semblables les unes par rapport aux autres dans le même groupe, sauf dans le premier groupe où l'on remarque quelques courbes à part des autres. Comme les enveloppes sont proches des courbes, ces dernières ont un sens. On peut une fois de plus mettre l'accent sur la différence d'amplitude entre ces courbes. La courbe correspondant au deuxième groupe (à l'intérieur du lagon, voir figure 20) a une variation saisonnière beaucoup plus forte que la courbe correspondant au premier groupe (dans l'océan). Et on constate la présence d'une "translation" pour la variation de ces courbes ; l'une (premier groupe) atteignant son maximum vers fin juin-début juillet et son minimum en décembre, et l'autre (deuxième groupe) atteignant son

maximum plutôt en avril et son minimum en août. Cela corrobore le premier résultat : le lagon a un cycle plus précoce que le large.

2.3.3.4 Perspectives suite à l'ACP Pour finir, on a vu que peu de pixels (en comparaison avec le nombre total) avaient très fortement contribué à la construction de l'axe 1, on pourrait donc refaire la même ACP en mettant les individus ayant (trop) fortement contribué à la construction de l'axe 1 en individus supplémentaires. On pourrait ainsi éventuellement avoir un deuxième, voire un troisième axe expliquant une part d'inertie plus grande. L'interprétation serait alors sûrement plus aisée et plus sûre. Ensuite, on avait remarqué deux groupes se différenciant dans le plan (1 ;2) des individus et on les a représentés en sélectionnant les points dans deux rectangles contenant chacun les points formant le groupe. Pour être plus rigoureux dans la sélection des individus constituant les groupes, on pourrait faire appel à un algorithme de classification ; si dans les résultats de l'ACP refaite avec les individus supplémentaires, on retrouve ce cas de figure.

Lorsque l'on aura accru l'étendue de la zone d'étude (en prenant toute la zone sud du lagon) et que l'on aura fait une ACP sur l'ensemble des pixels de cette zone plus étendue, on verra peut-être des différenciations de zones plus intéressantes en comparaison de celles qu'on a déjà pu mettre en évidence jusqu'à présent. Dans la suite, on va s'intéresser aux évènements pluvieux de Nouméa, pour les mettre en relation avec nos séries de concentration de chlorophylle sur des pixels plus ou moins proches des stations d'observation.

2.3.4 Évènements pluvieux à Nouméa

On dispose des précipitations journalières sur Nouméa du 1^{er} janvier 1965 au 31 décembre 2010. On fait donc une sélection des données se rapportant aux temps disponibles pour les concentrations de chlorophylle (du 7 juillet 2002 au 9 juillet 2010).

On vérifie tout d'abord la présence d'une saison des pluies grâce à une transformée de FOURIER. La procédure utilisée est appelée "Fast Fourier Transform" (FFT). La magnitude est le module de la transformée de FOURIER de la série. Voici le graphe des magnitudes obtenu :

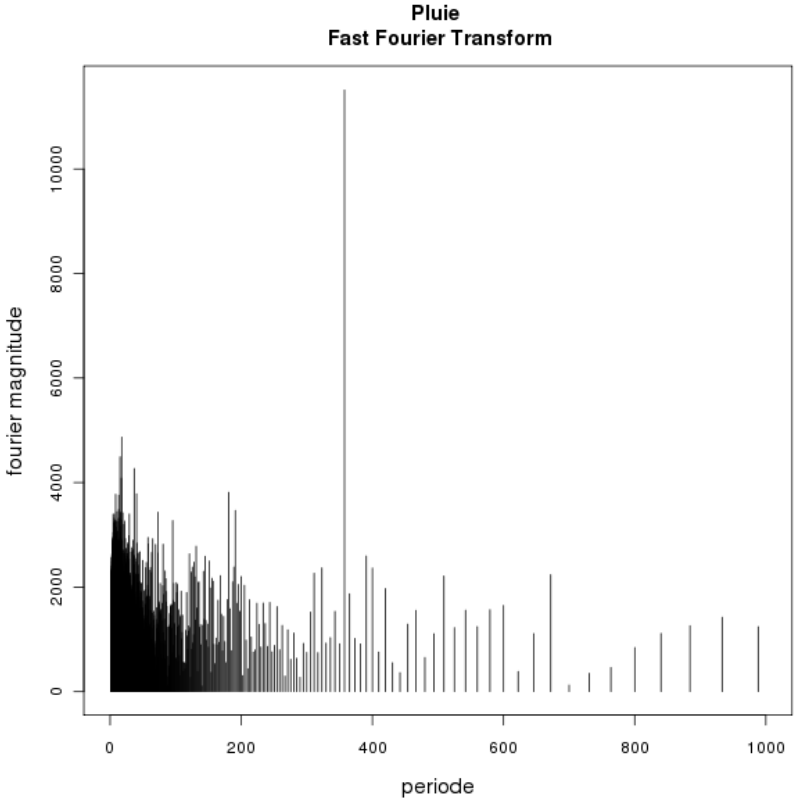


FIGURE 26 – Magnitude de la série des pluies pour chaque période

Une période se dégage nettement des autres. Cette période correspond à 357 jours, soit environ un an. On déduit donc une saison d'un an aussi bien pour les pluies que pour la concentration de chlorophylle [25].

2.3.4.1 Décalage entre la série de précipitation et des séries de concentration de chlorophylle Pour voir une éventuelle liaison entre les évènements pluvieux et la concentration de chlorophylle, le coefficient de corrélation de Spearman a été calculé entre les deux séries pour nos stations. Seulement, une intuition nous fait penser que l'effet de la pluie d'une journée n'est certainement pas instantané et qu'il faut attendre quelques jours avant de voir l'impact de cet évènement sur la concentration. On ne s'est donc pas contenté de calculer le coefficient de corrélation uniquement pour les deux séries, mais aussi le coefficient de corrélation entre la série de concentration de chlorophylle et la série des précipitations translatée. On a choisi des translations entre -14 jours et 5 jours, c'est-à-dire qu'on a calculé la corrélation entre la série de concentration de chla et la série des précipitations pour n jours avant lorsque la série est translatée de -n jours, et pour n jours plus tard lorsque la série est translatée de n jours [26]. Normalement, les précipitations de n jours à venir n'impactent pas la concentration de chlorophylle du jour même, mais ces calculs sont réalisés pour vérifier la validité du résultat sur les stations.

25. La saison pour la concentration de chlorophylle a déjà était déterminée mais cette méthode par FFT a été malgré tout appliquée avec un résultat similaire à celui des précipitations.

26. Lorsque n vaut 0, le coefficient calculé est celui évoqué au début de ce paragraphe.

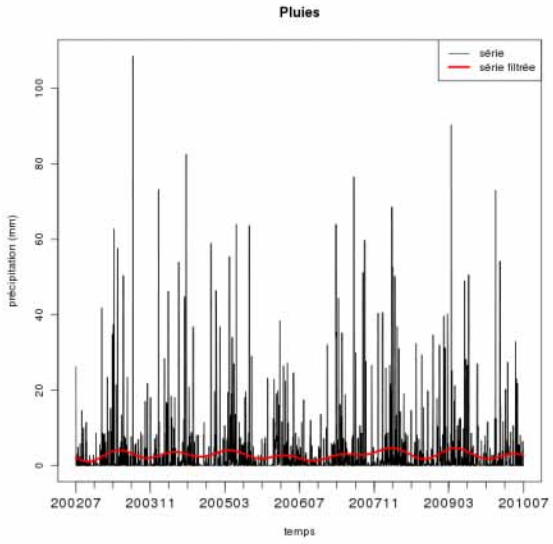
Une fois les coefficients de corrélation calculés, on a sélectionné les jours ayant le coefficient le plus fort pour la station en particulier. On n'a pas sélectionné de point trop proche de la côte. Par contre, on a choisi un pixel de l'océan lointain pour lequel l'impact de la pluie sur la concentration de chlorophylle devrait être très décalé et très faible, voire inexistant. Ceci permet également l'observation d'un signal océanique pur. Voici le tableau résumant les résultats :

TABLE 6 – Coefficients de corrélation des décalages entre séries de précipitations et de concentration de chlorophylle

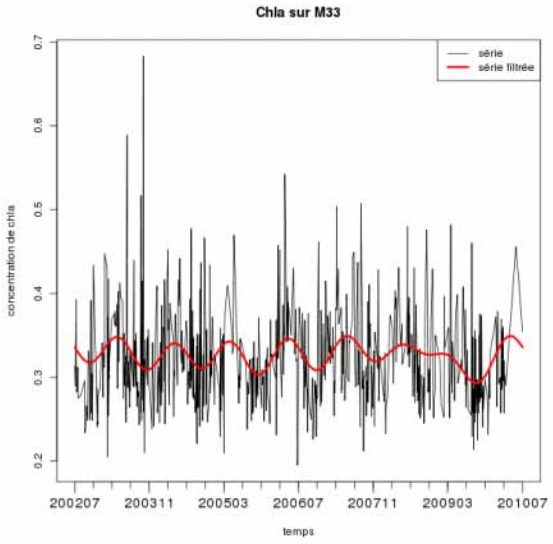
Stations	Jours les plus corrélés	Coefficients de corrélation	p-valeurs
B03	5	0,03	0,18
B08	-4	0,05	0,01
G003	-2	0,08	$8,6.10^{-6}$
GD10	-4	0,23	$5,7.10^{-35}$
Île aux canards	-2	0,09	$5,1.10^{-6}$
M33	4	0,13	$2,2.10^{-12}$
OC1	-4	0,15	$1,4.10^{-16}$
Océan lointain	-11	0,12	$1,9.10^{-10}$

Pour la station B03, toutes les p-valeurs étaient supérieures à 5%, ce qui veut dire qu'aucune liaison entre les séries de précipitations décalées et la série de concentration de chlorophylle pour cette station n'est significative. On a néanmoins indiqué le nombre de jours de décalage pour lequel le coefficient de corrélation est le plus fort. Concernant les autres stations, l'hypothèse de nullité du coefficient de corrélation est rejeté. Le coefficient le plus fort est pour la station GD10 et une translation de série de -4 jours, avec une valeur de 0,23. On remarque une bizarrerie avec M33. Sur cette station, la corrélation est plus forte avec la série décalée de 4 jours. Ceci n'a sûrement pas de sens... Signalons tout de même qu'à cet endroit, le renouvellement de l'eau est très fréquent, à cause d'un courant important (ce pixel est donc peut-être à mettre à part). Par contre, on remarque que la liaison avec les évènements pluvieux est plus décalée pour l'océan que pour le lagon, ce qui semble logique et qui peut s'expliquer par l'éloignement à la côte.

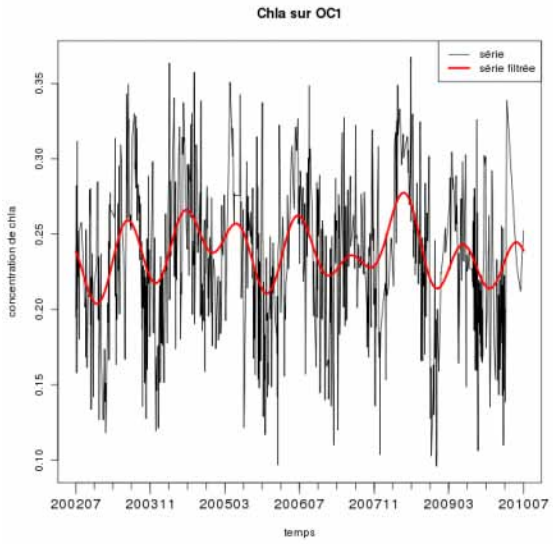
2.3.4.2 Séries filtrées et résidus La procédure de "Fast Fourier Transform" permet également d'établir un filtre sur des séries, par une transformée de FOURIER inverse. On établit un filtre passe bas sur chacune des séries (aussi bien pour les pluies que pour la chlorophylle, afin d'avoir le même traitement pour chacune des séries), ce qui permet de reconstruire les séries sans les hautes fréquences. Les séries sont alors lissées.



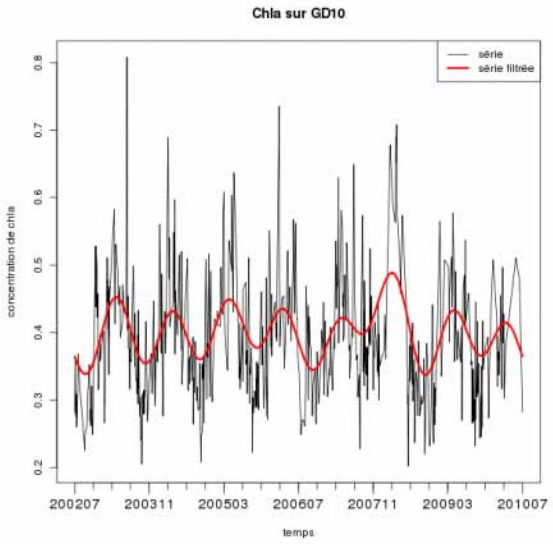
(a) Série et série filtrée des précipitations



(b) Série et série filtrée de la concentration de chlorophylle sur M33



(c) Série et série filtrée de la concentration de chlorophylle sur OC1



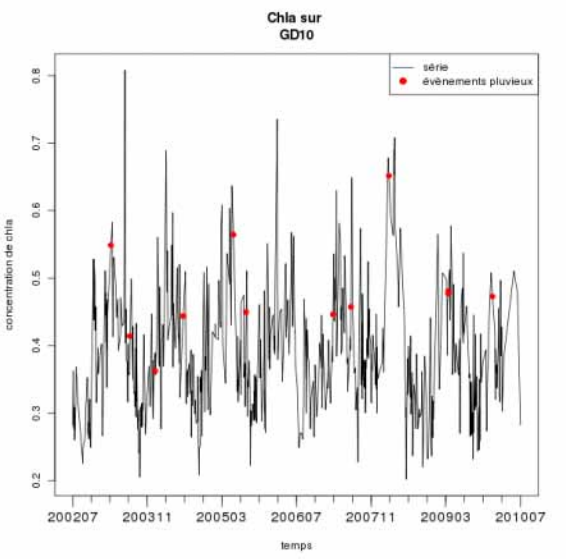
(d) Série et série filtrée de la concentration de chlorophylle sur GD10

FIGURE 27 – Séries et séries filtrées par un filtre passe bas

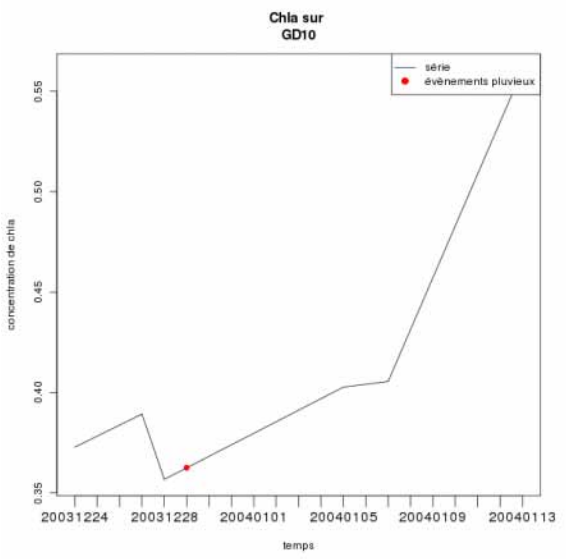
On a voulu appliquer un travail similaire à celui décrit dans le 2.3.4.1 sur les séries filtrées et surtout sur les séries résiduelles (différences entre les séries originelles et les séries filtrées). Le traitement sur les séries filtrées aurait servi essentiellement à déterminer quel est le décalage entre les saisons de pluie et les saisons de la concentration de chlorophylle ; et le traitement sur les séries résiduelles, essentiellement en vue de mettre en évidence que des anomalies d'évènements pluvieux sont liées à des anomalies dans la concentration de chlorophylle. Cependant, les résultats obtenus ne sont pas simples à interpréter (bien souvent, en ce qui concerne les séries résiduelles les liaisons ne sont pas significatives) ; on décide donc d'aborder le problème d'une autre manière. On s'intéresse aux évènements particulièrement pluvieux (précipitations en une journée supérieures à 60 mm), puis gra-

phiquement on visualisera la réaction de la chlorophylle.

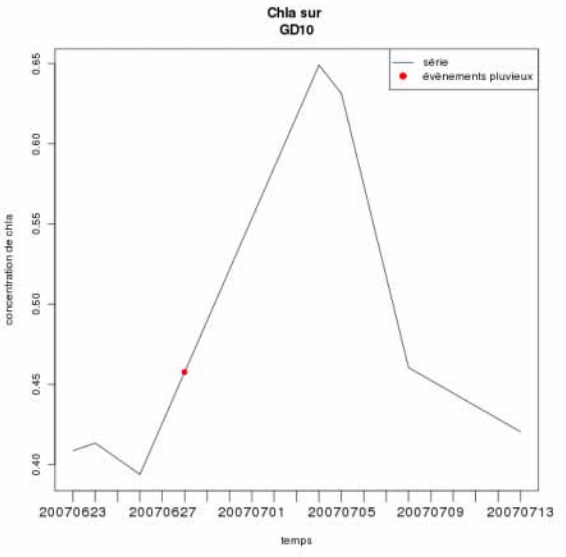
2.3.4.3 Évènements pluvieux très importants Les observations sont d'abord faites sur les séries originelles, c'est-à-dire qu'on s'intéresse aux évènements de pluies importants. Voici des exemples pour la station GD10 :



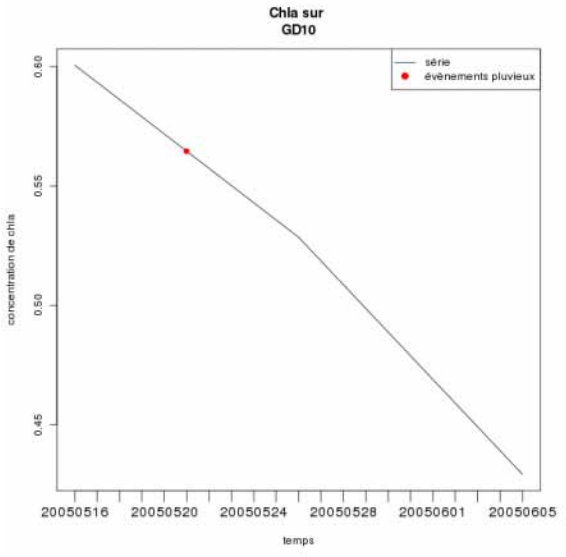
(a) Concentration de chlorophylle sur GD10



(b) Zoom sur le 29 décembre 2003 à GD10



(c) Zoom sur le 28 juin 2007 à GD10



(d) Zoom sur le 21 mai 2005 à GD10

FIGURE 28 – Série de chlorophylle sur GD10 avec les évènements pluvieux importants

Les points rouges représentent les jours où les évènements (précipitations supérieures à 60 mm dans la journée) se sont produits. Les trois derniers graphes ont pour but de mettre en évidence la réaction à l'évènement dans la série. N'oublions pas que nous avons affaire à une série interpollée linéairement. Il est simple de voir où se situent les données réelles et les données interpollées par les "décrochements" de la courbe. Regardons le graphe

qui correspond au 29 décembre 2003. Le jour le plus proche pour lequel on dispose de la donnée est la veille. Mais le jour le plus proche suivant l'évènement est le 5 janvier 2004, soit une semaine plus tard. On ne sait donc pas comment évolue la concentration de chlorophylle durant cette semaine. On peut juste souligner que la concentration est plus forte le 5 janvier 2004 que la veille de l'évènement. Les graphes ne sont pas présentés ici mais sur les stations OC1, Île aux canards, M33 et B03, le point indiquant l'évènement se situe aussi sur un segment de droite de coefficient directeur positif, c'est-à-dire que la concentration de chlorophylle suite à l'évènement est plus forte qu'avant. On peut aussi réfléchir à la raison pour laquelle on ne dispose pas de données durant ces jours (une semaine, voire plus). Les pixels sont masqués sans doute à cause de la présence de nuages, il n'est donc pas improbable que la pluie se soit perpétuée durant cette période. Il se peut donc que l'accroissement de chlorophylle sur ces stations soit dû non seulement à l'évènement mais aussi aux pluies éventuelles (moins fortes) qui l'ont suivi. En ce qui concerne le 28 juin 2007, on fait grosso modo les mêmes observations : la concentration suite à l'évènement est plus forte qu'avant l'évènement et on ne dispose pas de données pendant plus d'une semaine. Enfin, le graphe correspondant à l'évènement du 21 mai 2005 sert en quelques sortes de contre-exemple. La concentration est plus faible suite à l'évènement, mais on voit aussi qu'on n'a pas de données pendant plus de 10 jours consécutifs.

Jusqu'à présent, on s'intéressait à la relation possible entre une forte pluie et une forte chlorophylle. Mais on peut aussi s'intéresser aux anomalies, puisque chacune de ces séries a une tendance saisonnière. On s'est penché sur ce problème en étudiant les séries résiduelles ; cependant les résultats sont similaires à ceux présentés dans le paragraphe précédent, ils ne sont donc pas présentés.

2.3.4.4 Perspectives concernant les évènements pluvieux Dans cette partie, ce sont de forts évènements ponctuels qui nous intéressaient et il est difficile de conclure à une forte liaison entre ces fortes précipitations et des pics de concentration de chlorophylle puisque la plupart du temps, les données autour de ces évènements ne sont pas disponibles (sans doute à cause de nuages). Toutefois, sur les stations considérées, les concentrations sont généralement plus fortes suite à l'évènement important (dans une proportion de 60% environ). Dans la suite, il serait intéressant de traiter la série de pluie de la même façon que les séries de concentration de chlorophylle dans la partie 2.3.2. On disposerait ainsi de la composante saisonnière pour une année type. Une confrontation entre les séries des bruits serait alors envisageable. Des classes d'anomalies (par exemple, quatre classes de précipitations de très faibles à très fortes et des classes semblables pour la concentration de chlorophylle avec des seuils à déterminer) pourraient être constituées, puis la relation entre ces classes pourrait se faire par comparaison des dates présentes dans chaque classe. Une autre manière de faire les choses serait de regarder, à partir des classes de précipitations, la "réaction" de la chlorophylle sur les périodes correspondant aux classes plus ou moins décalées.

3 Extension de la zone d'étude : le sud du lagon de Nouvelle-Calédonie

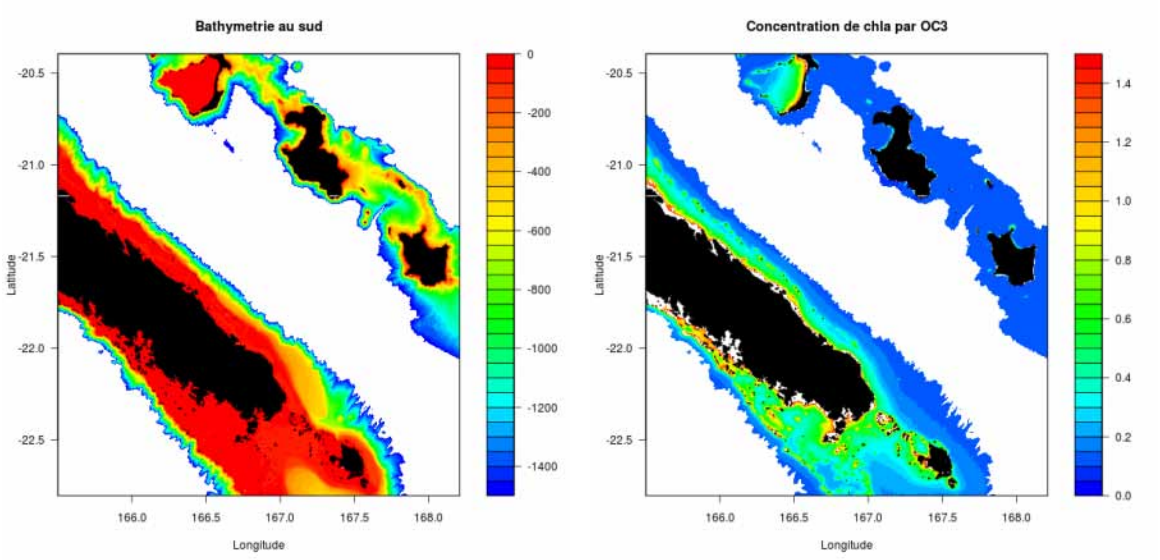
Jusqu'à présent on a travaillé sur la zone du lagon autour de Nouméa afin d'avoir un jeu de données plus restreint, permettant d'avoir plus de rapidité dans la récupération des données, le traitement de ces données et l'exécution des méthodes pour les analyser. Dans la suite du stage, on traitera les données pour l'ensemble de la zone sud du lagon de Nouvelle-Calédonie. On se servira donc du travail déjà réalisé sur la "petite zone" autour de Nouméa en l'adaptant quelque peu. On travaillera également sur une machine à distance plus puissante, avec plus de mémoire vive, qui pourra stocker toutes les données pour les traiter. Cependant, les dimensions des donnés sont encore trop importantes (malgré une machine plus puissante) pour pouvoir être exploitables toutes à la fois. Une réduction des données est nécessaire.

3.1 Réduction du nombre de données

Afin de s'affranchir de ce problème, une piste est envisageable : retirer, une fois pour toutes, les pixels qui ne nous concernent pas et les moins intéressants. On peut enlever les pixels terrestres masqués par le satellite et les pixels de l'océan assez éloignés des bords du lagon. Pour cela, la forme des données est modifiée. On n'utilisera plus des tableaux empilés les uns sur les autres (matrices à trois dimensions), mais des tableaux (à deux dimensions) dans lesquels se trouveront en lignes les données d'un même pixel pour tous les temps disponibles et en colonnes les données des pixels gardés (pixels du lagon et océaniques proches du lagon) pour un jour. On transforme en fait le tableau de données pour un jour en vecteur pour lequel on retire les pixels correspondant aux pixels non intéressants (pixels terrestres ou océaniques éloignés du lagon).

On s'aide du tableau de bathymétrie pour faire cette transformation de nos données. Les pixels ayant une bathymétrie positive (pixels terrestres) et les pixels ayant une bathymétrie sous une certaine profondeur, fixée à -1500 m, sont retirés. Ensuite, pour avoir une représentation d'un certain résultat sous forme de carte, on remodifiera le vecteur correspondant à ce que l'on veut montrer, pour le mettre sous forme de tableau (à deux dimensions) en s'aidant toujours du tableau de bathymétrie et en inscrivant dans les pixels retirés "donnée manquante" (NA sous R). Les fonctions permettant ces transitions sont inscrites dans l'annexe E.4.

Voici un exemple de cartes résultats fournies par cette méthode avec la bathymétrie et la moyenne de concentration de chlorophylle calculée par OC3.



(a) Bathymétrie sur la zone sud du lagon de Nouvelle-Calédonie

(b) Moyenne de chlorophylle calculée par OC3

FIGURE 29 – Exemples de cartes de sorties sur la sud du lagon de Nouvelle-Calédonie suite à la suppression de données inutiles

Les zones blanches sur la carte de bathymétrie sont les zones non prises en compte pour les traitements. Cela permet de montrer un fort allègement des données.

3.2 Perspectives sur la zone du sud du lagon

On aimerait appliquer le modèle SVM sur cette zone. Pour cela, on appliquerait le traitement décrit dans la partie précédente sur les données de réflectance (443, 488 et 555 nm) et sur la bathymétrie (traitement déjà présenté). Puis le modèle serait appliqué sur les tableaux.

Ensuite, le traitement appliqué sur la zone autour de Nouméa serait étendu sur cette zone plus grande :

1. Décomposer les séries (tendance, composante saisonnière et bruit)
2. Observer d'éventuelles différences comportementales selon les zones (pour des pixels en particulier). On pourrait s'intéresser à des points de la Côte Ouest, de la Côte Est, des Îles Loyautés plus ou moins proches des terres
3. Réaliser une ACP afin de :
   - vérifier si le même aspect de saison observé autour de Nouméa se retrouve sur l'ensemble du sud du lagon calédonien
   - trouver des évènements ou périodes particuliers
   - retrouver des zones différentiables par le comportement de leurs séries et des différenciations non observables sur la zone restreinte
4. Confronter des données de précipitations avec les données de chlorophylle en choisissant les stations d'observations en accord avec les points du lagon sélectionnés (par exemple selon les bassins versants)

4 Conclusion

L'objectif était de traiter les données satellites de la base OPeNDAP Modis Aqua afin de déterminer comment la concentration de chlorophylle, dans la partie sud du lagon calédonien, évolue.

L'étude s'est surtout concentrée sur la zone de Nouméa jusqu'à présent, à cause d'une masse de données très importante. Des idées et des outils ont donc, dans un premier temps, étaient développés sur cette zone restreinte.

4.1 Résumé des résulats

Le problème a été abordé par une première approche des données avec des statistiques de base sur des algorithmes d'estimations de concentration de chlorophylle. On a remarqué que les algorithmes utilisés en télédétection pour ces estimations sont peu adaptées aux particularités du lagon de Nouvelle-Calédonie (eaux essentiellement oligotrophes). Cette observation avait déjà était mise en évidence lors d'un stage précédent. Une adaptation des algorithmes a donc été envisagée. Finalement, un autre type d'algorithme, basé sur une régression SVM, avec les variables disponibles les plus explicatives (bathymétrie et réflectances 443, 488 et 555 nm), a été construit. On a vérifié que cette méthode donnait bien de meilleurs résultats sur nos données que les algorithmes habituels.

Disposant d'un modèle fiable sur nos données, une analyse du comportement de la concentration de chlorophylle a été possible. Tout d'abord, une interpolation linéaire a été mise en œuvre afin d'avoir des données journalières. Ensuite, à l'aide de cartes moyennes, espace-temps, etc. sur les stations, on a eu une vision globale de l'application du modèle SVM à la donnée. La présence d'un cycle annuel a pu déjà être observé. Puis, les séries de chaque pixel ont été décomposées (tendance linéaire, composante saisonnière et bruit) et on a considéré les séries temporelles de variation de concentration sur les stations. Les pixels correspondant à ces stations semblaient représentatifs avec des points océaniques et dans le lagon à des profondeurs et des distances de la côte différentes. À partir des observations, plusieurs résultats :

1. La concentration de chlorophylle moyenne (sur une année) dans l'océan est généralement plus faible que dans des zones lagonaires.
2. Les composantes saisonnières sont également différentes suivant l'emplacement des pixels considérés. On relève des différences :
   - d'amplitude (qui nous renseignent sur la variabilité intra-annuelle) : un pixel du lagon a généralement une amplitude plus forte qu'un pixel océanique.
   - de phase : il semble que le point maximum du cycle pour un pixel proche de la terre soit plus précoce que le point maximum du cycle d'un pixel éloigné de la côte.
3. Les variations interannuelles sont également plus fortes dans le lagon que dans l'océan ; et on ne peut pas généraliser une composante saisonnière pour un pixel trop proche de la côte.

Ensuite, à l'aide d'une ACP sur les pixels, deux groupes ont été extraits dans notre zone à partir de leurs positions dans le premier plan factoriel : des pixels océaniques dans un premier groupe et des pixels lagonaires de forte bathymétrie dans le second groupe. La différence notable entre ces deux groupes, outre leur position spatiale, est leur composante saisonnière. Comme décrit plus haut, des différences de phase et d'amplitude sont relevées. La variabilité des composantes saisonnières intra-groupe est effectivement faible. Grâce à cette ACP, on a également eu confirmation du fait que la forme de la composante saisonnière est différente selon la zone du pixel étudié.

Enfin, avec un jeu de données de précipitations journalières, une mise en relation entre la pluie et la concentration de chlorophylle a été tentée, en tenant compte d'un éventuel décalage de quelques jours. Cependant, la liaison des deux séries sur toute la période ne semble pas très forte. Néanmoins, on remarque que souvent plus le pixel est éloigné de la côte, plus le nombre de jours de décalage est important. L'eau ayant ruisselé atteint effectivement un pixel éloigné de la terre (si elle l'atteint) après un pixel plus proche de la terre. On s'est finalement intéressé à des évènements pluvieux exceptionnels (plus de 60 mm de précipitation en une journée) afin de voir une influence de ces évènements sur l'évolution de la concentration de chlorophylle. Bien souvent, les jours autour de ces évènements ne sont pas renseignés sur nos stations (les résultats disponibles suivants, sont souvent environ une semaine après la forte précipitation), il est donc difficile de mesurer l'éventuelle liaison entre un évènement exceptionnel et la concentration de chlorophylle quelques jours plus tard.

4.2 Perspectives pour la suite

Sur l'ACP réalisée, un effet taille était visible et l'axe 1 représentait un peu moins de 80% de l'inertie totale. On pourrait donc refaire cette ACP en mettant en individus supplémentaires les pixels ayant le plus fortement contribué à la construction de cet axe (souvent des pixels du lagon avec une faible profondeur). Cela permettrait peut-être d'avoir une explication pour d'autres axes que les deux premiers.

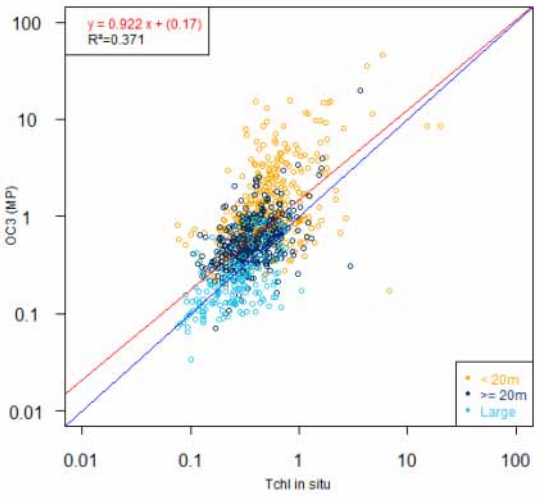
Concernant la relation entre précipitations et chlorophylle, au lieu de calculer un coefficient de corrélation sur les séries entières (décalées ou non), on pourrait effectuer le même travail sur les séries mais pour des périodes plus restreintes. Cela pourrait se faire par un découpage temporel subjectif (par année ou par mois par exemple) ou en catégorisant les dates suivant les anomalies de précipitation et de concentration de chlorophylle. Au lieu de mettre en relation un évènement ponctuel et une éventuelle "réponse" de la chlorophylle (comme fait dans le 2.3.4.3), on mettrait en relation un contexte de pluies anormalement élevées ou faibles avec un contexte de concentration chlorophyllienne plus ou moins forte.

Enfin, bien sûr, il faudrait appliquer toutes ces méthodes à l'ensemble de la zone sud du lagon de Nouvelle-Calédonie.

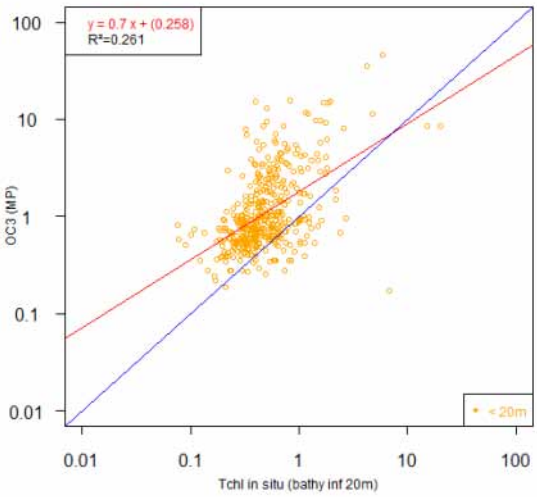
Annexe :

A Régressions des algorithmes en fonction des valeurs mesurées

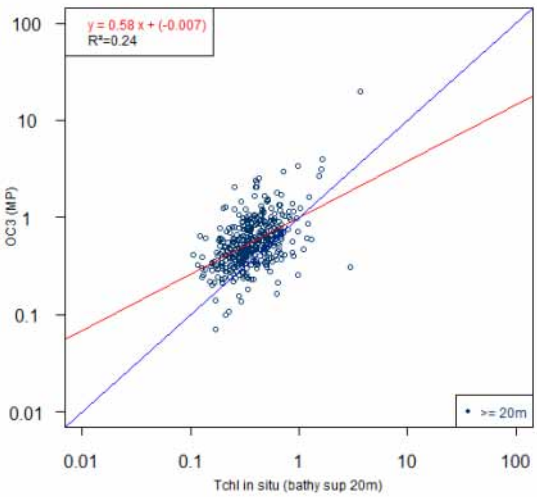
A.1 Estimations d'OC3 en fonction des valeurs *in situ*



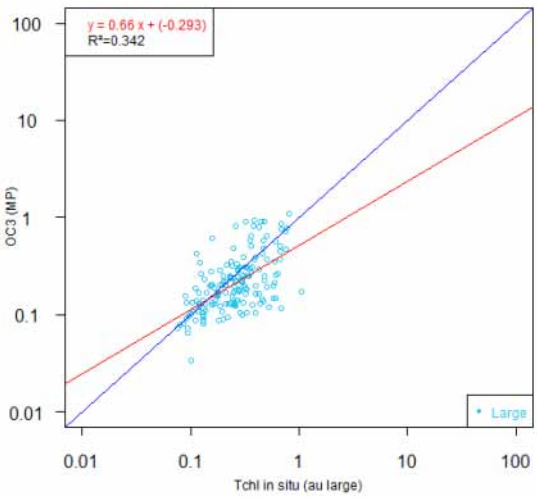
(a) Estimations par OC3 en fonction des mesures *in situ* pour tous les points



(b) Estimations par OC3 en fonction des mesures *in situ* pour les points dont la bathymétrie est inférieure à 20m



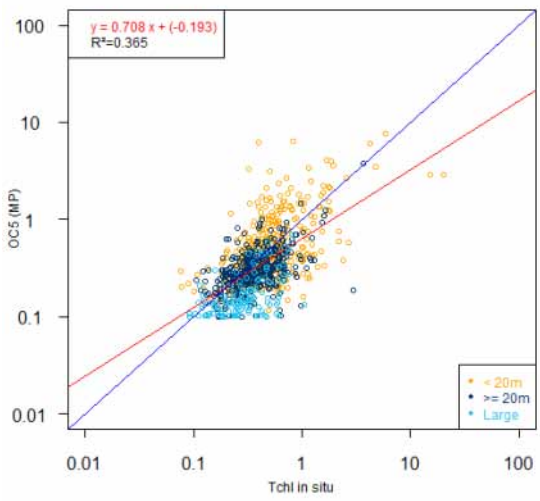
(c) Estimations par OC3 en fonction des mesures *in situ* pour les points dont la bathymétrie est supérieure à 20m



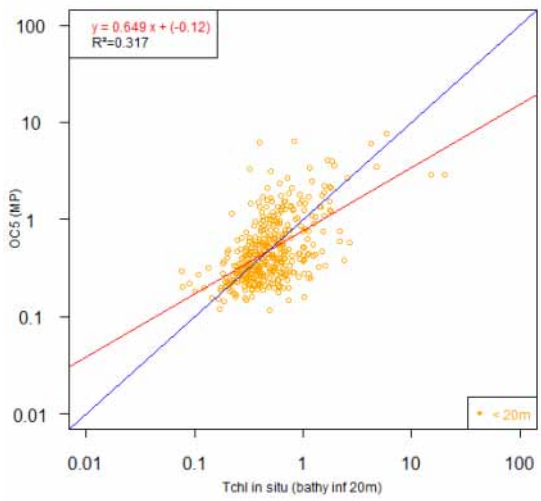
(d) Estimations par OC3 en fonction des mesures *in situ* pour les points du large

FIGURE 30 – Estimations d'OC3 en fonction des mesures *in situ*

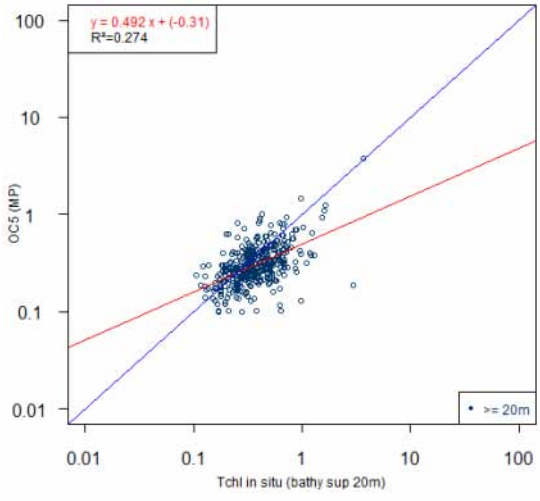
A.2 Estimations d'OC5 en fonction des valeurs *in situ*



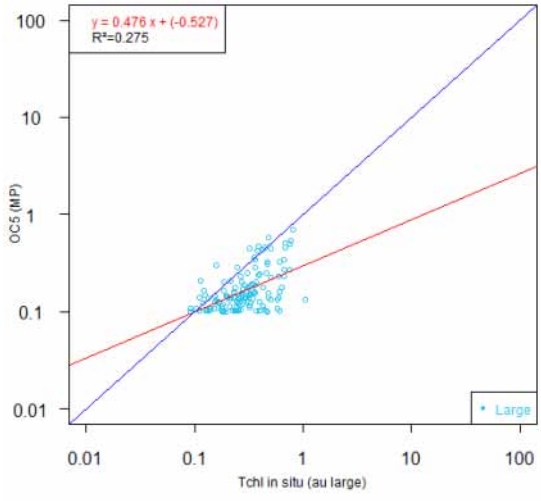
(a) Estimations par OC5 en fonction des mesures *in situ* pour tous les points



(b) Estimations par OC5 en fonction des mesures *in situ* pour les points dont la bathymétrie est inférieure à 20m



(c) Estimations par OC5 en fonction des mesures *in situ* pour les points dont la bathymétrie est supérieure à 20m



(d) Estimations par OC5 en fonction des mesures *in situ* pour les points du large

FIGURE 31 – Estimations d'OC5 en fonction des mesures *in situ*

B Tableaux comparatifs entre les algorithmes et les algorithmes corrigés par le premier axe d'ACP

TABLE 7 – Comparaison entre les algorithmes et les algorithmes corrigés par le premier axe d'ACP

OC3CL	MNB	MNB corrigé	$RMSE$	$RMSE$ corrigée
Lagon : bathymétrie inférieure à 20m	317,62	27,40	534,73	90,61
Lagon : bathymétrie entre 20m et 70m	118,98	9,72	224,66	64,93
Large : bathymétrie supérieure à 70m	16,09	45,79	92,64	102,51
Tous les points	135,99	11,46	319,77	84,87

OC5MP	MNB	MNB corrigé	$RMSE$	$RMSE$ corrigée
Lagon : bathymétrie inférieure à 20m	15,01	13,34	110,10	95,36
Lagon : bathymétrie entre 20m et 70m	-9,70	14,43	44,42	62,18
Large : bathymétrie supérieure à 70m	-41,27	-0,09	26,71	47,15
Tous les points	3,18	20,19	78,94	82,60

OC5CL	MNB	MNB corrigé	$RMSE$	$RMSE$ corrigée
Lagon : bathymétrie inférieure à 20m	69,53	28,36	264,63	123,46
Lagon : bathymétrie entre 20m et 70m	-8,54	-3,16	55,58	46,49
Large : bathymétrie supérieure à 70m	-28,31	26,90	49,34	93,84
Tous les points	19,67	24,44	101,29	79,47

Pour l'algorithme OC5, l'amélioration est moins flagrante[27]. En fait, on n'améliore que pour les zones du lagon dont la bathymétrie est inférieure à 20m.

27. Les améliorations sont surlignées en vert et les faibles améliorations sont en jaune.

C L'algorithme SVM

Le but est de décrire le fonctionnement des SVM et on pourra se référer à la bibliographie pour plus de détails.

L'algorithme SVM a été introduit essentiellement pour de la classification mais il a aussi été adapté pour des problèmes de régression. On s'intéresse à un phénomène h qui à partir d'un certain jeu d'entrées X, produit une sortie $T = h(X)$. Le but est de retrouver h à partir de la seule observation d'un certain nombre de couples entrées-sorties $\{(X_i, T_i) : i = 1, ..., n\}$.

C.1 Principe du SVM

On considère un couple (X, T) de variables aléatoires à valeurs dans $\mathcal{X} \times \mathcal{T}$. Dans un premier temps, on s'intéresse au cas où $\mathcal{T} = \{-1, 1\}$ (c'est-à-dire de la classification) mais on verra par la suite qu'on peut étendre au cas où $Card(\mathcal{T}) > 2$ et aussi à $\mathcal{T} = \mathbb{R}$ (pour la régression notamment).

But : On dispose d'un échantillon de n réalisations des couples de variables aléatoires $(X, T) : S = \{(X_1, T_1); ...; (X_n, T_n)\}$. Et on désire construire une fonction $h : \mathcal{X} \to \mathcal{T}$ telle que la probabilité $P(h(X) \neq T)$ soit minimale.

Hyperplan optimal : Plaçons-nous tout d'abord en dimension 2 (c'est-à-dire que la variable aléatoire X est de dimension 2) et supposons que les deux classes soient linéairement séparées. Mais dans ce cas, il existe parfois beaucoup de droites réalisant une séparation correcte des deux classes (comme représenté sur la FIGURE 32).

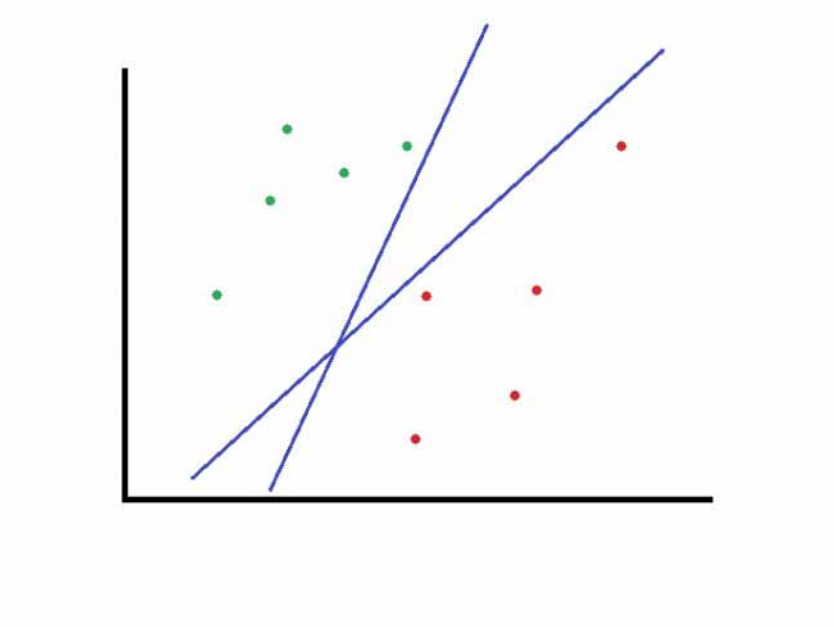


FIGURE 32 – Plusieurs droites pour la séparation

On cherche donc à séparer au mieux nos deux classes par un **hyperplan optimal** (qui en dimension 2 se trouve être une droite optimale), c'est-à-dire un hyperplan qui maximise la marge [28]. Le problème est donc un problème d'optimisation.

Notons la droite qui sépare nos classes par $f_{w,b}$. Elle est caractérisée par l'équation : $a_x x + a_y y + b = 0$ ce qui équivaut à $< w; z > +b = 0$ avec $<;>$ le produit scalaire dans le plan, $w = (a_x; a_y)$ et $z = (x; y)$. Comme les deux classes sont linéairement séparées, il existe w et b tels que :

$$\forall z \in S_- : f_{w,b}(z) < 0 \text{ et } \forall z \in S_+ : f_{w,b}(z) > 0$$

avec S_- l'ensemble des points du nuage pour lesquels $T = -1$ et S_+ l'ensemble des points du nuage pour lesquels $T = 1$. On cherche w et b maximisant la marge. Mais la marge ne dépend que d'une fraction des points du nuage (ceux contraignant la marge), ces points sont appelés **vecteurs de support**. Ils sont représentés sur la figure ci-dessous.

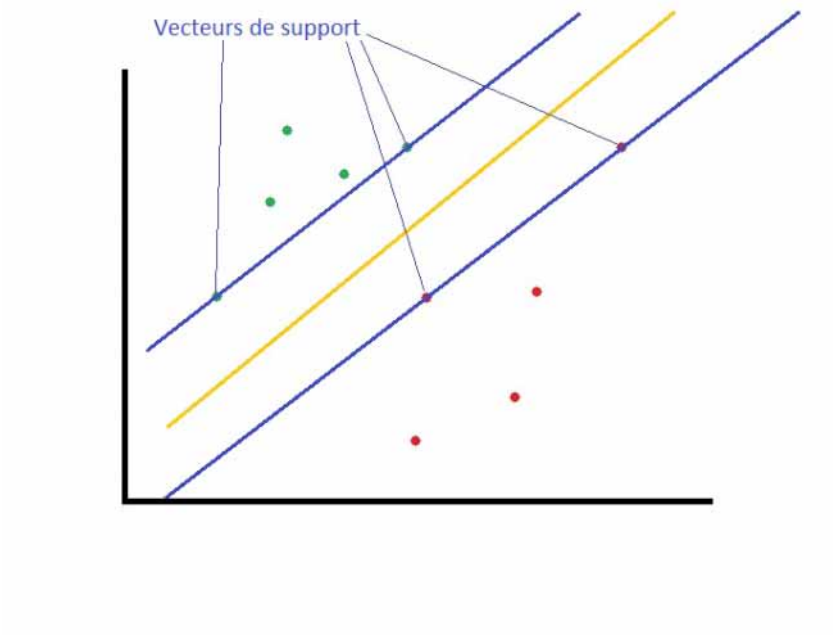


FIGURE 33 – Vecteurs de support

Rappel : Soit la droite (D) ayant pour équation $ax + by + c = 0$ alors la distance entre cette droite et un point du plan $A(x_A, y_A)$ est donnée par :

$$d(A, (D)) = \frac{|ax_A + by_A + c|}{\sqrt{a^2 + b^2}}$$

Appliquons la formule à notre cas, la distance de la droite $f_{w,b}$ à un point $A(x_A, y_A)$ du plan est donnée par :

28. La marge est la distance minimale entre l'hyperplan considéré et l'ensemble des points représentant nos données.

$$\begin{aligned} d(A, f_{w,b}) &= \frac{|a_x x_A + a_y y_A + b|}{\sqrt{a_x^2 + a_y^2}} \\ &= \frac{|f_{w,b}(x_A, y_A)|}{\|w\|} \\ &= \frac{| < w; z_A > +b|}{\|w\|} \end{aligned}$$

où, comme précédemment, $w = (a_x; a_y)$ et $z_A = (x_A; y_A)$.

Considérons $A(x_A, y_A)$ et $B(x_B, y_B)$ deux points du nuage de classes différentes. Alors, par exemple, en se ramenant au cas où les vecteurs supports sont sur les courbes de niveau -1 et 1 [29], on obtient :
$f_{w,b}(x_A, y_A) =< w; z_A > +b = 1 > 0$ et
$f_{w,b}(x_B, y_B) =< w; z_B > +b = -1 < 0$ donc
la marge est donnée dans ce cas par $d(f_{w,b}, A) = \frac{1}{\|w\|} = d(f_{w,b}, B)$. Et maximiser la marge revient finalement à minimiser $\|w\|$ sous certaines contraintes.

En effet, on remarque qu'une réalisation (z, t) de (X, T) est bien classée si et seulement si $tf_{w,b}(z) > 0$. De plus, comme le couple (w, b) est défini à un coefficient multiplicatif près, on peut imposer que : $tf_{w,b}(z) \geqslant 1$. Voilà qui nous donne des contraintes pour notre problème d'optimisation :

$$\begin{cases} \min \frac{1}{2}\|w\|^2 \\ \forall i \in [1; n] : T_i \times (< w; X_i > +b) \geqslant 1 \end{cases} \tag{C.1}$$

On passe finalement au problème dual pour résoudre le problème en introduisant des multiplicateurs de LAGRANGE [30]. Voici le problème dual final à résoudre :

$$\begin{cases} \max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j T_i T_j < X_i; X_j > \\ \forall i \in [1; n] : \alpha_i \geqslant 0 \\ \sum_{i=1}^{n} \alpha_i T_i = 0 \end{cases} \tag{C.2}$$

Seuls les α_i correspondant aux vecteurs de support sont non nuls. En résolvant ce problème et en notant α_i^* les solutions, on trouve l'expression de la fonction de décision (droite de séparation des classes) :

$$f_{w,b}(z) = \sum_{i=1}^{n} \alpha_i^* T_i < X_i; z > +b$$

c'est-à-dire qu'un vecteur directeur de la droite (hyperplan pour une dimension supérieure à 2), est :

$$w^* = \sum_{i=1}^{n} \alpha_i^* T_i X_i$$

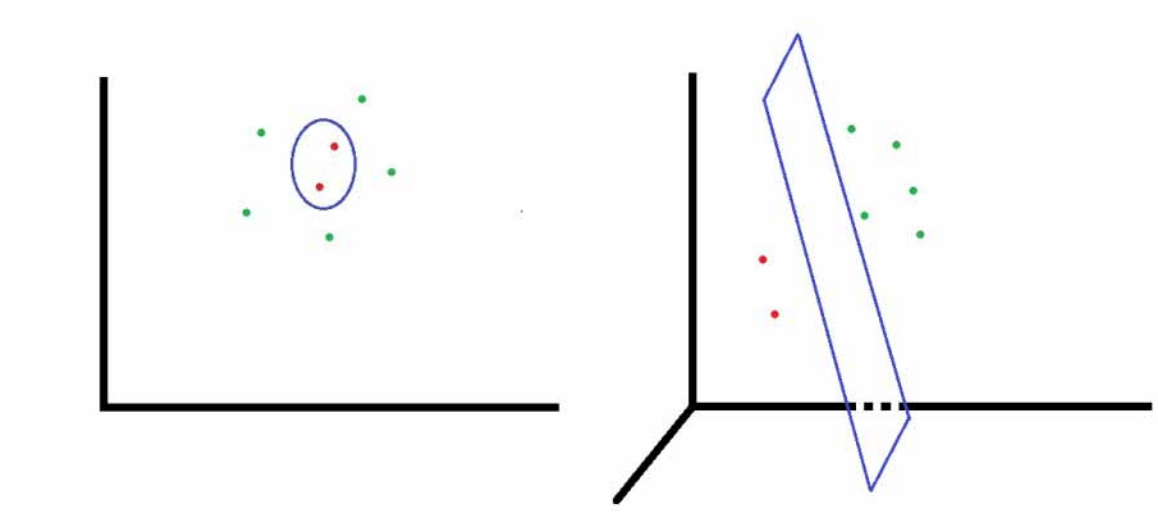
29. On peut se ramener à ce cas plus simple grâce à un tranformation affine du plan.
30. Des détails de la résolution sont donnés dans [J-MAR : SVM] et [MOH-FOM : SVM].

Remarque : Pour une dimension d supérieure à 2, la démarche est la même et on cherche à maximiser la distance des vecteurs de support avec un hyperplan.

Passons au cas où les classes ne sont pas linéairement séparables dans un espace de dimension d. L'idée est de projeter les données dans un espace $\mathcal{F}$ de dimension p plus grande que d (si nécessaire de dimension infinie) par une projection ϕ :

$$\begin{aligned} \phi : \mathbb{R}^d &\longrightarrow \mathcal{F} \\ x &\longmapsto \phi(x) \end{aligned}$$

L'espace de dimension plus grande est appelé **espace des caractéristiques**. Puis, pour poursuivre l'idée, il faudrait résoudre le problème de maximisation de marge dans ce nouvelle espace "plus grand", pour une séparation linéaire.



(a) Séparation quadratique dans le plan (b) Séparation linéaire dans l'espace

FIGURE 34 – Passer d'une séparation quadratique en petite dimension à une séparation linéaire en dimension plus grande

Pour ce faire, on repart du problème primal (C.1) et on introduit de nouvelles variables pour assouplir les contraintes [31] :

$$\begin{cases} \min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \\ \forall i \in [1;n] : T_i \times (< w; X_i > + b) \geqslant 1 - \xi_i \\ \forall i \in [1;n] : \xi_i \geqslant 0 \end{cases} \tag{C.3}$$

Le paramètre C doit être positif. C'est un paramètre déterminant la tolérance du SVM aux exemples mal séparés. Et on déduit, comme dans le cas de la séparation linéaire, le problème dual associé :

31. On autorise certains exemples à avoir une marge < 1 et on pénalise par le dépassement de la contrainte.

$$\begin{cases} \max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j T_i T_j < X_i; X_j > \\ \forall i \in [1;n] : 0 \leqslant \alpha_i \leqslant C \\ \sum_{i=1}^{n} \alpha_i T_i = 0 \end{cases} \tag{C.4}$$

C'est alors qu'on passe dans un espace $\mathcal{F}$ de plus grande dimension par une fonction ϕ. Et on cherche à résoudre le problème dual (C.4) dans ce nouvel espace $\mathcal{F}$:

$$\begin{cases} \max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j T_i T_j < \phi(X_i); \phi(X_j) > \\ \forall i \in [1;n] : 0 \leqslant \alpha_i \leqslant C \\ \sum_{i=1}^{n} \alpha_i T_i = 0 \end{cases} \tag{C.5}$$

et la solution f du problème a la forme :

$$f(x) = \sum_{i=1}^{n} \alpha_i^* T_i < \phi(X_i); \phi(x) > +b$$

On remarque que le problème et la solution ne dépendent que du produit scalaire $< \phi(X); \phi(.) >$; alors plutôt que de choisir la transformation non linéaire ϕ, l'astuce est de choisir une **fonction noyau** :

$$\begin{aligned} k : \mathcal{X} \times \mathcal{X} &\longrightarrow \mathbb{R} \\ (x, x') &\longmapsto < \phi(x); \phi(x') > \end{aligned}$$

telle que la projection ϕ correspondante existe vraiment (sans calculer explicitement ϕ). Pour cette condition, on dispose du :

Théorème C.1. (Théorème de MERCER)
Si k est un noyau défini positif (c'est-à-dire que pour tout ensemble d'exemples, la matrice de terme général $k(X_i, X_j)$ est définie positive) sur un espace χ, alors il existe un espace de HILBERT $\mathcal{H}$ muni du produit scalaire $< .;. >_{\mathcal{H}}$ et une application $\phi : \chi \to \mathcal{H}$ tels que :

$$\forall (x, x') \in \chi^2 : k(x, x') = < \phi(x), \phi(x') >_{\mathcal{H}}$$

Finalement, dans le problème, on peut remplacer les produits scalaires par l'évaluation d'un noyau défini positif. Le séparateur s'exprime alors par :

$$f(x) = \sum_{i=1}^{n} = \alpha_i T_i k(X_i, x) + b$$

Le choix du noyau est très important pour la résolution du problème.

Exemples de noyaux :
- Polynomial : $k_d(x, y) = < x; y >^d$ ou $k_d(x, y) = (< x; y > +c)^d$
- Gaussien : $k(x, y) = exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right)$
- Laplacien : $k(x, x') = exp\left(-\frac{\|x-y\|_1}{2\sigma}\right)$

On peut construire des fonctions noyaux à partir de fonctions noyaux déjà connues. D'ailleurs, en pratique, on combine des noyaux simples (comme ceux présentés dans l'exemple) pour en obtenir des plus complexes. Des exemples de construction sont montrés dans [J-MAR : SVM].

Remarque : En terme de complexité, elle est comprise entre dn^2 et dn^3 avec n le nombre d'exemples d'apprentissage et d la dimension des entrées.

C.2 Le SVM pour la régression

On se place désormais dans le cas où les étiquettes de l'ensemble $\mathcal{T}$ peuvent prendre n'importe quelle valeur réelle. On cherche encore une fonction $f_{w,b}$ linéaire, définie par : $f_{w,b}(x) =< w; x > +b$, telle que pour nos données d'apprentissage $\{(X_i, T_i); i = 1, ..., n\}$ l'erreur soit contrôlée :

$$\forall i \in [1; n] : |T_i - < w; X_i > -b| < \varepsilon$$

On déduit donc le problème d'optimisation suivant, en assouplissant les contraintes :

$$\begin{cases} \min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi'_i) \\ \forall i \in [1; n] : T_i - (< w; X_i > +b) \leqslant \varepsilon + \xi_i \\ \forall i \in [1; n] : (< w; X_i > +b) - T_i \leqslant \varepsilon + \xi'_i \\ \forall i \in [1; n] : \xi_i, \xi'_i \geqslant 0 \end{cases} \tag{C.6}$$

avec $C > 0$. Et comme dans la section C.1, on résout le problème dual de ce problème d'optimisation. Le lagrangien est donné par :

$$\begin{aligned} L = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi'_i) - \sum_{i=1}^{n}(\eta_i\xi_i + \eta'_i\xi'_i) \\ + \sum_{i=1}^{n}\alpha_i(\varepsilon + \xi_i - T_i + < w; X_i > +b) \\ + \sum_{i=1}^{n}\alpha'_i(\varepsilon + \xi'_i + T_i - < w; X_i > -b) \end{aligned}$$

avec les α_i, α'_i, η_i et η'_i les multiplicateurs de LAGRANGE positifs. Par annulation des dérivées partielles, on déduit :

$$\begin{cases} \sum_{i=1}^{n}(\alpha_i - \alpha'_i) = 0 \\ w = \sum_{i=1}^{n}(\alpha_i - \alpha'_i)X_i \\ \forall i \in [1; n] : \eta_i = C - \alpha_i \\ \forall i \in [1; n] : \eta'_i = C - \alpha'_i \end{cases}$$

La fonction de prédiction se déduit du produit scalaire et des multiplicateurs de LAGRANGE, dans le cas d'une relation linéaire. Pour $\widehat{t}$, déduit d'une réalisation x, on a finalement :

$$\widehat{t} = \sum_{i=1}^{n}(\alpha_i - \alpha'_i) < X_i; x > +b$$

avec b choisi de façon à ce que l'hyperplan maximise effectivement la marge. Ce coefficient peut être déterminé de plusieurs façons différentes ; l'une d'elle est de choisir en fonction des vecteurs de support :

$$b = -\frac{1}{2} < w; X_r + X_s >$$

avec X_r et X_s les vecteurs de support (c'est-à-dire des réalisations ayant une valeur non nulle respectivement pour α_r et α'_s).

Dans le cas d'une relation non linéaire, le produit scalaire est remplacé par la fonction noyau k :

$$\widehat{t} = \sum_{i=1}^{n} (\alpha_i - \alpha'_i) k(X_i, x) + b$$

Remarque : Plus de détails sont donnés dans [FAR-MOH : SVM] et [LAA-MAR : SVM]. Dans ce dernier, le cas multivarié est également traité.

D Compléments de sorties de l'ACP sur les pixels

D.1 Tableaux des temps (variables) pour l'axe 3

TABLE 8 – Tableaux récapitulatifs des résultats de l'ACP pour les dates dans l'axe 3

Nombre de jours du mois ayant une bonne contribution pour la construction de l'axe 3 :

Années — Mois	01	02	03	04	05	06	07	08	09	10	11	12
2002							25	31	30	31	30	31
2003	28	2	0	1	0	0	0	0	1	0	0	0
2004	0	0	0	0	2	22	11	0	0	0	0	0
2005	0	0	6	30	30	21	0	0	0	0	0	0
2006	0	0	0	31	16	21	0	0	0	0	0	0
2007	0	0	6	23	23	8	4	0	0	0	0	0
2008	0	0	24	24	31	3	2	0	0	0	8	0
2009	0	0	0	0	0	1	0	0	19	31	30	31
2010	31	28	31	30	17	26	9					

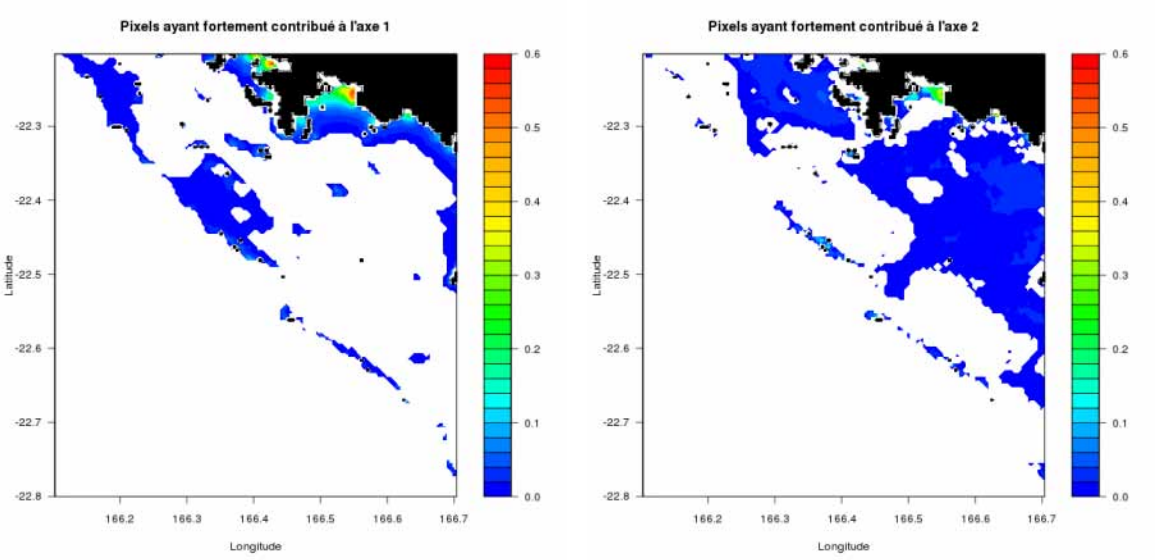
Nombre de jours du mois ayant une coordonnée négative dans l'axe 3 :

Années — Mois	01	02	03	04	05	06	07	08	09	10	11	12
2002							0	0	0	0	0	0
2003	0	0	4	27	31	30	31	29	0	0	0	8
2004	9	22	31	30	31	30	31	31	28	24	23	31
2005	31	28	31	30	31	30	31	29	19	20	9	12
2006	26	28	31	30	31	31	31	27	29	22	30	7
2007	12	19	31	30	31	30	26	31	27	30	26	25
2008	26	29	31	30	31	30	31	16	8	1	0	0
2009	0	1	31	30	31	30	3	0	0	0	0	0
2010	0	0	0	0	0	0	0					

Nombre de jours du mois ayant une coordonnée positive dans l'axe 3 :

Années — Mois	01	02	03	04	05	06	07	08	09	10	11	12
2002							25	31	30	31	30	31
2003	31	28	27	3	0	0	0	2	30	31	30	23
2004	22	7	0	0	0	0	0	0	2	7	7	0
2005	0	0	0	0	0	0	0	2	11	11	21	19
2006	5	0	0	0	0	0	0	4	1	9	0	24
2007	19	9	0	0	0	0	5	0	3	1	4	6
2008	5	0	0	0	0	0	0	15	22	30	30	31
2009	31	27	0	0	0	0	28	31	30	31	30	31
2010	31	28	31	30	31	30	9					

D.2 Cartes des contribution des pixels (individus) pour les premier et deuxième axes d'ACP



(a) Pixels ayant bien contribué à la construction de l'axe 1

(b) Pixels ayant bien contribué à la construction de l'axe 2

FIGURE 35 – Contribution des pixels pour la construction des axes

Les pixels ayant bien contribué sont en couleur (avec une échelle indiquant leur degré de contribution), les autres sont en blancs.

E Éléments de codes R utilisés

E.1 Fonctions basiques pour traiter nos données

```
#############
# Fonctions #
#############

# Fonction pour transformer la matrice et la mettre dans "le bon sens" #
# Matr [matrice] : la matrice à transformer
# Renvoie la matrice transformée

transforme = function(Matr)
{
    return(Matr[,ncol(Matr):1])
}



# Fonction pour tracer les graphes
# Matr [matrice] : matrice contenant les valeurs pour faire le graphique
# lon [vecteur] : contient les abscisses des valeurs à disposition
# lat [vecteur] : contient les ordonnées des valeurs à disposition
# zmin [nombre] : valeur la plus petite pour les valeurs
# zmax [nombre] : valeur la plus grande pour les valeurs
# titre [chaîne de caractères : titre du graphique
trace = function(Matr,lon,lat,zmin,zmax,titre)
{
    # Préparation des paramètres #
    nb_lign=nrow(Matr)
    nb_col=ncol(Matr)

    # Graphique #
    filled.contour(lon, lat, Matr, xlim=c(min(lon),max(lon)),
ylim=c(min(lat), max(lat)), zlim=c(zmin,zmax), main=titre,
color = colorRampPalette(c("blue", "cyan", "green", "yellow",
"orange", "red"), interpolate=c("linear","spline")),
levels = pretty(c(zmin,zmax), 30), nlevels = 30)
}



# Fonction comme trace mais on ajoute les terres #
# Bathy [matrice] : contient les valeurs de la bathymétrie #
trace_avec_bathy = function(Matr,Bathy,lon,lat,zmin,zmax,titre)
{
    # Préparation des paramètres #
    nb_lign=nrow(Matr)
    nb_col=ncol(Matr)
```

```
    # Graphique #
    filled.contour(lon, lat, Matr, xlim=c(min(lon),max(lon)),
ylim=c(min(lat), max(lat)), zlim=c(zmin,zmax), xlab="Longitude",
ylab="Latitude", main=titre, color = colorRampPalette(c("blue", "cyan",
"green", "yellow", "orange", "red"), interpolate=c("linear","spline")),
levels = pretty(c(zmin,zmax), 30), nlevels = 30, plot.axes={box();axis(1);
axis(2);contour(lon, lat, Bathy,col="black", xlim = c(min(lon), max(lon)),
level = c(0), add=TRUE);image(lon, sort(lat), Bathy, zlim=c(0,max(Bathy)),
col="black", add=TRUE)})
}


# Fonction qui calcule la moyenne d'une matrice par rapport au temps #
# Tab [tableau à 3 dimensions] : contient les données en fonction
# masque [tableau à 3 dimensions] : contient des 1 et des <0 selon la validité
# de la mesure de la longitude, la latitude et le temps
# Renvoie la matrice "moyenne"
moyenne_temps = function(Tab,masque)
{
    Tab[Tab<0 | masque<=0]=NA  # On met valeur non attribuée pour les endroits
# où il y a des drapeaux #
    Matr_Moy = apply(Tab,c(1,2), mean, na.rm = TRUE)
    return(Matr_Moy)
}


# Fonction qui calcule l'écart-type d'une matrice par rapport au temps #
# Tab [tableau à 3 dimensions] : contient les données en fonction
# masque [tableau à 3 dimensions] : contient des 1 et des <0 selon la validité
# de la mesure de la longitude, la latitude et le temps
# Renvoie la matrice "écart-type"
ecart_temps = function(Tab,masque)
{
    Tab[Tab<0 | masque<=0]=NA  # On met valeur non attribuée pour les endroits
# où il y a des drapeaux #
    Matr_SD = apply(Tab,c(1,2), sd, na.rm = TRUE)
    return(Matr_SD)
}


# Fonction pour calculer l'approximation linéaire de valeurs manquantes s'il y a
# assez de données, sinon elle renvoit le vecteur lui-même
na.approx2 = function(x)
{
    if (length(x[!is.na(x)])> 2)
    {
```

```
        return(na.approx(x, na.rm = FALSE))
    }
    else
    {
        return(x)
    }
}
```

E.2 Création du modèle SVM à partir des données *in situ*

```
library(e1071)
# Import des données #

l_data = data.frame(
Tchl.ins = log(1 + c(inf20m$Tchl.ins,sup20m$Tchl.ins,large$Tchl.ins)),
bat.ins = log(1 + c(inf20m$bat.ins,sup20m$bat.ins,large$bat.ins)),
Rrs_412.satMoyP = log(1 + c(inf20m$Rrs_412.satMoyP,
sup20m$Rrs_412.satMoyP,large$Rrs_412.satMoyP)),
Rrs_443.satMoyP = log(1 + c(inf20m$Rrs_443.satMoyP,
sup20m$Rrs_443.satMoyP,large$Rrs_443.satMoyP)),
Rrs_488.satMoyP = log(1 + c(inf20m$Rrs_488.satMoyP,
sup20m$Rrs_488.satMoyP,large$Rrs_488.satMoyP)),
Rrs_531.satMoyP = log(1 + c(inf20m$Rrs_531.satMoyP,
sup20m$Rrs_531.satMoyP,large$Rrs_531.satMoyP)),
Rrs_555.satMoyP = log(1 + c(inf20m$Rrs_555.satMoyP,
sup20m$Rrs_555.satMoyP,large$Rrs_555.satMoyP)),
Rrs_620.satMoyP = log(1 + c(inf20m$Rrs_620.satMoyP,
sup20m$Rrs_620.satMoyP,large$Rrs_620.satMoyP)),
Rrs_667.satMoyP = log(1 + c(inf20m$Rrs_667.satMoyP,
sup20m$Rrs_667.satMoyP,large$Rrs_667.satMoyP))
)
# Données en logarithme #

################
# Modélisation #
################

# En log #
formule = "Tchl.ins ~ bat.ins + Rrs_443.satMoyP + Rrs_488.satMoyP
+ Rrs_555.satMoyP"
data2 = l_data[,c("Tchl.ins","bat.ins","Rrs_443.satMoyP",
"Rrs_488.satMoyP","Rrs_555.satMoyP")]
l_model = svm(as.formula(formule), data2, type = "eps-regression",
kernel = "radial")

# Représentation graphique des points #
pal = colorRampPalette(c("blue","yellow","red"))
```

```
color = pal(100)
bat = log(data2$bat.ins)
bat_norm = round(100*(bat-min(bat))/(max(bat)-min(bat)))
plot(l_model$fitted, l_data$Tchl.ins, xlim = c(0,2), ylim = c(0,2),
col = color[bat_norm], pch = 19)
abline(0,1,col="green")
cor(l_model$fitted, l_data$Tchl.ins)
# Correlation 0.68 #
```

E.3 Application du modèle SVM sur nos données satellites

```
library(ncdf4)
library(e1071)

######################
# Données satellites #
######################

# Lecture des fonctions
source(paste(racine,"codes/simple_test/fonctions.r", sep = ""))

# Données de bathymétrie #
fichier_fond = paste(Donnees_cn, "Bathy.cdf", sep="")
# Utilisation du package ncdf4 pour ouvrir les fichiers *.nc
# ou *.cdf et extraire leurs informations #
nc = nc_open(fichier_fond)
bathy = ncvar_get(nc,"BATH")
lat.b = ncvar_get(nc,"LAT")
lon.b = ncvar_get(nc,"LON")
nc_close(nc)


# Nom du fichier sur le quel on va travailler #
fichier = paste(Donnees_cn, "MYDL2_map3_Agg_tout_region2.nc", sep="")
# On prend les données disponibles dans le fichier de la zone étudiée #
nc = nc_open(fichier)
lon = ncvar_get(nc,"lon")
lat = ncvar_get(nc,"lat")
Rrs_443 = ncvar_get(nc,"Rrs_443")  # Slope : 2.E-6 Intercept : 0.05
Rrs_488 = ncvar_get(nc,"Rrs_488")  # Slope : 2.E-6 Intercept : 0.05
Rrs_555 = ncvar_get(nc,"Rrs_555")  # Slope : 2.E-6 Intercept : 0.05
temps = ncvar_get(nc,"time")
masque = ncvar_get(nc,"valhysat_mask")
nc_close(nc)

# Correction #
Rrs_443 = Rrs_443 * 0.000002 + 0.05
```

```
Rrs_488 = Rrs_488 * 0.000002 + 0.05
Rrs_555 = Rrs_555 * 0.000002 + 0.05


# Transformation #
for (j in 1:dim(masque)[3])
{
    Rrs_443[,,j] = transforme(Rrs_443[,,j])
    Rrs_488[,,j] = transforme(Rrs_488[,,j])
    Rrs_555[,,j] = transforme(Rrs_555[,,j])
    masque[,,j] = transforme(masque[,,j])
}

# Limites de l'espace pour la zone #
minlat = min(lat)
maxlat = max(lat)
minlon = min(lon)
maxlon = max(lon)


# On ne prend les données de bathymétrie que pour notre zone #
in_lon = indices_proches(lon.b,minlon, maxlon)
in_lat = indices_proches(lat.b,minlat,maxlat)
Bath_Z2 = bathy[in_lon[1]:in_lon[2],in_lat[1]:in_lat[2]]
nb_lign = dim(Bath_Z2)[1]
nb_col = dim(Bath_Z2)[2]

nb_temps = dim(masque)[3]


dat_1950 = ISOdate(1950,1,1)
# On met les dates au format "habituel" #
dates = dat_1950 + 60 * temps
jours=substr(dates,1,10)

# On nomme les lignes lignes et colonnes des tableaux extraits #
dimnames(Rrs_443)[[1]]=lon
dimnames(Rrs_443)[[2]]=lat
dimnames(Rrs_443)[[3]]=substr(dates,1,19)

dimnames(Rrs_488)[[1]]=lon
dimnames(Rrs_488)[[2]]=lat
dimnames(Rrs_488)[[3]]=substr(dates,1,19)

dimnames(Rrs_555)[[1]]=lon
dimnames(Rrs_555)[[2]]=lat
dimnames(Rrs_555)[[3]]=substr(dates,1,19)
```

```
dimnames(masque)[[1]]=lon
dimnames(masque)[[2]]=lat
dimnames(masque)[[3]]=substr(dates,1,19)

names(temps)=dates


# On enlève les valeurs masquées #
Rrs_443[Rrs_443<0] = NA
Rrs_488[Rrs_488<0] = NA
Rrs_555[Rrs_555<0] = NA
Rrs_443[masque<=0] = NA
Rrs_488[masque<=0] = NA
Rrs_555[masque<=0] = NA

# On passe au log #
Rrs_443 = log(1 + Rrs_443)
Rrs_488 = log(1 + Rrs_488)
Rrs_555 = log(1 + Rrs_555)


# Prédiction pour les données satellites #
l_Chl_sat = Rrs_443
l_Chl_sat[,,] = NA
for (i in 1:nb_lign)
{
    for (j in 1:nb_col)
    {
        l_donnees_sat = data.frame(
        bat.ins = array(log(1+log(1-Bath_Z2[i,j])), dim = c(nb_temps)),
        Rrs_443.satMoyP = log(1 + Rrs_443[i,j,]),
        Rrs_488.satMoyP = log(1 + Rrs_488[i,j,]),
        Rrs_555.satMoyP = log(1 + Rrs_555[i,j,])
        )
        # Attention au "-" dans la bathy : dans le fichier, la bathy
        # est négative mais dans le fichier de départ (qu'on a pris
        # pour trouver la "formule"), elle était positive

        # Prédiction pour un jeu de données à partir du modèle svm #
        if ((length(Rrs_443[i,j,is.na(Rrs_443[i,j,])]) == nb_temps &&
length(Rrs_488[i,j,is.na(Rrs_488[i,j,])]) == nb_temps &&
length(Rrs_555[i,j,is.na(Rrs_555[i,j,])]) == nb_temps) || Bath_Z2[i,j] >=0)
        {
            l_Chl_sat[i,j,] = NA
        }
        else
```

```
        {
            l_pred = predict(l_model, l_donnees_sat)
            l_Chl_sat[i,j,names(l_pred)] = l_pred
        }
    }
}

# On passe à l'exonentielle car on a une estimation du log #
Chl_sat = exp(l_Chl_sat) - 1
```

E.4 Fonctions pour transformer une matrice en vecteur et vice-versa, tout en éliminant des données inutiles

```
# Fonction pour généraliser la transformation matrice-vecteur avec une
# matrice donnant la bathymétrie. On donne les seuils #
# Mat [Matrice] : matrice d'entrée. Matrice à transformer en vecteur
# Bath [Matrice] : matrice donnant la bathymétrie. Les seuils sont
#   déterminés par cette matrice
# Mat et Bath doivent avoir les mêmes dimensions
# mini [nombre] : nombre pour le seuil minimal (minorant pour Bath)
# maxi [nombre] : nombre pour le seuil maximal (majorant pour Bath)
Matrice_vecteur = function(Mat, Bath, mini, maxi)
{
    vect = Mat[Bath<=maxi & Bath>=mini]
    return(vect)
}


# Fonction qui fait l'inverse de Matrice_vecteur : elle transforme le
# vecteur en matrice à l'aide de la matrice qui nous a servi à écarter
# les valeurs non utiles7
# Vect [Vecteur] : vecteur d'entrée. Vecteur à retransformer en matrice
# Bath [Matrice] : matrice donnant la bathymétrie. Les seuils sont
#   déterminés par cette matrice
# Le nombre de Bath entre mini et maxi doit être le même que la longueur
# du vecteur Vect
# mini [nombre] : nombre pour le seuil minimal (minorant pour Bath)
# maxi [nombre] : nombre pour le seuil maximal (majorant pour Bath)
Vecteur_matrice = function(Vect, Bath, mini, maxi)
{
    nb_lign = nrow(Bath)
    nb_col = ncol(Bath)
    new_mat = array(c(NA), dim=c(nb_lign, nb_col))
    new_mat[Bath_Z2<=maxi & Bath_Z2>=mini] = Vect

    return(new_mat)
}
```

Références

[Site de l'IRD] sites de l'IRD : http ://www.ird.fr/ et http ://www.espace.ird.fr

[Rap-2009-IRD] Rapport d'activité de 2009 de l'IRD

[C-BIE : Cours AD] Cours d'Analyse de données en M2ISN de M. BIERNACKI

[J-MAR : SVM] Jérémie MARY équipe TAO LRI, "Méthodes d'Apprentissage Avancées, SVM", 30 janvier 2006

[MOH-FOM : SVM] Hasan MOHAMADALLY et Boris FOMANI, "SVM : Machines à Vecteurs de Support ou Séparateurs à Vastes Marges", 16 janvier 2006

[O-BOU : SVM] Olivier BOUSQUET du Centre de Mathématiques Appliquées École Polytechnique Palaiseau, "Introduction aux *Support Vector Machines* (SVM)", 15 novembre 2001

[RAK-CAN : SVM] Alain RAKOTOMAMONJY et Stéphane CANU de l'INSA de Rouen, "Estimation de la concentration en ozone par SVM", 2001

[LAA-MAR : SVM] Hicham LAANAYA, Arnaud MARTIN, Driss ABOUTADJINE et Ali KHENCHAF de l'Université Mohammed V-Agdal de Rabat et ENSIETA de Brest, "Régression floue et crédibiliste par SVM pour la classification des images sonar"

[FAR-MOH : SVM] Aly FARAG et Refaat M MOHAMED de l'Université de Louisville, "Regression Using Support Vector Machines : Basic Foundations", décembre 2004

[S-OUI] Sylvain OUILLON, Pascal DOUILLET, Anne PETRENKO, Jacques NEVEUX, Cécile DUPOUY, Jean-Marie FROIDEFOND, Serge ANDRÉFOUËT et Alain MUÑOZ-CARAVACA, "Optical Algorithms at Satellite Wavelengths for Total Suspended Matter in Tropical Coastal Waters", article sensors 2008

[S-OUI2] Sylvain OUILLON, Pascal DOUILLET, Renaud FICHEZ, Jean-Yves PANCHÉ, "Enhancement of regional variations in salinity and temperature in a coral reef lagoon, New Caledonia", Geoscience 337 2005