2014 - Domaine Outils pour la surveillance environnementale
Action 9 – FlowCAM / ZooPhytoImage

Version évolutive de l'outil opérationnel de numérisation et d'analyse semi-automatique d'images de phytoplancton, utilisant le matériel FlowCAM et le logiciel ZooPhytoImage. Nouvelles perspectives

Action 9 – FlowCAM / ZooPhytoImage – Livrable 1

Rapport final, février 2015

Guillaume WACQUET (Université de Mons)

Philippe GROSJEAN (Université de Mons)

Florent COLAS (Ifremer)

Denis HAMAD (ULCO)

Luis Felipe ARTIGAS (ULCO)

Février 2015

AUTEURS

Guillaume WACQUET (Université de Mons) – Guillaume.Wacquet@umons.ac.be

Philippe GROSJEAN (Université de Mons) – Philippe.Grosjean@umons.ac.be

Florent COLAS (Ifremer) – Florent.Colas@ifremer.fr

Denis HAMAD (ULCO) – Denis.Hamad@lisic.univ-littoral.fr

Luis Felipe ARTIGAS (ULCO) – Felipe.Artigas@univ-littoral.fr

CORRESPONDANTS

Onema : Marie Claude XIMENES (Onema), marie-claude.ximenes@onema.fr

Ifremer : Catherine BELIN (Ifremer), Catherine.Belin@ifremer.fr

AUTRES CONTRIBUTEURS

Kevin DENIS (Université de Mons) – Kevin.Denis@umons.ac.be

Nour ALI (Université de Mons) – Nour.Ali@umons.ac.be

Morgan TARDIVEL (Ifremer) – Morgan.Tardivel@ifremer.fr

Bertrand FOREST (Ifremer) – Bertrand.Forest@ifremer.fr

Marie-Pierre CRASSOUS (Ifremer) – Marie.Pierre.Crassous@ifremer.fr

Michel LUNVEN (Ifremer) – Michel.Lunven@ifremer.fr

Marie-Madeleine DANIELOU (Ifremer) – Marie.Madeleine.Danielou@ifremer.fr

Alain LEFEBVRE (Ifremer) – Alain.Lefebvre@ifremer.fr

Droits d'usage : libre accès
Niveau géographique : national
Couverture géographique : nationale
Niveau de lecture : experts

TITRE

Version évolutive de l'outil opérationnel de numérisation et d'analyse semi-automatique d'images de phytoplancton, utilisant le matériel FlowCAM et le logiciel Zoo/PhytoImage. Nouvelles perspectives.

RESUME

Le système couplé FlowCAM/ZooPhytoImage est devenu un outil véritablement opérationnel en 2014. Cependant, pour qu'il soit totalement adapté aux observations du phytoplancton réalisées dans le cadre du réseau d'observation REPHY, et afin de mieux répondre aux sollicitations présentes et futures concernant l'évaluation de la qualité des eaux littorales et marines dans le cadre des exigences européennes, telles que la DCE et la DCSMM, des nouvelles fonctionnalités doivent être intégrées aux outils existants. C'est pourquoi, différents axes d'évolution ont été proposés par l'UMONS et Ifremer pour adapter, à la fois l'appareil de numérisation et le logiciel ZooPhytoImage aux contraintes définies par le REPHY.

Premièrement, la version 5 de Zoo/PhytoImage contient de récentes nouveautés telles que le développement de routines pour importer et analyser les données automatiquement, et la refonte de l'interface graphique pour une meilleure ergonomie et une simplicité d'utilisation. Cette dernière a été étudiée du point de vue de son adéquation pour des traitements spécifiques liés aux besoins des principaux partenaires. En effet, il apparaissait désirable d'augmenter considérablement l'interactivité visuelle avec le logiciel, notamment au niveau de la visualisation des vignettes, des données brutes et de l'automatisation de certaines tâches. A cette fin, dans cette version du logiciel, les outils de traitement d'images et de transformation des données brutes sont appliqués implicitement.

Deuxièmement, la dernière version de Zoo/PhytoImage permet d'obtenir des identifications automatiques pertinentes du phytoplancton mais sans distinguer une cellule d'une colonie. Or, même si les colonies contribuent en grande partie à la productivité annuelle, l'ensemble des estimateurs de la biomasse sont calibrés essentiellement sur l'abondance en termes de cellules par unité de volume. Dans ce rapport, la méthode proposée consiste à construire des modèles prédictifs permettant d'estimer le nombre de cellules par colonie, en se basant sur les comptages manuels réalisés sur les particules du set d'apprentissage. Dans cette étude, des scores élevés de performance obtenus par différentes méthodes prédictives ont été mis en évidence sur six groupes taxinomiques de phytoplancton colonial de la Manche Orientale et du Sud de la Mer du Nord.

Enfin, le module de correction de l'erreur, intégré à Zoo/PhytoImage depuis la version 4, permet d'obtenir des identifications avec un faible pourcentage d'erreur par groupe, pour chacun des échantillons analysés. Nous proposons ici, d'utiliser l'information liée à la validation manuelle des vignettes par l'expert, afin d'ouvrir la voie à l'apprentissage actif. Dans ce rapport, nous montrons les intérêts de ce processus pour la reconnaissance semi-automatisée du phytoplancton, à savoir : la construction et l'adaptation automatique du set d'apprentissage permettant de partir d'un set « global » au niveau national ; l'amélioration des performances de classification automatique de nouveaux échantillons ; un gain de temps lors de la validation des prédictions automatiques dans le cadre de la correction de l'erreur.

MOTS CLES (THEMATIQUE ET GEOGRAPHIQUE)

Plancton, Analyse automatisée, Analyse d'image, Classification supervisée, Apprentissage actif, Dénombrement de cellules.

TITLE

Evolutive version of the operational tool for digitization and semi-automated analysis of phytoplankton images, using the FlowCAM device and the Zoo/PhytoImage software. New perspectives.

ABSTRACT

The coupled system FlowCAM/ZooPhytoImage has become a real operational tool in 2014. However, to be fully adapted to the observations of phytoplankton performed in the context of the REPHY observation network and in order to better respond to present and future requests concerning the evaluation of quality of coastal and marine waters within the European requirements, such as the WFD and MSFD, new functionalities must be integrated into existing tools. Therefore, different axis of development have been proposed by UMONS and Ifremer to adapt both the digitization device and the Zoo/PhytoImage software to the constraints defined by the REPHY.

First, version 5 of Zoo/PhytoImage contains recent innovations such as the development of routines to automatically import and analyze data, and the redesign of the graphical user interface to improve the ergonomics and the ease of use. The latter has been studied from the point of view of its fit for specific treatments related to the needs of major partners. Indeed, it appears desirable to significantly increase the visual interaction with the software, particularly in terms of exploration of vignettes and raw data, and the automation of some tasks. For this, in this version of the software, the image processing tools and raw data transformation are implicitly applied.

Second, the latest version of Zoo/PhytoImage provides relevant automatic identifications of phytoplankton but without distinction of a cell from a colony. However, even if the colonies greatly contribute to the annual productivity, all the estimators of biomass are essentially calibrated on the abundance in terms of cells per volume unit. In this report, the proposed method consists in building predictive models to estimate the number of cells per colony, based on manual counts performed on the particles of the training set. In this study, high performance scores obtained by different predictive methods, were highlighted on six taxonomic groups of colonial phytoplankton from eastern Channel and Southern North Sea.

Finally, the error correction module, integrated to Zoo/PhytoImage since version 4, provides identifications with a low percentage of error per group for each analyzed sample. We propose here to use the information related to the manual validation of vignettes by expert to open the way to the active learning. In this report, we show the interest of this process for the semi-automated recognition of phytoplankton, such as the building and automatic adaptation of the training set allowing to use a "global" training set at national level; the improvement of clustering performances for new samples; a time saving during the validation of the automatic predictions in the context of the error correction.

KEY WORDS (THEMATIC AND GEOGRAPHICAL AREA)

Plankton, Automated analysis, Image processing, Supervised classification, Active learning, Cells counting.

SYNTHESE POUR L'ACTION OPERAT IONNELLE

Dans le cadre des réseaux de surveillance de l'IFREMER tels que le REPHY, l'analyse d'échantillons phytoplanctoniques est traditionnellement associée à de longues et fastidieuses séances de comptage des particules fixées de plancton sous binoculaire. Cependant, aujourd'hui, de nouvelles méthodes d'acquisition de données sur le phytoplancton sont disponibles, et s'appuient principalement sur l'analyse assistée par ordinateur d'images numériques. Depuis quelques années, l'IFREMER mène des actions pour intégrer ces équipements au REPHY. Suite au travail de post-doctorat de A. Tunin-Ley (LER Arcachon puis Université de Bordeaux) puis de G. Wacquet (LER Boulogne-sur-Mer), différents axes d'évolution ont été proposés par l'Université de Mons et l'IFREMER pour adapter le FlowCAM couplé à Zoo/PhytoImage aux exigences du REPHY qui sont la justesse et la répétabilité de la mesure. A cela s'ajoute la rapidité : l'objectif de ces nouveaux outils étant d'offrir un gain de temps tangible aux observateurs REPHY par rapport aux observations au microscope optique.

Dans ce contexte, le Laboratoire Détection, Capteurs et Mesures (LDCM) de l'IFREMER (centre de Bretagne) a développé un prototype de matériel appelé « FastCAM », permettant d'améliorer certaines technologies du FlowCAM, et en particulier l'optique et le fluidique. Les premières expérimentations ont permis de mettre en évidence un gain important en terme de temps de numérisation (11 fois plus rapide) grâce notamment à l'installation d'une caméra haute résolution permettant l'acquisition de 340 images par seconde. Cependant, ces résultats préliminaires doivent être validés sur des échantillons naturels en comparaison avec les lectures au microscope optique.

D'un point de vue logiciel, le FlowCAM (et peut-être, à terme, le FastCAM) est couplé à Zoo/PhytoImage. Ce logiciel est spécialisé dans le traitement d'image numériques de plancton. Il est « open source » et existe depuis 2004. Actuellement, ses droits sont partagés entre l'UMONS, l'IFREMER et la politique scientifique belge (BelSpo) qui ont tous les trois contribué financièrement à son développement durant ces dernières années. La version initiale publique 1 a permis de mettre en place les concepts du traitement d'images numériques de plancton. Il s'agit, en effet, du premier logiciel public dédié spécifiquement à la classification supervisée du plancton sur base d'images numériques.

Aujourd'hui, la version 5 de Zoo/PhytoImage fournit une solution puissante en fonctionnalités logicielles dans un système revisité, et est distribué sur le site du CRAN (http://cran.r-project.org). Les principales nouveautés sont les suivantes :

- Refonte du code pour l'exécuter sur la dernière version de R (version 3),
- Refonte de l'interface graphique pour une meilleure ergonomie et une simplicité d'utilisation,
- Développement de routines pour importer et analyser les données automatiquement.

Parmi ces changements, le plus important pour les utilisateurs finaux est probablement la nouvelle interface graphique utilisateur. Cette dernière a été étudiée du point de vue de son adéquation pour des traitements spécifiques liés aux besoins des principaux partenaires. En effet, il apparaissait désirable d'augmenter

considérablement l'interactivité visuelle avec le logiciel, notamment au niveau de la visualisation des vignettes, des données brutes et de l'automatisation de certaines tâches. Pour cela, une refonte complète de l'interface graphique utilisateur de Zoo/PhytoImage sur base de définition des "use cases" dans le cadre d'une perspective d'exploitation en routine, était nécessaire. Le présent rapport détaille les fonctionnalités du logiciel, telles que disponibles dans sa version publique 5.

Cependant, jusqu'à présent, cette dernière version de Zoo/PhytoImage permet d'obtenir des identifications semi-automatiques pertinentes du phytoplancton mais sans distinguer une cellule d'une colonie. Or, même si les colonies contribuent en grande partie à la productivité annuelle, l'ensemble des estimateurs de la biomasse sont calibrés essentiellement sur l'abondance en termes de cellules par unité de volume. La méthode proposée dans ce rapport consiste à construire des modèles prédictifs permettant d'estimer le nombre de cellules par colonie dans tous les échantillons étudiés, en se basant sur les comptages manuels réalisés sur les particules du set d'apprentissage. A cette fin, des outils visuels et statistiques ont été développés : outils d'aide au comptage manuel sur ordinateur, régressions linéaires et non linéaires, classification supervisée de type « machine learning » et estimation de la qualité de prédiction à l'aide de la validation croisée. Dans cette étude, des scores élevés de performance obtenus par les différentes méthodes prédictives ont été mis en évidence sur six groupes taxinomiques de phytoplancton colonial de la Manche Orientale et du Sud de la Mer du Nord. Ce module de dénombrement constitue un axe d'évolution prioritaire qui devra être intégré à Zoo/PhytoImage.

La seconde grande évolution consiste à utiliser l'information liée à la validation manuelle des vignettes par l'expert, afin d'ouvrir la voie à l'apprentissage actif. Cette méthode présente un triple intérêt : (i) construction et adaptation automatique du set d'apprentissage permettant de partir d'un set « global » au niveau national ; (ii) amélioration des performances de classification automatique de nouveaux échantillons ; (iii) gain de temps lors de la validation des prédictions automatiques dans le cadre de la correction de l'erreur. En effet, l'outil de classification est obtenu par apprentissage d'un algorithme de type « machine learning » qui établit un lien entre les attributs des particules et les groupes taxonomiques sur base d'un ensemble de particules d'identité connue (le set d'apprentissage qui est élaboré manuellement par l'opérateur sur base de quelques centaines ou milliers de particules d'exemple). Actuellement, cette étape de création manuelle du set d'apprentissage s'avère être une tâche fastidieuse et coûteuse en temps, mais également subjective. Une solution possible pour remédier à ce problème tient dans le set d'apprentissage adaptatif localement. En d'autres termes, il serait envisageable de constituer un set « global » à l'échelle nationale, constitué d'images disparates (provenant de la numérisation par différents appareils, de différentes zones géographiques, à différentes saisons, etc.), et de lui rajouter automatiquement les vignettes validées localement, géographiquement, et temporellement (échantillons des semaines précédentes et/ou échantillons des années précédentes à la même période de l'année) avec élimination progressive des vignettes du set d'apprentissage global au fur et à mesure que des données locales viennent compléter le set. Ce type de manipulation du set d'apprentissage devrait alors améliorer également la reconnaissance automatique des particules contenues dans un nouvel échantillon, et ainsi limiter les erreurs induites par l'algorithme de classification supervisée. Dans ce contexte, le nombre de vignettes à valider lors de

l'étape de correction de l'erreur devrait baisser considérablement, ce qui impliquerait un gain de temps pour l'utilisateur.

Le dernier point étudié dans ce rapport concerne la mise en forme des résultats en sortie de Zoo/PhytoImage, en vue de l'intégration dans Quadrige². Dans un premier temps, il a été décidé d'utiliser le format « Quadrilabo ». En effet, en 2013, le processus d'intégration des données dans le système d'information de l'IFREMER, peut passer, pour les utilisateurs ne pouvant ou ne souhaitant pas utiliser l'interface de saisie classique, par l'envoi du fichier de données au format « Quadrilabo » à la cellule d'administration qui vérifie alors techniquement le contenu du fichier et lance le programme automatisé d'intégration. Ce fichier se présente sous la forme d'une matrice de type EXCEL regroupant en ligne l'ensemble des résultats à intégrer (une ligne par résultat). Chaque ligne est décrite en colonne par des champs de 3 types :

- Champs EDILABO (Standard) : regroupant les informations permettant l'identification des résultats (Paramètre, Support, Fraction, Méthode, …). Ils reposent généralement sur des codes SANDRE standardisés. Ces codes SANDRE doivent faire l'objet d'une identification préalable en s'appuyant sur le site du SANDRE.
- Champs Quadrige² : regroupant les informations propres à Quadrige² relatives aux métadonnées (caractéristiques des passages, prélèvements, échantillons). Ils sont indispensables pour faire le lien entre les résultats de mesures et les données *in situ* associées de Q².
- Champ "NIVEAU SAISIE" : Un résultat dans Q² peut être rattaché à un passage, à un prélèvement ou un échantillon. Cette information, propre à Q², est indispensable au programme de reprise automatique et doit figurer clairement dans un champ spécifique.

Dans ce rapport, nous présentons brièvement les différentes méthodes et fonctions envisagées pour la mise en forme des résultats Zoo/PhytoImage. Les conclusions et perspectives dégagées au terme d'une réflexion commune sur la bancarisation des données, sont présentées dans le livrable 2 (« Mise en œuvre opérationnelle de l'outil FlowCAM/ZooPhytoImage dans le cadre de la surveillance REPHY »).

Perspectives

Le module de dénombrement des cellules dans les colonies est une évolution prioritaire du logiciel Zoo/PhytoImage. Cependant, dans ce rapport, les scores de performance présentés ont été calculés par validation croisée. Dans un premier temps, il serait donc pertinent de comparer les résultats obtenus par les modèles prédictifs retenus avec les observations et comptages réalisés par lectures au microscope optique. Puis, dans un second temps, ce module devrait, à terme, permettre d'obtenir des mesures de biovolume, de biomasse et d'équivalent carbone pour chacun des groupes taxonomiques identifiés dans un échantillon d'eau de mer. En effet, dans la littérature, la majorité des formules mathématiques de conversion permettant d'obtenir ces critères écologiques, sont basées principalement sur des mesures de taille d'une cellule pour chacune des espèces recensées.

De plus, le processus de correction d'erreur étant réalisé de manière itérative (l'utilisateur valide plusieurs sets de vignettes « suspectes », l'un après l'autre), il

serait intéressant de pouvoir évaluer l'apport de l'apprentissage actif appliqué après chacune des étapes de validation des vignettes. En effet, l'injection des vignettes précédemment validées dans le set d'apprentissage devrait améliorer la prédiction automatique des vignettes suspectes dans l'étape suivante, et ainsi limiter le nombre de vignettes à corriger pour l'utilisateur.

Le présent livrable est composé des rapports et annexes suivants :

- « Le FastCAM : une alternative au FlowCAM ? ». F. Colas & M. Tardivel, 2014.
- « Zoo/PhytoImage version 5 : Manuel Utilisateur ». Ph. Grosjean, K. Denis & G. Wacquet, Décembre 2014.
- « Dénombrement des cellules dans les colonies ». G. Wacquet & Ph. Grosjean, Novembre-Décembre 2014.
- « Apprentissage actif, training set adaptatif ». G. Wacquet & Ph. Grosjean, Juin-Août 2014.
- « Mise en forme des résultats Zoo/PhytoImage, en vue de l'intégration dans Quadrige² ». G. Wacquet & Ph. Grosjean, Novembre-Décembre 2014.
- Annexe 1 : « Compte-rendu du Comité de Pilotage du projet FlowCAM/ZooPhytoImage – VisioConférence », 28 mai 2014.
- Annexe 2 : « Compte-rendu du Comité de Pilotage du projet FlowCAM/ZooPhytoImage – Réunion et Formation », 2-3 décembre 2014.
- Annexe 3 : « Diapositives – Le FastCAM : une alternative au FlowCAM ? ». F. Colas & M. Tardivel, 2014.
- Annexe 4 : « Diapositives – Le système FlowCAM/ZooPhytoImage, évolution de l'observation et de la surveillance du phytoplancton ». A. Lefebvre, G. Wacquet, Ph. Grosjean, N. Neaud-Masson, D. Maurer & C. Belin, Journées REPHY 2014 - Nantes.
- Annexe 5 : « Diapositives – Zoo/PhytoImage version 5. Simplification de l'interface pour utilisation REPHY en routine ? ». Ph. Grosjean, Journées REPHY 2014 - Nantes.

Ce livrable est complémentaire d'un autre livrable fourni pour cette action intitulé : « Mise en œuvre opérationnelle de l'outil FlowCAM/ZooPhytoImage dans le cadre de la surveillance REPHY ».

Unité de Recherches et Développements Technologiques
Laboratoire Détection, Capteurs et Mesures
IFREMER

Le FastCAM : une alternative au FlowCAM ?

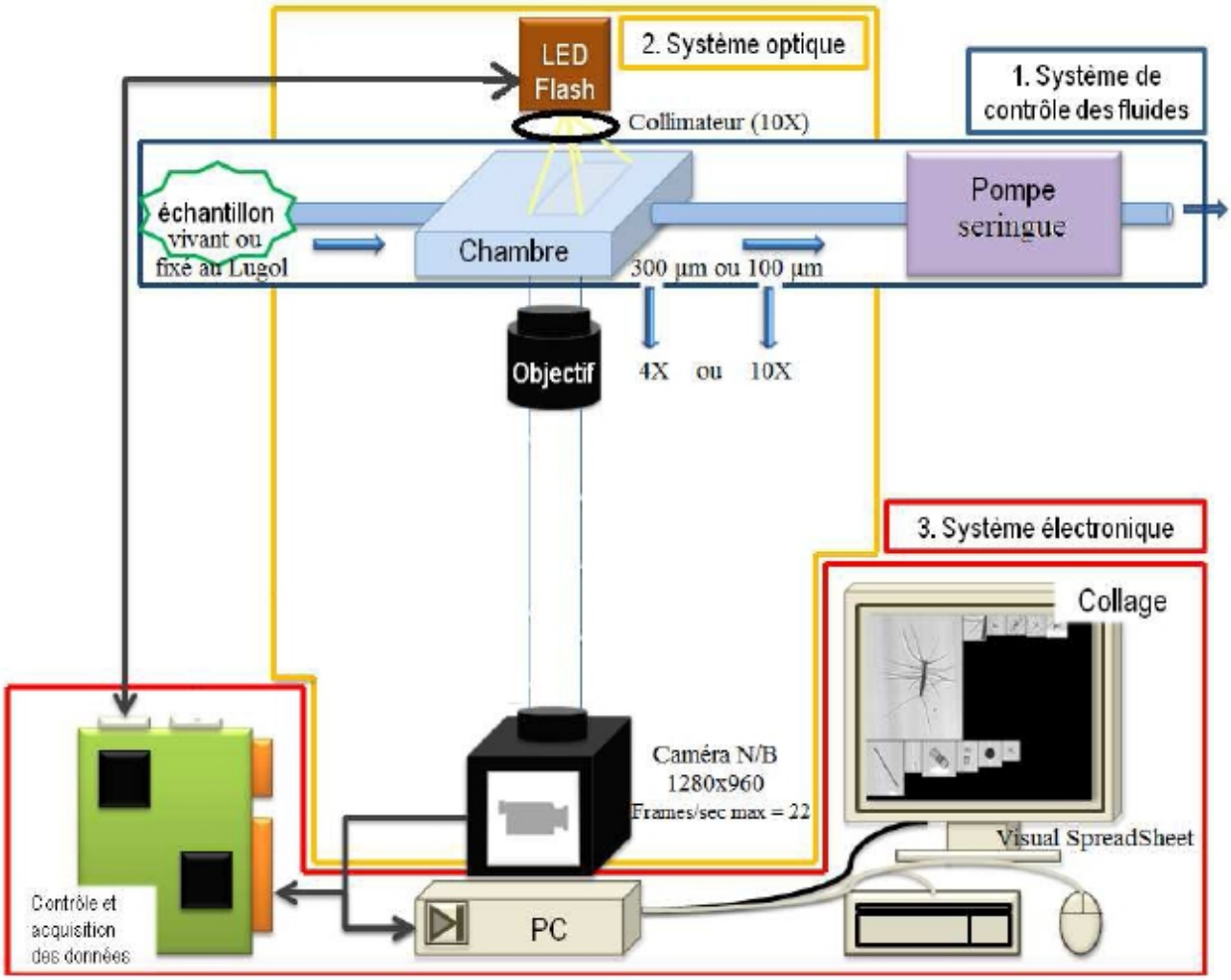
Morgan TARDIVEL, Bertrand FOREST, Marie-Pierre CRASSOUS,
Michel LUNVEN, Marie-Madeleine DANIELOU & Florent COLAS

2014

Table des matières

Introduction

Le FlowCAM (développé par la société Fluid Imaging Technologies®) est un système d'imagerie en flux de particules (figure 2) dont la taille varie de quelques micromètres à quelques millimètres. Les clichés automatiquement enregistrés permettent une classification des particules suivant des critères morphologiques. Cette opération peut être réalisée directement par le logiciel fourni avec le système (Visual SpreadSheet) ou par d'autres, tel que Zoo/PhytoImage, logiciel développé par l'Université de Mons.



Le temps d'analyse d'un échantillon est un point clef de ce système. En effet, pour l'analyse du phytoplancton deux grandissements sont possibles : 10X et 4X. Le premier permet une meilleur description morphologique mais avec un temps d'acquisition, bien plus long que le second. Un gain considérable serait de pouvoir réaliser une acquisition au grandissement 10X avec un temps d'analyse plus court, comparable au 4X. Des pistes d'évolution du système seront étudiées dans ce sens.

Vitesse de numérisation

Le volume à numériser doit être choisi pour que la mesure soit représentative du milieu. Historiquement, les analyses du REPHY reposent sur l'observation de 10 mL au microscope optique. La transposition de ce protocole au FlowCAM peut se faire en considérant que le volume à numériser doit être de 10 mL. Cependant, l'analyse est alors très longue. Par ailleurs, l'approche

FlowCAM/ZooPhytoImage impose une contrainte supplémentaire : l'analyse d'un échantillon est considérée statistiquement représentative si au moins 2000 particules sont comptées et analysées.
Le débit de numérisation du FlowCAM est donné par :

$$\oint_{num} = V_i . N$$

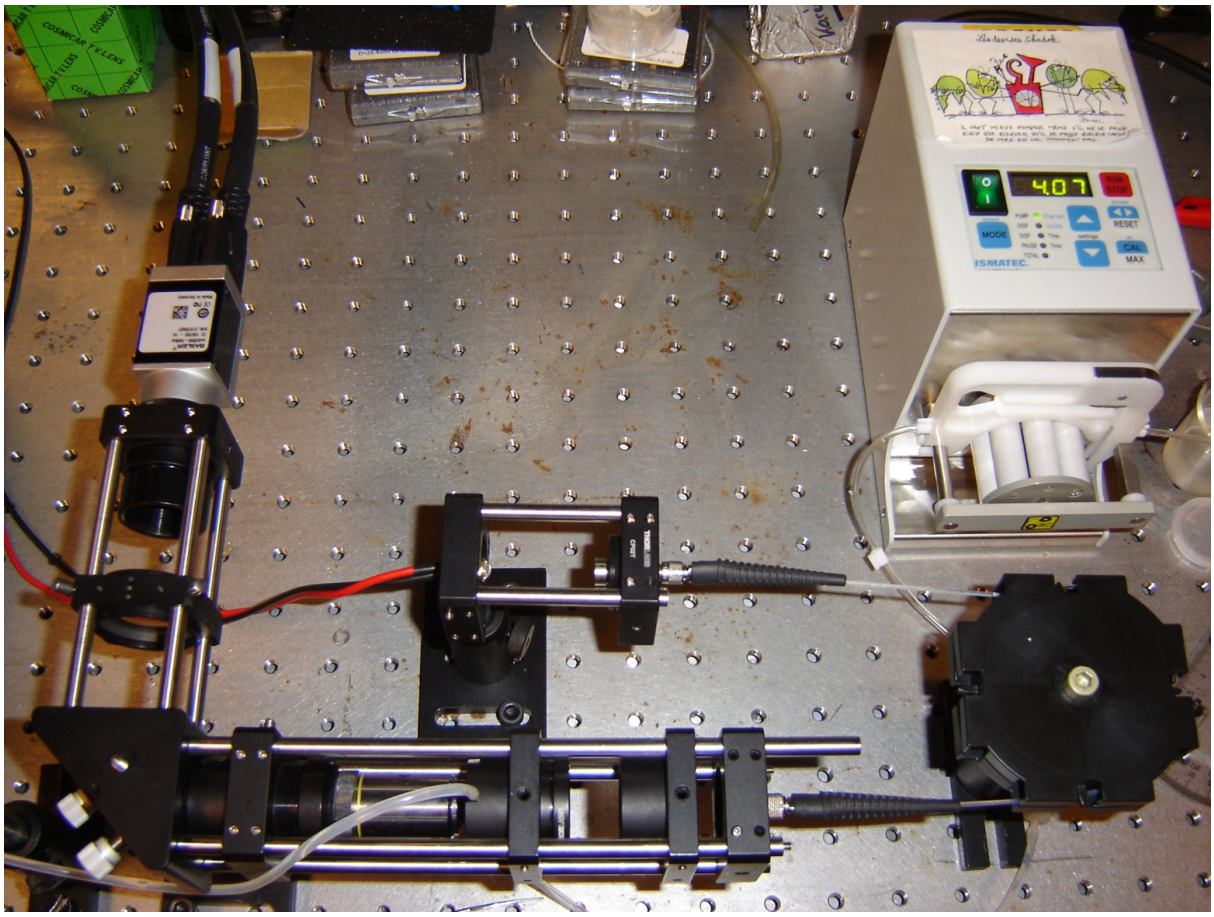
avec V_i correspondant au volume imagé et N le nombre de trames par minute.

	4X	10X
$V_i\ (\mu L)$	0,9	0,05
$\oint_{num}\ (mL/min)$	1,2	0,07
t_{10mL} (min)	8,3	143

De cette équation, le temps de numérisation minimal de 10 mL peut être calculé pour les deux grandissements. Il faut donc environ 143 min avec l'objectif 10X et seulement 8,3 min avec le 4X à la fréquence d'acquisition maximale du FlowCAM de 22 images/s.

Le FastCAM

Le FlowCAM ne permet pas une numérisation rapide avec un grandissement 10X, ce qui nuit à la résolution taxonomique de la classification. C'est pourquoi le Laboratoire Détection, Capteurs et Mesures de l'IFREMER a développé un prototype de matériel appelé « FastCAM », beaucoup plus rapide en acquisition que le FlowCAM.



Vitesse de numérisation

Pour accélérer la numérisation des échantillons un moyen efficace est de travailler avec une caméra plus rapide. En effet, le FastCAM est équipé d'une caméra haute résolution (1024x2048 pixels) capable d'acquérir des images à une fréquence de 340 image/s. Dans ce contexte, le temps d'acquisition d'un échantillon avec le couplage 10X/100µm est de l'ordre de 13 minutes soit une réduction du temps par un facteur 11.

Cependant, pour le moment, l'utilisation d'un tel dispositif posent alors d'autres problèmes techniques :

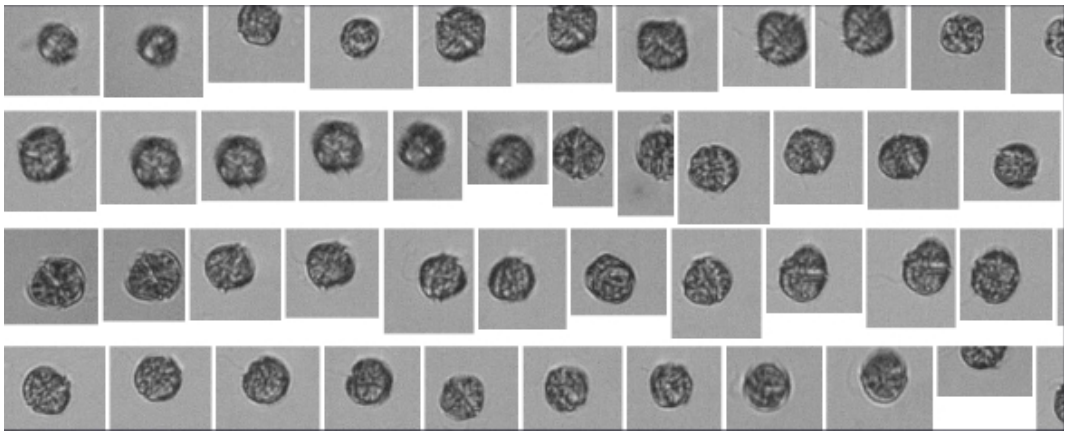
- le flux brut de données à sauvegarder est de l'ordre de 700Mo/s, ce qui est actuellement assuré en sauvegardant les données brutes sur la mémoire vive de l'ordinateur, puis en les transférant sur un disque dur rapide,
- le post-traitement des données demande un temps conséquent (environ 20 à 30 minutes).

Qualité des images

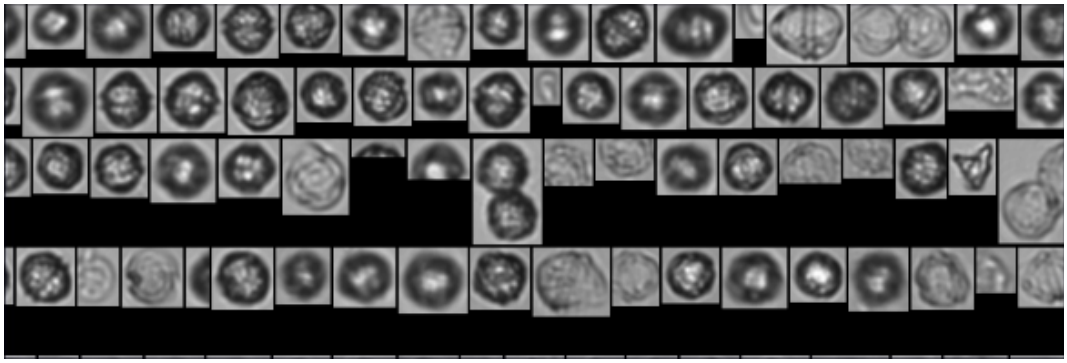
En parallèle de la réduction du temps de numérisation, une étude a été menée sur l'optimisation de la qualité des vignettes acquises. Un travail a alors été fourni sur l'optique (et en particulier sur l'illumination) afin d'obtenir une ouverture numérique (c'est-à-dire une résolution) optimale. La difficulté, ici, est de garantir une image nette pour toutes les cellules passant dans la cuve tout en assurant une bonne résolution.

Voici une comparaison visuelle des vignettes acquises avec le FlowCAM et le FastCAM, pour une espèce donnée (ici, *Alexandrium minutum*) :

FastCAM



FlowCAM



Conclusion

L'avancement du projet FastCAM, est le suivant :

- au niveau optique : les travaux effectués semblent donner des résultats satisfaisants avec des espèces dont le diamètre est de l'ordre de 20-30 µm.
- Au niveau informatique : actuellement, en ce qui concerne la gestion des paramètres et des vignettes, les fichiers générés en sortie du FastCAM sont de format ZID (format associé au logiciel Zoo/PhytoImage).
- Au niveau mécanique : le packaging est actuellement celui d'un prototype de laboratoire. Avant de continuer le développement, il est nécessaire de valider l'ensemble du système mais également le processus de classification avec Zoo/PhytoImage.

Perspectives

Dans un premier temps, il apparaît important de comparer les systèmes afin de tirer les avantages et inconvénients du FastCAM par rapport au FlowCAM. Il faudra notamment comparer le temps de calcul nécessaire (par l'ordinateur) par échantillon. En effet, dans le cas du FastCAM/ZooPhytoImage, tous les calculs seront réalisés après l'acquisition (post-traitement des données acquises). Dans le cas du FlowCAM/ZooPhytoImage, une partie des calculs est réalisée pendant le passage de l'échantillon et une seconde partie après le passage de l'échantillon.

Les premiers tests de classification ont commencé en octobre 2014 comme prévu. Certaines espèces, comme celles du genre *Pseudo-Nitzschia,* ont montré une mauvaise reconnaissance avec le couplage 10X/100µm. En effet, les images sont apparues de mauvaise qualité suivant le petit axe des cellules. Il est donc nécessaire d'améliorer l'optique pour permettre une meilleure reconnaissance de ces espèces (et en particulier, en comparaison avec les résultats obtenus par le FlowCAM équipé d'un objectif 4X).

Il est également recommandé de tester le FastCAM non plus uniquement sur des cultures ou des échantillons mono-spécifiques, mais sur des échantillons ou cultures pluri-spécifiques afin de se mettre dans les conditions d'applications futures.

Écologie Numérique des Milieux Aquatiques
UMONS
Faculté des Sciences

Zoo/PhytoImage Version 5
Manuel Utilisateur

Philippe GROSJEAN & Guillaume WACQUET

Zoo/PhytoImage version 5

Analyse d'Images de Plancton Assistée par Ordinateur

Manuel Utilisateur

L'équipe de développement de ZooImage
Décembre 2014

Ph. Grosjean, K. Denis & G. Wacquet: Écologie Numérique des Systèmes Aquatiques, UMONS, Belgique
X. Irigoien, G. Boyra & I. Arregi: AZTI Tecnalia, Espagne
A. Lopez-Urrutia: Centro Oceanográfico de Gijón, IEO, Espagne
M. Sieracki & B. Tupper (FlowCAM plugin)

1. INTRODUCTION

L'analyse d'échantillons zooplanctoniques ou phytoplanctoniques est traditionnellement associée à de longues et fastidieuses séances de comptage des particules fixées de plancton sous binoculaire et avec des vapeurs de formaldéhyde flottant autour. Bien que cette image du planctonologiste restera probablement pendant un certain temps, il semble y avoir une autre façon de recueillir des données sur le zooplancton : l'analyse assistée par ordinateur d'images numériques de plancton. Toute une gamme de matériel pour prendre des photos de nos animaux, à la fois *in situ* et/ou à partir d'échantillons fixés, est maintenant disponible : FlowCAM, OPC laser, VPR, Zooscan, ... (plus, à venir, l'holocam, Sipper, Zoovis, bouée HAB, ...), sans oublier l'utilisation d'un appareil photo numérique sur binoculaire ou avec un macro objectif. Cependant, les images numériques de zooplancton sont à peine utilisables en tant que telles : elles doivent être analysées de manière à extraire des attributs biologiquement et écologiquement significatifs à partir des pixels. Un logiciel permettant de réaliser une telle analyse est donc indispensable.

Zoo/PhytoImage a pour objectif de fournir une solution puissante et riche en fonctionnalités logicielles pour utiliser les images de zooplancton ou phytoplancton provenant d'origines diverses et les transformer en une table de mesures utilisables (c'est-à-dire, les abondances, les spectres de taille totaux et partiels, les biomasses totales et partielles, .. .). Zoo/PhytoImage n'est pas fermé à l'un des dispositifs cités précédemment, et n'est pas un produit commercial. Il est distribué gratuitement (licence GPL, distribuée à travers son site web, http://www.sciviews.org/zooimage) et est ouvert, ce qui signifie qu'il fournit un cadre général pour importer des images, les analyser et exporter les résultats à partir et vers un grand nombre de systèmes. Donc, tout le monde peut utiliser Zoo/PhytoImage... mais mieux encore, chaque développeur peut également y contribuer! L'approche Open Source de câblage de nombreux développeurs à travers le monde dans un projet commun a déjà montré son efficacité : Linux, Apache, mais aussi R ou ImageJ dans le domaine des statistiques et de l'analyse d'image respectivement, sont de bons exemples. Zoo/PhytoImage est basé sur ImageJ et R, et fonctionne sur Linux ... mais il peut aussi être exécuté sur Windows, Mac OS ou diverses Unixes[1]. La meilleure qualification de Zoo/PhytoImage est sa "réutilisation". Il est né en réutilisant diverses caractéristiques de logiciels existants comme ImageJ, ou R, et fournit lui-même des composants réutilisables, au bénéfice des utilisateurs et des développeurs.

Zoo/PhytoImage peut être utilisé sur des images acquises dans différentes situations : *in situ* (comme le VPR ou la bouée HAB) ou dans un laboratoire (échantillons fixés numérisés avec le Zooscan, par exemple). Le cadre général de Zoo/PhytoImage est conçu de manière à ce que le logiciel soit capable de traiter efficacement des images de caractéristiques et d'origines diverses. Par conséquent, ce n'est pas un système rationalisé et rigide. Il est plutôt constitué d'un ensemble d'applications différentes et personnalisables rassemblées en un seul système. Ce manuel utilisateur vous guidera dans votre première utilisation de Zoo/PhytoImage.

*Ce manuel décrit la version actuelle de ZooImage (5.1-0), qui est une version publique! Il est adapté aux besoins de nos partenaires: UMONS, IFREMER, Belspo, ULCO et LISIC. 4/5 du code est commun avec la version 3, qui est publique et téléchargeable à partir du site du CRAN (*http://cran.r-project.org*).*

1 La version courante est développée principalement sur MacOS X, mais a été également testée sur Windows et Linux.

2. CHANGEMENTS PAR RAPPORT AUX VERSIONS 3 ET 4

La version 3.0 de Zoo/PhytoImage est la dernière version publique distribuée sur http://www.sciviews.org/zooimage jusqu'à présent. La version 4 du logiciel n'était pas publique et contenait plusieurs développements réalisés pour nos besoins (université UMONS) et pour nos principaux partenaires : l'IFREMER en France et Belspo (Politique Scientifique Belge) en Belgique.

La version 5 de Zoo/PhytoImage contient la plupart de ces développements dans un système revisité, et est distribué sur le site du CRAN (http://cran.r-project.org). Enfin, les récentes nouveautés apportées dans la version 5 complète l'ensemble des fonctionnalités. Les principales modifications sont les suivantes :

- Refonte du code pour l'exécuter sur la dernière version de R (version 3),
- Refonte de l'interface graphique pour une meilleure ergonomie et une simplicité d'utilisation,
- Développement de routines pour importer et analyser les données automatiquement.

Parmi ces changements, le plus important pour les utilisateurs finaux est probablement la nouvelle interface graphique utilisateur. Cette dernière a été étudiée du point de vue de son adéquation pour des traitements spécifiques liés aux besoins des principaux partenaires, et des modifications ont déjà été apportées dans la version 3. Il apparaissait cependant désirable d'augmenter considérablement l'interactivité visuelle avec le logiciel, notamment au niveau de la visualisation des vignettes, des données brutes et de l'automatisation de certaines tâches. Pour cela, une refonte complète de l'interface graphique utilisateur de Zoo/PhytoImage sur base de définition des "use cases" dans le cadre d'une perspective d'exploitation en routine, est nécessaire.

Un des objectifs principal de la refonte de l'interface graphique utilisateur de Zoo/PhytoImage réside dans la réduction du nombre de tâches de l'utilisateur pour l'importation et la mise en forme des données. A cette fin, les outils de traitement d'images et de transformation des données brutes sont appliqués implicitement.

Dans Zoo/PhytoImage version 5, l'interface graphique est construite à l'aide du package Shiny. Celui-ci permet de créer facilement des applications interactives web avec R. C'est pourquoi, il est nécessaire d'avoir un navigateur internet installé sur votre machine afin de lancer l'interface graphique utilisateur.

3. INSTALLATION ET EXECUTION

3.1. Exigences matérielles

L'analyse d'images et la classification automatique des images sont des processus informatiquement intenses, et vous aurez probablement à analyser beaucoup d'objets (généralement des centaines de milliers, voire des millions). Ainsi, vous aurez besoin d'un ordinateur récent et puissant pour exécuter Zoo/PhytoImage décemment. En particulier :

- Un microprocesseur multi-coeur récent et rapide, et un processeur multithreads.
- **4Gb de mémoire RAM** ou plus. Selon la taille des images que vous voulez analyser, vous pourrez avoir besoin de plus de mémoire. Les très grandes images issues d'un scanner à plat requièrent au moins 1Gb de RAM. Les images du Zooscan peuvent en requérir plus! Aujourd'hui, il est très facile d'utiliser 16Gb ou 32Gb de RAM sur des systèmes 64-bits, donc envisagez sérieusement cette option.

- Après la vitesse de processeur et la RAM, la partie suivante la plus importante de l'ordinateur pour travailler sur les images, est **la carte graphique et l'écran**. Choisissez une carte graphique rapide et optimisée permettant l'affichage de 1280×1024, ou 1600×1200 pixels ou plus avec une profondeur de couleur 24/32 bit (millions de couleurs), associée à un écran haute qualité de pas moins de 19'. Une configuration double-écran peut également aider car il permet d'avoir plus d'espace pour afficher côte-à-côte les images et les graphiques.
- Bien que Zoo/PhytoImage optimise l'espace disque en compressant tous les fichiers, traiter un nombre important d'images haute résolution consomme beaucoup d'espace sur le disque. Vous avez donc besoin d'un **disque dur rapide d'une capacité d'au moins 2-4Tb**. Un petit disque SSD augmente considérablement la vitesse d'analyse lorsqu'il est utilisé pour stocker les quelques échantillons qui sont en cours d'analyse.
- Finalement, un bon **système de sauvegarde** est également requis, sauf si vous utilisez un système RAID.

3.2. Installation de Zoo/PhytoImage sous Windows

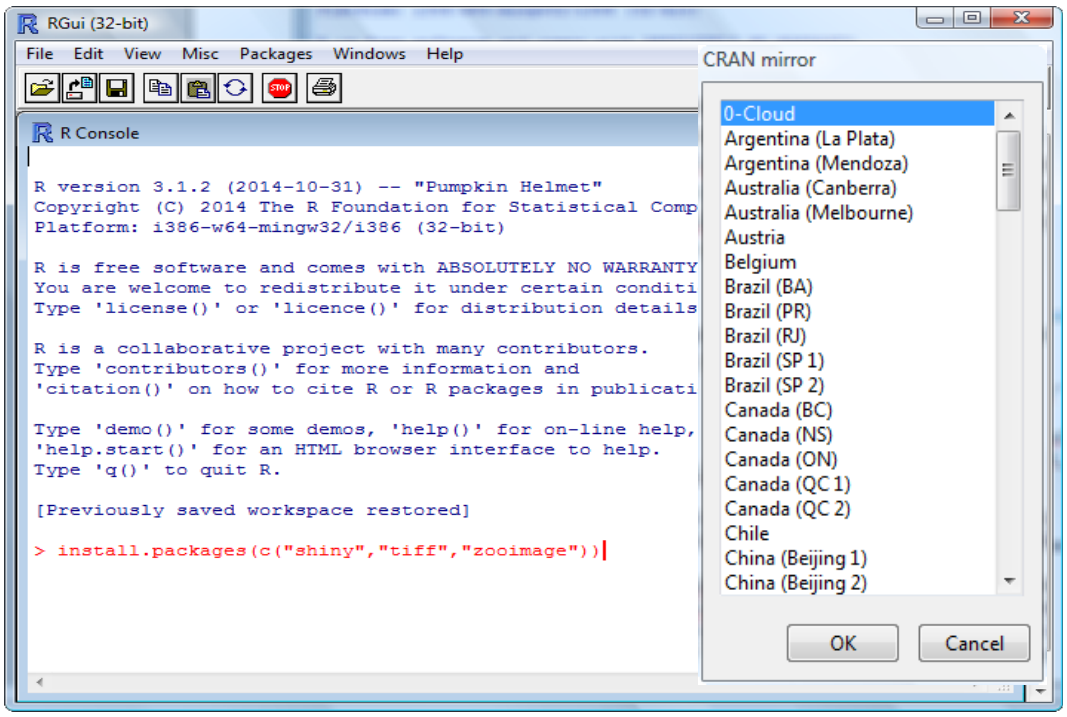
La version 5 de Zoo/PhytoImage nécessite une version récente de R[2] (version 3.0.x ou plus). Elle peut être téléchargée directement sur le site du CRAN (http://cran.r-project.org).

Pour le lancement de l'interface graphique utilisateur interactive, il est également nécessaire d'installer un navigateur internet compatible avec votre système d'exploitation. Nous conseillons vivement d'utiliser Safari ou Google Chrome et de le définir en tant que navigateur par défaut (pour que l'interface s'affiche automatiquement dans ce navigateur).

Lorsque vous double-cliquez sur l'icône de R sur le bureau, ou en sélectionnant l'entrée R dans le menu de démarrage, une fenêtre apparaît à l'écran : la console R. Cette dernière vous permet de contrôler R directement par lignes de commande. Vous ne devez pas vous soucier de cette fenêtre, sauf si vous êtes familier avec le langage R. Cependant, il enregistre les résultats et les messages importants de vos actions dans Zoo/PhytoImage.

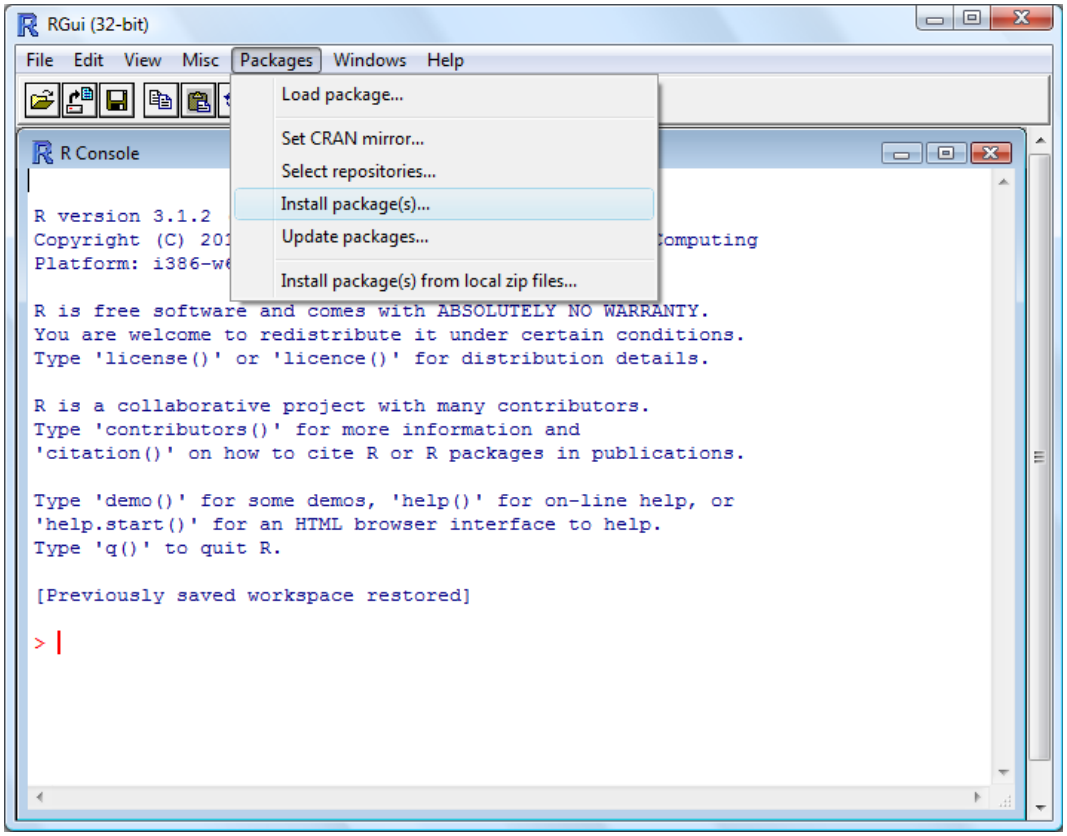
Les packages nécessaires (« shiny », « tiff » et « zooimage ») peuvent être installés directement à partir de la console R, en tapant : **install.packages(c("shiny","tiff","zooimage"))**. Choisissez ensuite un miroir (par défaut : 0-cloud) pour démarrer les téléchargements et les installations.

2 R est le logiciel de statistiques et l'environnement avec lequel ZooImage est développé.

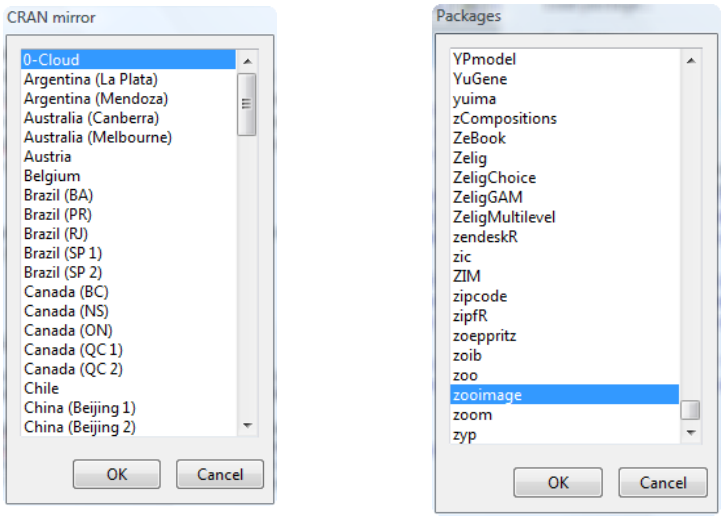


Capture d'écran de la console R pour l'installation de Zoo/PhytoImage version 5.

Il est également possible d'installer Zoo/PhytoImage manuellement. A partir du menu "Packages" → "Installer le(s) package(s)", sélectionnez un miroir de téléchargement (par défaut : 0-cloud), puis les packages "shiny", "tiff" et "zooimage".

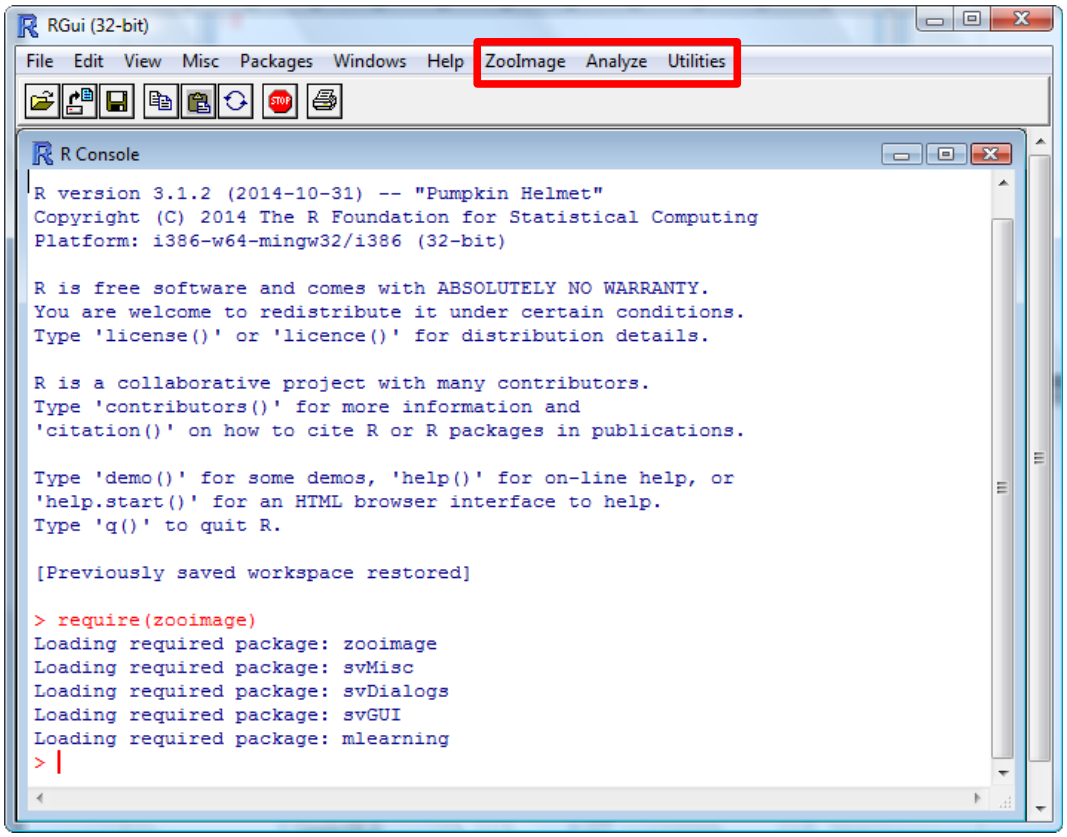


Capture d'écran pour l'installation manuelle de Zoo/PhytoImage version 5.



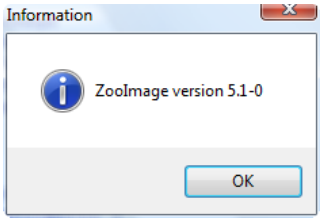
Sélection du miroir de téléchargement et des packages nécessaires à l'installation de Zoo/PhytoImage version 5.

Une fois l'installation des packages terminée, il est possible de s'assurer du bon déroulement des étapes précédentes en vérifiant que la version installée est bien 5.1-0. Pour cela, dans un premier temps, tapez dans la console R : **require(zooimage)**, pour lancer Zoo/PhytoImage. Trois nouvelles entrées dans la barre de menu de R sont alors ajoutées.



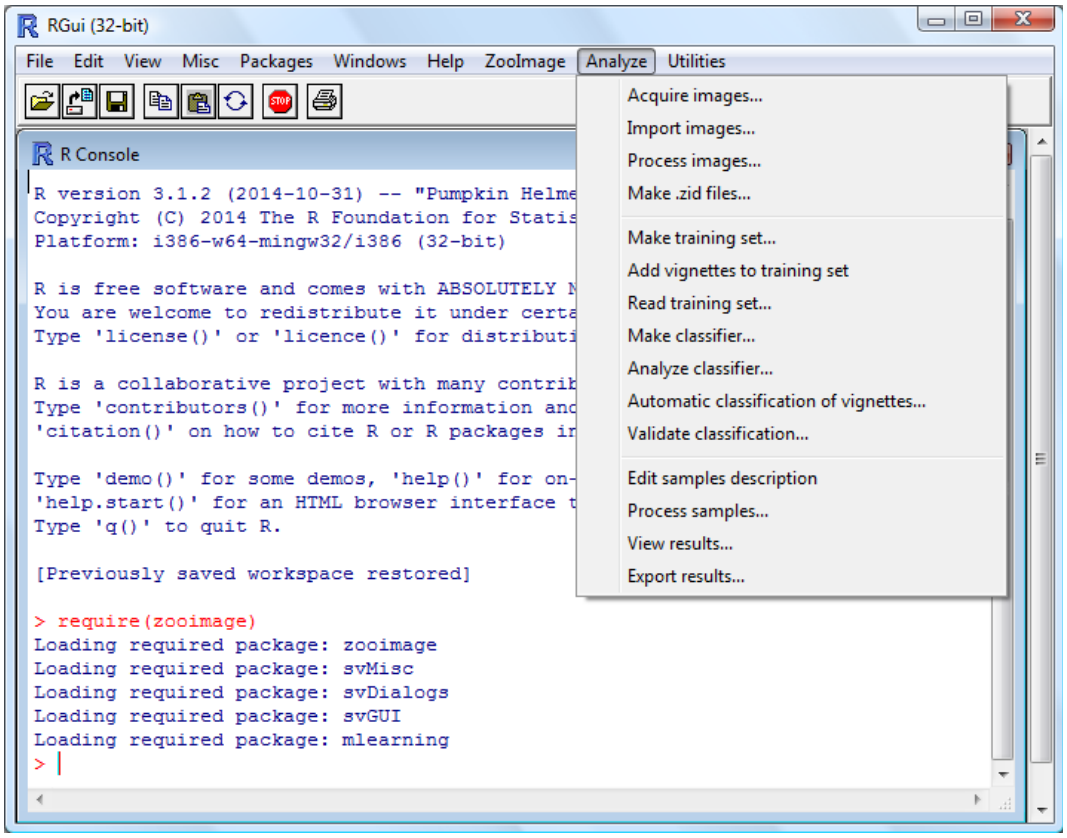
Lancement de Zoo/PhytoImage et ajout de nouvelles entrées dans la barre de menu de R.

En sélectionnant, dans le menu "ZooImage", "About...", une boîte de dialogue s'affiche et informe l'utilisateur de la version de Zoo/PhytoImage en cours d'exécution.



4. UTILISATION DES FONCTIONS A PARTIR DU MENU DE ZOO/PHYTOIMAGE

Zoo/PhytoImage peut être vu comme une boîte à outils permettant de réaliser toutes les étapes du processus de classification. Dans le menu "Analyze", toutes les fonctions nécessaires à la reconnaissance (semi-)automatisée des images sont disponibles.



Menu « Analyze » contenant toutes les fonctions nécessaires au processus de classification.

Une analyse Zoo/PhytoImage peut être subdivisée en trois parties :

- La première partie concerne l'importation et le traitement d'images.

  1. **Acquire images...** Lance un logiciel d'acquisition externe (Vuescan, ou tout autre programme).

  2. **Import images...** Possibilité de convertir le format des images et/ou de les renommer. Si les images sont déjà dans le format correct, cette fonction permet juste de s'assurer que des métadonnées adaptées leur sont associées.

  3. **Process images...** Fondamentalement, ImageJ est lancé. Vous êtes censés utiliser un des plugins ZooImage spécifiques dans ImageJ pour traiter vos images

  4. **Make .zid files...** Les fichiers « Zid » sont des fichiers « ZooImage Data ». Ils contiennent tout ce dont vous avez besoin pour le reste du traitement, c'est-à-dire les images de chaque individu[3], leurs mesures et les métadonnées. Toutes ces informations sont sont alors compressées[4].

3 Ces images particulières sont appelées « vignettes » dans la terminologie ZooImage.

4 Si vous commencez avec des images en niveau de gris 16 bits au format TIFF non compressées et haute résolution, vous obtenez généralement des fichiers .zid ayant un poids d'environ 100 fois moins que les images originales.

- La seconde partie vous permet de générer un outil de reconnaissance automatique et optimisé pour votre série planctonique.

  1. **Make a training set...** Cette fonction prépare un répertoire avec une hiérarchie de sous-répertoires représentant votre classification manuelle (vous pouvez modifier librement cette structure) et extrait les vignettes des échantillons que vous voulez utiliser pour construire votre ensemble d'apprentissage manuellement. Ensuite, vous devez les classer manuellement sur l'écran en les déplaçant dans leurs répertoires respectifs avec la souris.

  2. **Add vignettes to training set.** Cette fonction permet de compléter un ensemble d'apprentissage existant en y ajoutant des vignettes sans casser la structure de l'ensemble d'apprentissage.

  3. **Read training set...** Une fois les vignettes triées, cette fonction collecte et intègre cette information dans ZooImage. Des statistiques sur votre classification manuelle (nombre de vignettes dans chaque groupe) sont affichées.

  4. **Make classifier...** Utilisation d'un ensemble d'apprentissage manuel pour entraîner un outil de reconnaissance automatique. Vous avez le choix entre des algorithmes variés. Vous obtenez ensuite certaines statistiques à la fin du processus pour évaluer les performances de votre outil de reconnaissance (par validation croisée).

  5. **Analyze classifier...** Obtention d'autres analyses des performances de votre outil de reconnaissance. Actuellement, la matrice de confusion, les graphes de Précision/Recall, le F-Score ainsi que le dendrogramme montrant les différences entre la classification manuelle et automatique[5] sont calculés.

  6. **Automatic classification of vignettes...** Cette fonction permet de sélectionner un échantillon et de représenter la même hiérarchie de répertoires que celle utilisée dans l'ensemble d'apprentissage original, avec ses vignettes pré-triées selon la prédiction automatique fournie par l'outil de reconnaissance choisi. Cela peut être utile pour : (1) vérifier visuellement la qualité de l'outil de reconnaissance à travers les identifications des vignettes, et (2) permettre une correction manuelle (validation) de cette classification.

  - 7. **Validate classification...** Cet outil combine des outils statistiques avancés et une interface utilisateur ergonomique pour une validation simple (et partielle) de la classification. Les outils détectent des individus « suspects » et les présentent étape par étape, afin que la procédure d'optimisation soit la plus efficace possible. Typiquement , la validation de seulement un tiers de toutes les vignettes offre un rendement de même niveau qu'une validation aléatoire de 90-95% des vignettes ! Il est également combiné avec des outils de modélisation de l'erreur spécifique à l'échantillon, et de correction statistique selon ce modèle. La combinaison de la détection de suspects et de la correction d'erreur offre une amélioration de la rapidité de la validation : en validant manuellement 15-20% seulement des vignettes, il est possible d'obtenir des abondance par groupes avec typiquement moins de 10% d'erreur pour tous les groupes.

- La troisième partie utilise l'outil de reconnaissance et les mesures calculées sur tous les individus identifiés dans vos images (première partie) pour calculer automatiquement les abondances, les biomasses et les spectres de taille dans tous vos échantillons. Vous pouvez alors visualiser les résultats ou les exporter.

5 Les résultats de ces outils peuvent être affichés dans des représentations matricielles et graphiques.

1. **Edit samples description**. Des séries d'échantillons sont identifiées par une liste écrite dans un format Zoo/PhytoImage spécifique. Cette liste contient également de plus amples métadonnées à propos des séries, et vous avez l'opportunité d'ajouter de nombreuses autres mesures aux données des échantillons (température, salinité, fluorescence, etc.).

2. **Process samples...** C'est la fonction qui traite chaque échantillon d'une série donnée les uns après les autres, (1) en identifiant tous les individus en utilisant votre outil de reconnaissance automatique, (2) en calculant les abondances par taxon, (3) en calculant les classes de taille totale et par taxon pour les représentations et les études des spectres de taille, et (4) en calculant les biomasses totale et par taxon, en utilisant une table de conversion entre ECD[6] et la teneur en carbone, le poids sec, etc. Les données sont converties par m^3, si l'information « dilution » appropriée est disponible dans les métadonnées.

3. **View results...** Représentation graphique des résultats. Vous pouvez dessiner des graphique composites (jusqu'à 12 graphiques différents sur la même page) soit des séries temporelles des changements[7] d'abondances ou de biomasses, soit des spectres de taille d'échantillons données.

4. **Export results...** Les résultats sont écrits sur le disque dur dans un format ASCII. Ce format est lisible par d'autres logiciels (Excel, Matlab, etc.).

Bien que vous pouvez exporter vos résultats pour les analyser dans un logiciel différent, vous n'avez pas à le faire. Zoo/PhytoImage est exécuté dans une session R, et les milliers de fonctions de R sont disponibles pour produire des analyses statistiques et des graphiques plus sophistiqués sans quitter Zoo/PhytoImage/R.

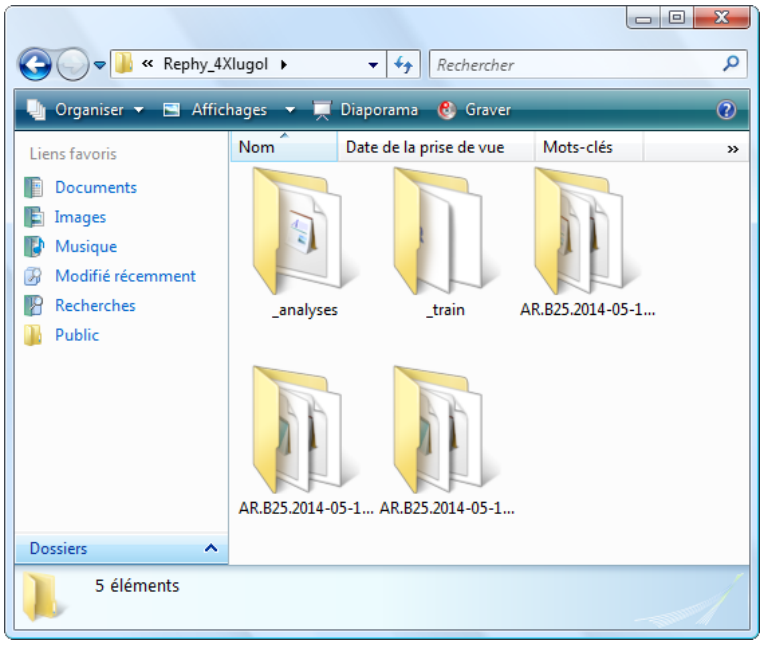
6 ECD = Equivalent Circular Diameter. (Diamètre Équivalent Circulaire)

7 Les représentation spatiales ne sont pas traitées dans cette version, mais sont prévues dans les versions futures.

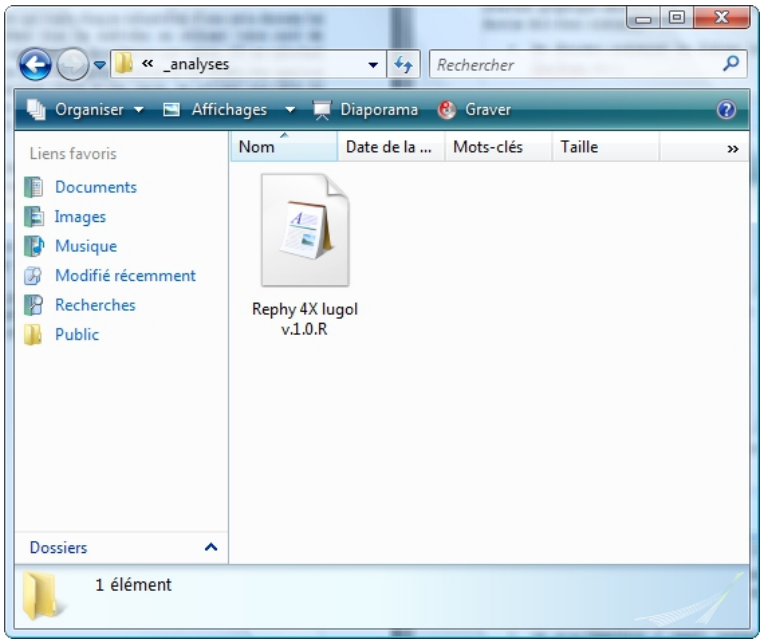
5. UTILISATION DE L'INTERFACE GRAPHIQUE UTILISATEUR DE ZOO/PHYTOIMAGE

Pour l'utilisation de Zoo/PhytoImage en routine, une interface graphique utilisateur interactive et ergonomique est disponible dans cette version. Cependant, l'utilisation de cette interface graphique nécessite une organisation spécifique des fichiers dans le répertoire de travail. Ce dernier doit donc contenir :

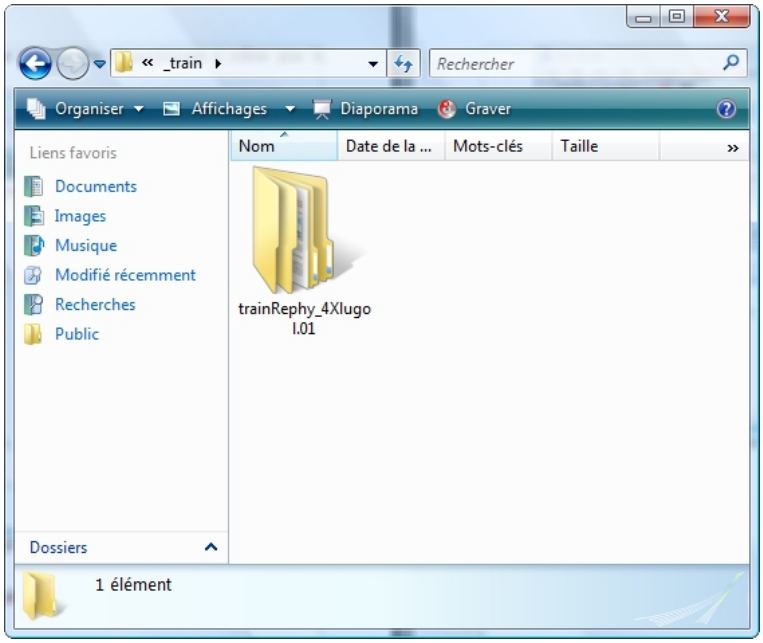
- les dossiers contenant les fichiers bruts en sortie de l'appareil d'acquisition (FlowCAM, ZooScan, etc.),



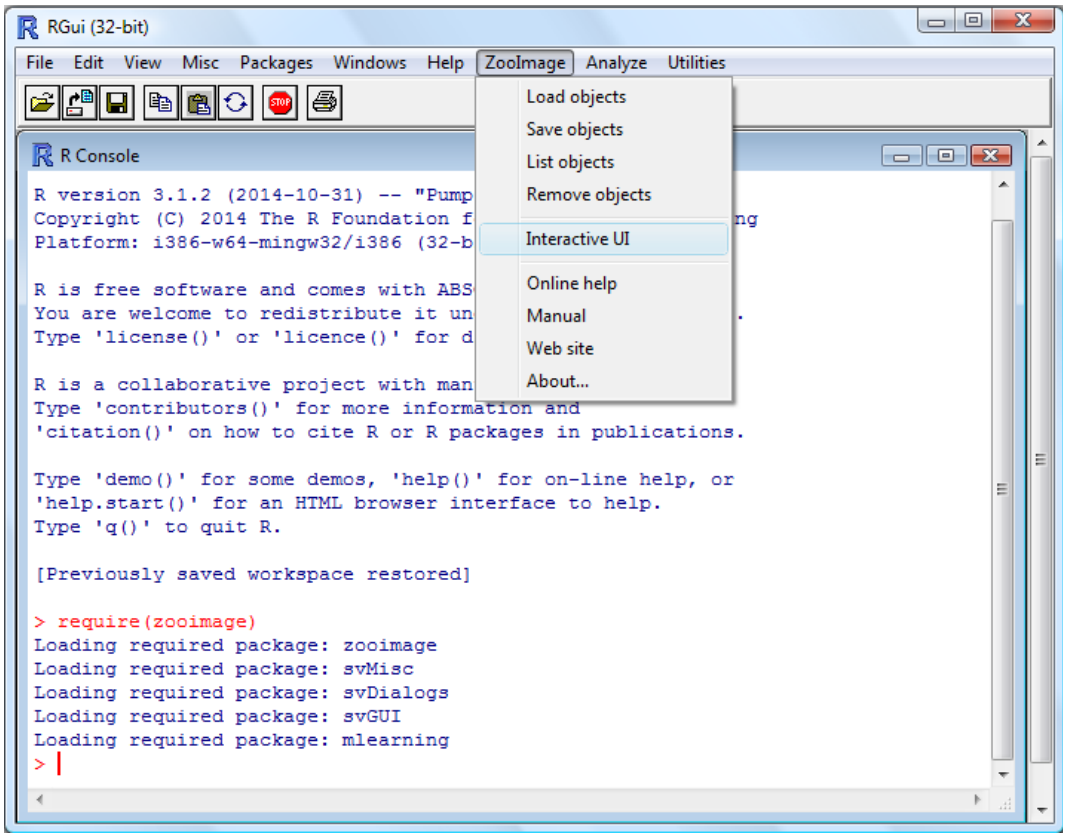
- un sous-répertoire « _analyses » contenant les fichiers méthodes R définissant les règles d'analyse des échantillons, et qui contiendra les résultats des analyses pour chacun des échantillons,



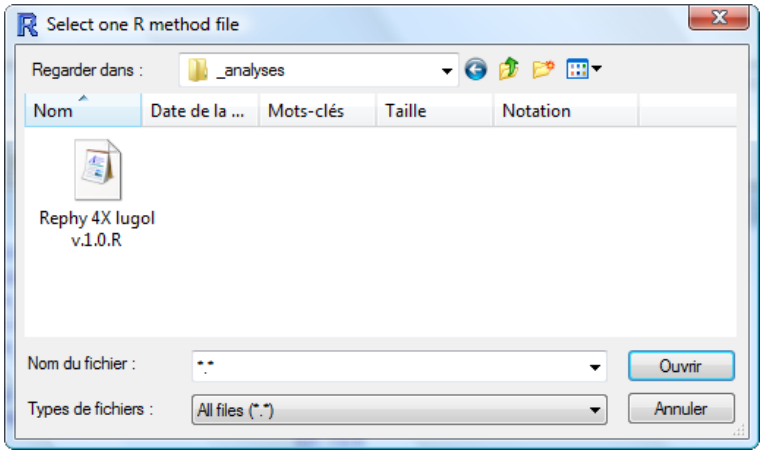
- un sous-répertoire « _train » contenant les ensembles d'apprentissages à utiliser pour la génération de l'outil de reconnaissance.



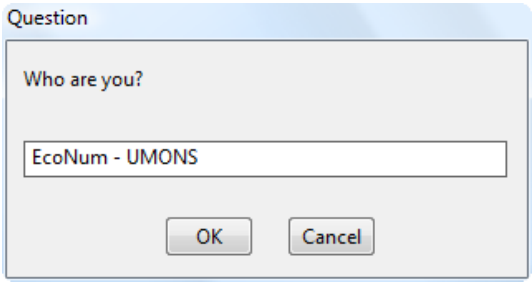
Une fois cette structure de dossiers établie, vous pouvez accéder à l'interface graphique utilisateur en sélectionnant le menu « ZooImage », « Interactive UI ».



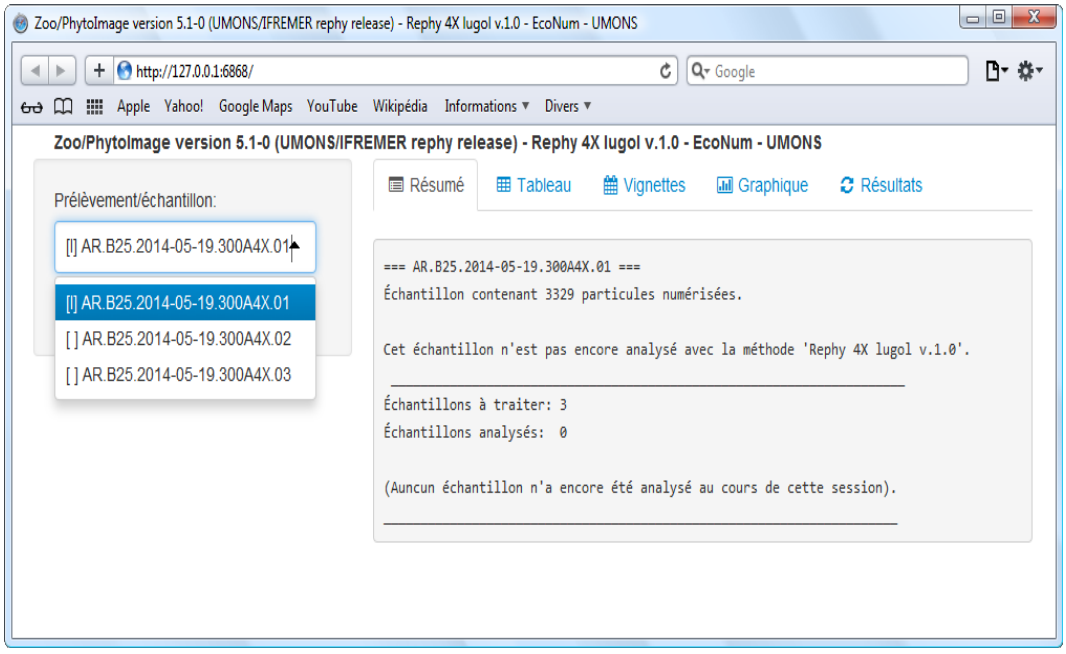
Choisissez alors le fichier méthode R (dans le sous-répertoire « _analyses ») que vous souhaitez utiliser pour le traitement des échantillons :



Sélectionnez le fichier puis cliquez sur « Ouvrir ». Le programme demande alors à l'utilisateur de rentrer son nom (ou le nom de l'organisme) dans une boîte de dialogue.



Entrez le nom d'utilisateur, puis cliquez sur « OK ». L'interface graphique utilisateur est alors lancée dans le navigateur internet installé par défaut sur votre machine (préférez Safari ou Google Chrome).



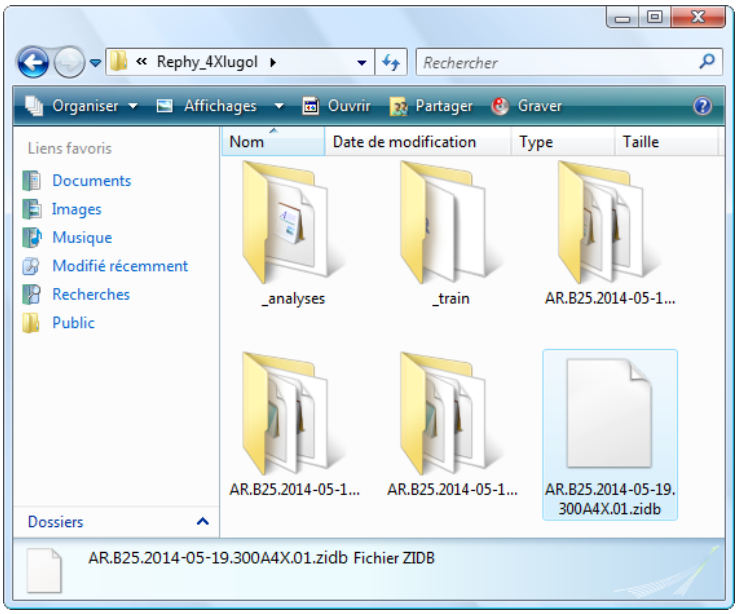
Différents modules se dégagent de cette interface :

- **Importation des données brutes dans Zoo/PhytoImage.**

  Prélèvements/Échantillons. Les échantillons présents dans le répertoire de travail sont listés ici. Chacun d'entre eux est précédé d'un code entre crochets.

  [] Échantillon non importé et donc non analysé,
  [I] Échantillon importé mais non analysé,
  [A] Échantillon importé et analysé.

  Dans cette nouvelle version de Zoo/PhytoImage, l'importation des données dans le logiciel est effectué de manière complètement automatique. Il vous suffit donc de cliquer sur l'échantillon que vous souhaitez importer afin de créer le fichier .zidb associé sur le disque (dans votre répertoire de travail). Le code précédant l'échantillon devient alors [I].



- **Visualisation des données importées.**

  Résumé. Dans cet onglet, vous pouvez retrouver des informations générales sur l'échantillon sélectionné et importé : nombre de particules numérisés, état de l'analyse avec la méthode sélectionnée, nombre d'échantillons à traiter, nombre d'échantillons analysés.

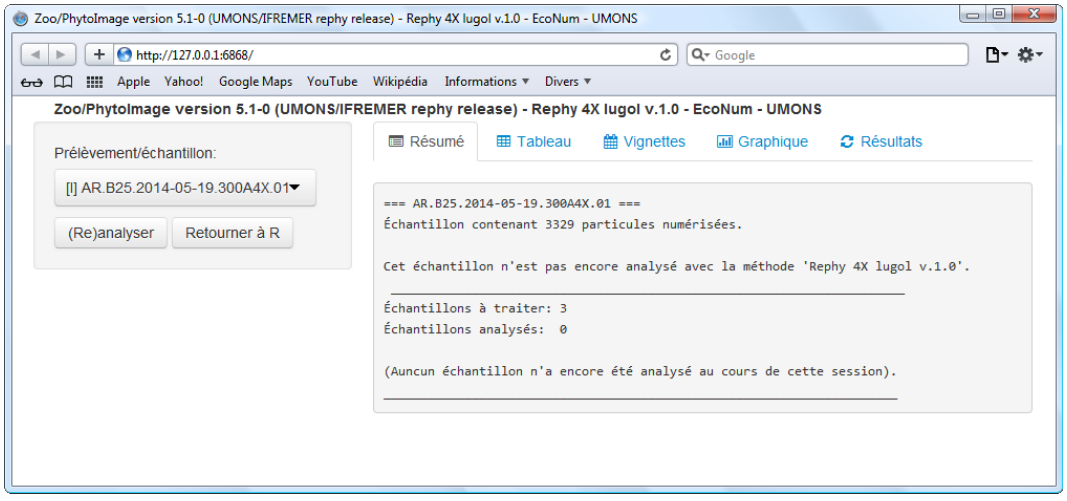
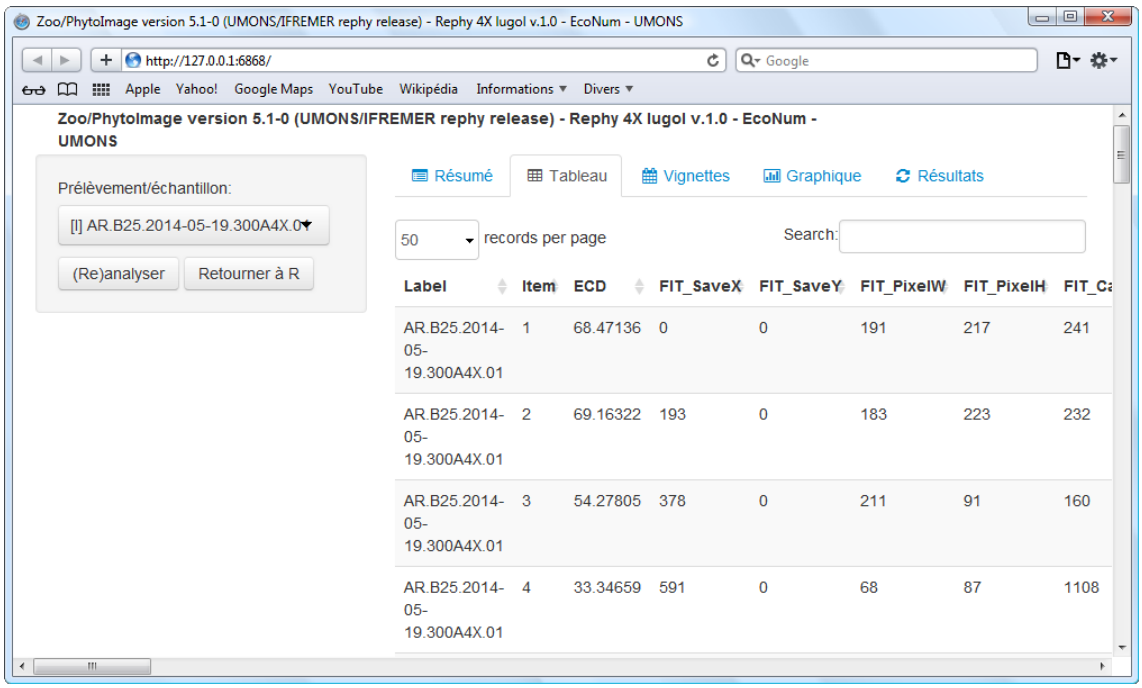
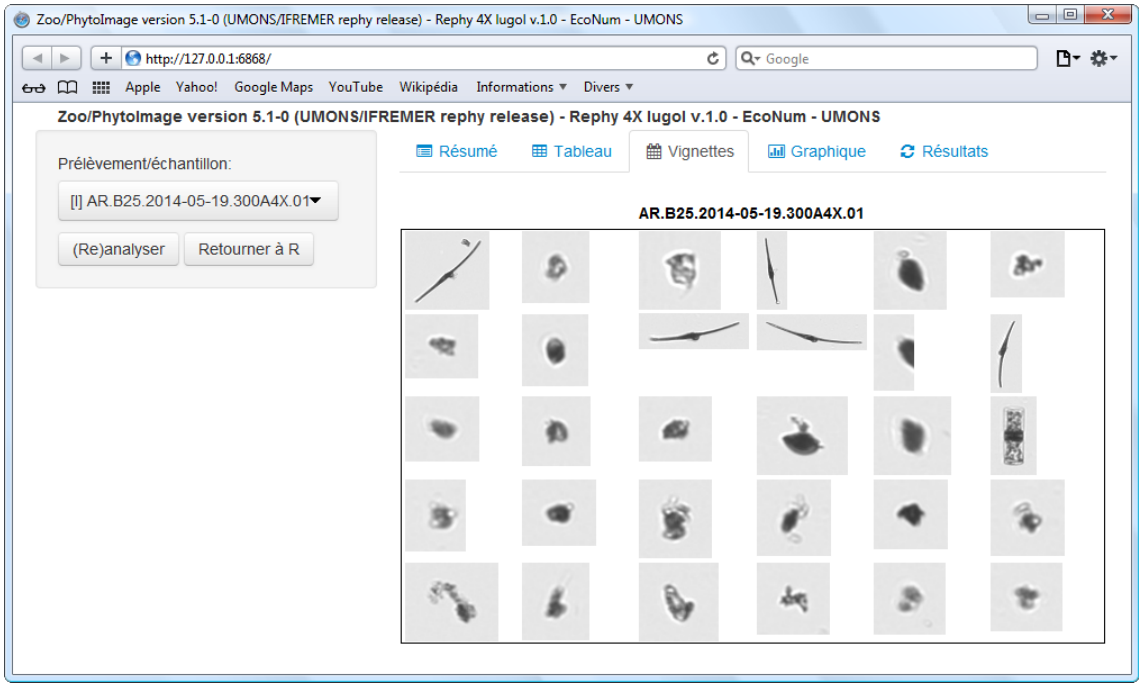


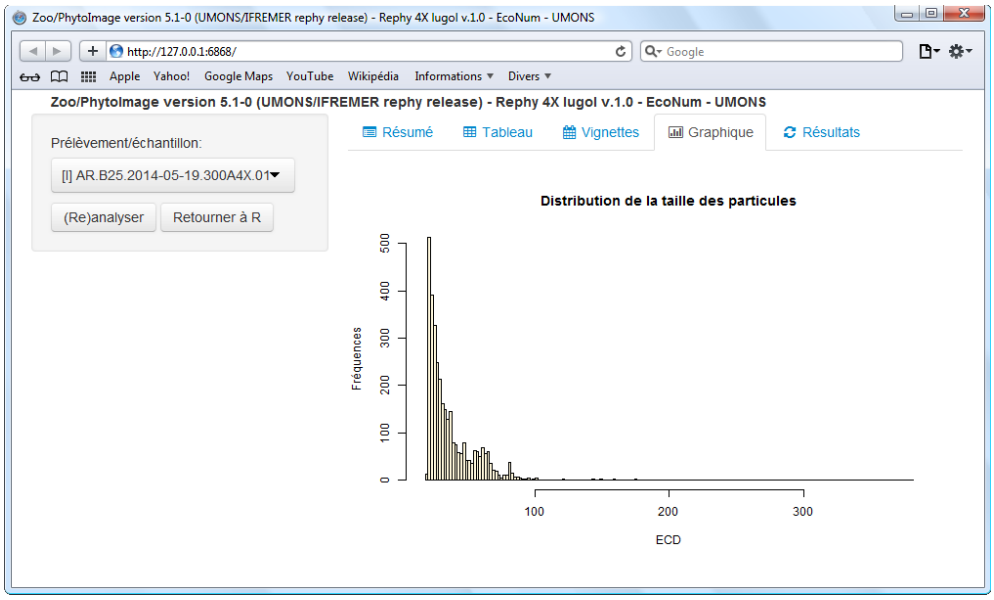
Tableau. Cet onglet permet une visualisation rapide et simple des mesures pour chacun des individus de l'échantillon. Il est également possible de trier par valeurs croissantes (ou décroissantes) les différentes colonnes du tableau.



Vignettes. Dans cet onglet sont représentées les 30 premières vignettes de l'échantillon.



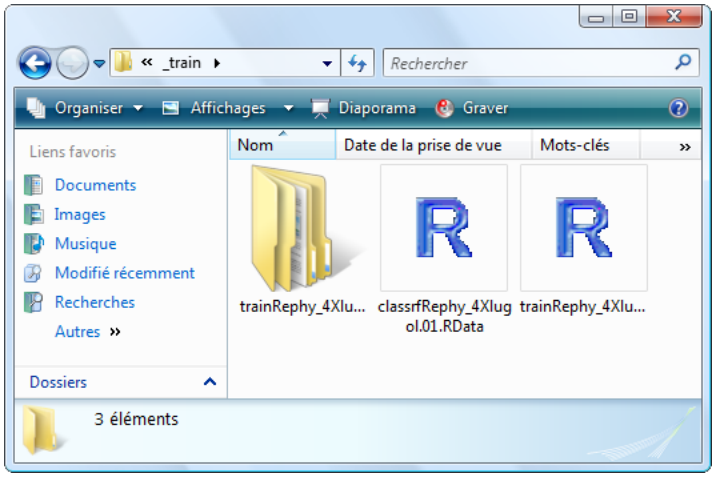
Graphique. Cet onglet permet de visualiser le graphique de distribution de la taille des particules. En abscisse est représentée la taille des particules (basée sur la mesure ECD de chacune des particules), et en ordonnée, la fréquence.



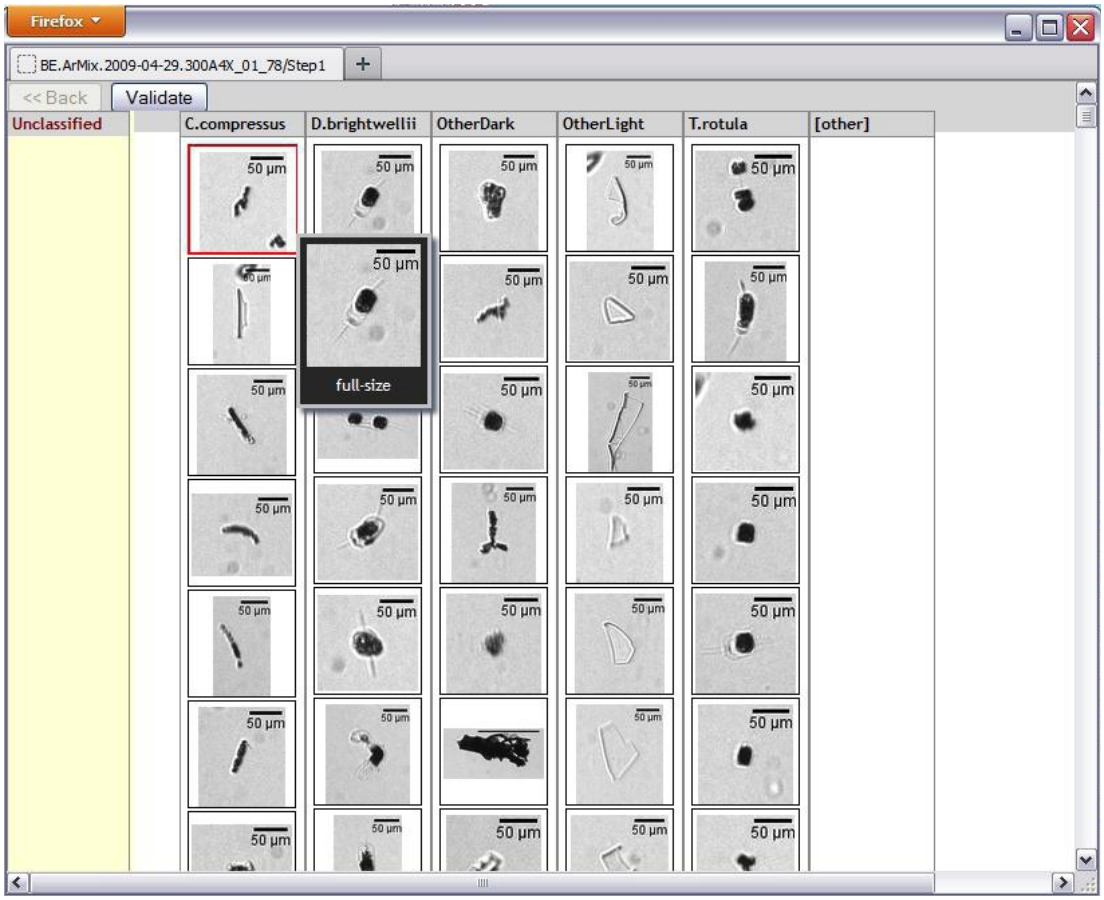
- **Analyse des échantillons et visualisation des résultats.**

  (Re)analyser. Lorsqu'un échantillon est importé, il est possible de l'analyser en le sélectionnant dans la liste, puis en cliquant sur le bouton « (Re)analyser ». Le processus de reconnaissance est alors exécuté sur base de la méthode R sélectionnée et de l'ensemble d'apprentissage présent dans le sous-répertoire « _train » de votre répertoire de travail. Deux remarques :

  - Il est possible de ré-analyser un échantillon déjà analysé avec une méthode différente. Ceci permet une comparaison des performances de reconnaissance entre différents algorithmes.
  - Lors d'une première analyse avec l'interface de Zoo/PhytoImage, deux objets R sont créés dans le sous-répertoire « _train » de votre répertoire de travail : le premier correspondant à l'ensemble d'apprentissage et le second correspondant à l'outil de reconnaissance automatique généré à partir de l'ensemble d'apprentissage. Pour utiliser un nouvel ensemble d'apprentissage, il est impératif de supprimer ces deux objets R afin que le programme puisse recréer deux nouveaux objets.



Correction erreur. Ici est décrit le fonctionnement de l'outil de validation de la classification. Lorsqu'une analyse est lancée, Zoo/PhytoImage créé une page web qui vous présente un premier ensemble de (par défaut) 1/20ème des vignettes dans l'échantillon.

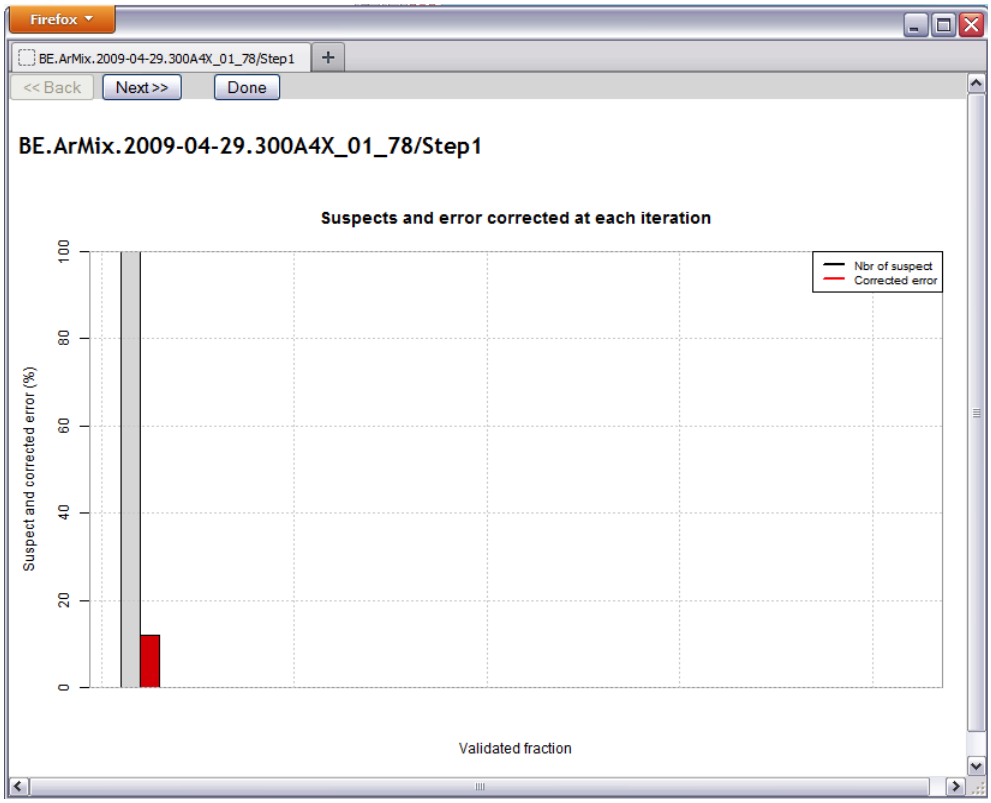


Cette page présente une première série de particules, sélectionnées aléatoirement dans l'échantillon, triées automatiquement par l'outil de reconnaissance choisi. Chaque classe est représentée par une colonne dans la page (e.g., *Ceratium_furca*, *Ceratium_fusus*, etc. dans l'exemple). Toutes les vignettes classées dans un groupe sont présentées dans la colonne correspondante. Déplacer le curseur au-dessus d'une vignette déclenche automatiquement une fenêtre flottante qui affiche la particule correspondante en pleine taille pour l'inspecter.

Toutes les vignettes peuvent être glissées et déposées librement partout. Ainsi, vous pouvez réorganiser les vignettes pour effectuer les corrections nécessaires. Pour de très longues grilles, avec des dizaines voire des centaines de colonnes, vous pouvez utiliser une zone spéciale sur la gauche nommée '**Unclassified**' pour stocker temporairement les objets que vous souhaitez déplacer dans une colonne distante dans la grille. Cependant, vous ne pouvez rien laisser dans cette zone spéciale lorsque vous voulez valider votre travail.

Pour toutes les particules que vous ne pouvez pas reconnaître, ou n'appartenant pas aux classes prédéfinies, vous pouvez les déposer dans une classe spéciale **[other]** à l'extrême droite de la grille.

Une fois la validation des vignettes effectuée, cliquez sur le bouton **Validate**. Un rapport sur le processus de validation réalisé pendant cette première étape est affiché.



Il présente un diagramme en bâtons avec des bâtons gris représentant la proportion d'objets suspects dans la fraction venant d'être validée. Pendant cette première étape, aucun modèle n'est calculé... donc, tous les objets sont considérés comme suspects. Une barre rouge à sa droite indique la fraction d'objets qui ont été incorrectement classés et que vous venez de corriger. Dans le cas présent, il s'élève à environ 15%. *C'est une très bonne indication de l'erreur globale de cette classification, puisque cette première fraction est purement choisie au hasard !* Ainsi, vous savez que vous avez un total d'environ 15% d'erreur et que vous avez déjà corrigé 1/20ème de cette erreur.

Si vous continuez à valider des sous-échantillons aléatoires, vous aurez encore à regarder les 19/20ème restant de votre échantillon. Si vous décidez d'accepter une erreur restante de moins de 5% du total, vous aurez encore besoin de valider 2/3, ce qui représente environ 12/20ème de l'ensemble de l'échantillon. Mais attendez... **faire cela ne garantit pas d'obtenir moins de 5% d'erreur dans *tous* les groupes.** Typiquement, vous laisserez beaucoup plus d'erreur dans les groupes les plus rares. Ainsi, il est préférable de *tout valider,* ou...

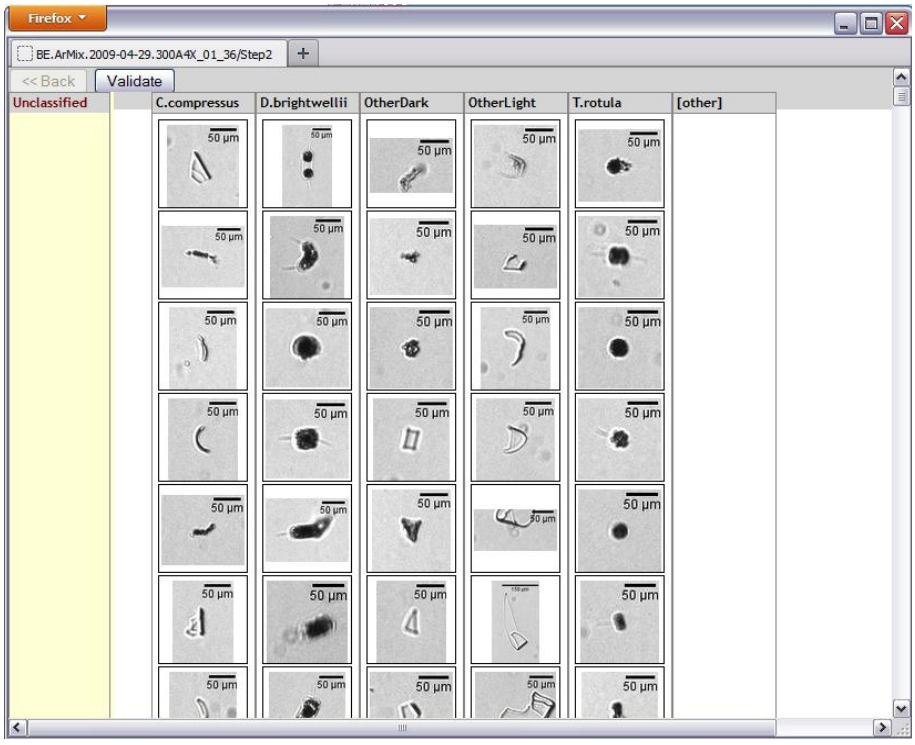
… Le validateur intelligent fournit un moyen beaucoup plus efficace de validation de votre échantillon en gardant cet objectif à l'esprit d'une erreur de moins de 5% dans *tous* les groupes. Pour atteindre cet objectif, un modèle statistique et une probabilité bayésienne sont calculés pour chacune des particules spécifiant si elle a une chance d'être suspecte (comprenez, understand, probablement classée à tort) ou non.

Le modèle considère également plusieurs autres aspects :

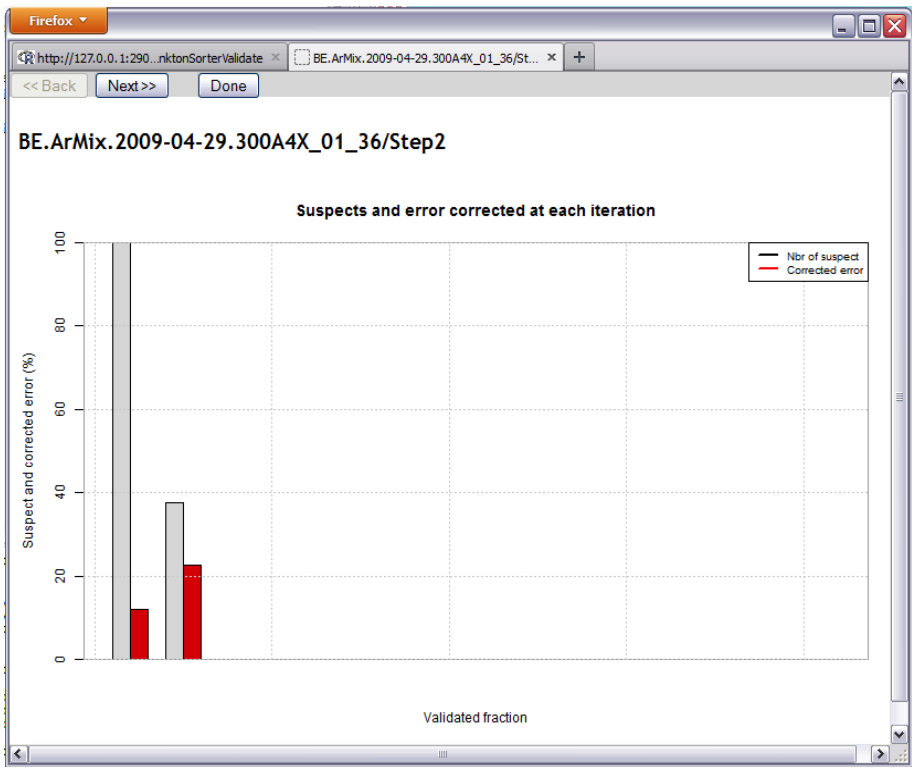
- La probabilité retournée par l'outil de reconnaissance pour la seconde classe prédite pour la particule est comparée avec la probabilité de la première classes sélectionnée. L'idée est que, si la différence entre ces deux probabilités est faible, nous devons considérer que la particule est proche de la frontière entre les deux classes et doit donc être vérifiée,
- Le nombre de particules classées dans la même classe pour l'ensemble de l'échantillon. S'il y en a peu, c'est un groupe rare. Ceci implique deux conséquences : (1) la probabilité de faux-positifs augmente, et (2) la classe a plus de probabilité de ne contenir aucune particule pour cet échantillon (car ce groupe taxonomique est absent là, à ce moment là). Ainsi, la probabilité d'être suspect augmente avec la rareté des particules classées dans la même classe,
- Les informations de la matrice de confusion sont utilisées pour déterminer quelles classes ont tendance à être moins bien discriminées. Encore une fois, cette information augmente la probabilité des particules correspondantes à être suspectes,
- Il est également possible de fournir une 'information biologique' (non pas à partir du menu/boîte de dialogue, mais en appelant la fonction **correctError()** directement dans la console R, voir sa page d'aide à **?correctError**). Cette information biologique doit indiquer si une classe donnée a des chances ou non d'être trouvée dans cet échantillon. Entrez ce que vous savez de la situation géographique, du moment de l'année, de la température de l'eau, ou simplement d'une inspection rapide de l'échantillon sous un microscope (classe A : très peu probable d'être présente, classe B : certainement présente). Indiquez juste une valeur faible (par exemple, 0.01) à la classe A et une valeur importante (par exemple, 0.99) à la classe B. Notez que les nombres que vous fournissez ne sont pas nécessairement limités entre 0 et 1, mais le concept est plus simple à considérer si vous voyez ces poids comme des pseudo-probabilités d'occurrence de la classe dans votre échantillon.

Zoo/PhytoImage utilise le premier ensemble de particules comme un ensemble d'apprentissage pour détecter les objets suspects, en utilisant tous les attributs mesurés sur ces particules, ainsi que les variables additionnelles décrites ci-dessus. Plusieurs algorithmes peuvent être utilisés, mais Random forest est utilisé par défaut.

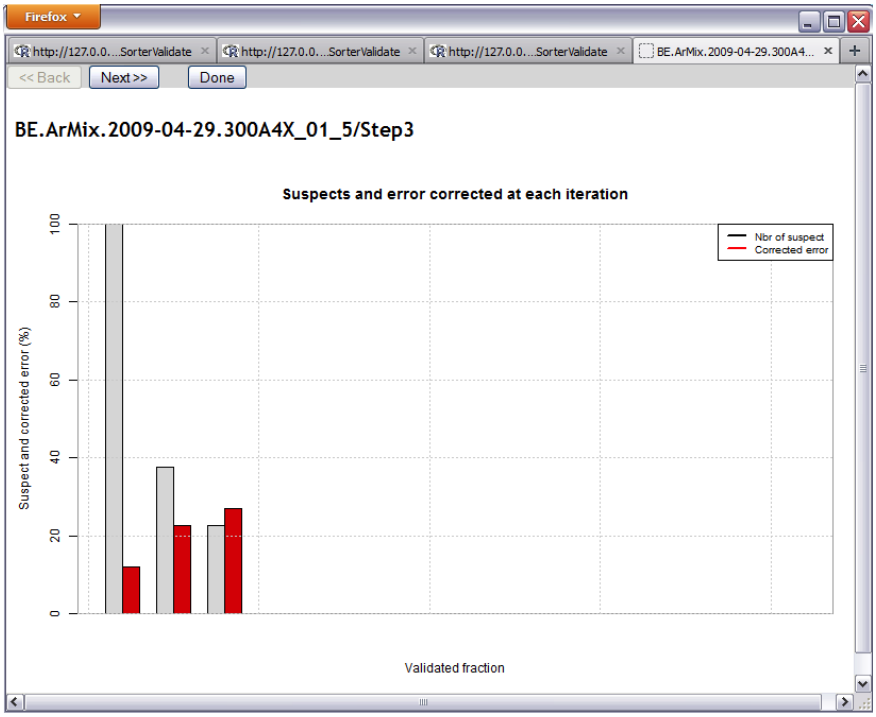
Ainsi, lorsque vous cliquez sur **Next**, Zoo/PhytoImage vous présente un autre sous-ensemble de particules dans l'échantillon. Mais cette fois, le sous-ensemble n'est pas choisi aléatoirement, mais principalement choisi parmi les objets suspects. En conséquence, la proportion d'erreur se trouve être supérieure. Ainsi, votre travail de validation est plus efficace car vous commencez désormais, à vous concentrer sur les particules réellement problématiques !



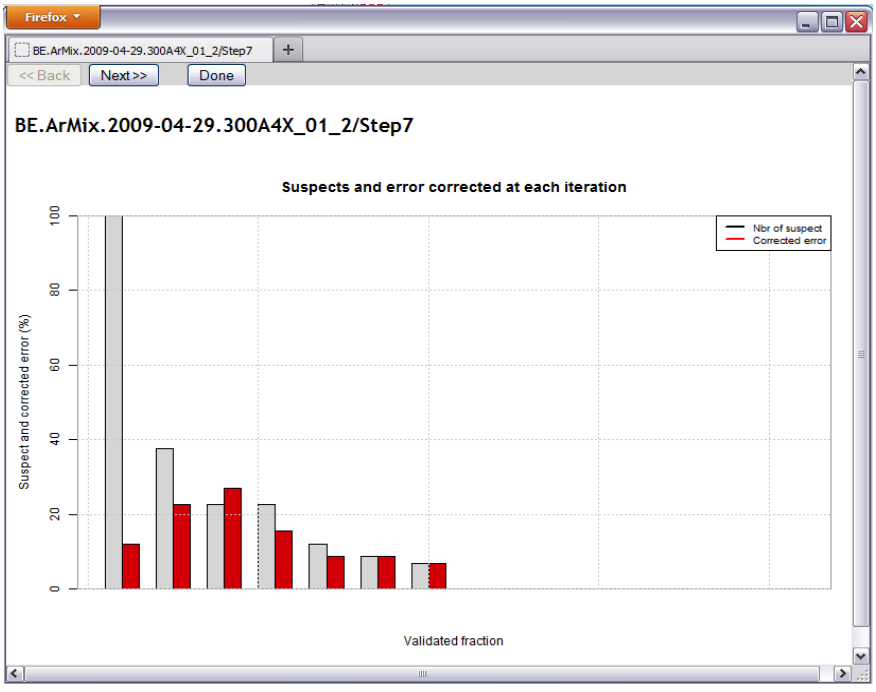
Généralement, il est assez clair que ce second ensemble présente beaucoup plus d'erreur que le précédent... et vous remarquerez également que, en effet, vous avez également beaucoup plus de particules « problématiques » (difficulté à reconnaître les particules, objets coupés, blobs avec une forme étrange, etc.). N'hésitez pas à utiliser le groupe **[other]** pour collecter ce que vous ne pouvez pas placer ailleurs (mais soyez cohérent avec ce que vous faîtes ici). Cliquez sur **Validate** lorsque vous en avez fini avec cette deuxième étape.



Dans le rapport, le diagramme en bâtons possède maintenant une seconde série de barres grises/rouges. Comme vous pouvez le voir ici, l'identification des objets suspects est légèrement différente (rappelons que l'ensemble d'apprentissage ne contient que très peu de particules... 1/20ème de l'échantillon total). Pourtant, vous avez presque doublé la fraction de particules erronées à cette étape. Lancez le une troisième fois.



Pour cet échantillon, l'algorithme prédit une quantité relativement faible d'objets suspects (sur d'autres échantillons, avec une proportion plus élevée de l'erreur initiale, cette fraction peut atteindre facilement 80 à 90%). Néanmoins, la fraction de particules erronées a augmenté un peu plus. Vous avez maintenant concentré l'erreur plus efficacement. Continuez avec quelques autres sous-ensembles :

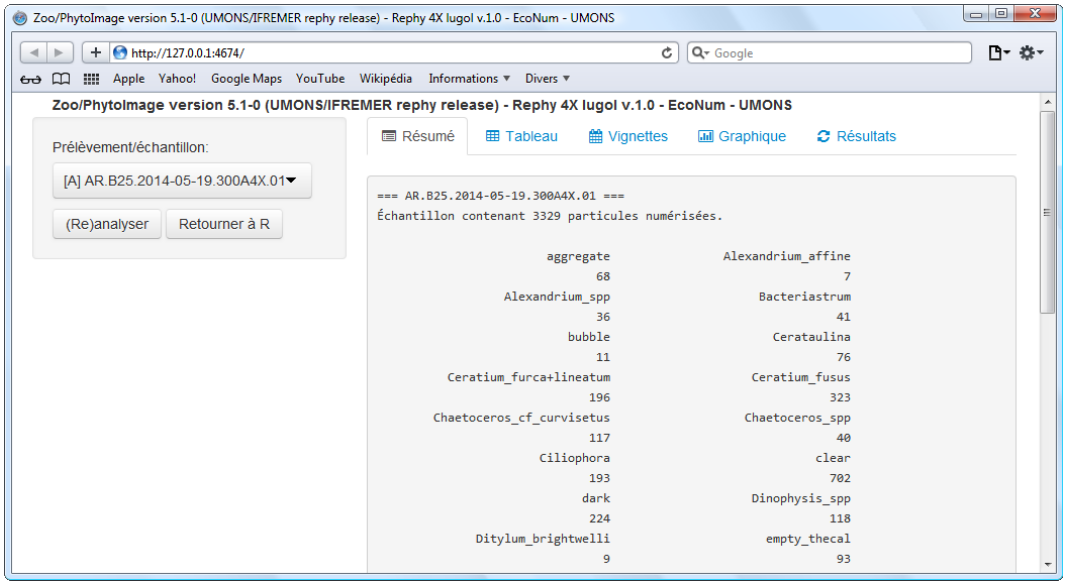


Ici, après l'étape 7, vous remarquez deux choses importantes. Premièrement, la détection des suspects correspond maintenant étroitement avec l'erreur réelle. La détection est améliorée avec la fraction de l'échantillon déjà validée qui peut être utilisée pour entraîner l'algorithme de détection. Deuxièmement, l'erreur résiduelle chute à moins de 10%.

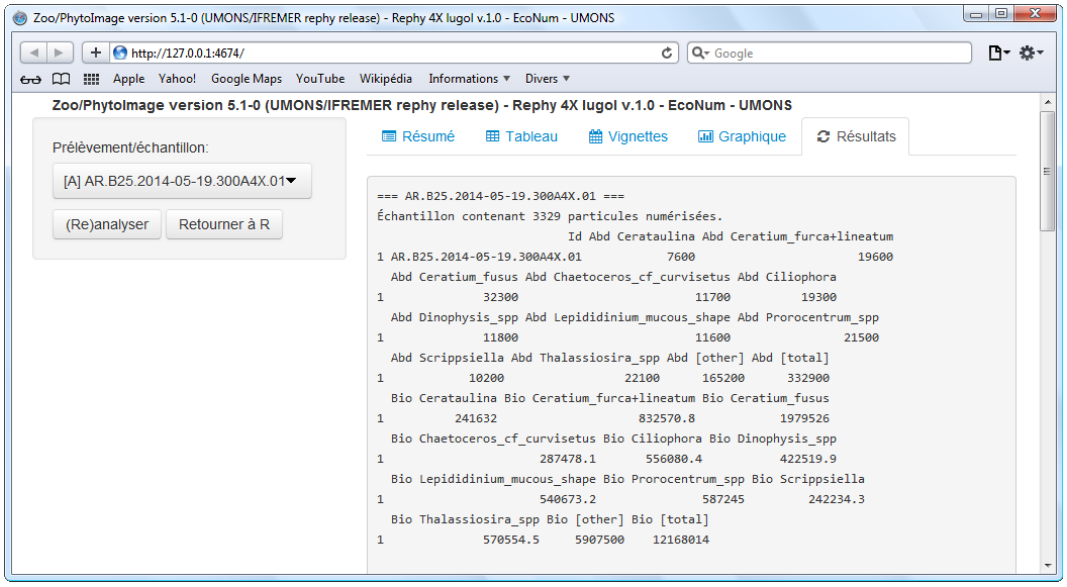
A partir de ce moment, vous savez que vous avez validé manuellement toutes les particules erronées jusqu'à environ 5%. Mais, puisque le modèle est utilisé pour calculer un *facteur de correction* pour les objets restants, le calcul des abondances par classe deviendra assez bon. Rappelez vous aussi que les particules des groupes rares ont été choisies de préférence dans les premiers ensembles. Ceci vous assure une bonne prédiction pour ces groupes rares qui sont souvent problématiques.

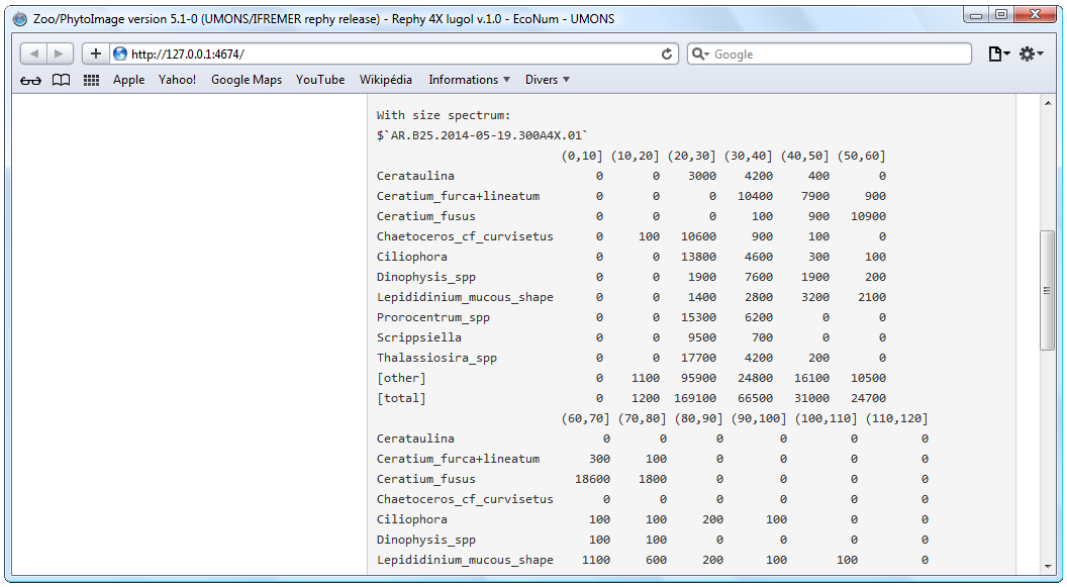
Donc, en gardant cela à l'esprit, vous pouvez raisonnablement considérer que la validation pourrait être arrêtée maintenant et que vous pourriez faire confiance à la correction introduite par cette validation partielle, et par la correction statistique grâce au modèle de détection des suspects. Cliquez alors sur le bouton **Done**.

Résultats. Une fois l'analyse terminée, retournez dans l'interface graphique utilisateur de Zoo/PhytoImage. Vous remarquerez alors que le code devant le nom de l'échantillon (dans la liste **Prélévement/Echantillons**) a changé (**[A]**) ; que le nombre de particules classées par groupe taxonomique est présentée dans l'onglet **Résumé** ; et qu'une colonne « Class » a été ajoutée dans l'onglet **Tableau**.



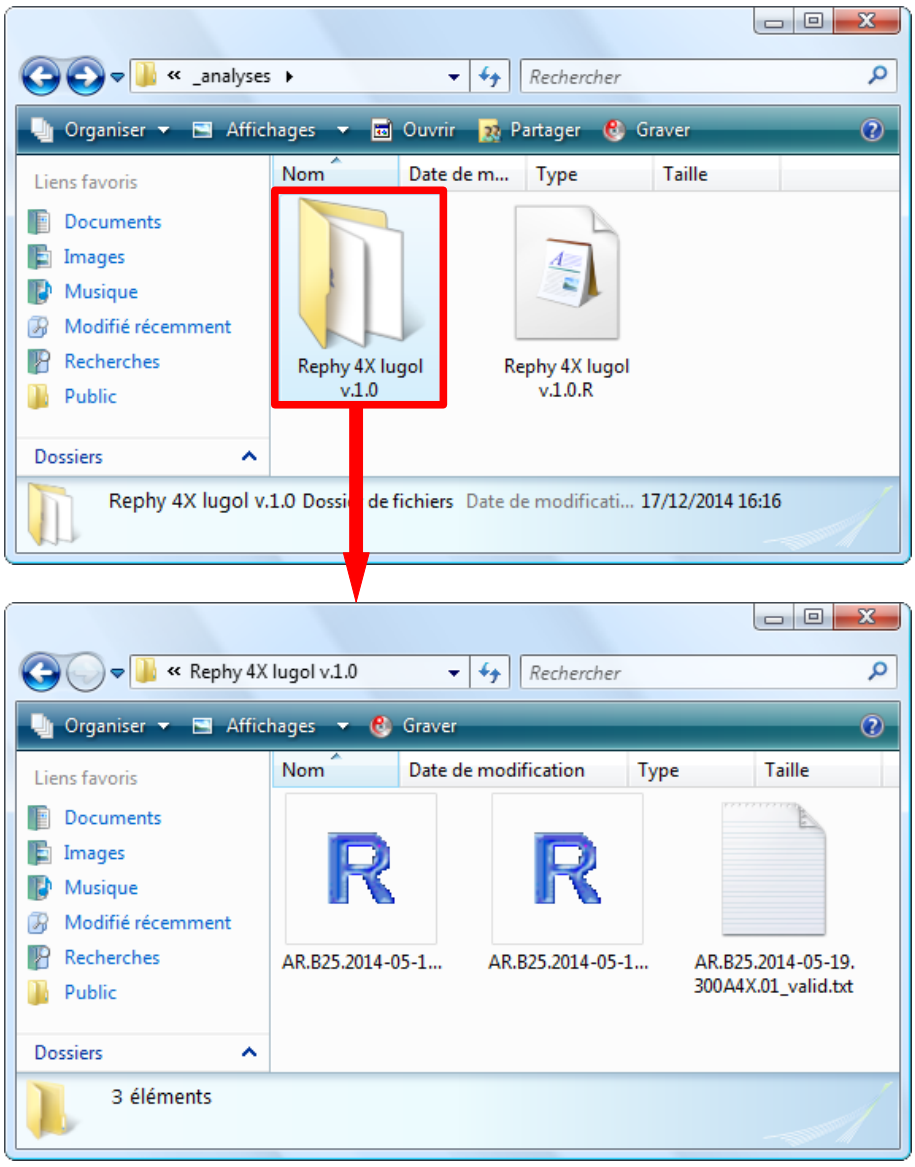
Si vous allez dans l'onglet Résultats, vous obtenez les abondance corrigées, mais également les biovolumes et les spectres de taille des différents groupes taxonomiques.





De plus, les résultats sont sauvegardés directement sur le disque. En effet, dans le sous-répertoire « _analyses » de votre répertoire de travail, un nouveau dossier a été créé. Ce dossier porte le nom du fichier méthode que vous avez utilisé. Puis, dans ce dossier, vous trouverez trois fichiers :

- <echantillon>_res.RData : ce fichier peut être chargé et manipulé dans R. Il contient les abondances, les biovolumes et les spectres de taille pour chaque groupe taxonomique,
- <echantillon>_valid.RData : ce fichier peut également être chargé et manipulé dans R. Il contient les mesures sur chacune ds particules, plus une colonne « Class » correspondant à la classification finale,
- <echantillon>_valid.txt : ce fichier contient les informations essentielles à l'interprétation des résultats, comme le nom de l'ensemble d'apprentissage, le nom de l'outil de reconnaissance, l'algorithme utilisé, etc., ainsi que des métadonnées telles que le nom de l'analyste, la date d'analyse, etc.


« _analyses
Rechercher
Organiser
Affichages
Ouvrir
Partager
Graver
Liens favoris
Documents
Images
Musique
Modifié récemment
Recherches
Public
Dossiers
Nom
Date de m...
Type
Taille
Rephy 4X lugol v.1.0
Rephy 4X lugol v.1.0.R
Rephy 4X lugol v.1.0 Dossi de fichiers
Date de modificati... 17/12/2014 16:16
« Rephy 4X lugol v.1.0
Rechercher
Organiser
Affichages
Graver
Liens favoris
Documents
Images
Musique
Modifié récemment
Recherches
Public
Dossiers
Nom
Date de modification
Type
Taille
AR.B25.2014-05-1...
AR.B25.2014-05-1...
AR.B25.2014-05-19.300A4X.01_valid.txt
3 éléments

6. UTILISATION DE ZOO/PHYTOIMAGE EN LIGNE DE COMMANDE R

Une description complète et détaillée de l'utilisation des fonctions zooimage dans la console R est décrite dans le Chapitre 12 du livre suivant :
Yanchang Zhao and Yonghua Cen (Eds.). Data Mining Applications with R. ISBN 978-0124115118, December 2013. Academic Press, Elsevier.

Nous encourageons les lecteurs intéressés à télécharger les fichiers d'accompagnement à partir du site : http://www.sciviews.org/zooimage/Data mining with R/. Ceux-ci contiennent un script R entièrement commenté, ainsi qu'un jeu de données exemples qui reprend les fonctionnalités disponibles en ligne de commande.

Voici un aperçu des outils les plus importants, en plus de ce que vous pouvez déjà réaliser en utilisant l'interface graphique utilisateur et le menu de ZooPhytoImage version 5.

- Les vignettes sont directement accessibles sous R et peuvent être inclues dans des affichages R, ou affichées comme une galerie. Le code à implémenter ressemble à cela :

```
## Chargement des données dans R à partir d'un fichier ZIDB
db1 <- zidbLink(chemin_du_zidb)
## Contient les données dans *_dat1 et les vignettes dans *_nn
items1 <- ls(db1)
vigs1 <- items1[-grep("_dat1", items1)]
## Affiche une planche 5*5 des 25 premières vignettes
zidbPlotNew("The 25 first vignettes in MTPS.2004-10-20.H1")
for (i in 1:25) zidbDrawVignette(db1[[vigs1[i]]], item = i, nx = 5, ny = 5)
```

- La méthode summary d'un objet ZIClass (un outil de reconnaissance) affiche u grand nombre de statistiques sommaires telles que Recall, Precision, Specificity, F-score, balanced accuracy, etc. Ces statistiques sont calculées groupe par groupe. Voir la page d'aide ZIClass (?ZIClass).
- L'objet ZIClass possède une méthode de confusion qui créé une matrice de confusion avec quatre modes d'affichage spécifique : image, diagramme en bâtons, étoiles et dendrogramme. Le diagramme en bâtons donne un nouvel aperçu du F-score par groupe. Voir ?confusion et l'exemple dans le script R. L'affichage en étoile peut également être utilisé pour comparer deux outils de reconnaissance appliqués au même ensemble de test.
- Il y a également des compléments sur la façon dont Zoo/PhytoImage calcule les abondances et les biomasses/biovolumes. Vous pouvez calculer ces quantités à différents niveaux de détail et indiquer quels groupes n'ont pas d'intérêt (e.g., neige marine et zooplancton si votre étude porte sur le phytoplancton).
- L'objet confusion peut être ajusté pour différentes probabilités *a priori* (abondances par groupe) en utilisant la fonction prior(). Cela vous permet alors de visualiser l'impact de la composition de différents échantillons sur les taux de faux positifs et faux négatifs par groupe.
- N'oubliez pas également tous les outils R disponibles pour manipuler des objets d'apprentissage machine. Voir les possibilités d'apprentissage machine à partir du site http://cran.r-project.org/web/views/MachineLearning.html.

Finalement, le chapitre 12 dans le livre « Data mining applications with R » présente une collection de références bibliographiques (64), la plupart d'entre eux pointent sur des publications dont les analyses ont été effectuées en utilisant Zoo/PhytoImage. C'est également une excellente source d'inspiration montrant concrètement comment Zoo/PhytoImage peut être utilisé.

Écologie Numérique des Milieux Aquatiques
UMONS
Faculté des Sciences

Dénombrement des cellules dans les colonies

Guillaume WACQUET & Philippe GROSJEAN

Novembre – Décembre 2014

Table des matières

Introduction

Jusqu'à présent, la version 5 de Zoo/PhytoImage permet d'obtenir des identifications semi-automatiques pertinentes du phytoplancton mais sans distinguer une cellule d'une colonie. Or, même si les colonies contribuent en grande partie à la productivité annuelle, l'ensemble des estimateurs de la biomasse sont calibrés essentiellement sur l'abondance en termes de cellules par unité de volume.

Une première étude a été menée en 2010 par P. Govaerts [3] sur des échantillons provenant de cultures, mais également du milieu naturel. Cependant, l'acquisition des données a été réalisée, en grande majorité, à l'aide du couplage objectif/cellule de flux 2X/600µm. Dans le cadre du projet FlowCAM/ZooPhytoImage, le couplage retenu est 4X/300µm, comme défini dans le livrable n°3 ONEMA pour l'année 2013. C'est pourquoi, dans ce rapport, une nouvelle étude est menée sur des taxa provenant d'échantillons naturels numérisés à l'aide du FlowCAM et du couplage 4X/300µm.

La méthode proposée dans ce rapport consiste à construire des modèles prédictifs permettant d'estimer le nombre de cellules par colonie dans tous les échantillons étudiés, en se basant sur les comptages manuels réalisés sur les particules du set d'apprentissage. A cette fin, des outils visuels et statistiques ont été développés : outils d'aide au comptage manuel sur ordinateur, régressions linéaires et non linéaires, classification supervisée de type « machine learning » et estimation de la qualité de prédiction à l'aide de la validation croisée.

Dans cette étude, les scores de performance obtenus par les différentes méthodes prédictives sont mis en évidence sur six groupes taxinomiques de phytoplancton colonial de la Manche Orientale et du Sud de la Mer du Nord.

Présentation des données

Outil d'aide au dénombrement des cellules

Les cellules agencées en colonies ne sont pas encore identifiées individuellement par le logiciel Zoo/PhytoImage. Une colonie est donc considérée comme un objet unique. Il n'est, de ce fait, pas possible de comparer les comptages cellulaires avec les comptages automatiques. Les espèces coloniales étant fréquentes dans les échantillons observés et les mesures d'abondance en termes de cellules représentant une information capitale, ce travail est donc nécessaire afin de pouvoir comparer les résultats obtenus par le système couplé FlowCAM/PhytoImage avec ceux des dénombrements REPHY.

Pour cela, un outil d'aide au dénombrement des cellules dans les vignettes a été développé sous R. Ce module de dénombrement interactif pourra être intégré, à terme, dans Zoo/PhytoImage. Il consiste en :

- l'affichage de la vignette et la proposition automatique d'une estimation du nombre de cellules dans la colonie, basée sur des algorithmes de segmentation d'image ou sur des abaques, comme illustrée sur la figure 1,
- la correction manuelle (par clics souris) ou la validation (cf. figure 2).

Une fois validé, une nouvelle entrée est créée directement et automatiquement pour chacune des particules dans le set d'apprentissage.

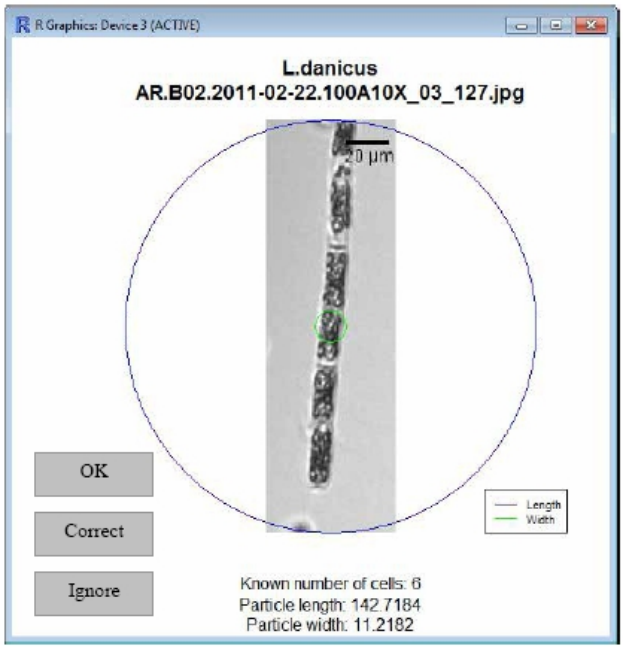


Fig. 1 *: Estimation du nombre de cellules*

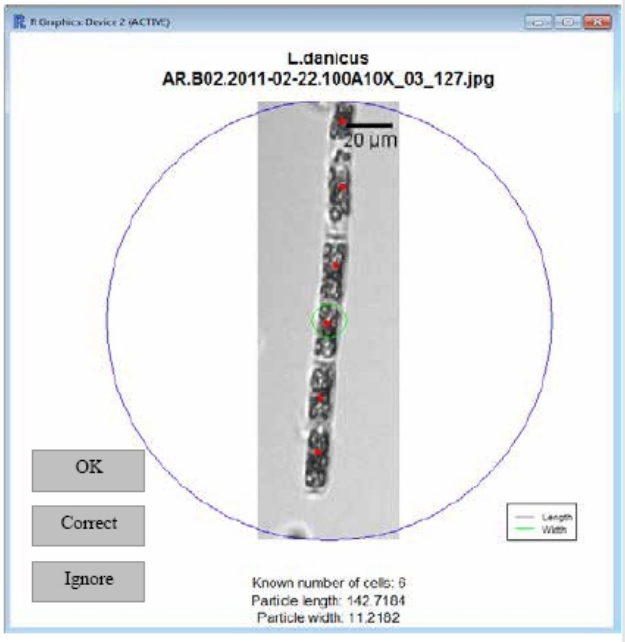


Fig. 2 *: Correction manuelle du nombre de cellules*

Comptages manuels des vignettes

Les particules constituant le set d'apprentissage utilisé dans cette étude, proviennent d'échantillons naturels. Ces derniers ont été prélevés et numérisés à l'aide du FlowCAM durant l'année 2013 dans le cadre du réseau de surveillance REPHY. Onze points sont représentées dans l'échantillonnage global :

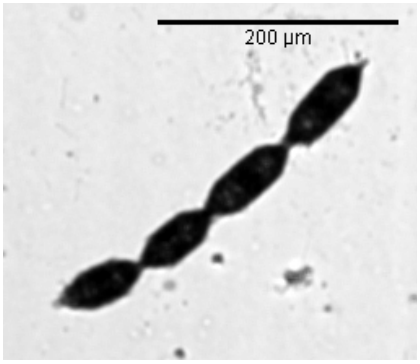
- 3 points à Boulogne-sur-Mer,
- 3 points à Dunkerque,
- 5 points en Baie de Somme.

Le couple objectif/cellule de flux choisi pour l'analyse au FlowCAM est la combinaison 4X/300µm. Dans cette étude, un sous-ensemble de vignettes a été sélectionné afin de construire un set d'apprentissage dans lequel sont repris différentes espèces coloniales (cf. table 1).

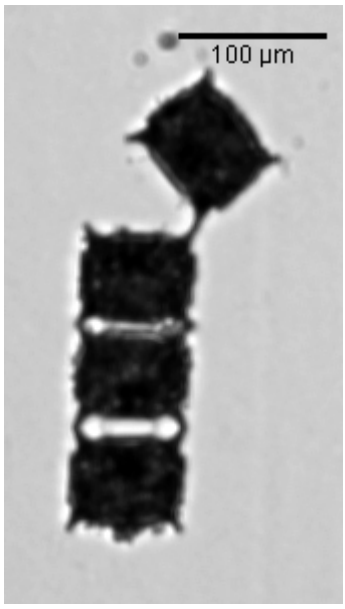
Table 1 : Table représentant l'ensemble des groupes taxinomiques étudiés, provenant d'échantillons naturels (300µm/4X).

Groupes taxinomiques (300µm/4X)	**Nombre de vignettes**	**Nombre total de cellules dénombrés**
Biddulphia rhombus	54	75
Biddulphia sinensis	64	162
Ditylum brightwellii	66	95
Thalassiosira rotula	57	199
Asterionella glacialis	61	268
Thalassionema nitzschoides	70	286

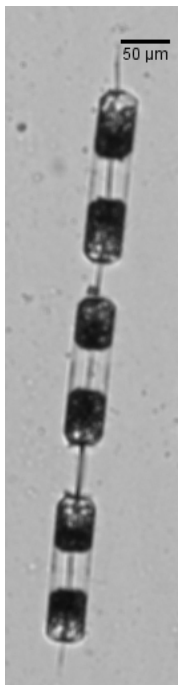
6 groupes taxinomiques contenant des colonies (morphologies et tailles différentes) :



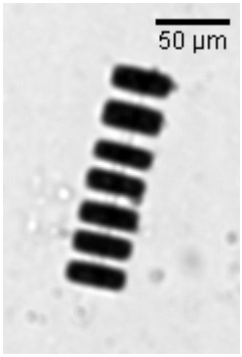
Biddulphia rhombus



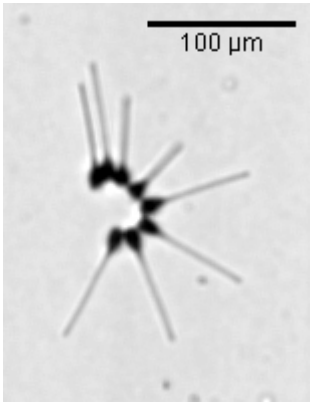
Biddulphia sinensis



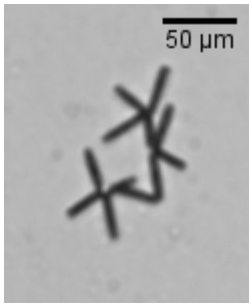
Ditylum brightwellii



Thalassiosira rotula



Asterionella glacialis



Thalassionema nitzschoides

Utilisation des variables FlowCAM

Dans un premier temps, seules les variables issues du FlowCAM sont utilisées pour construire des modèles prédictifs. Ces variables sont au nombre de 13 :

- **FIT_Area_ABD, FIT_Diameter_ABD, FIT_Volume_ABD** : surface en µm², diamètre en µm et volume en µm^3, basés sur l'Area Based Diameter.
- **FIT_Diameter_ESD, FIT_Volume_ESD** : diamètre en µm et volume en µm^3, basés sur l'Equivalent Spherical Diameter.
- **FIT_Aspect_Ratio** : rapport entre la longueur et la largeur ESD (Length/Width).
- **FIT_Length, FIT_Width** : longueur et largeur ESD.
- **FIT_Perimeter** : nombre de pixel comptabilisé sur la silhouette de la particule.
- **FIT_Convex_Perimeter** : approximation du périmètre de la particule en reliant l'ensemble des distances décrites autour de l'objet par le processus récursif des distances de Ferret.
- **FIT_Roughness** : mesure de la rugosité de la silhouette, grâce au rapport Perimeter/Convex_Perimeter.
- **FIT_Compactness** : dérivé du périmètre et de l'aire (« ABD »). Cette variable équivaut à 1 pour une particule parfaitement circulaire. Formule : Perimeter² / (4 x pi x Area_ABD).
- **FIT_Elongation** : basé sur le périmètre et la surface.

Dans le cadre de cette étude, nous nous intéressons à l'estimation du nombre de cellules par colonies. Les performances des méthodes prédictives sont comparées par validations croisées (10-folds) sur des modèles de régressions linéaires et non linéaires et par quelques algorithmes de « machine learning ».

Grâce à la validation croisée, il est possible d'évaluer les taux de reconnaissance des différentes méthodes prédictives. Ici, trois scores sont évalués :

- **%VP vignettes** : le pourcentage de vignettes correctement prédites (Vrais-Positifs) dans les différentes classes (une classe étant représentée par un nombre de cellules par particule).
- **%VP cells/colonie** : le pourcentage de cellules correctement prédites (Vrais-Positifs) dans les différentes classes (une classe étant représentée par un nombre de cellules par particule).
- **Estimation totale** : le pourcentage de l'abondance du nombre de cellules total prédit sur l'abondance du nombre de cellules total mesurée manuellement dans l'échantillon.

Modèle linéaire à un prédicteur (basé sur le critère R²)

Table 2 : Table représentant les meilleures variables au sens du R² pour l'ensemble des groupes taxinomiques étudiés, provenant d'échantillons naturels (300µm/4X).

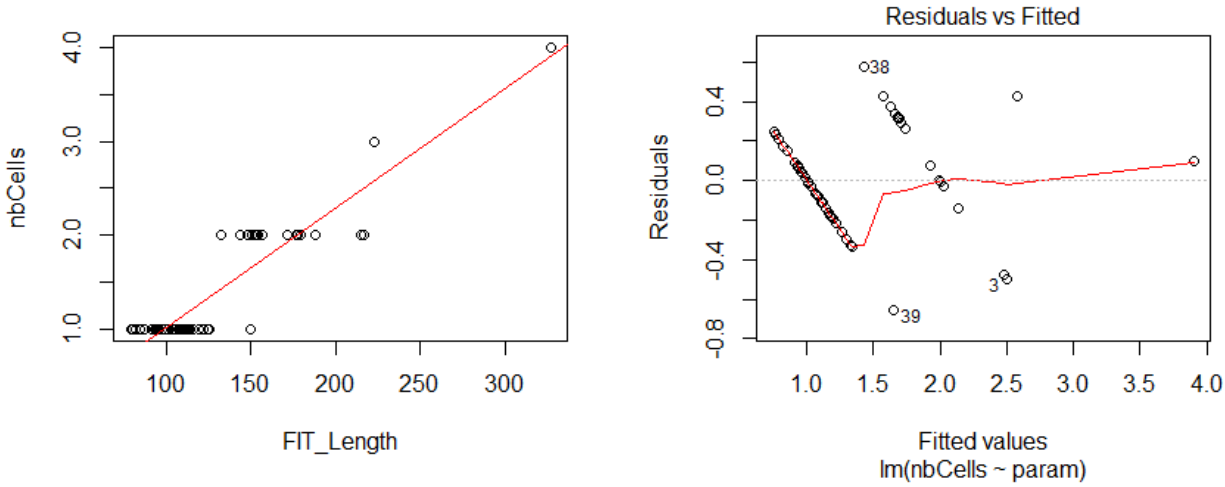
Groupes taxinomiques (300µm/4X)	Variable retenue	Relation	R²	p-value	BIC (log)
Biddulphia rhombus	FIT_Length	0.0127 V – 0.2535	0.8240	2.20e-16	25.9728
Biddulphia sinensis	FIT_Perimeter	0.0034 V + 0.2471	0.7254	2.20e-16	72.5959
Ditylum brightwellii	FIT_Diameter_ESD	0.0088 V – 0.0734	0.9106	2.20e-16	17.5924
Thalassiosira rotula	FIT_Perimeter	0.0092 V + 0.2816	0.9348	2.20e-16	163.157
Asterionella glacialis	FIT_Area_ABD	0.0033 V + 1.1481	0.7191	2.20e-16	178.1119
Thalassionema nitzschoides	FIT_Diameter_ABD	0.1773 V – 1.3180	0.6521	2.20e-16	67.4785

La table 2 présente les variables retenues au sens du R² pour chacune des espèces. Les résultats détaillés sont présentés ci-dessous.

Détails des résultats (régression et résidus)

Table 3 : Matrice de confusion (représentation de la régression linéaire par les moindres carrés VS le comptage manuel) pour *Biddulphia rhombus* (%VP vignettes : 96.30%)

	Linear model with FIT_Length			
Manual	1 cell/col	2 cells/col	3 cells/col	4 cells/col
1 cell/col	**35**	1	0	0
2 cells/col	1	**15**	0	0
3 cells/col	0	0	**1**	0
4 cells/col	0	0	0	**1**

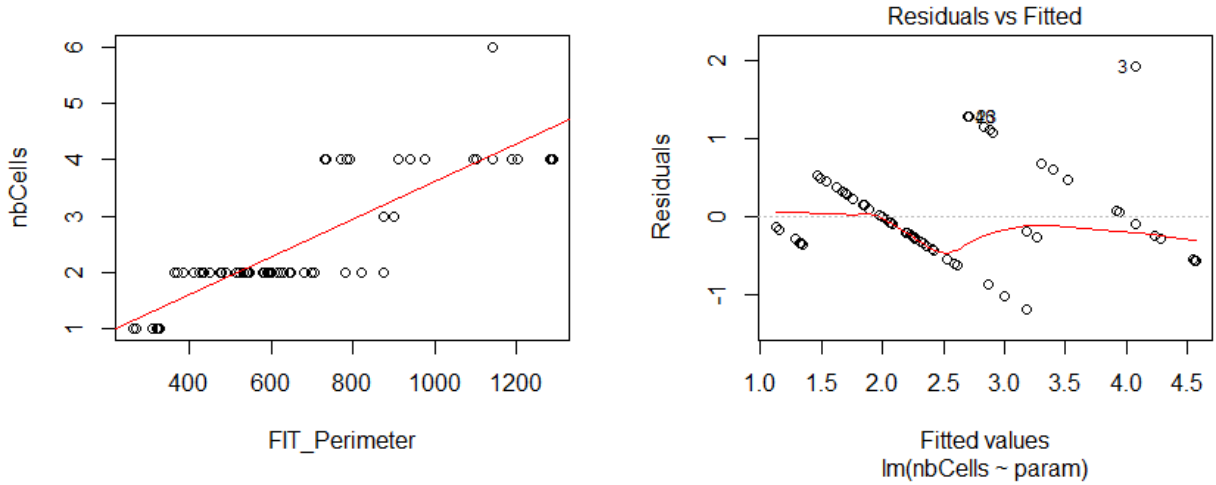


Régression linéaire simple par les moindres carrés (à gauche) et graphique de la distribution des résidus (à droite) de *Biddulphia rhombus*.

Les vignettes de colonies disponibles pour cette espèce sont majoritairement composées de 2 cellules. Une seule colonie composée de 3 et 4 cellules a été comptée manuellement. Pour cette raison, l'écart des résidus est naturellement d'une cellule.

Table 4 : Matrice de confusion (représentation de la régression linéaire par les moindres carrés VS le comptage manuel) pour *Biddulphia sinensis* (%VP vignettes : 70.31%)

	Linear model with FIT_Perimeter				
Manual	1 cell/col	2 cells/col	3 cells/col	4 cells/col	5 cells/col
1 cell/col	**6**	0	0	0	0
2 cells/col	2	**30**	6	0	0
3 cells/col	0	0	**2**	0	0
4 cells/col	0	0	7	**7**	3
6 cells/col	0	0	0	1	**0**

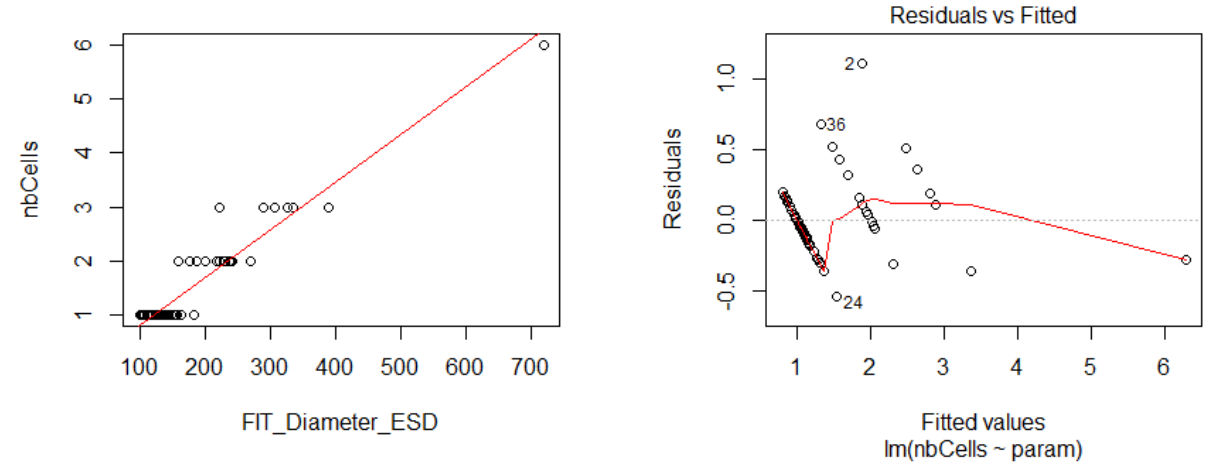


Régression linéaire simple par les moindres carrés (à gauche) et graphique de la distribution des résidus (à droite) de *Biddulphia sinensis*.

Sur le graphique de régression linéaire pour l'espèce *Biddulphia sinensis*, nous pouvons observer un chevauchement entre les mesures du prédicteur et des niveaux du nombre de cellules par colonie. Ici, l'écart maximal des résidus est de l'ordre de -/+2 cellules.

Table 5 : Matrice de confusion (représentation de la régression linéaire par les moindres carrés VS le comptage manuel) pour *Ditylum brightwellii* (%VP vignettes : 92.42%)

	Linear model with FIT_Diameter_ESD			
Manual	1 cell/col	2 cells/col	3 cells/col	6 cells/col
1 cell/col	**46**	1	0	0
2 cells/col	2	**10**	0	0
3 cells/col	0	2	**4**	0
6 cells/col	0	0	0	**1**

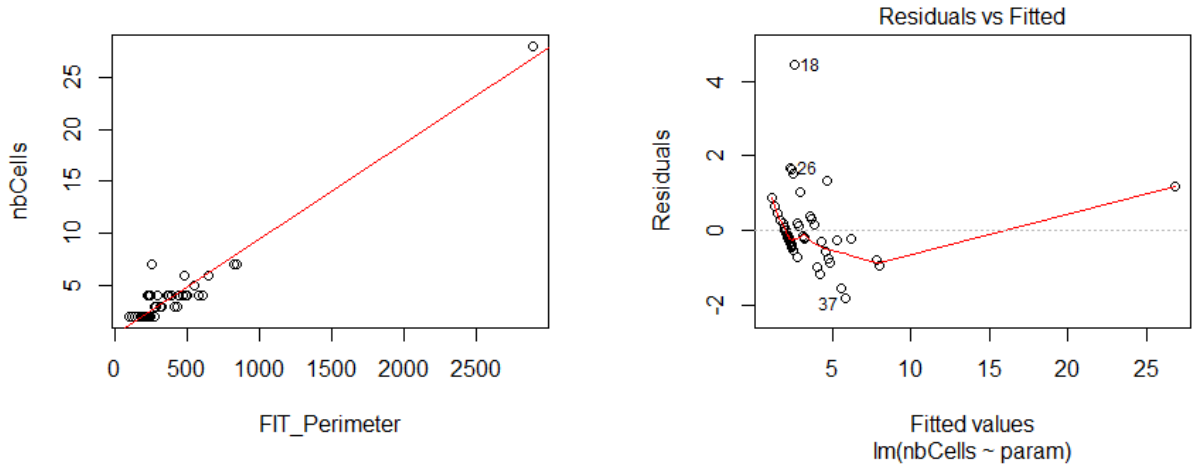


Régression linéaire simple par les moindres carrés (à gauche) et graphique de la distribution des résidus (à droite) de *Ditylum brightwellii*.

Pour l'espèce *Ditylum brightwellii*, les résidus oscillent de -0.5 à 0.5. L'écart maximal est donc de l'ordre de -/+1 cellule.

Table 6 : Matrice de confusion (représentation de la régression linéaire par les moindres carrés VS le comptage manuel) pour *Thalassiosira rotula* (%VP vignettes : 68.42%)

	Linear model with FIT_Perimeter						
Manual	2 cells/col	3 cells/col	4 cells/col	5 cells/col	6 cells/col	8 cells/col	27 cells/col
2 cells/col	**28**	2	0	0	0	0	0
3 cells/col	0	**5**	2	0	0	0	0
4 cells/col	3	1	**4**	3	2	0	0
5 cells/col	0	0	0	**1**	0	0	0
6 cells/col	0	0	0	1	**1**	0	0
7 cells/col	0	1	0	0	0	2	0
28 cells/col	0	0	0	0	0	**0**	**1**

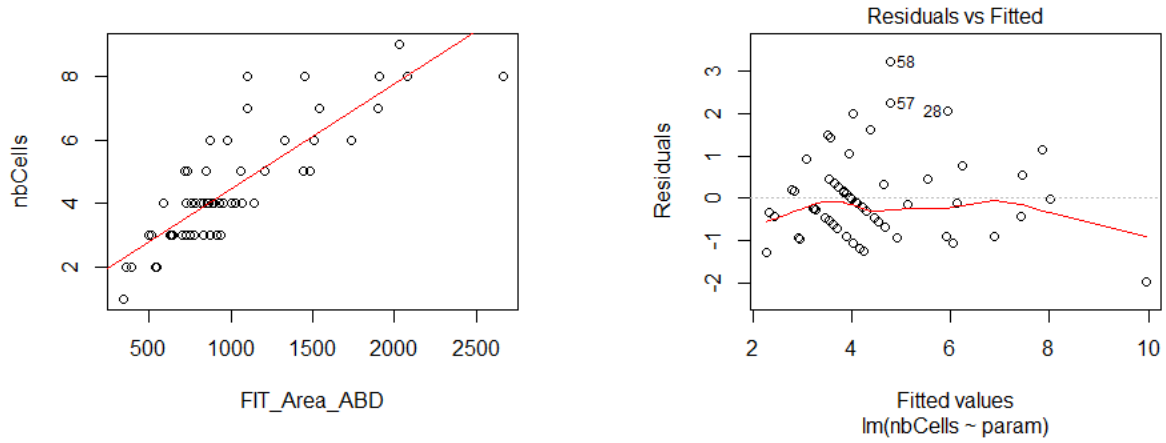


Régression linéaire simple par les moindres carrés (à gauche) et graphique de la distribution des résidus (à droite) de *Thalassiosira rotula*.

Ici, un chevauchement important entre les mesures du prédicteur et les niveaux du nombre de cellules par colonie peut être observé. La variable sélectionnée ne permet donc pas d'expliquer de manière précise le nombre de cellules par colonie. De plus, l'écart maximal des résidus est important.

Table 7 : Matrice de confusion (représentation de la régression linéaire par les moindres carrés VS le comptage manuel) pour *Asterionella glacialis* (%VP vignettes : 50.82%)

	Linear model with FIT_Area_ABD							
Manual	2 cells/col	3 cells/col	4 cells/col	5 cells/col	6 cells/col	7 cells/col	8 cells/col	9 cells/col
1 cell/col	1	0	0	0	0	0	0	0
2 cells/col	**3**	2	0	0	0	0	0	0
3 cells/col	0	**7**	7	0	0	0	0	0
4 cells/col	0	1	**15**	4	0	0	0	0
5 cells/col	0	0	3	**2**	2	0	0	0
6 cells/col	0	0	1	1	**2**	1	0	0
7 cells/col	0	0	0	1	1	**1**	0	0
8 cells/col	0	0	0	1	1	1	**1**	1
9 cells/col	0	0	0	0	0	0	1	**0**

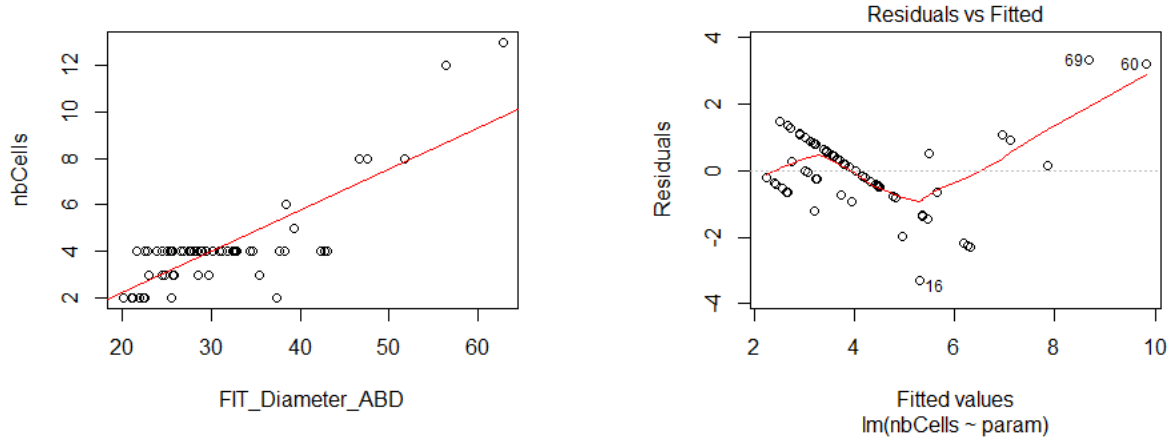


Régression linéaire simple par les moindres carrés (à gauche) et graphique de la distribution des résidus (à droite) de *Asterionella glacialis*.

De même que pour l'espèce *Thalassiosira rotula*, un chevauchement important entre les mesures du prédicteur et les niveaux du nombre de cellules par colonie peut être observé pour l'espèce *Asterionella glacialis*. De plus, l'écart maximal des résidus est important.

Table 8 : Matrice de confusion (représentation de la régression linéaire par les moindres carrés VS le comptage manuel) pour *Thalassionema nitzschoides* (%VP vignettes : 41.54%)

	Linear model with FIT_Diameter_ABD				
Manual	2 cells/col	3 cells/col	4 cells/col	5 cells/col	6 cells/col
2 cells/col	**3**	5	0	1	0
3 cells/col	0	**5**	2	1	0
4 cells/col	0	18	**19**	6	3
5 cells/col	0	0	0	**0**	1
6 cells/col	0	0	0	1	**0**



Régression linéaire simple par les moindres carrés (à gauche) et graphique de la distribution des résidus (à droite) de *Thalassionema nitzschoides*.

Pour *Thalassionema nitzschoides*, il existe un chevauchement entre les mesures du prédicteur et les niveaux du nombre de cellules par colonie. La variable sélectionnée seule ne permet donc pas d'expliquer de manière précise le nombre de cellules par colonie.

Sur ces trois derniers exemples, nous observons que les meilleurs prédicteurs sélectionnés au sens du R^2, ne permettent pas de prédire précisément le nombre de cellules dans les colonies. C'est pourquoi, nous nous tournons vers la sélection de modèle linéaire multivarié.

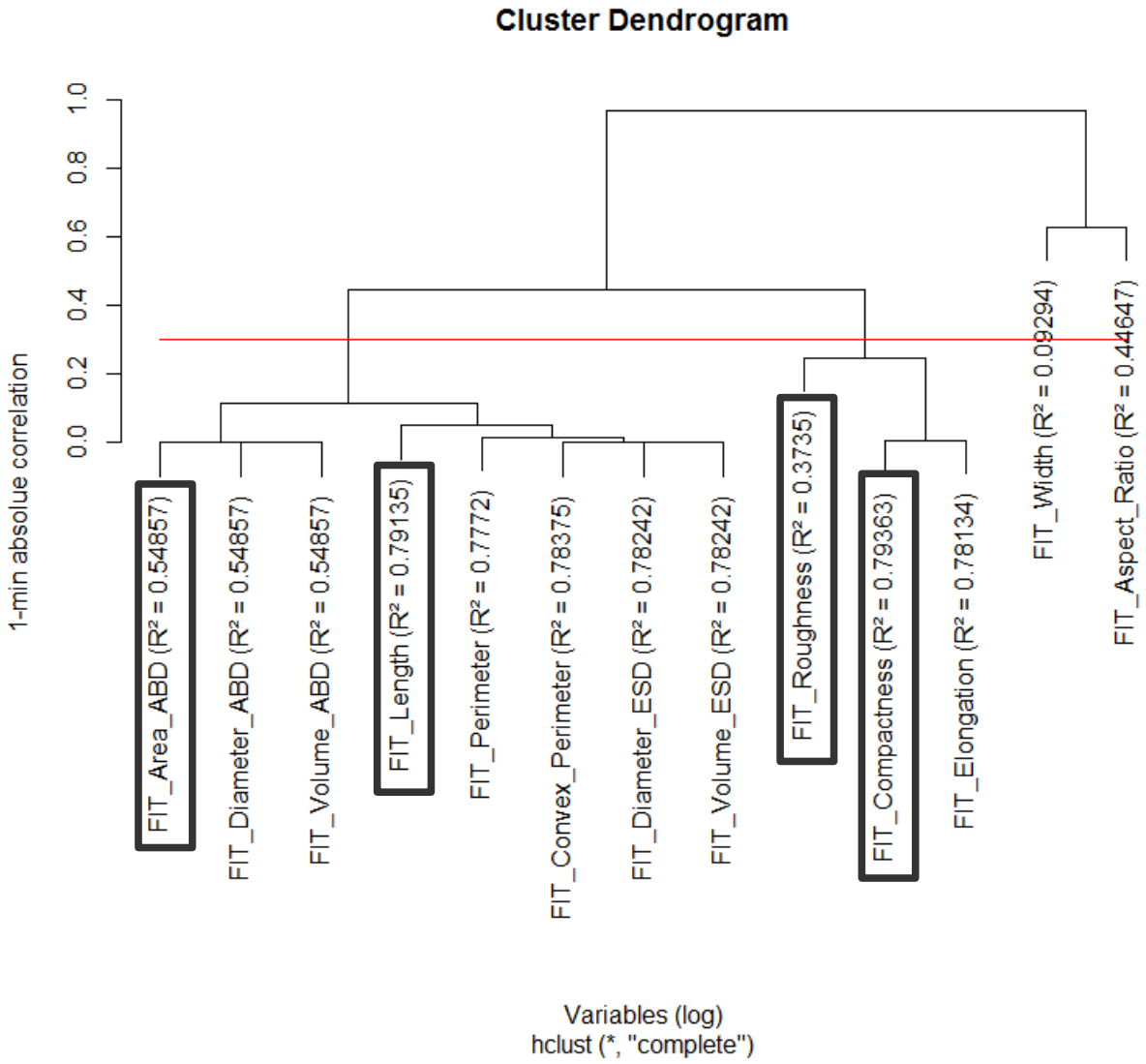
Modèle linéaire multivarié (basé sur le critère BIC)

Afin de sélectionner les « meilleures variables » pour construire les modèles prédictifs, nous analysons ici les relations entre variables par deux méthodes :

Analyse des relations entre variables

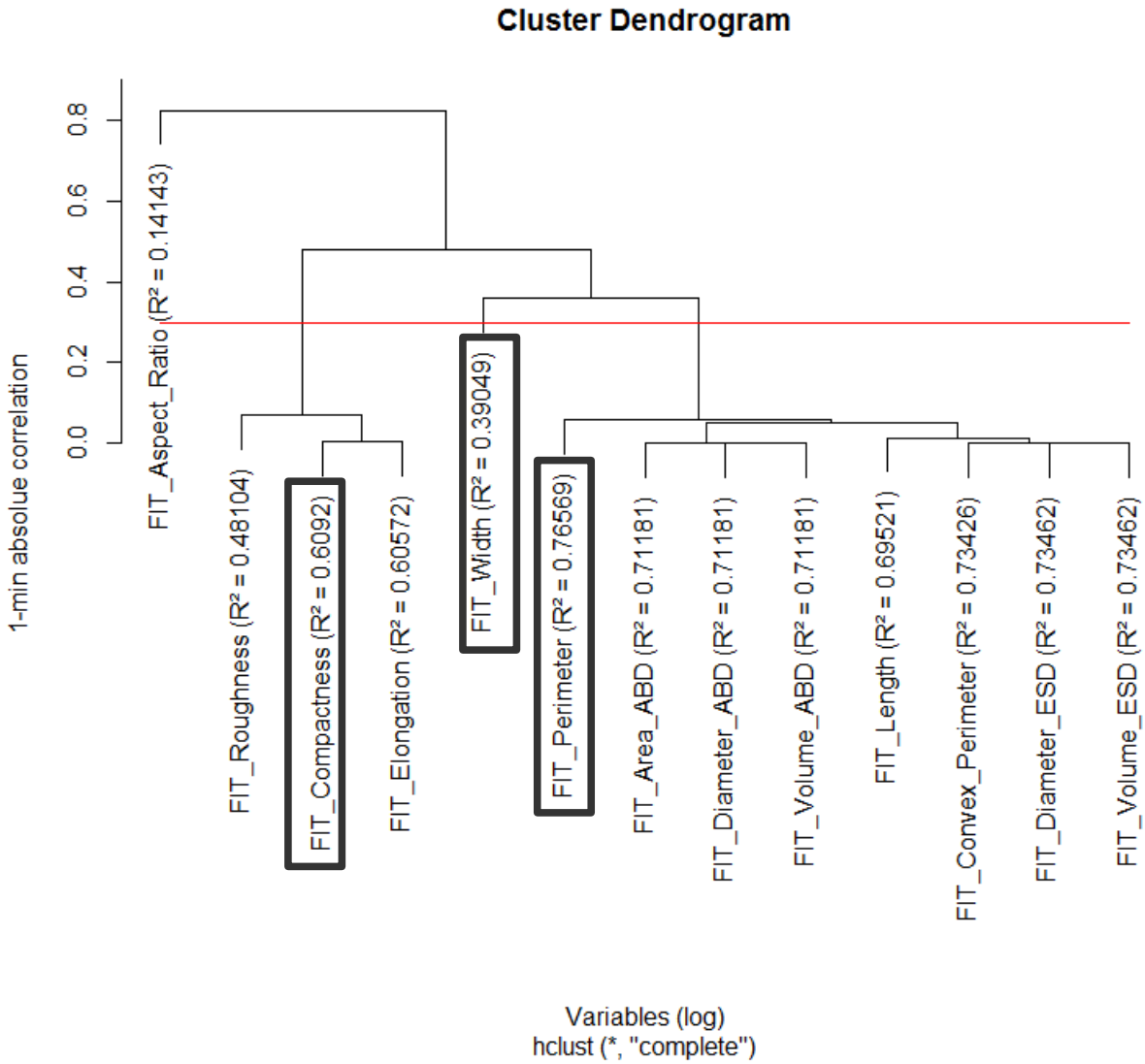
- Analyse **MANUELLE** des relations entre variables :
  - regroupement des variables grâce à une représentation par classification hiérarchique utilisant les liens complets unissant les distances relatives de la matrice de corrélation (1 - minimum absolu de corrélation) des variables représentées en logarithme.
  - utilisation du R^2 pour exprimer l'intensité de la relation entre le nombre de cellules et les différents prédicteurs.
- Analyse **AUTOMATIQUE** des relations entre variables :
  - choix de modèle par sélection de variables : recherche de modèles parcimonieux (avec un nombre restreint de variables explicatives)
  - lorsque le nombre de variables (p) est grand, il n'est pas raisonnable de penser explorer tous les modèles possibles (2^p) afin de sélectionner le "meilleur" au sens du critère BIC. Ici, la stratégie envisagée est la sélection pas-à-pas « mixte » introduisant une étape d'élimination de variable après chaque étape de sélection afin de retirer du modèle d'éventuels variables qui seraient devenues moins indispensables du fait de la présence de celles nouvellement introduites.

Biddulphia rhombus



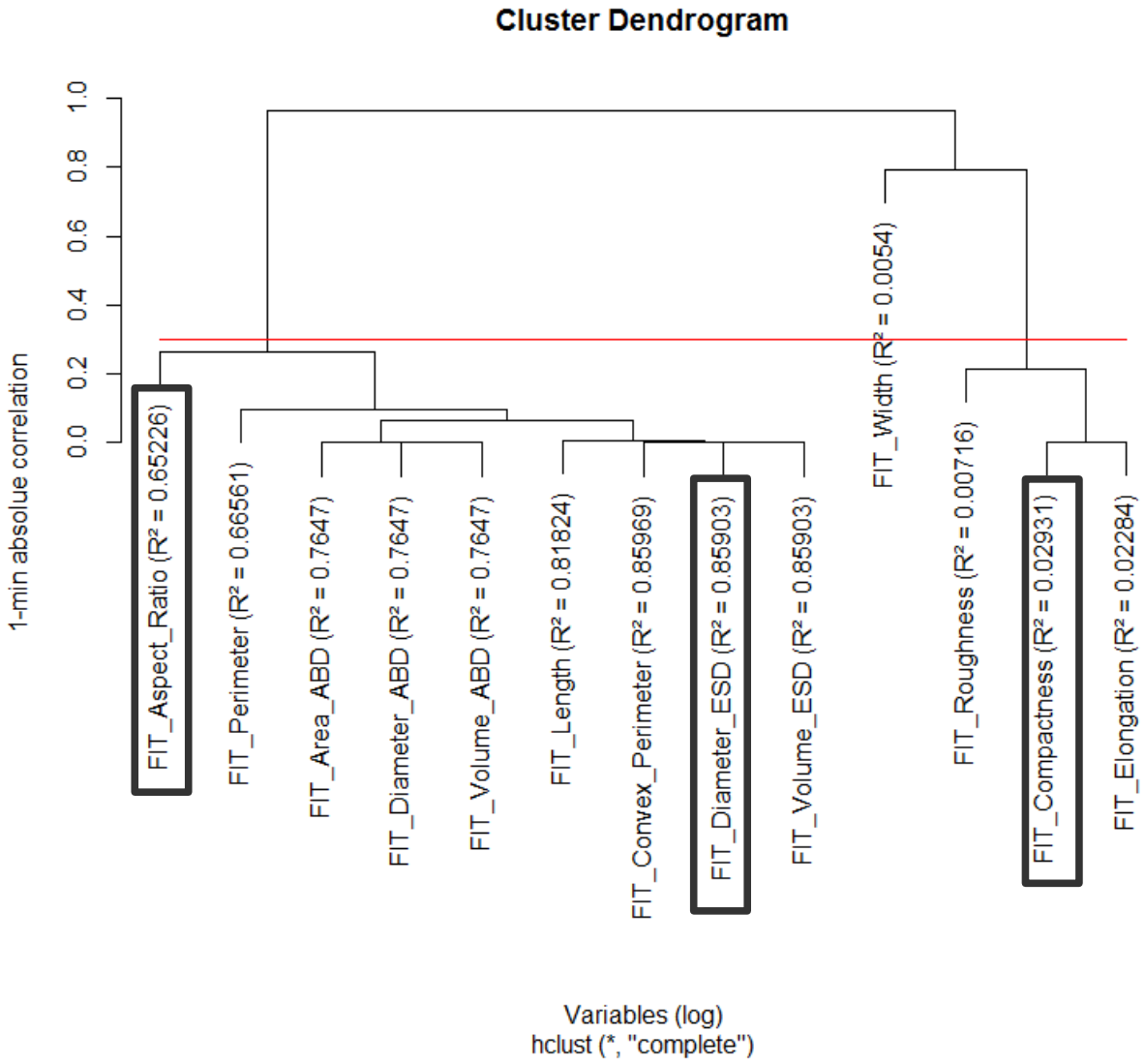
Représentation par classification hiérarchique utilisant des liens complets unissant les distances relatives de la matrice de corrélation (1-minimum absolu de corrélation) des variables transformées en logarithme et de la relation suivant le R^2 entre chaque prédicteurs et le nombre de cellule par colonie pour *Biddulphia rhombus*. Seuils de dissimilarité représentés (lignes horizont.) : 0.3. || 4 prédicteurs potentiels : log.Cells ~ log.Area_ABD + log.Length + log.Roughness + log.Compactness

Biddulphia sinensis



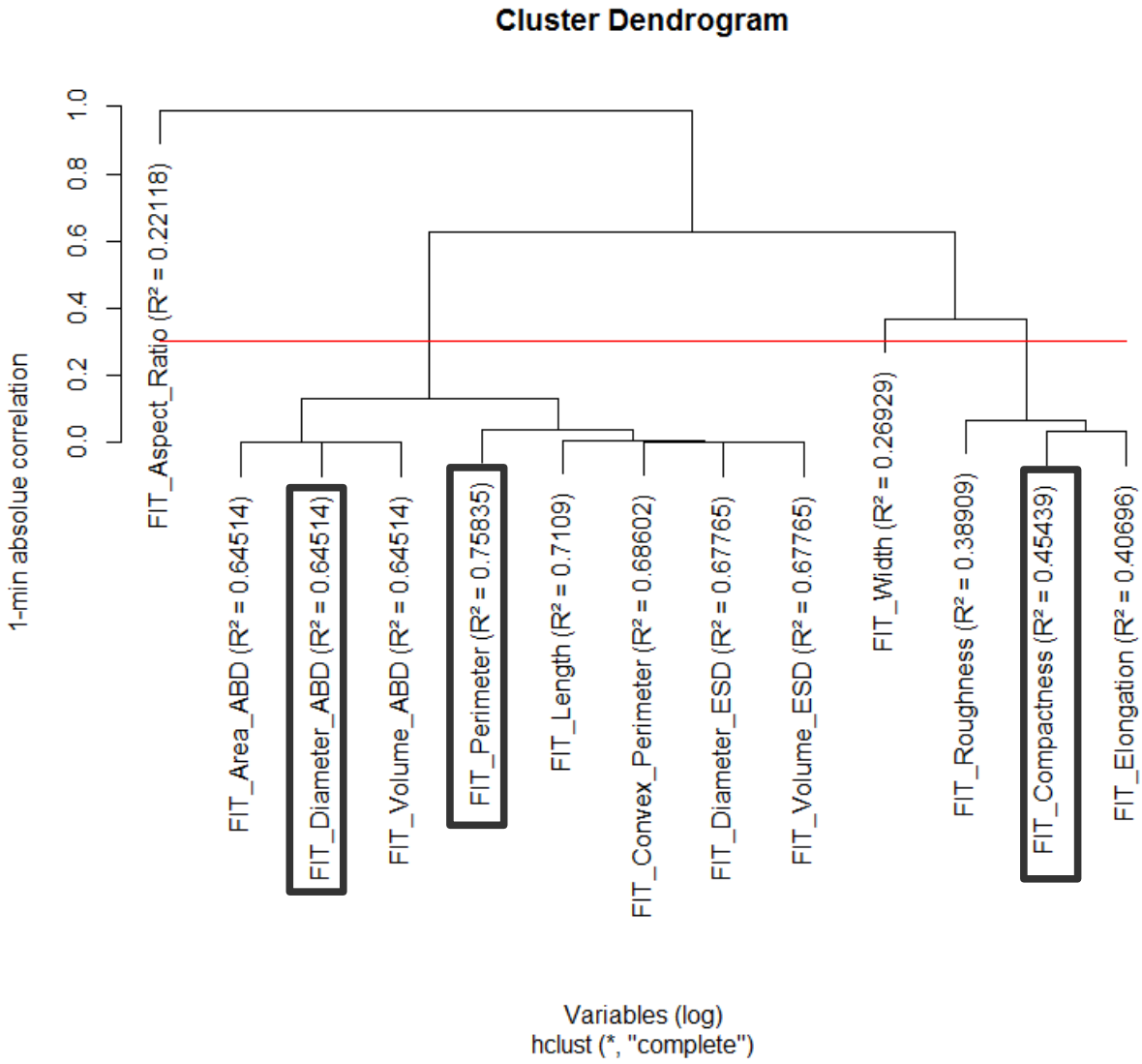
Représentation par classification hiérarchique utilisant des liens complets unissant les distances relatives de la matrice de corrélation (1-minimum absolu de corrélation) des variables transformées en logarithme et de la relation suivant le R^2 entre chaque prédicteurs et le nombre de cellule par colonie pour *Biddulphia sinensis*. Seuils de dissimilarité représentés (lignes horizont.) : 0.3. || 3 prédicteurs potentiels : log.Cells ~ log.Compactness + log.Width + log.Perimeter

Ditylum brightwellii



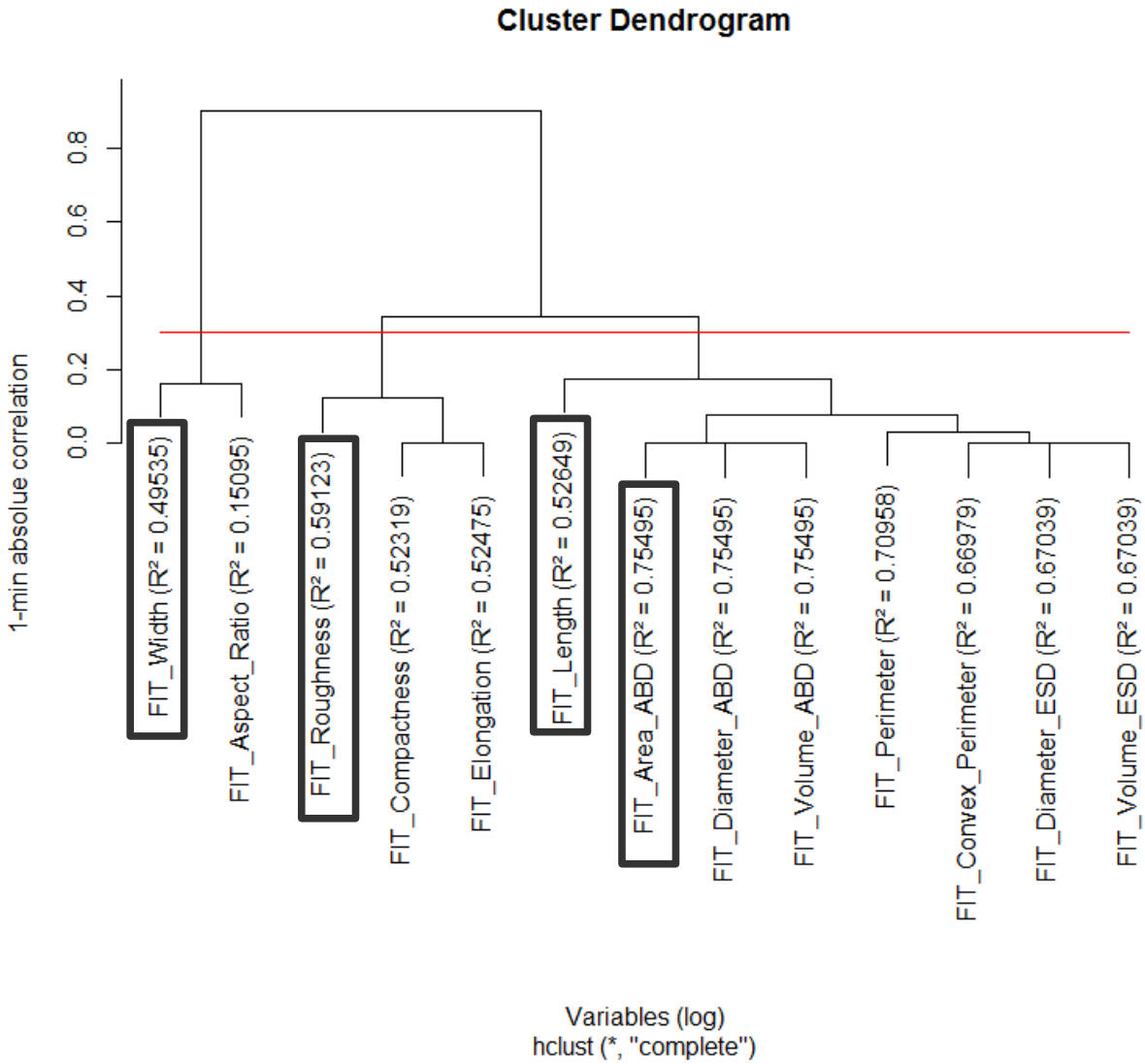
Représentation par classification hiérarchique utilisant des liens complets unissant les distances relatives de la matrice de corrélation (1-minimum absolu de corrélation) des variables transformées en logarithme et de la relation suivant le R^2 entre chaque prédicteurs et le nombre de cellule par colonie pour *Ditylum brightwellii*. Seuils de dissimilarité représentés (lignes horizont.) : 0.3. || 3 prédicteurs potentiels : log.Cells ~ log.Aspect_Ratio + log.Diameter_ESD + log.Compactness

Thalassiosira rotula



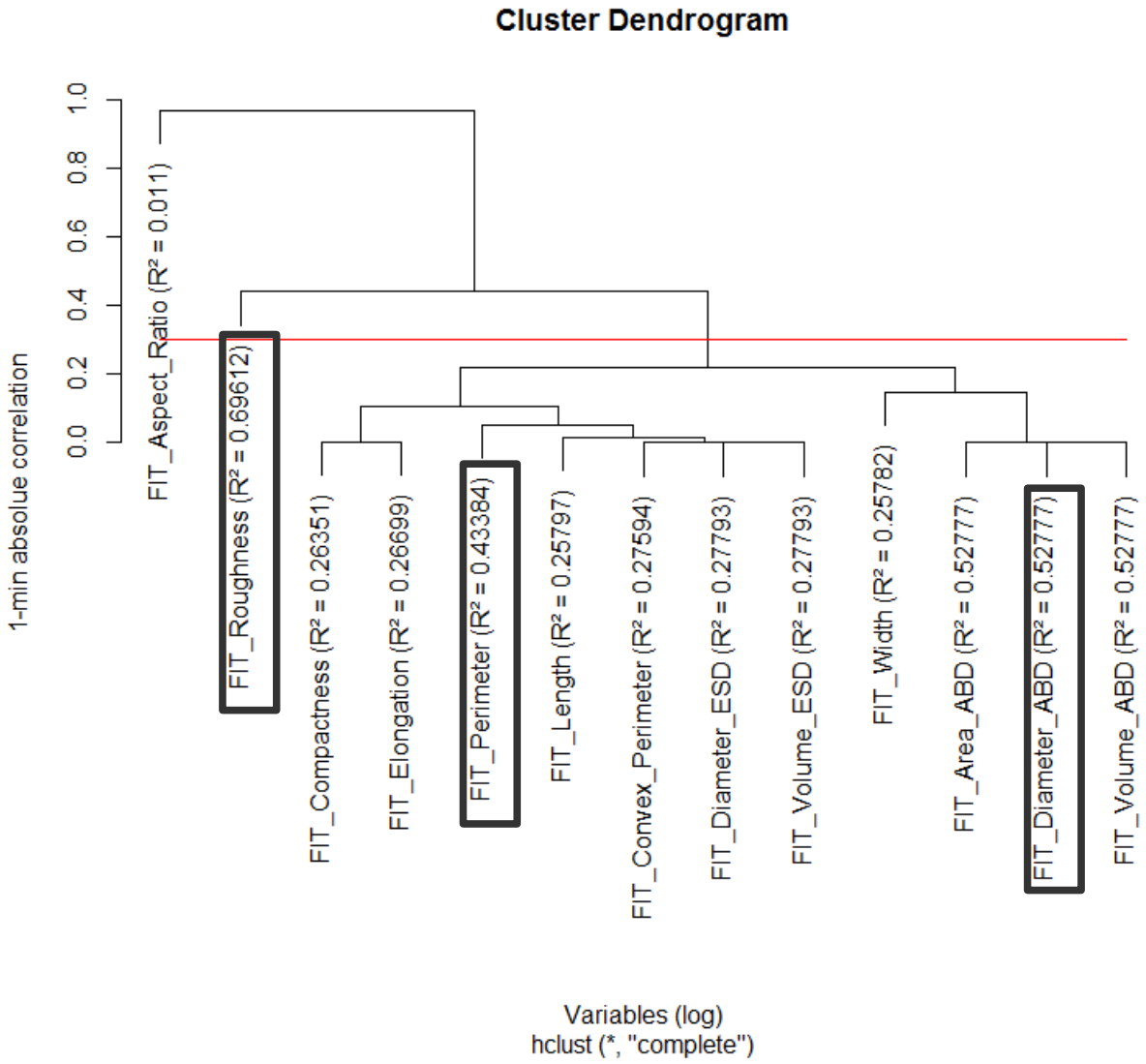
Représentation par classification hiérarchique utilisant des liens complets unissant les distances relatives de la matrice de corrélation (1-minimum absolu de corrélation) des variables transformées en logarithme et de la relation suivant le R^2 entre chaque prédicteurs et le nombre de cellule par colonie pour *Thalassiosira rotula*. Seuils de dissimilarité représentés (lignes horizont.) : 0.3. || 3 prédicteurs potentiels : log.Cells ~ log.Diameter_ABD + log.Perimeter + log.Compactness

Asterionella glacialis



Représentation par classification hiérarchique utilisant des liens complets unissant les distances relatives de la matrice de corrélation (1-minimum absolu de corrélation) des variables transformées en logarithme et de la relation suivant le R^2 entre chaque prédicteurs et le nombre de cellule par colonie pour *Asterionella glacialis.* Seuils de dissimilarité représentés (lignes horizont.) : 0.3. || 4 prédicteurs potentiels : log.Cells ~ log.Width + log.Length + log.Roughness + log.Area_ABD

Thalassionema nitzschoides



Représentation par classification hiérarchique utilisant des liens complets unissant les distances relatives de la matrice de corrélation (1-minimum absolu de corrélation) des variables transformées en logarithme et de la relation suivant le R^2 entre chaque prédicteurs et le nombre de cellule par colonie pour *Thalassionema nitzschoides*. Seuils de dissimilarité représentés (lignes horizont.) : 0.3. || 3 prédicteurs potentiels : log.Cells ~ log.Diameter_ABD + log.Perimeter + log.Roughness

A la vue de ces résultats, et après sélection des prédicteurs, nous choisissons d'utiliser 3 algorithmes classiques de machine learning :

- LDA : Linear Discriminant Analysis (Analyse Discriminante Linéaire),
- Tree : Arbre de décision (méthode de partitionnement par arbre décisionnel),
- RF : Random Forest (méthode de partitionnement par forêt d'arbres décisionnels).

Table 9 : Table représentant les %VP (pour la reconnaissance des vignettes) pour l'ensemble des groupes taxinomiques étudiés, provenant d'échantillons naturels (300µm/4X), et en fonction de la méthode utilisée. A = méthode Automatique, M = méthode Manuelle.

Groupes taxinomiques	Variables (log)	R^2	BIC	LM (%VP)	LDA (%VP)	Tree (%VP)	RF (%VP)
Biddulphia rhombus	FIT_Length	0.7914	-26.7136	94.44%	X	94.44%	X
	(M) FIT_Length+FIT_Area_ABD+FIT_Compactness+FIT_Roughness	0.8652	-42.3409	94.44%	96.30%	94.44%	94.44%
	(A) FIT_Length + FIT_Compactness + FIT_Area_ABD + FIT_Width + FIT_Diameter_ESD + FIT_Roughness	0.9162	-56.0341	96.30%	96.30%	94.44%	94.44%
	ALL VARIABLES	0.9192	-46.0202	100.00%	100.00%	94.44%	92.59%
Biddulphia sinensis	FIT_Perimeter	0.7657	-10.5145	70.31%	X	82.81%	X
	(M) FIT_Perimeter + FIT_Width + FIT_Compactness	0.7702	-3.4399	71.78%	87.50%	87.50%	79.69%
	(A) FIT_Perimeter	0.7657	-10.5145	70.31%	X	82.81%	X
	ALL VARIABLES	0.7991	+12.9169	76.56%	90.62%	87.50%	79.69%
Ditylum brightwellii	FIT_Diameter_ESD	0.8590	-41.9986	92.42%	X	93.94%	X
	(M) FIT_Diameter_ESD + FIT_Aspect_Ratio + FIT_Compactness	0.8745	-41.2869	92.42%	92.42%	93.94%	84.85%
	(A) FIT_Length+FIT_Aspect_Ratio+FIT_Roughness+FIT_Compactness	0.8561	-28.0694	93.94%	92.42%	93.94%	84.33%
	ALL VARIABLES	0.9154	-42.1839	93.94%	96.97%	95.45%	90.91%
Thalassiosira rotula	FIT_Perimeter	0.7583	+14.1891	68.42%	X	63.16%	X
	(M) FIT_Diameter_ABD + FIT_Perimeter + FIT_Compactness	0.7722	+18.9184	66.67%	75.44%	59.65%	64.91%
	(A) FIT_Diameter_ABD + FIT_Length + FIT_Diameter_ESD + FIT_Convex_Perimeter + FIT_Compactness	0.8341	+8.9089	77.19%	75.44%	68.42%	66.67%
	ALL VARIABLES	0.8409	+26.7460	73.68%	63.16%	59.65%	64.91%
Asterionella glacialis	FIT_Area_ABD	0.7550	-6.8314	52.46%	X	52.46%	X
	(M) FIT_Area_ABD + FIT_Width + FIT_Length + FIT_Roughness	0.7593	+4.4098	52.46%	68.85%	60.66%	54.10%
	(A) FIT_Area_ABD + FIT_Width + FIT_Length + FIT_Perimeter + FIT_Convex_Perimeter + FIT_Roughness	0.8162	-3.8078	55.74%	75.41%	65.57%	50.82%
	ALL VARIABLES	0.8245	+9.7895	59.02%	77.05%	54.10%	54.10%
Thalassionema nitzschoides	FIT_Diameter_ABD	0.5278	+15.7334	38.57%	X	70.00%	X
	(M) FIT_Diameter_ABD + FIT_Perimeter + FIT_Roughness	0.7483	-19.8047	58.57%	75.71%	80.00%	68.57%
	(A) FIT_Diameter_ABD + FIT_Width + FIT_Roughness	0.7568	-22.2303	57.14%	80.00%	74.29%	58.57%
	ALL VARIABLES	0.7761	+1.7166	54.29%	85.71%	75.71%	69.14%

Globalement, pour les trois dernières espèces (*Thalassiosira rotula, Asterionella glacialis, Thalassionema* nitzschoides), un modèle multivarié permet d'améliorer la prédiction par rapport au modèle univarié. De plus, nous pouvons noter que les taux d'erreur les plus faibles sont, pour la plupart du temps, obtenus par l'analyse discriminante linéaire (LDA) représentés en bleu dans la table 9, et en particulier, en utilisant les prédicteurs retenus au sens du critère BIC. Par exemple, pour *Thalassiosira rotula*, l'ensemble des méthodes prédictives testées estiment correctement le nombre de cellules par colonies dans une fourchette comprise entre 66% et 77% de %VP, avec les variables sélectionnées automatiquement.

Cependant, les scores de reconnaissance obtenus pour certaines espèces (comme *Asterionella glacialis* et *Thalassionema nitzschoides*), étant relativement faibles, il est nécessaire de se tourner vers des techniques de sélection de modèles plus complexes.

Dans la suite du rapport, nous étudions l'impact de l'ajout de variables pour la sélection de prédicteurs pour la construction de modèles prédictifs.

Utilisation des attributs FlowCAM et ZooPhytoImage v.5

Dans cette section, les variables calculés dans la version 5 de Zoo/PhytoImage sont ajoutées aux variables issues du FlowCAM. Elles sont alors utilisées pour construire des modèles prédictifs. Ces nouvelles variables sont les suivantes : **ECD**, **FeretRoundness**, **MeanFDia**, **Perim_Ratio**, **Transp2** et **CV**.

Ici, les mêmes scores de performances sont évalués :

- **%VP vignettes** : le pourcentage de vignettes correctement prédites (Vrais-Positifs) dans les différentes classes (une classe étant représentée par un nombre de cellules par particule).
- **%VP cells/colonie** : le pourcentage de cellules correctement prédites (Vrais-Positifs) dans les différentes classes (une classe étant représentée par un nombre de cellules par particule).
- **Estimation totale** : le pourcentage de l'abondance du nombre de cellules total prédit sur l'abondance du nombre de cellules total mesurée manuellement dans l'échantillon.

Modèle linéaire à un prédicteur (basé sur le critère R²)

Table 11 : Table représentant les meilleures variables au sens du R² pour l'ensemble des groupes taxinomiques étudiés, provenant d'échantillons naturels (300µm/4X).

Groupes taxinomiques (300µm/4X)	Variable retenue (log)	Relation	R²	p-value	LM (%VP)
Biddulphia rhombus	FIT_Compactness	1.1445 V – 0.5525	0.7936	2.20e-16	90.74%
Biddulphia sinensis	FIT_Perimeter	0.9057 V - 4.9944	0.7657	2.20e-16	73.44%
Ditylum brightwellii	FIT_Convex_Perimeter	1.0796 V – 6.4721	0.8597	2.20e-16	90.91%
Thalassiosira rotula	MeanFDia	0.9637 V - 2.8283	0.7585	2.20e-16	61.41%
Asterionella glacialis	ECD	1.6775 V - 4.5041	0.7550	2.20e-16	52.46%
Thalassionema nitzschoides	FIT_Roughness	2.6928 V + 0.4905	0.6961	2.20e-16	54.29%

Pour les trois derniers exemples, nous observons que les meilleurs prédicteurs sélectionnés au sens du R², ne permettent pas de prédire précisément le nombre de cellules dans les colonies. C'est pourquoi, nous nous tournons, comme auparavant, vers la sélection de modèle linéaire multivarié.

Modèle linéaire multivarié (basé sur le critère BIC)

De la même manière que dans la section précédente, nous analysons les relations entre variables afin de mettre en évidence des prédicteurs qui seront utilisés pour la construction de modèles prédictifs par régression linéaire et non linéaire. L'analyse de ces relations est effectuée par deux méthodes :

- Analyse manuelle des relations entre variables,
- Analyse automatique des relations entre variables.

Dans cette première expérimentation, nous reprenons les 3 algorithmes classiques de machine learning cités précédemment :

- LDA : Linear Discriminant Analysis (Analyse Discriminante Linéaire),
- Tree : Arbre de décision (méthode de partitionnement par arbre décisionnel),
- RF : Random Forest (méthode de partitionnement par forêt d'arbres décisionnels).

Groupes taxinomiques	Variables (log)	R^2	BIC	LM (%VP)	LDA (%VP)	Tree (%VP)	RF (%VP)
Biddulphia rhombus	FIT_Compactness	0.7936	-27.31	90.74%	X	94.44%	X
	(M) FIT_Compactness + FIT_Length + FIT_Aspect_Ratio + FIT_Roughness + ECD	0.9075	-54.71	98.15%	100.00%	94.44%	90.74%
	(A) FIT_Width + FIT_Diameter_ESD + ECD + MeanFDia	0.9217	-67.68	100.00%	100.00%	94.44%	92.59%
	ALL VARIABLES	0.9253	-38.27	100.00%	100.00%	94.44%	92.59%
Biddulphia sinensis	FIT_Perimeter	0.7657	-10.51	73.44%	X	85.94%	X
	(M) FIT_Compactness + FIT_Width + FIT_Perimeter + ECD	0.7703	+0.69	73.44%	87.50%	87.50%	79.69%
	(A) FIT_Perimeter	0.7657	-10.51	73.44%	X	87.50%	X
	ALL VARIABLES	0.8043	+23.72	71.88%	90.62%	87.50%	76.56%
Ditylum brightwellii	FIT_Convex_Perimeter	0.8597	-42.31	90.91%	X	93.94%	X
	(M) FIT_Convex_Perimeter+FIT_Aspect_Ratio+FIT_Compactness+ECD	0.8784	-39.19	90.91%	95.45%	95.45%	92.42%
	(A) FIT_Width + FIT_Diameter_ESD + FIT_Convex_Perimeter + FIT_Compactness + FIT_Roughness + ECD + MeanFDia	0.9176	-52.30	93.94%	93.94%	95.45%	89.39%
	ALL VARIABLES	0.9231	-35.88	93.94%	96.97%	95.45%	89.39%
Thalassiosira rotula	MeanFDia	0.7585	+14.16	61.41%	X	68.42%	X
	(M) FIT_Width + FIT_Compactness + FIT_Aspect_Ratio + FIT_Length + ECD + MeanFDia	0.7974	+20.31	66.67%	64.91%	70.18%	61.40%
	(A) FIT_Length + FIT_Width + FIT_Diameter_ESD + FIT_Convex_Perimeter + CV + Transp2	0.8523	+6.34	80.70%	80.70%	70.18%	63.16%
	ALL VARIABLES	0.8633	+30.24	82.46%	82.46%	61.40%	61.40%
Asterionella glacialis	ECD	0.7550	-6.83	52.46%	X	52.46%	X
	(M) FIT_Roughness + FIT_Length + FIT_Perimeter + FIT_Width + ECD	0.7596	+8.44	55.74%	70.49%	65.57%	54.10%
	(A) FIT_Compactness+FIT_Elongation+FIT_Roughness+ECD+Transp2	0.8156	-7.75	57.48%	73.77%	62.30%	44.26%
	ALL VARIABLES	0.8307	+19.96	54.10%	73.77%	57.78%	52.46%
Thalassionema nitzschoides	FIT_Roughness	0.6961	-15.13	54.29%	X	62.86%	X
	(M) FIT_Elongation + FIT_Perimeter + FeretRoundness	0.6581	+1.63	52.86%	75.71%	68.57%	52.86%
	(A) FIT_Convex_Perimeter + FIT_Compactness + FIT_Volume_ABD	0.7514	-20.69	60.00%	75.71%	70.00%	55.71%
	ALL VARIABLES	0.7945	+8.46	51.63%	88.57%	81.43%	67.14%

Table 12 : Table représentant les %VP (pour la reconnaissance des vignettes) pour l'ensemble des groupes taxinomiques étudiés, provenant d'échantillons naturels (300µm/4X), et en fonction de la méthode utilisée. A = méthode Automatique, M = méthode Manuelle.

Groupes taxinomiques	Variables (log)	R^2	BIC	LM (%VP)	LDA (%VP)	Tree (%VP)	RF (%VP)
Biddulphia rhombus	FIT_Compactness	0.7936	-27.31	86.67%	X	88.00%	X
	(M) FIT_Compactness + FIT_Length + FIT_Aspect_Ratio + FIT_Roughness + ECD	0.9075	-54.71	94.67%	100.00%	88.00%	84.00%
	(A) FIT_Width + FIT_Diameter_ESD + ECD + MeanFDia	0.9217	-67.68	100.00%	100.00%	88.00%	86.67%
	ALL VARIABLES	0.9253	-38.27	100.00%	100.00%	88.00%	86.67%
Biddulphia sinensis	FIT_Perimeter	0.7657	-10.51	66.67%	X	80.25%	X
	(M) FIT_Compactness + FIT_Width + FIT_Perimeter + ECD	0.7703	+0.69	64.20%	83.95%	83.95%	72.22%
	(A) FIT_Perimeter	0.7657	-10.51	66.67%	X	80.25%	X
	ALL VARIABLES	0.8043	+23.72	60.49%	84.57%	81.48%	70.99%
Ditylum brightwellii	FIT_Convex_Perimeter	0.8597	-42.31	82.11%	X	86.32%	X
	(M) FIT_Convex_Perimeter+FIT_Aspect_Ratio+FIT_Compactness+ECD	0.8784	-39.19	82.11%	90.53%	88.42%	84.21%
	(A) FIT_Width + FIT_Diameter_ESD + FIT_Convex_Perimeter + FIT_Compactness + FIT_Roughness + ECD + MeanFDia	0.9176	-52.30	90.53%	88.42%	88.42%	78.95%
	ALL VARIABLES	0.9231	-35.88	90.53%	91.58%	88.42%	78.95%
Thalassiosira rotula	MeanFDia	0.7585	+14.16	47.24%	X	48.74%	X
	(M) FIT_Width + FIT_Compactness + FIT_Aspect_Ratio + FIT_Length + ECD + MeanFDia	0.7974	+20.31	51.76%	58.29%	51.26%	47.24%
	(A) FIT_Length + FIT_Width + FIT_Diameter_ESD + FIT_Convex_Perimeter + CV + Transp2	0.8523	+6.34	65.83%	67.84%	48.74%	46.73%
	ALL VARIABLES	0.8633	+30.24	66.83%	67.84%	43.22%	45.73%
Asterionella glacialis	ECD	0.7550	-6.83	47.76%	X	49.25%	X
	(M) FIT_Roughness + FIT_Length + FIT_Perimeter + FIT_Width + ECD	0.7596	+8.44	50.37%	64.55%	57.09%	48.13%
	(A) FIT_Compactness+FIT_Elongation+FIT_Roughness+ECD+Transp2	0.8156	-7.75	50.00%	69.03%	57.84%	41.79%
	ALL VARIABLES	0.8307	+19.96	47.01%	70.15%	50.37%	49.25%
Thalassionema nitzschoides	FIT_Roughness	0.6961	-15.13	52.80%	X	58.74%	X
	(M) FIT_Elongation + FIT_Perimeter + FeretRoundness	0.6581	+1.63	48.25%	75.17%	61.54%	49.65%
	(A) FIT_Convex_Perimeter + FIT_Compactness + FIT_Volume_ABD	0.7514	-20.69	59.44%	69.23%	63.99%	52.10%
	ALL VARIABLES	0.7945	+8.46	51.75%	79.02%	72.03%	62.94%

Table 13 : Table représentant les %VP (pour la reconnaissance du nombre de cellules par colonies) pour l'ensemble des groupes taxinomiques étudiés, provenant d'échantillons naturels (300µm/4X), et en fonction de la méthode utilisée. A = méthode Automatique, M = méthode Manuelle.

Groupes taxinomiques	Variables (log)	R^2	BIC	LM	LDA	Tree	RF
Biddulphia rhombus	FIT_Compactness	0.7936	-27.31	99.95%	X	94.67%	X
	(M) FIT_Compactness + FIT_Length + FIT_Aspect_Ratio + FIT_Roughness + ECD	0.9075	-54.71	101.33%	100.00%	94.67%	94.67%
	(A) FIT_Width + FIT_Diameter_ESD + ECD + MeanFDia	0.9217	-67.68	100.00%	100.00%	94.67%	96.00%
	ALL VARIABLES	0.9253	-38.27	100.00%	100.00%	94.67%	96.00%
Biddulphia sinensis	FIT_Perimeter	0.7657	-10.51	95.06%	X	97.53%	X
	(M) FIT_Compactness + FIT_Width + FIT_Perimeter + ECD	0.7703	+0.69	96.30%	100.00%	100.62%	98.77%
	(A) FIT_Perimeter	0.7657	-10.51	95.06%	X	97.53%	X
	ALL VARIABLES	0.8043	+23.72	98.15%	96.30%	96.91%	96.91%
Ditylum brightwellii	FIT_Convex_Perimeter	0.8597	-42.31	97.89%	X	93.68%	X
	(M) FIT_Convex_Perimeter+FIT_Aspect_Ratio+FIT_Compactness+ECD	0.8784	-39.19	97.89%	97.89%	94.74%	94.74%
	(A) FIT_Width + FIT_Diameter_ESD + FIT_Convex_Perimeter + FIT_Compactness + FIT_Roughness + ECD + MeanFDia	0.9176	-52.30	100.00%	98.95%	94.74%	94.74%
	ALL VARIABLES	0.9231	-35.88	100.00%	96.84%	94.74%	94.74%
Thalassiosira rotula	MeanFDia	0.7585	+14.16	91.46%	X	88.94%	X
	(M) FIT_Width + FIT_Compactness + FIT_Aspect_Ratio + FIT_Length + ECD + MeanFDia	0.7974	+20.31	95.48%	87.44%	94.47%	85.93%
	(A) FIT_Length + FIT_Width + FIT_Diameter_ESD + FIT_Convex_Perimeter + CV + Transp2	0.8523	+6.34	97.49%	88.94%	88.44%	83.92%
	ALL VARIABLES	0.8633	+30.24	96.98%	86.43%	92.96%	87.44%
Asterionella glacialis	ECD	0.7550	-6.83	98.51%	X	100.37%	X
	(M) FIT_Roughness + FIT_Length + FIT_Perimeter + FIT_Width + ECD	0.7596	+8.44	99.25%	97.39%	97.39%	97.76%
	(A) FIT_Compactness+FIT_Elongation+FIT_Roughness+ECD+Transp2	0.8156	-7.75	98.51%	98.51%	100.00%	97.39%
	ALL VARIABLES	0.8307	+19.96	99.25%	98.88%	99.63%	98.13%
Thalassionema nitzschoides	FIT_Roughness	0.6961	-15.13	99.30%	X	98.95%	X
	(M) FIT_Elongation + FIT_Perimeter + FeretRoundness	0.6581	+1.63	96.85%	97.20%	98.25%	96.15%
	(A) FIT_Convex_Perimeter + FIT_Compactness + FIT_Volume_ABD	0.7514	-20.69	98.60%	98.25%	103.85%	96.15%
	ALL VARIABLES	0.7945	+8.46	99.65%	95.45%	97.90%	100.35%

Table 14 : Table représentant les % d'estimation du nombre de cellules par colonies pour l'ensemble des groupes taxinomiques étudiés, provenant d'échantillons naturels (300µm/4X), et en fonction de la méthode utilisée. A = méthode Automatique, M = méthode Manuelle.

Globalement, à la vue de ces trois tableaux (tables 12, 13 et 14), il apparaît que la méthode LDA donne quasiment toujours les meilleures scores de reconnaissance (table 12, pour les vignettes et table 13, pour le nombre de cellules par colonies). Cependant, le nombre de vrais positifs pour les trois dernières espèces (*Thalassiosira rotula, Asterionella glacialis, Thalassionema* nitzschoides), reste assez faible. C'est pourquoi, il est nécessaire de se tourner vers des techniques de sélection de modèles plus complexe.

Modèles de régression non linéaires

La méthode linéaire permettant d'obtenir les meilleurs scores de reconnaissance est l'analyse discriminante linéaire (LDA). C'est pourquoi, nous souhaitons appliquer des méthodes non linéaires dérivées de la LDA, ainsi qu'un algorithme de machine learning basé sur le principe des réseaux de neurones. Nous utilisons donc ici :

- MDA : Mixture Discriminant Analysis. L'idée est de modéliser chaque classe par un mélange de deux ou plusieurs gaussiennes avec différents centroïdes, mais avec chaque composante gaussienne, intra et inter classes, partageant la même matrice de covariance. Cela permet de définir des frontières plus complexes. Cette extension est appelée MDA.
- FDA : Flexible Discriminant Analysis. Il s'agit d'une généralisation de la régression linéaire pour des formes de régression plus flexibles et non paramétriques. Cette forme plus flexible d'analyse discriminante est appelée FDA .
- NNET : Neural Network. Réseau de neurones à une couche cachée (initialisation par défaut : weights = 1, size = 1, maxit = 100).

Groupes taxinomiques	Variables (log)	R^2	BIC	LM (%VP)	MDA (%VP)	FDA (%VP)	NNET (%VP)
Biddulphia rhombus	FIT_Compactness	0.7936	-27.31	90.74%	96.30%	94.44%	96.30%
	(M) FIT_Compactness + FIT_Length + FIT_Aspect_Ratio + FIT_Roughness + ECD	0.9075	-54.71	98.15%	100.00%	100.00%	98.15%
	(A) FIT_Width + FIT_Diameter_ESD + ECD + MeanFDia	0.9217	-67.68	100.00%	100.00%	100.00%	98.15%
	ALL VARIABLES	0.9253	-38.27	100.00%	100.00%	100.00%	98.15%
Biddulphia sinensis	FIT_Perimeter	0.7657	-10.51	73.44%	90.63%	85.94%	77.19%
	(M) FIT_Compactness + FIT_Width + FIT_Perimeter + ECD	0.7703	+0.69	73.44%	95.31%	85.94%	77.19%
	(A) FIT_Perimeter	0.7657	-10.51	73.44%	90.63%	85.94%	77.19%
	ALL VARIABLES	0.8043	+23.72	71.88%	95.31%	90.63%	71.88%
Ditylum brightwellii	FIT_Convex_Perimeter	0.8597	-42.31	90.91%	93.94%	92.42%	92.42%
	(M) FIT_Convex_Perimeter+FIT_Aspect_Ratio+FIT_Compactness+ECD	0.8784	-39.19	90.91%	93.94%	95.45%	96.97%
	(A) FIT_Width + FIT_Diameter_ESD + FIT_Convex_Perimeter + FIT_Compactness + FIT_Roughness + ECD + MeanFDia	0.9176	-52.30	93.94%	95.45%	93.94%	98.48%
	ALL VARIABLES	0.9231	-35.88	93.94%	98.48%	96.97%	96.97%
Thalassiosira rotula	MeanFDia	0.7585	+14.16	61.41%	78.95%	59.65%	61.40%
	(M) FIT_Width + FIT_Compactness + FIT_Aspect_Ratio + FIT_Length + ECD + MeanFDia	0.7974	+20.31	66.67%	73.69%	75.44%	63.16%
	(A) FIT_Length + FIT_Width + FIT_Diameter_ESD + FIT_Convex_Perimeter + CV + Transp2	0.8523	+6.34	80.70%	87.72%	77.19%	63.16%
	ALL VARIABLES	0.8633	+30.24	82.46%	89.47%	85.96%	64.91%
Asterionella glacialis	ECD	0.7550	-6.83	52.46%	59.02%	52.46%	32.79%
	(M) FIT_Roughness + FIT_Length + FIT_Perimeter + FIT_Width + ECD	0.7596	+8.44	55.74%	88.52%	68.85%	32.79%
	(A) FIT_Compactness+FIT_Elongation+FIT_Roughness+ECD+Transp2	0.8156	-7.75	57.38%	85.25%	70.49%	32.79%
	ALL VARIABLES	0.8307	+19.96	54.10%	98.36%	75.41%	32.79%
Thalassionema nitzschoides	FIT_Roughness	0.6961	-15.13	54.29%	74.29%	72.86%	65.71%
	(M) FIT_Elongation + FIT_Perimeter + FeretRoundness	0.6581	+1.63	52.86%	81.43%	74.29%	65.71%
	(A) FIT_Convex_Perimeter + FIT_Compactness + FIT_Volume_ABD	0.7514	-20.69	60.00%	85.71%	75.71%	65.71%
	ALL VARIABLES	0.7945	+8.46	51.43%	88.57%	87.14%	65.71%

Table 16 : Table représentant les %VP (pour la reconnaissance des vignettes) pour l'ensemble des groupes taxinomiques étudiés, provenant d'échantillons naturels (300µm/4X), et en fonction de la méthode utilisée. A = méthode Automatique, M = méthode Manuelle.

Groupes taxinomiques	Variables (log)	R^2	BIC	LM (%VP)	MDA (%VP)	FDA (%VP)	NNET (%VP)
Biddulphia rhombus	FIT_Compactness	0.7936	-27.31	86.67%	96.00%	92.00%	92.00%
	(M) FIT_Compactness + FIT_Length + FIT_Aspect_Ratio + FIT_Roughness + ECD	0.9075	-54.71	94.67%	100.00%	100.00%	94.67%
	(A) FIT_Width + FIT_Diameter_ESD + ECD + MeanFDia	0.9217	-67.68	100.00%	100.00%	100.00%	94.67%
	ALL VARIABLES	0.9253	-38.27	100.00%	100.00%	100.00%	94.67%
Biddulphia sinensis	FIT_Perimeter	0.7657	-10.51	66.67%	88.89%	81.48%	50.62%
	(M) FIT_Compactness + FIT_Width + FIT_Perimeter + ECD	0.7703	+0.69	64.20%	82.59%	82.72%	50.62%
	(A) FIT_Perimeter	0.7657	-10.51	66.67%	88.89%	81.48%	50.62%
	ALL VARIABLES	0.8043	+23.72	60.49%	93.83%	84.57%	54.32%
Ditylum brightwellii	FIT_Convex_Perimeter	0.8597	-42.31	82.11%	87.37%	85.26%	85.26%
	(M) FIT_Convex_Perimeter+FIT_Aspect_Ratio+FIT_Compactness+ECD	0.8784	-39.19	82.11%	89.47%	10.53%	91.58%
	(A) FIT_Width + FIT_Diameter_ESD + FIT_Convex_Perimeter + FIT_Compactness + FIT_Roughness + ECD + MeanFDia	0.9176	-52.30	10.53%	10.53%	88.42%	93.68%
	ALL VARIABLES	0.9231	-35.88	10.53%	93.68%	91.58%	91.58%
Thalassiosira rotula	MeanFDia	0.7585	+14.16	47.24%	63.32%	41.71%	38.69%
	(M) FIT_Width + FIT_Compactness + FIT_Aspect_Ratio + FIT_Length + ECD + MeanFDia	0.7974	+20.31	51.76%	60.30%	59.80%	39.70%
	(A) FIT_Length + FIT_Width + FIT_Diameter_ESD + FIT_Convex_Perimeter + CV + Transp2	0.8523	+6.34	65.83%	72.36%	62.81%	39.70%
	ALL VARIABLES	0.8633	+30.24	66.83%	74.87%	71.86%	40.70%
Asterionella glacialis	ECD	0.7550	-6.83	47.76%	54.10%	48.13%	25.77%
	(M) FIT_Roughness + FIT_Length + FIT_Perimeter + FIT_Width + ECD	0.7596	+8.44	50.37%	85.45%	62.31%	25.77%
	(A) FIT_Compactness+FIT_Elongation+FIT_Roughness+ECD+Transp2	0.8156	-7.75	50.00%	83.21%	66.04%	25.77%
	ALL VARIABLES	0.8307	+19.96	47.01%	97.76%	72.39%	25.77%
Thalassionema nitzschoides	FIT_Roughness	0.6961	-15.13	52.80%	70.28%	68.88%	28.39%
	(M) FIT_Elongation + FIT_Perimeter + FeretRoundness	0.6581	+1.63	48.25%	76.22%	73.78%	28.39%
	(A) FIT_Convex_Perimeter + FIT_Compactness + FIT_Volume_ABD	0.7514	-20.69	59.44%	75.52%	69.23%	28.39%
	ALL VARIABLES	0.7945	+8.46	51.75%	79.02%	77.62%	28.39%

Table 17 : Table représentant les %VP (pour la reconnaissance du nombre de cellules par colonies) pour l'ensemble des groupes taxinomiques étudiés, provenant d'échantillons naturels (300µm/4X), et en fonction de la méthode utilisée. A = méthode Automatique, M = méthode Manuelle.

Groupes taxinomiques	Variables (log)	R^2	BIC	LM	MDA	FDA	NNET
Biddulphia rhombus	FIT_Compactness	0.7936	-27.31	99.95%	100.00%	98.67%	97.33%
	(M) FIT_Compactness + FIT_Length + FIT_Aspect_Ratio+FIT_Roughness + ECD	0.9075	-54.71	101.33%	100.00%	100.00%	98.67%
	(A) FIT_Width + FIT_Diameter_ESD + ECD + MeanFDia	0.9217	-67.68	100.00%	100.00%	100.00%	98.67%
	ALL VARIABLES	0.9253	-38.27	100.00%	100.00%	100.00%	98.67%
Biddulphia sinensis	FIT_Perimeter	0.7657	-10.51	95.06%	103.70%	98.15%	89.51%
	(M) FIT_Compactness + FIT_Width + FIT_Perimeter + ECD	0.7703	+0.69	96.30%	98.15%	99.38%	89.51%
	(A) FIT_Perimeter	0.7657	-10.51	95.06%	103.70%	98.15%	89.51%
	ALL VARIABLES	0.8043	+23.72	98.15%	100.62%	96.30%	90.12%
Ditylum brightwellii	FIT_Convex_Perimeter	0.8597	-42.31	97.89%	96.84%	97.89%	94.74%
	(M) FIT_Convex_Perimeter+FIT_Aspect_Ratio+FIT_Compactness+ECD	0.8784	-39.19	97.89%	97.89%	97.89%	96.84%
	(A) FIT_Width + FIT_Diameter_ESD + FIT_Convex_Perimeter + FIT_Compactness + FIT_Roughness + ECD + MeanFDia	0.9176	-52.30	100.00%	97.89%	98.95%	96.84%
	ALL VARIABLES	0.9231	-35.88	100.00%	97.89%	96.84%	97.89%
Thalassiosira rotula	MeanFDia	0.7585	+14.16	91.46%	84.92%	85.93%	70.85%
	(M) FIT_Width + FIT_Compactness + FIT_Aspect_Ratio + FIT_Length + ECD + MeanFDia	0.7974	+20.31	95.48%	89.45%	87.94%	70.85%
	(A) FIT_Length + FIT_Width + FIT_Diameter_ESD + FIT_Convex_Perimeter + CV + Transp2	0.8523	+6.34	97.49%	89.45%	88.95%	71.36%
	ALL VARIABLES	0.8633	+30.24	96.98%	89.95%	86.93%	70.85%
Asterionella glacialis	ECD	0.7550	-6.83	98.51%	96.27%	98.51%	68.28%
	(M) FIT_Roughness + FIT_Length + FIT_Perimeter + FIT_Width + ECD	0.7596	+8.44	99.25%	99.25%	97.76%	68.28%
	(A) FIT_Compactness+FIT_Elongation+FIT_Roughness+ECD+Transp2	0.8156	-7.75	98.51%	97.39%	99.25%	68.28%
	ALL VARIABLES	0.8307	+19.96	99.25%	98.89%	100.00%	68.28%
Thalassionema nitzschoides	FIT_Roughness	0.6961	-15.13	99.30%	100.00%	97.90%	73.43%
	(M) FIT_Elongation + FIT_Perimeter + FeretRoundness	0.6581	+1.63	96.85%	97.55%	96.50%	73.43%
	(A) FIT_Convex_Perimeter + FIT_Compactness + FIT_Volume_ABD	0.7514	-20.69	98.60%	96.15%	98.25%	73.43%
	ALL VARIABLES	0.7945	+8.46	99.65%	95.45%	94.76%	73.43%

Table 18 : Table représentant les % d'estimation du nombre de cellules par colonies pour l'ensemble des groupes taxinomiques étudiés, provenant d'échantillons naturels (300µm/4X), et en fonction de la méthode utilisée. A = méthode Automatique, M = méthode Manuelle.

Biddulphia rhombus

Pour *Biddulphia rhombus*, la « meilleure » méthode prédictive est la MDA multivariée (avec les variables sélectionnées automatiquement et manuellement, ainsi qu'avec toutes les variables).

Biddulphia sinensis

Pour *Biddulphia sinensis*, la « meilleure » méthode prédictive est la MDA multivariée (avec toutes les variables).

Ditylum brightwellii

Pour *Ditylum brightwellii*, la « meilleure » méthode prédictive est la MDA multivariée (avec toutes les variables).

Thalassiosira rotula

Pour *Thalassiosira rotula*, la « meilleure » méthode prédictive est la LM multivariée (avec toutes les variables).

Asterionella glacialis

Pour *Asterionella glacialis*, la « meilleure » méthode prédictive est la MDA multivariée (avec toutes les variables).

Thalassionema nitzschoides

Pour *Thalassionema nitzschioides*, la « meilleure » méthode prédictive est la MDA multivariée (avec toutes les variables).

Les graphes présentés ci-dessus pour chacune des espèces étudiées représentent :

- les scores de performance obtenus par les différentes méthodes prédictives (linéaires et non linéaires) UNIVARIEES,
- les scores de performance obtenus par les différentes méthodes prédictives (linéaires et non linéaires) MULTIVARIEES (avec sélection MANUELLE des prédicteurs),
- les scores de performance obtenus par les différentes méthodes prédictives (linéaires et non linéaires) MULTIVARIEES (avec sélection AUTOMATIQUE des prédicteurs),
- les scores de performance obtenus par les différentes méthodes prédictives (linéaires et non linéaires) MULTIVARIEES (avec TOUS les prédicteurs).

Le dernier graphe permet de comparer visuellement les performances obtenues par les « meilleures » méthodes prédictives dans chacun des cas cités précédemment (**rouge : %VP vignettes**, **vert : %VP cells/colonie**, **bleu : Estimation totale**).

Globalement, à la vue de ces différentes figures, il apparaît que la méthode MDA donne la plupart du temps les meilleurs scores de reconnaissance en utilisant la totalité des variables fournies par le FlowCAM et par la version 5 de Zoo/PhytoImage.

De plus, cette méthode non linéaire permet d'améliorer les performances de reconnaissance obtenues avec la méthode LDA. En effet, pour l'espèce *Asterionella glacialis*, l'amélioration est d'environ 27% en %VP pour le nombre de cellules par colonie.

Conclusion

La prise en compte de la spécificité des taxons en colonie pour leur dénombrement par Zoo/PhytoImage représente une évolution prioritaire du logiciel Zoo/PhytoImage, dans la mesure où cela concerne de très nombreuses espèces.

Les outils présentés dans ce rapport et en particulier ceux basés sur la régression non linéaire permettent d'obtenir des résultats satisfaisants sur les espèces testées. Néanmoins, il serait intéressant que les observateurs travaillant sur les FlowCAMs réalisent, en parallèle, un comptage des particules pour chaque colonie pour améliorer la partie quantitative de l'outil de reconnaissance. De plus, dans ce rapport, les scores de performance présentés ont été calculés par validation croisée. Dans un premier temps, il serait donc pertinent de comparer les résultats obtenus par les modèles prédictifs retenus avec les observations et comptages réalisés par lectures au microscope optique.

Dans un second temps, ce module devrait, à terme, permettre d'obtenir des mesures de biovolume, de biomasse et d'équivalent carbone pour chacun des groupes taxonomique identifiés dans un échantillon d'eau de mer. En effet, dans la littérature, la majorité des formules mathématiques de conversion permettant d'obtenir ces critères écologiques, sont basées principalement sur des mesures de taille d'une cellule pour chacune des espèces recensées.

Bibliographie

[1] Alcaraz M., Saiz E. et al. (2003). **Estimating zooplankton biomass through image analysis**, MarineBiology, 143:307–315.

[2] Benfield M.C., Grosjean P., Culverhouse P.F., Irigoien X., Sieracki M.E., Lopez-Urrutia A., Dam H.G., Hu Q., Davis C.S., Hansen A., Pilskaln C.H., Riseman E.M., Schultz H., Utgoff P.E. and G. Gorsky, (2007). **RAPID Research on Automated Plankton Identification**. Oceanography 20(2): 172- 187.

[3] Govaerts P., (2010). **Comptage automatique du nombre de cellules par colonies de phytoplancton de la Mer du Nord à l'aide du FlowCAM et de PhytoImage**. Rapport de projet de fin d'année, Université de Mons, 71 pp.

[4] Grosjean Ph., Picheral M., Warembourg C. and Gorsky G., (2004). **Enumeration, measurement, and identification of net zooplankton samples using the Zooscan digital imaging system**. ICES J. Mar. Sci., 61:518-525.

[5] Lund J.W.G, Kipling. C, Le Cren.E.D. (1958). **The inverted microscope method of estimating algal numbers and the statistical basis of estimations by counting**. Hydrobiol.. 11, 143-170.

[6] Sieracki C.K., Sieracki M.E. and Yentsch C.S., (1998). **An imaging in-flow system for automated analysis of marine microplankton**. Mar. Ecol. Prog. Ser. 168:285–96.

[7] Sieracki M. E., Benfield M. et al (2009). **Optical plankton imaging and analysis systems for ocean observation.** OceanObs'09 Symposium White Paper.

[8] Tunin-Ley A., Maurer D., (2011). **Mise en oeuvre opérationnelle d'un système couplé de numérisation (FlowCAM) et de traitement d'images (ZooPhytoImage), pour l'analyse automatisée, ou semi-automatisée, de la composition phytoplanctonique d'échantillons d'eau de mer.** *Rapport RST/LER/AR/11/002*.

Écologie Numérique des Milieux Aquatiques
UMONS
Faculté des Sciences

Apprentissage actif, Training set adaptatif

Guillaume WACQUET & Philippe GROSJEAN

Juin – Août 2014

Table des matières

Introduction

Contexte

L'outil FlowCAM/ZooPhytoImage est constitué du dispositif FlowCAM destiné à numériser les images de particules planctoniques, et du logiciel ZooPhytoImage qui permet d'identifier automatiquement et de dénombrer le plancton à partir de ces images. Ce dernier est un logiciel d'analyse d'images et de classification automatique, basé sur le principe du "*machine learning*". Il permet de réaliser les différentes étapes du processus qui conduit à la classification automatisée, ou semi-automatisée, d'un ensemble d'objets, à partir d'un jeu d'images donné, en utilisant des algorithmes d'apprentissage supervisé.

L'étape préliminaire indispensable est donc celle de l'apprentissage à la reconnaissance des particules étudiées. Cet apprentissage (réalisé à travers une interface) s'effectue en deux phases [4] :

- La première phase concerne la réalisation d'un set d'apprentissage constitué à partir d'une banque d'images issues de l'appareil de numérisation (ici, le FlowCAM) et représentatives des particules rencontrées dans les échantillons à analyser ultérieurement. Ces images sont identifiées VISUELLEMENT par l'observateur puis classées MANUELLEMENT dans autant de dossiers ou sous-dossiers que nécessaires pour représenter les niveaux d'identification souhaités.
- La seconde phase concerne la réalisation d'un outil de reconnaissance automatique, ou semi-automatique, du plancton en utilisant le set d'apprentissage précédemment créé pour « entraîner » un algorithme de classification supervisée à reconnaître la nature des particules sur la base des mesures obtenues sur les images numériques. Cet outil, nommé *classifier* en anglais, peut être assimilé à une boîte noire capable de déterminer le groupe d'une image uniquement sur la base des paramètres mesurés sur celle-ci.

Une fois l'outil de reconnaissance automatique optimisé [7], celui-ci peut alors être utilisé sur un jeu d'images nouvelles pour classer automatiquement les particules dans les catégories qui auront été définies manuellement dans le set d'apprentissage initial. Cependant, les abondances obtenues ne peuvent pas toujours être fiables en routine. Il faut donc valider les données, grâce aux options nouvellement ajoutées au logiciel (étape de vérification manuelle des classements effectués par l'ordinateur) [5]. Cette opération est accélérée par le fait que l'ordinateur a effectué un classement correct à 65-75%, par rapport à une classification purement manuelle des vignettes. Le module de correction statistique de l'erreur intégré à ZooPhytoImage permet alors d'obtenir une estimation moins biaisée de l'abondance sans avoir à valider manuellement la totalité des vignettes [6][8].

Problématique

Pour l'élaboration et l'optimisation manuelle d'un set d'apprentissage, plusieurs points doivent être pris en compte [9] :

- Le nombre d'images par groupe taxinomique doit être compris entre 100 et 200. Si certains groupes (de forme très constante) sont très bien reconnus avec un faible nombre d'images, parfois une douzaine seulement, dans tous les cas, on observe une très nette amélioration des taux de reconnaissance lorsque le nombre d'images par catégorie est supérieur à 100. Cependant, ce seuil peut être plus élevé lorsque la variabilité morphologique des particules est importante, comme c'est le cas pour une partie des diatomées, notamment pour les espèces coloniales. Les images du set d'apprentissage doivent ainsi représenter de manière significative la variabilité naturelle d'un taxon donné, que celle-ci soit intrinsèque (taille des

cellules, état physiologique, forme des extensions, longueur des colonies), ou extrinsèque (position de la cellule par rapport à l'objectif au moment de l'acquisition d'image, netteté de l'image).

- Idéalement, il faudrait acquérir un nombre équivalent d'images entre les groupes. En effet, les groupes dotés du plus grand nombre de vignettes sont aussi les mieux reconnus. Il semblerait que le nombre d'images donne proportionnellement plus de poids à un groupe au niveau de la classification des taxons par l'algorithme.
- Il est nécessaire de constituer un set d'apprentissage le plus détaillé possible au départ, même si des regroupements sont effectués ultérieurement. Pour les groupes non ciblés, il est préférable de créer des catégories assez larges, n'ayant pas forcément de cohérence d'un point de vue taxinomique (cela s'applique particulièrement pour la création d'un outil de reconnaissance spécialisé).
- Les débris sont souvent une source de confusion importante avec les cellules planctoniques de petite taille. Il est donc primordial de soigneusement définir ces catégories. Une catégorisation des débris selon leur teinte (niveau de gris), même si elle est forcément subjective et imprécise pour l'observateur humain, semble être la plus pertinente. En revanche, une catégorisation selon la forme ou la texture donne rarement de bons résultats. Il peut également s'avérer nécessaire de regrouper avec les débris d'autres particules, comme certains groupes du zooplancton.

Actuellement, dans ce contexte, l'étape de création manuelle d'un set d'apprentissage s'avère être une tâche fastidieuse et coûteuse en temps (due aux nombre d'items nécessaires dans chacun des groupes taxonomiques souhaités), mais également subjective (car subordonnée aux connaissances et au choix de l'utilisateur). De plus, la variabilité spatio-temporelle des populations planctoniques étant importante, les performances de classification de l'outil généré peuvent être variables [7]. Une solution possible pour remédier à ce problème tient dans le set d'apprentissage adaptatif localement. En d'autres termes, nous partons d'un set d'apprentissage « global » à l'échelle nationale, et nous lui rajoutons les vignettes validées localement, géographiquement, et temporellement (échantillons des semaines précédentes et/ou échantillons des années précédentes à la même période de l'année), avec élimination progressive des vignettes du set d'apprentissage global au fur et à mesure que des données locales viennent compléter le set.

Dans ce cadre, il est donc envisageable de constituer un set d'apprentissage « initial » constitué d'images disparates (provenant de la numérisation par différents appareils, de différentes zones géographiques, à différentes saisons, etc.), et de se tourner vers l'apprentissage actif. De cette façon, il serait possible de compléter (semi-)automatiquement le set « initial », de l'adapter à la zone étudiée, mais également de supprimer l'effet de « machine-dépendance ».

Objectifs du travail

L'objectif de ce travail est d'étudier l'impact de l'ajout et de la suppression automatiques de nouveaux items dans un set d'apprentissage « initial ». Pour cela, plusieurs expérimentations doivent être menées afin de garantir une construction automatique et adaptative du set d'apprentissage tout en conservant ou en améliorant les performances de classification de l'outil de reconnaissance associé.

L'objet de ce document est donc de définir et d'étudier une **liste de règles de décision** concernant l'ajout et la suppression des items afin de disposer d'un outil permettant de gérer complètement le module d'apprentissage actif.

Démarche méthodologique

Set d'apprentissage

Le set d'apprentissage « initial » est composé de 40 groupes (dont 8 correspondent à des particules détritiques et 32 à des particules planctoniques). Ce set a été conçu afin d'avoir **30** items par groupe (à l'exception de 5 catégories sous-représentées), comme illustré sur la FIGURE 1.

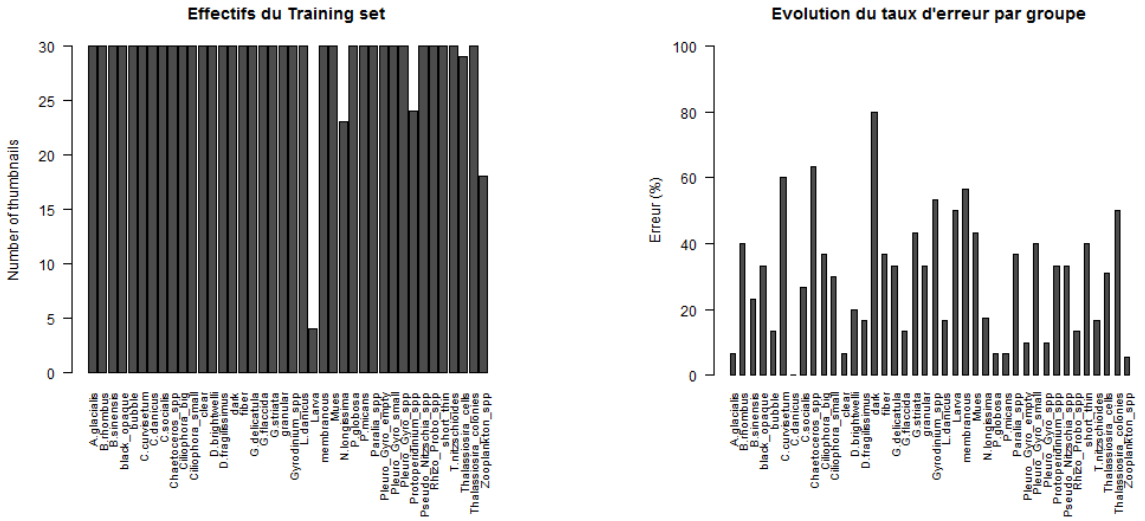


FIGURE 1 - Effectifs du set d'apprentissage « initial » et taux d'erreur par groupe (Random Forest).

L'objectif est de compléter ce set d'apprentissage de manière automatisée au fur et à mesure que des nouveaux échantillons sont analysés puis validés, tout en garantissant des performances de reconnaissance supérieures ou égales à celles obtenues initialement. Pour cela, quelques règles ont été mises en place : le nombre maximal d'items par groupe doit être de **300** ; si le seuil n'est pas atteint, on ajoute des nouveaux items ; si le seuil est atteint, on supprime un pourcentage d'items du set d'apprentissage (dans nos expérimentations, **5%**) puis on ajoute la même quantité de nouveaux items.

Stratégies

Pour l'ajout et la suppression des items, trois types de stratégie ont été étudiés :

- **SV** : Ajout des vignettes suspectes validées (jusqu'à l'itération 5 du processus de correction d'erreur)

SV : **S**uspects **V**alidés

- **%SV+NSV** : Ajout d'un pourcentage de vignettes suspectes validées, puis complétion avec les vignettes non suspectes validées (jusqu'à l'itération 5 du processus de correction d'erreur)

SV : **S**uspects **V**alidés **NSV** : **N**on **S**uspects **V**alidés

- **%SV+NSV+NSNV** : Ajout d'un pourcentage de vignettes suspectes validées, puis complétion avec les vignettes non suspectes validées (jusqu'à l'itération 5 du processus de correction d'erreur), et les vignettes non suspectes non validées (AVEC ou SANS post-validation par l'utilisateur)

SV : **S**uspects **V**alidés **NSV** : **N**on **S**uspects **V**alidés **NSNV** : **N**on **S**uspects **N**on **V**alidés

Méthodes de sélection

Deux méthodes de sélection des vignettes à ajouter/supprimer du set d'apprentissage ont été évaluées afin de mettre en évidence des performances de classification supérieures ou égales à celles obtenues avec le set d'apprentissage initial :

- **RANDOM** : ajout et suppression aléatoires d'items dans le training set,
- **OUTLIER** : ajout des items les « moins aberrants » et suppression des items les « plus aberrants » (basés sur un facteur local d'aberration pour chaque item [3]).

Dans la partie suivante, une étude comparative des performances de classification selon les stratégies utilisées, est menée. Le set d'apprentissage « initial » est alors complété grâce à de nouveaux items validées à partir de l'analyse de 11 nouveaux échantillons (deux points de prélèvements en Manche Orientale entre Janvier et Juin 2014).

Résultats expérimentaux

Échantillonnage aléatoire (mode = « random »)

Dans un premier temps, nous choisissons d'ajouter/supprimer aléatoirement des items aux différents groupes du set d'apprentissage. Les FIGURES 2 et 3 représentent les scores de performance des classifieurs au fur et à mesure de la complétion du set d'apprentissage en termes de taux d'erreur globale par validation croisée, de taux d'erreur pour un nouvel échantillon, mais également de nombre d'items suspects à valider, pour chacune des stratégies envisagées.

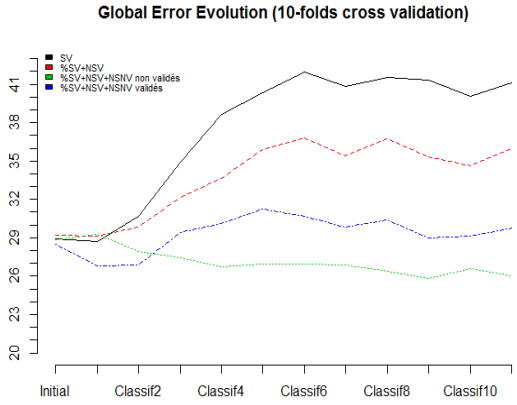


FIGURE 2 – Mode Random - Évolution du taux d'erreur globale par validation croisée (10 folds) à chaque itération (**Initial** = classifieur initial; **Classif1** = *Initial+Ech1;* **Classif2** = Classif1+Ech2; etc.).

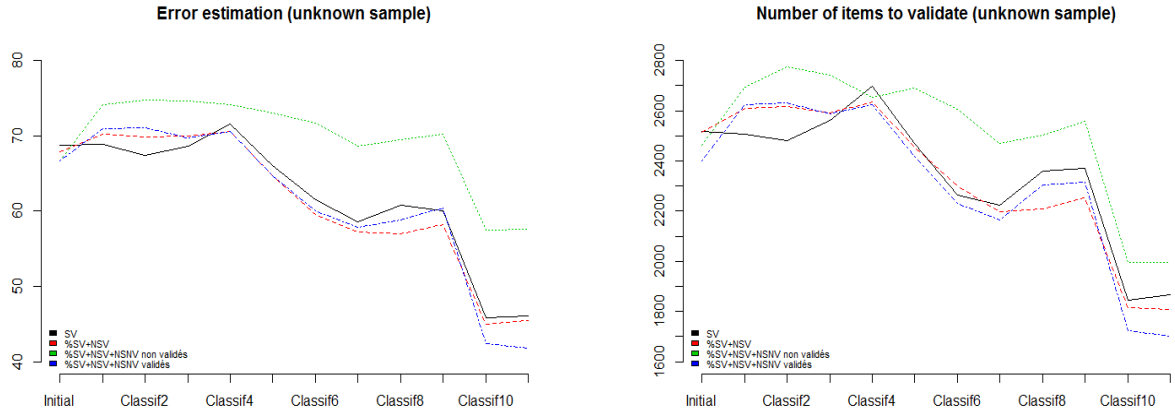


FIGURE 3 – Mode Random - Évolution du taux d'erreur globale sur un nouvel échantillon et évolution du nombre d'items suspects à valider, en fonction du classifieur (**Initial** = classifieur initial; **Classif1** = *Initial+Ech1;* **Classif2** = Classif1+Ech2; etc.).

Comme nous pouvons l'observer :

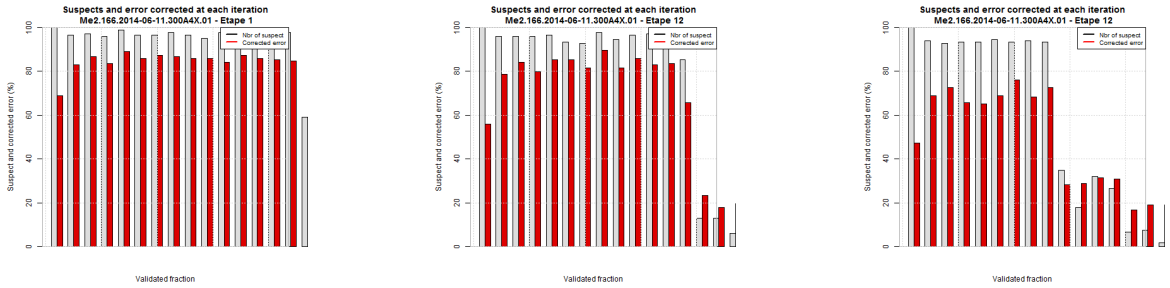
- sur la FIGURE 2, globalement, l'ajout aléatoire d'items non suspects non validés (**%SV+NSV+NSNV** post-validés par l'utilisateur ou non, en bleu et vert respectivement sur le graphe) permet d'obtenir des performances de reconnaissance par validation croisée proches de celles obtenues grâce au classifieur généré à partir du set d'apprentissage « initial » (correspondant à l'abscisse « Initial »). De plus, nous pouvons constater que, dans ce cas, l'unique ajout d'items suspects validés (méthode **SV** en noir) et d'items non suspects validés (méthode **%SV+NSV** en rouge) dégradent fortement les scores de reconnaissance (augmentation considérable et rapide du taux d'erreur). En effet, l'ajout d'items suspects dans

les groupes du set d'apprentissage renforce les erreurs de classification et les confusions de par leur proximité avec d'autres particules. L'ajout d'items non suspects non validés paraît donc pertinent.

- sur la FIGURE 3, pour un nouvel échantillon, les meilleurs résultats sont obtenus pour les méthodes **SV** (en noir), **%SV+NSV** (en rouge) et **%SV+NSV+NSNV** AVEC post-validation des items (en bleu). Les erreurs de classification estimées à chaque itération pour ces méthodes sont proches et évoluent de manière similaire. Contrairement à cela, la méthode **%SV+NSV+NSNV** SANS post-validation des items (en vert), montre des performances inférieures, notamment avec une augmentation rapide du nombre d'items à valider. Dans ce cas, la reconnaissance n'est ici améliorée qu'à partir de la 10ème itération (contrairement aux autres méthodes qui montrent une amélioration nette de l'erreur à partir de la 5ème itération). La chute brutale de l'erreur observée à cette itération peut alors s'expliquer principalement par la similitude entre les items de l'échantillon 10 ajoutés au set d'apprentissage et ceux de l'échantillon étudié (bloom de *Pseudo-Nitzschia* dans les 2 cas).

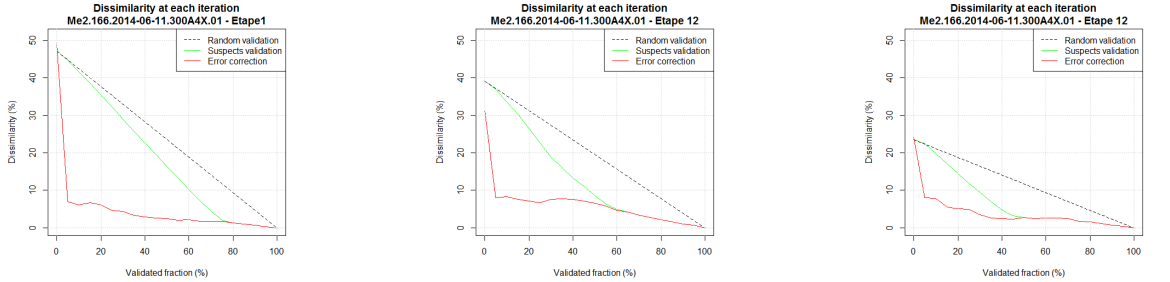
Discussion

Même si la méthode **%SV+NSV+NSNV** SANS post-validation permet une amélioration globale de la reconnaissance par validation croisée, elle ne contribue pas à obtenir des scores de performance de classification aussi élevés que la méthode **%SV+NSV+NSNV** AVEC post-validation, sur un échantillon inconnu, tout en garantissant une diminution du nombre d'items à valider (comme illustré sur les FIGURES 4 et 5). Ceci peut s'expliquer par le fait que les items non suspects, non validés et sans post-validation, ajoutés aux différents groupes du set d'apprentissage, sont considérés par le classifieur, comme devant appartenir réellement aux groupes ciblés. Ces ajouts ne font, en réalité, que renforcer les erreurs commises par le classifieur.



Avec classifieur initial | ***Avec classifieur à l'étape 11 (méthode %SV+NSV+NSNV sans post-validation)*** | ***Avec classifieur à l'étape 11 (méthode %SV+NSV+NSNV avec post-validation)***

FIGURE 4 – Mode Random - Proportion d'items suspects et fraction d'items incorrectement classés à chaque étape du processus de correction d'erreur.

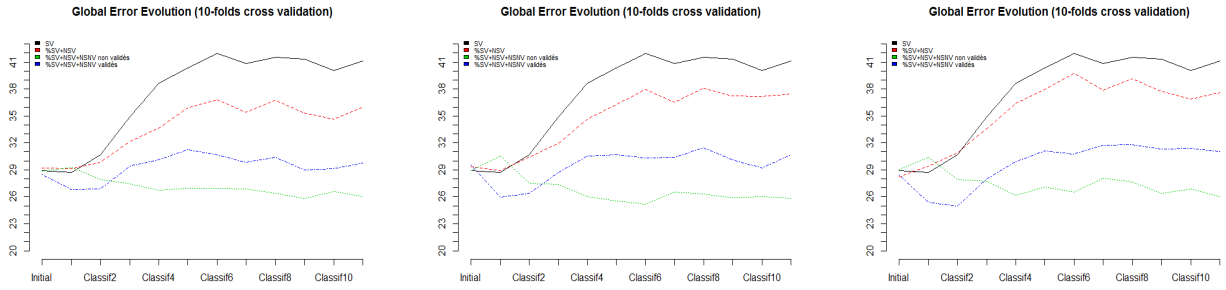


Avec classifieur initial | ***Avec classifieur à l'étape 11 (méthode %SV+NSV+NSNV sans post-validation)*** | ***Avec classifieur à l'étape 11 (méthode %SV+NSV+NSNV avec post-validation)***

FIGURE 5 – *Mode Random* – Correction de l'erreur en fonction de la fraction validée.

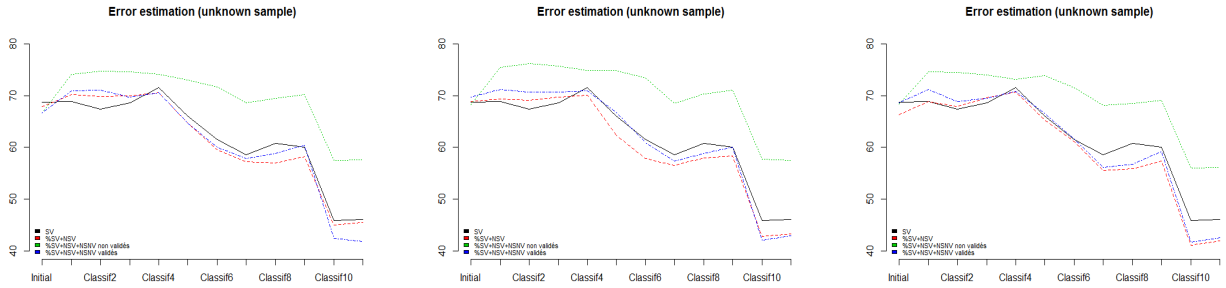
De plus, il paraît inconcevable d'utiliser cette stratégie en pratique (en particulier dans le cadre d'un réseau national de surveillance du plancton, à but sanitaire) de par l'incertitude des items aléatoires non validés ajoutés au différents groupes taxonomiques du set d'apprentissage. Il paraît donc judicieux d'utiliser la méthode **%SV+NSV+NSNV** AVEC post-validation des items par l'utilisateur.

Dans nos expérimentations, la proportion d'items à remplacer dans les groupes du set d'apprentissage ayant atteint le seuil maximal d'items (ici, **300**) a été fixée à **5%** (soit 15 vignettes à remplacer). Les FIGURES 6 et 7 montrent les résultats obtenus par validation croisée et sur un nouvel échantillon respectivement, selon la proportion d'items à remplacer (5%, 10% et 20%).



5% ***10%*** ***20%***

FIGURE 6 – Mode Random - Évolution du taux d'erreur globale par validation croisée (10 folds) à chaque itération, selon la proportion d'items à remplacer (**Initial** = classifieur initial; **Classif1** = *Initial+Ech1;* **Classif2** = Classif1+Ech2; etc.).



5% ***10%*** ***20%***

FIGURE 7 – Mode Random - Évolution du taux d'erreur globale sur un nouvel échantillon et évolution du nombre d'items suspects à valider, en fonction du classifieur, selon la proportion d'items à remplacer (**Initial** = classifieur initial; **Classif1** = *Initial+Ech1;* **Classif2** = Classif1+Ech2; etc.).

Comme le montrent ces figures, l'impact de la proportion d'items à remplacer est négligeable. Il apparaît donc que la quantité n'est pas aussi importante et déterminante que la qualité de ce que l'on remplace dans le set d'apprentissage. C'est pourquoi, au vu de ces résultats, et pour la suite des expérimentations, nous choisissons d'utiliser et d'évaluer la méthode **%SV+NSV+NSNV** AVEC post-validation (en bleu) en fixant la proportion d'items à ajouter/supprimer à **5%**.

Néanmoins, il est important de noter que ces expérimentations ont été réalisées sur un nombre retreint d'échantillons s'étalant sur une courte période de 6 mois. Ces résultats devront donc être validés sur un nombre d'échantillons plus important appartenant à une série temporelle relativement longue.

Échantillonnage basé sur la détection d'outliers (mode = « outlier »)

Les résultats obtenus précédemment montrent l'efficacité de la méthodologie utilisée. Grâce à celle-ci, il est possible de construire un set d'apprentissage de manière automatisée, tout en garantissant des performances de classification supérieures ou égales à celles obtenues initialement, et en réduisant considérablement le nombre d'items suspects à valider manuellement après l'étape de prédiction automatique.

Dans cette partie, nous testons une nouvelle approche de sélection des items à ajouter/supprimer basée sur la détection de données aberrantes. En effet, par rapport à l'approche de sélection aléatoire, une sélection intelligente des items à ajouter/supprimer devrait permettre d'améliorer les performances du système d'apprentissage actif. Pour cela, nous avons choisi d'utiliser une méthode de détection d'outliers basée sur le k-voisinage [2]. Cette approche repose sur l'idée proposée par Breunig et al. [1][3] et relative au facteur local d'aberration (ou Local Outlier Factor, noté LOF) de chaque item, qui est dépendant de la densité locale de son voisinage. Ce voisinage est défini par la distance entre l'item observé et son k-ième plus proche voisin (k étant une variable définissant un nombre minimal d'items dans le voisinage et fixé à **5** dans nos expérimentations).

Pourquoi la méthode LOF ?

Cette méthode permet d'analyser des groupes de densités différentes, en comparant la densité locale d'un item avec la densité moyenne de ses k-plus proches voisins. De plus, elle ne fait aucune hypothèse sur la distribution des données.

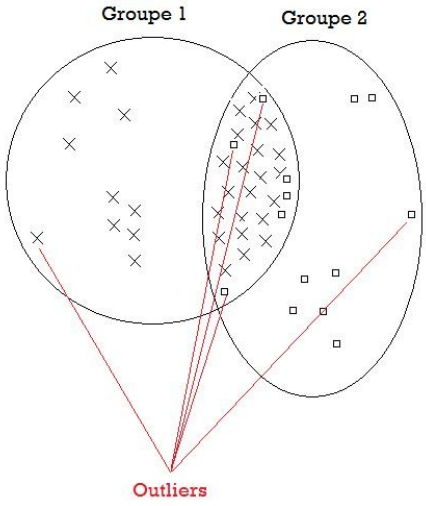


FIGURE 8 – Exemple illustratif du fonctionnement de la méthode LOF dans le cas de groupes (et sous-groupes) de densités différentes.

Dans notre cas, comme illustré sur la FIGURE 8, il est possible que certains groupes taxonomiques soient scindé en plusieurs sous-groupes ayant des densités différentes. Or, les modèles de détection d'outliers basés sur des calculs de distance ne permettent pas de gérer ce type de données. Afin de comparer le voisinage des items appartenant à des régions de densités différentes, il est donc nécessaire de considérer la densité relative.

Propriétés de la méthode LOF.

Un des avantages que nous allons exploiter ici, réside dans l'affectation d'une valeur LOF à chacun des items :

- si LOF ≈ 1, alors l'item est dans le cluster (autrement dit, la région de densité homogène autour de l'item et de ses voisins) et est considéré comme « inlier »,
- si LOF >> 1, alors l'item est considéré comme « outlier ».

Résultats

Les FIGURES 9 et 10 représentent les scores de performance des classifieurs au fur et à mesure de la complétion du set d'apprentissage en termes de taux d'erreur globale par validation croisée, de taux d'erreur pour un nouvel échantillon, mais également de nombre d'items suspects à valider, pour chacune des stratégies envisagées.

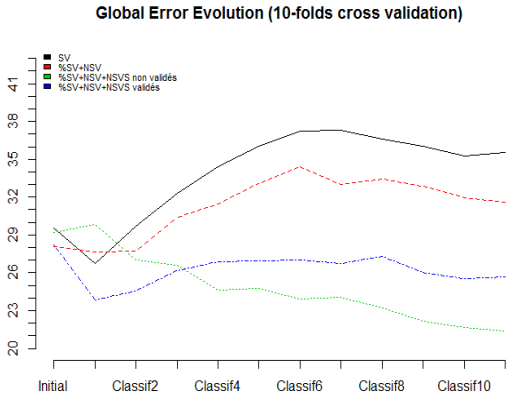


FIGURE 9 – Mode Outlier (LOF avec k=5) - Évolution du taux d'erreur globale par validation croisée (10 folds) à chaque itération (**Initial** = classifieur initial; **Classif1** = *Initial+Ech1;* **Classif2** = Classif1+Ech2; etc.).

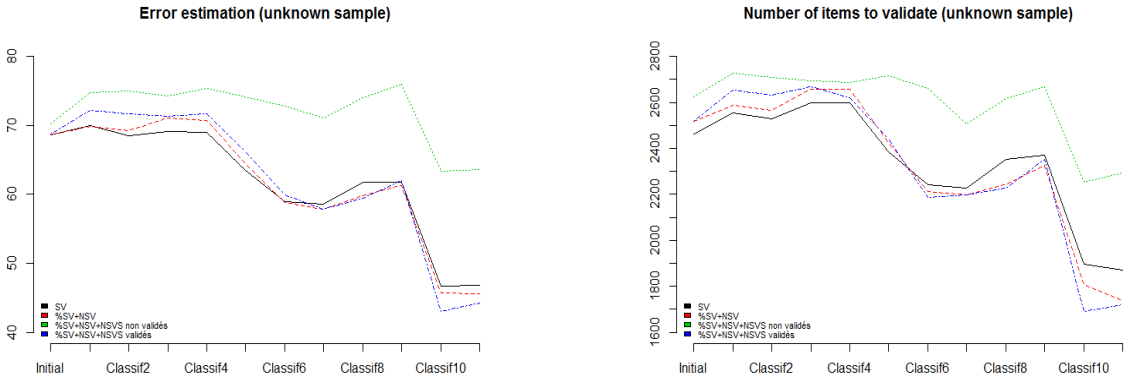


FIGURE 10 – Mode Outlier (LOF avec k=5) - Évolution du taux d'erreur globale sur un nouvel échantillon et évolution du nombre d'items suspects à valider, en fonction du classifieur (**Initial** = classifieur initial; **Classif1** = *Initial+Ech1;* **Classif2** = Classif1+Ech2; etc.).

Comme nous pouvons l'observer :

- sur la FIGURE 9, les mêmes tendances que pour le mode « RANDOM » se dégagent. En effet, l'ajout d'items non suspects non validés **%SV+NSV+NSNV** post-validés (en bleu), permet également d'obtenir des performances de reconnaissance par validation croisée proches de celles obtenues grâce au classifieur généré à partir du set d'apprentissage « initial ». De plus, comme constaté auparavant, l'unique ajout d'items suspects validés (méthode **SV** en noir) et d'items non suspects validés (méthode **%SV+NSV** en rouge) dégradent fortement les scores de reconnaissance.
- sur la FIGURE 10, pour un nouvel échantillon, les résultats obtenus pour les méthodes **SV** (en noir), **%SV+NSV** (en rouge) et **%SV+NSV+NSNV** AVEC post-validation des items (en bleu), sont proches et évoluent de manière similaire. Contrairement à cela, la méthode **%SV+NSV+NSNV** SANS post-validation des items (en vert), montre de nouveau des performances inférieures. Les mêmes conclusions que pour le mode « RANDOM » peuvent donc être formulées.

Comparaison Random VS Outlier

Les FIGURES 11 et 12, 13 et 14 présentent les résultats obtenus par validation croisée et sur un nouvel échantillon, en terme d'erreur globale, d'erreur estimée, de proportion d'items suspects à valider et d correction d'erreur, pour les deux types de sélection des items à ajouter/supprimer (RANDOM vs OUTLIER).

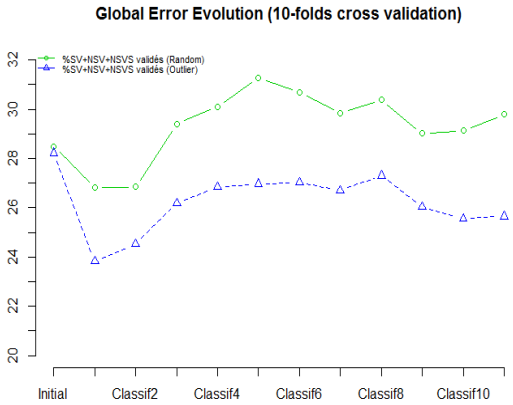


FIGURE 11 – Comparaison des modes Random et Outlier - Évolution du taux d'erreur globale par validation croisée (10 folds), à chaque itération (**Initial** = classifieur initial; **Classif1** = Initial+Ech1; **Classif2** = Classif1+Ech2; etc.).

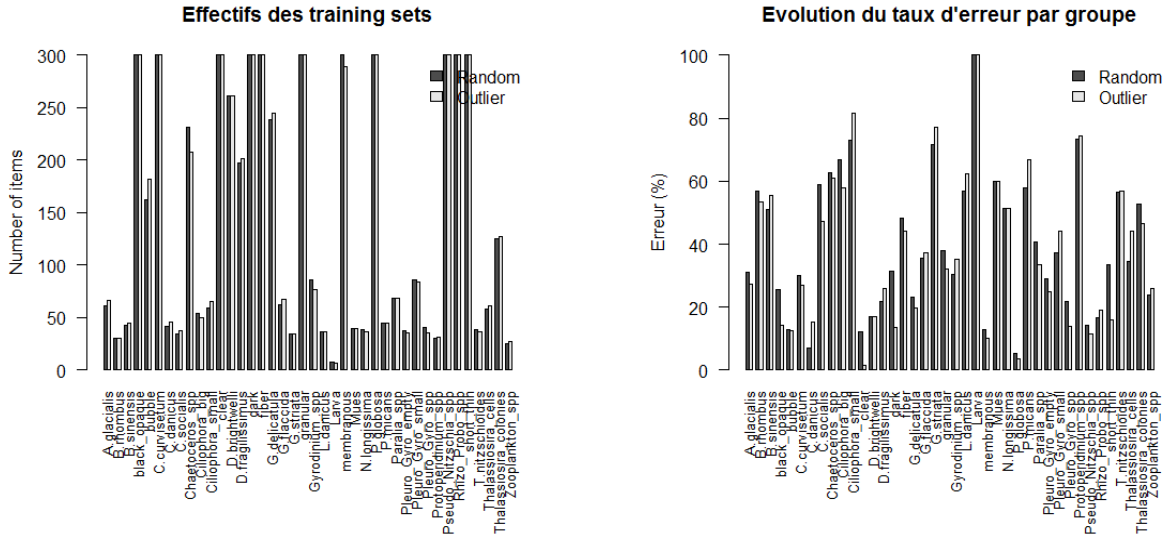


FIGURE 12 – Comparaison des modes Random et Outlier - Effectifs des sets d'apprentissage à l'étape 11 et taux d'erreur par groupe associés (algorithme Random Forest).

La comparaison des erreurs globales de classification par validation croisée (FIGURE 11) donne un net avantage au mode de sélection OUTLIER (en bleu) par rapport au mode RANDOM (en vert). En effet, il est possible d'observer une différence de 3 à 4% d'erreur entre ces deux méthodes, et ce à partir de l'étape 1 jusqu'à l'étape 11. Grâce à la FIGURE 12, il est possible d'affirmer que cette différence est à mettre en relation, en majeure partie, avec les groupes contenant des particules détritiques. En effet, on peut observer que les écarts d'erreur les plus importants (pour un même nombre d'items dans les groupes, en l'occurrence 300) concernent les groupes « black_opaque », « clear », « dark » et « short_thin » qui sont des groupes rassemblant des particules inorganiques.

Les FIGURES 13 et 14 présentent les résultats obtenus par validation croisée et sur un nouvel échantillon, en terme d'erreur globale, d'erreur estimée, de proportion d'items suspects à valider et d correction d'erreur, pour les deux types de sélection des items à ajouter/supprimer (RANDOM vs OUTLIER).

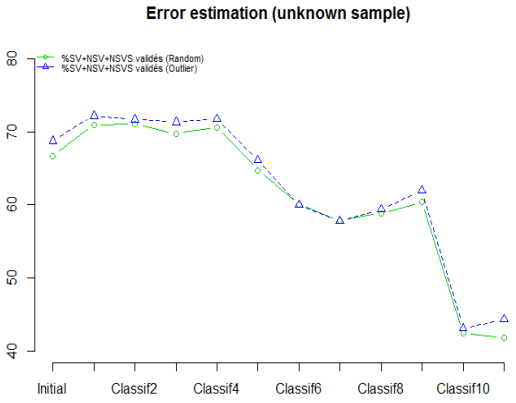
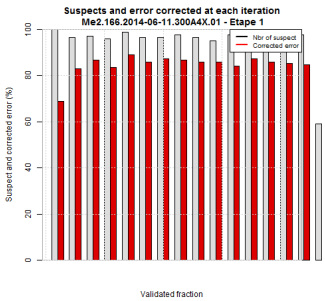
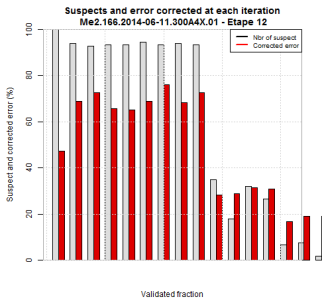


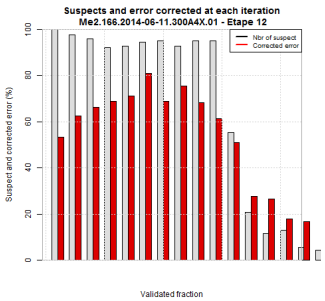
FIGURE 13 – Comparaison des modes Random et Outlier - Évolution du taux d'erreur globale sur un nouvel échantillon, à chaque itération (**Initial** = classifieur initial; **Classif1** = Initial+Ech1; **Classif2** = Classif1+Ech2; etc.).



Avec classifieur initial

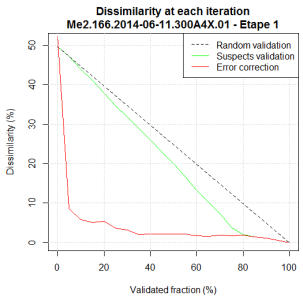


Avec classifieur à l'étape 11 Méthode RANDOM

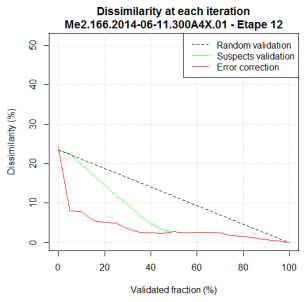


Avec classifieur à l'étape 11 Méthode OUTLIER

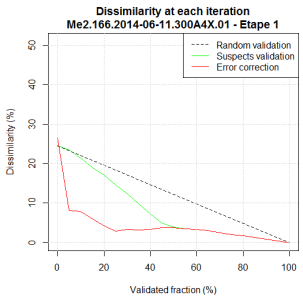
FIGURE 14 – Comparaison des modes Random et Outlier - Proportion d'items suspects et fraction d'items incorrectement classés à chaque étape du processus de correction d'erreur.



Avec classifieur initial



Avec classifieur à l'étape 11 Méthode RANDOM



Avec classifieur à l'étape 11 Méthode OUTLIER

FIGURE 15 – Comparaison des modes Random et Outlier - Correction de l'erreur en fonction de la fraction validée.

Les résultats présentés sur ces 3 figures tendent à montrer une similitude entre les performances obtenues par les méthodes RANDOM et OUTLIER. En effet, les erreurs globales de reconnaissance au fur et à mesure des étapes sont quasi-identiques, et la correction d'erreur permet d'obtenir des scores similaires avec des nombres d'items suspects à valider proches.

Discussion

Les résultats présentés sur la FIGURE 11 montrent une nette amélioration de l'erreur de classification en validation croisée pour le mode OUTLIER (et par rapport au mode RANDOM). Cependant, ceci peut être expliqué par le fait que l'on supprime des groupes du set d'apprentissage, des items que l'on peut considérer comme aberrants pour y ajouter des nouveaux items qui le sont moins. La variabilité intra-groupe peut donc se retrouver réduite (notamment dans les groupes contenant des particules détritiques, qui sont également ceux présentant le plus de variabilité) et ainsi limiter les confusions entre groupes taxonomiques.

Toutefois, cette baisse de l'erreur ne se reflète pas sur les scores de reconnaissance pour un nouvel échantillon (FIGURES 13, 14 et 15). En effet, ici, les performances obtenues pour les deux types de sélection des items, apparaissent comme similaires. Sachant que les confusions les plus importantes se retrouvent généralement avec les particules détritiques, il est donc envisageable de formuler l'hypothèse que la baisse de la variabilité dans les groupes contenant ce type de particules dans le set d'apprentissage ne permet pas une meilleure discrimination pour un nouvel échantillon (à démontrer et à confirmer...!?!).

Même si l'utilisation du mode de sélection des items à ajouter/supprimer basée sur la détection d'outliers, peut sembler intéressante intuitivement, son apport réel n'est pas démontré dans le cadre de nos expérimentations. Néanmoins, elle peut s'avérer utile dans le cadre d'une recherche de degré d'appartenance d'un item à un groupe de par l'opportunité d'avoir une valeur d'aberration pour chaque observation. De cette manière, il est possible de ranger les items selon leur valeur LOF afin de mettre en évidence les « inliers » et « outliers » de chacun des groupes ou sous-groupes.

Fonctions R

Fonction « activeLearning »

Cette fonction est la fonction principale pour l'apprentissage actif.

Fonction 1 : activeLearning

Inputs : *smpdir* (répertoire contenant les échantillons)
ZItrain (set d'apprentissage initial)
mode (mode de sélection des items - « *random* » ou « *outlier* »)
Outputs : *ZItrain* (set d'apprentissage final)
for i in 1:nombre d'échantillons dans *smpdir*
| *smp* ← échantillon
| *classifier* ← générer un classifieur à l'aide de *ZItrain*
| *ec* ← créer un objet *errorCorrection* en mode « *stat* » à l'aide de *smp* et *classifier*
| *env* ← récupérer environnement de *ec*
| *ZItrain* ← mettre à jour le set d'apprentissage (fonction « *addItems* »)
à l'aide de *smp*, *ZItrain, env* et *mode*
return *ZItrain*

Fonction « addItems »

En réalité, deux fonctions : addItems et dropItems. La première (addItems) prend en argument la seconde (dropItems). Ces deux fonctions personnalisables permettent de définir les règles d'ajout et d'élimination des items.

Fonction 2 : addItems

Inputs : *smp* (échantillon)
ZItrain (set d'apprentissage initial)
env (environnement)
mode (mode de sélection des items - « *random* » ou « *outlier* »)
dropItems() (fonction de suppression des items dans *ZItrain*)
Outputs : *ZItrain* (set d'apprentissage final)
nbItems ← seuil maximal d'items dans chaque groupe de *ZItrain*
for i in groupes prédits dans *env/smp*
| **if** *mode* == « *random* »
| | **if** nombre d'items dans *ZItrain* < *nbItems*
| | | *ZItrain* ← ajout aléatoire d'items dans *ZItrain*
| | **if** nombre d'items dans *ZItrain* == *nbItems*
| | | *nbDrop* ← nombre d'items à supprimer dans *ZItrain*
| | | *dropItems()* ← suppression aléatoire de *nbDrop* items dans *ZItrain*
| | | *ZItrain* ← ajout aléatoire de *nbDrop* items dans *ZItrain*
| **if** *mode* == « *outlier* »
| | **if** nombre d'items dans *ZItrain* < *nbItems*
| | | *ZItrain* ← ajout des items les moins aberrants dans *ZItrain*
| | **if** nombre d'items dans *ZItrain* == *nbItems*
| | | *nbDrop* ← nombre d'items à supprimer dans *ZItrain*
| | | *dropItems()* ← suppression des *nbDrop* items les plus aberrants dans *ZItrain*
| | | *ZItrain* ← ajout des *nbDrop* items les moins aberrants dans *ZItrain*
return *ZItrain*

Fonction 3 : dropItems

Inputs : *ZItrain* (set d'apprentissage initial)
mode (mode de sélection des items - « *random* » ou « *outlier* »)
nbDrop (nombre d'items à ajouter/supprimer dans *ZItrain*)
Outputs : *ZItrain* (set d'apprentissage final)
if *mode* == « *random* »
| *ZItrain* ← suppression aléatoire de *nbDrop* items dans *ZItrain*
if *mode* == « *outlier* »
| *ZItrain* ← suppression des *nbDrop* items les plus aberrants dans *ZItrain*
return *ZItrain*

Fonction « plotGraph »

Cette fonction permet d'afficher les différents graphes de performance.

Fonction 2 : plotGraph

Inputs : *dat* (données)
type (type d'affichage - «*items*», «*errors*», «*global*», «*dissimilarity*», «*partition*», ...)
name (titre du graphe)
if *type* == « *items* »
| affichage de l'effectif du set d'apprentissage
| title(*name*)
if *type* == « *errors* »
| affichage de l'évolution du taux d'erreur par groupe par validation croisée
| title(*name*)
if *type* == « *global* »
| affichage de l'évolution du taux d'erreur global par validation croisée
| title(*name*)
if *type* == « *dissimilarity* »
| affichage du graphe de dissimilarité (correction d'erreur)
| title(*name*)
if *type* == « *partition* »
| affichage du graphe de partition (correction d'erreur)
| title(*name*)
if *type* == « *barplot* »
| affichage du graphe d'erreur (correction d'erreur)
| title(*name*)

Utilisation (fichier « testAL.R »)

```
require(zooimage)          # Chargement de ZooImage
library(DMwR)              # Bibliothèque nécessaire pour la méthode LOF
setwd("C:/Users/Desktop/UMONS – EcoNum/activeLearning/")      # Répertoire de travail

# Fichiers nécessaires à l'apprentissage actif
source("sampleInfoNew.R")      # Fonction de tri des échantillons par date
source("activeLearning.R")     # Fonction principale d'apprentissage actif
source("addItems.R")           # Fonction de remaniement du training set
source("plotGraph.R")          # Fonction d'affichage des performances et résultats

load("./Training/_train_BL_4X_Rephy_degrade.RData")      # Training set initial
activeLearning(smpdir = "./Data/", ZItrain, mode = "random") # Fonction d'apprentissage actif
```

Conclusions et Perspectives

Conclusions

Ce travail se situe dans le cadre de l'apprentissage actif. Celui-ci reçoit de plus en plus d'intérêt dans la communauté scientifique vu qu'il apporte des solutions intéressantes à de nombreux problèmes réels de classification automatique pour lesquels il est nécessaire de construire et d'optimiser un ensemble d'apprentissage. En effet, ce dernier représente une étape cruciale pour l'obtention de résultats de reconnaissance pertinents. Cependant, cette tâche peut s'avérer fastidieuse, coûteuse en temps et subjective.

Dans ce contexte, nous avons mis l'accent sur la complétion (semi-)automatique d'un set d'apprentissage basique (basée sur diverses stratégies), en émettant l'hypothèse que l'intégration de nouvelles données dans les différents groupes du set devrait permettre d'augmenter considérablement ses performances de reconnaissance. Nous nous sommes plus particulièrement intéressés à la méthode de sélection des items à ajouter/supprimer des groupes.

Pour cela, nous avons distingué deux types de sélection : une méthode aléatoire (mode RANDOM) et une méthode basée sur la détection de données aberrantes (mode OUTLIER). Les performances des différentes stratégies et modes de sélection sont mises en évidence à l'aide de 11 échantillons pour la complétion du set d'apprentissage, et un douzième échantillon-test. Les critères de comparaison utilisés sont l'erreur globale, les taux d'erreur par groupe, ainsi que la portion d'items à valider. Une étude comparative a enfin été menée pour déterminer les avantages de chacune des stratégies utilisées.

Au vu des résultats, nous avons pu montrer que la méthode **%SV+NSV+NSNV** AVEC post-validation des items à ajouter au set d'apprentissage offrait les performances de reconnaissance les plus élevés tout en garantissant une diminution du nombre d'items à valider lors de la correction d'erreur (pour un nouvel échantillon) et une stabilité pour les scores obtenus par le classifieur en validation croisée. Néanmoins, même si la méthode de sélection des items basée sur le facteur local d'aberration semble être la plus pertinente pour améliorer les performances en validation croisée, l'étude comparative avec le mode aléatoire n'a pas permis de mettre en évidence un avantage net.

Toutefois, il est important de noter que ces expérimentations ont été réalisées sur un nombre retreint d'échantillons s'étalant sur une période courte allant de janvier à juin 2014. Ces résultats devront donc être validés sur un nombre d'échantillons plus important appartenant à une série temporelle relativement longue.

Perspectives

Aux niveaux fondamental et applicatif, les perspectives de ce travail se situent sur plusieurs points :

- Il est possible que des changements liés au protocole de numérisation (avec modification matériel comme par exemple, l'illumination) ou à la composition planctonique des échantillons d'une année à l'autre (par exemple, début et fin de bloom plus tôt/tard que les années précédentes), puissent intervenir. C'est pourquoi, une bibliothèque d'images importante (par exemple, 1000 à 2000 items par groupe taxonomique) sur une longue série temporelle pourrait être construite, afin de caractériser au mieux la variabilité temporelle des particules. En mettant au point une méthode d'optimisation du temps de calcul, il serait alors envisageable de calculer les distances entre les items de l'échantillon et les items de la bibliothèque. Les items présentant les distances les plus faibles, pourrait alors être conservés pour former un set d'apprentissage. De cette façon, chaque échantillon pourrait être analysé grâce à un set d'apprentissage adapté.

- Dans ce travail, nous n'avons pas pris en compte les données contextuelles (date de prélèvement, zone géographique, etc.). Il serait également intéressant d'introduire ces informations pour la construction du set d'apprentissage (dans notre travail ou comme défini au point précédent), afin d'aider l'algorithme dans sa sélection d'items à ajouter/supprimer (remplacement des items du set d'apprentissage pour la même période de l'année et la même zone géographique que les nouveaux items). De cette façon, il serait possible de limiter l'impact d'un changement de la structure des communautés planctoniques au fil des années.
- En routine, l'analyse des échantillons peut être réalisé chronologiquement (comme dans nos expérimentations). Dans ce contexte, la plupart du temps, la variabilité de la composition de plusieurs échantillons successifs est négligeable. Il est donc envisageable d'analyser ces échantillons avec le même set d'apprentissage, jusqu'à l'analyse d'un nouvel échantillon totalement différent. Dans ce cadre, il serait intéressant de comparer les classes prédites initialement (sans correction de l'erreur) et les classes prédites après validation des items suspects. En se fixant un seuil de variation, l'utilisateur aurait deux possibilités :
    - s'il y a dépassement du seuil, alors tous les suspects doivent être validés et l'apprentissage actif (comme défini dans ce travail) est activé,
    - sinon, l'utilisateur ne valide qu'une fraction de suspects (10-15%) et fait alors confiance à la correction d'erreur.
- Il est important d'accorder une attention toute particulière aux données aberrantes (notamment dans un nouvel échantillon). Dans ce travail, nous avons commencé à étudier cette problématique en utilisant une méthode de détection basée sur le voisinage des items (LOF). En effet, ces outliers peuvent impacter différemment les résultats de la classification. Il est donc important de déterminer les items qui seront ajoutés/supprimés au set d'apprentissage et de vérifier leurs cohérences par rapport aux items présents.
- Bien entendu, toutes ces méthodes doivent être testés et validés sur des séries temporelles assez longues (plusieurs années) et contenant un nombre important d'échantillons.

Bibliographie

[1] Al Hasan M., Chaoji V., Salem S., Zaki M., 2009. **Robust partitional clustering by outlier and density insensitive seeding**. *Pattern Recognition Letters*, 30:994–1002.

[2] Breunig M., Kriegel H.P., Ng R., Sander J., 1999. **OPTICS-OF: identifying local outliers**. *Proc. European Conf. on Principles of Data Mining and Knowledge Discovery (PKDD)*, Prague, Czech Republic.

[3] Breunig M., Kriegel H.P., Ng R., Sander J., 2000. **Lof: Identifying density-based local outliers**. *ACM SIGMOD 2000 International Conference on Management of Data,* pp. 93—104.

[4] Denis K., Grosjean P., 2006. **Reconnaissance automatique du phytoplancton par analyse d'images numériques (PhytoImage). Première approche en utilisant la banque d'images IFREMER**. *Rapport Université de Mons Hainaut*. Laboratoire d'Ecologie Numérique des Milieux Aquatiques : 88 p.

[6] Dereume-Hancart F., 2013. **Correction statistique de l'erreur dans le cadre de la classification automatique du plancton**. Projet UMONS – EcoNum, Directeur : Grosjean Ph.

[7] Grosjean P. & Denis K., 2010. **Zoo/PhytoImage – Optimisation du set d'apprentissage REPHY en 2010.** *Rapport Université de Mons.*

[8] Solow A., Davis C., Hu Q., 2001. **Estimating the taxonomic composition of a sample when individuals are classified with error**. *Marine Ecology Progress Series*, 295:21-31.

[9] Tunin-Ley A., Maurer D., 2011. **Mise en oeuvre opérationnelle d'un système couplé de numérisation (FlowCAM) et de traitement d'images (ZooPhytoImage), pour l'analyse automatisée, ou semi-automatisée, de la composition phytoplanctonique d'échantillons d'eau de mer.** *Rapport RST/LER/AR/11/002*.

Écologie Numérique des Milieux Aquatiques
UMONS
Faculté des Sciences

Mise en forme des résultats Zoo/PhytoImage, en vue de l'intégration dans Quadrige²

Guillaume WACQUET & Philippe GROSJEAN

Novembre – Décembre 2014

Table des matières

Introduction

Dans ce rapport, nous abordons le problème de la mise en forme des résultats en sortie de Zoo/PhytoImage, en vue de l'intégration dans Quadrige². Dans un premier temps, il a été décidé d'utiliser le format « Quadrilabo ». En effet, en 2013, le processus d'intégration des données dans le système d'information de l'IFREMER, peut passer, pour les utilisateurs ne pouvant ou ne souhaitant pas utiliser l'interface de saisie classique, par l'envoi du fichier de données au format « Quadrilabo » à la cellule d'administration qui vérifie alors techniquement le contenu du fichier et lance le programme automatisé d'intégration. Ce fichier se présente sous la forme d'une matrice de type EXCEL regroupant en ligne l'ensemble des résultats à intégrer (une ligne par résultat). Chaque ligne est décrite en colonne par des champs de 3 types :

- Champs EDILABO (Standard) : regroupant les informations permettant l'identification des résultats (Paramètre, Support, Fraction, Méthode, …). Ils reposent généralement sur des codes SANDRE standardisés. Ces codes SANDRE doivent faire l'objet d'une identification préalable en s'appuyant sur le site du SANDRE.
- Champs Quadrige² : regroupant les informations propres à Quadrige² relatives aux métadonnées (caractéristiques des passages, prélèvements, échantillons). Ils sont indispensables pour faire le lien entre les résultats de mesures et les données *in situ* associées de Q².
- Champ "NIVEAU SAISIE" : Un résultat dans Q² peut être rattaché à un passage, à un prélèvement ou un échantillon. Cette information, propre à Q², est indispensable au programme de reprise automatique et doit figurer clairement dans un champ spécifique.

Dans cette étude, nous étudions plusieurs packages R disponibles sur le site du CRAN afin de créer et de mettre en forme automatiquement le fichier EXCEL au format Quadrilabo. Les fonctions R ainsi que les résultats associés sont présentés dans les sections suivantes.

Packages testés

gdata package

- Offre de bonnes solutions multiplateformes (disponible pour Windows, Mac et Linux).
- Requiert l'installation de librairies additionnelles Perl (notamment pour Windows).

```
require(gdata)
df = read.xls ("myfile.xlsx"), sheet = 1, header = TRUE)
```

XLConnect package

- Basée sur Java, multiplateforme.
- Ne nécessite pas l'installation d'Excel.
- Pour les grands ensembles de données, il peut être très lent.

```
require(XLConnect)
wb = loadWorkbook("myfile.xlsx")
df = readWorksheet(wb, sheet = "Sheet1", header = TRUE)
```

xlsx package

- Basé sur Java (utilisation de rJava), multiplateforme.
- Cependant, la fonction read.xlsx() peut être lente, lors de l'ouverture de fichiers Excel volumineux. La fonction read.xlsx2() est beaucoup plus rapide, mais nécessite de définir des classes de colonne manuellement.

```
require(xlsx)
read.xlsx("myfile.xlsx", sheetName = "Sheet1")
read.xlsx2("myfile.xlsx", sheetName = "Sheet1")
```

Ces trois packages ont été retenus car il permettent d'offrir des solutions multiplateformes (contrairement, par exemple, au package ***RODBC***, dédié exclusivement à Windows). Finalement, ici, la nécessité d'installer un interpréteur Perl ainsi que des librairies additionnelles pour l'utilisation du package ***gdata*** sous Windows (environ une centaine de Mo sur le disque), nous a poussé à nous tourner vers les packages ***XLConnect*** et ***xlsx***. De plus, ces derniers ne nécessitent pas l'installation d'outils Excel, et utilisent le même code Java (projet Apache POI), ce qui permet d'avoir le même niveau élevé d'inter-opérabilité entre les systèmes d'exploitation.

Champs Obligatoires (EDILABO)

(détaillés dans le manuel « Intégration des données QUADRILABO »)

- ligne : numéro de ligne (généré automatiquement)
- lieu : lieu de surveillance (code SANDRE) → cf. table de correspondance
- reseau : réseau de surveillance (code SANDRE)
  "REPHY", "REPHYO", "RNO", "ROCCH", "REPOM"
- saisisseur : organisme saisisseur (code SANDRE)
  "LER_Arcachon", "LER_Normandie", "LER_Bretagne-Nord", "LER_Boulogne-sur-Mer", "LER_Languedoc-Roussillon", "LER_Provence-Azur-Corse", "LER_Morbihan-Pays-de-Loire", "LER_Bretagne-Occidentale"
- date : date de prélèvement au format JJ/MM/AAAA → extraction automatique
- enginPrel : engin de prélèvement (code SANDRE)
- preleveur : organisme prèleveur (code SANDRE)
  "LER_Arcachon", "LER_Normandie", "LER_Bretagne-Nord", "LER_Boulogne-sur-Mer", "LER_Languedoc-Roussillon", "LER_Provence-Azur-Corse", "LER_Morbihan-Pays-de-Loire", "LER_Bretagne-Occidentale"
- supportEch : support échantillon (code SANDRE)
  "Eau", "Phytoplancton"
- numEch : numéro de l'échantillon → extraction automatique
- nivSaisieRes : niveau saisie résultat
  "PASS", "PREL", "ECHANT"
- parametre : paramètre (code SANDRE)
  "FLORTOT", "FLORPAR", "FLORIND"
- libelleParametre : libellé du paramètre (code SANDRE)
  "FLORTOT", "FLORPAR", "FLORIND"
- support : support (code SANDRE)
  "Eau", "Phytoplancton"
- fraction : fraction (code SANDRE)
  "Eau", "Eau_brute"
- methode : méthode (code SANDRE)
  "Microscope", "FlowCAM"
- taxonRes OU groupeTaxonRes (codes SANDRE) → cf. table de correspondance
- resultat : résultats → valeurs numériques
- unite : unité (code SANDRE)
  "nb/mL", "nb/L", "nb/m^3", "nb/10mL", "cellules/mL", "colonies/mL"

- analyste : organisme qui fait l'analyse (code SANDRE)
  "LER_Arcachon", "LER_Normandie", "LER_Bretagne-Nord", "LER_Boulogne-sur-Mer", "LER_Languedoc-Roussillon", "LER_Provence-Azur-Corse", "LER_Morbihan-Pays-de-Loire", "LER_Bretagne-Occidentale"
- remarque : valeur du résultat
  "Valeurs_valides"

Fonctions R

Fonction « createResObjXLS »

Cette fonction permet de créer un objet résultat sous la forme d'un data.frame, en vue d'être exporté dans un fichier XLS. Les champs QUADRILABO (au total, 60 paramètres) regroupent 3 types de données : champs EDILABO (données de référence relatives aux paramètres, fractions, support, méthodes, résultats, codes SANDRE, valeurs numériques), champs QUADRIGE² (métadonnées, informations relatives aux passages, prélèvements, échantillons, texte, date) et les champs SAISIE (niveau de saisie des résultats) → cf. manuel « *Intégration des données QUADRILABO* ».

Fonction 1 : createResObjXLS

Inputs : *dat* (objet résultat obtenu après validation : « smpName_valid.RData »)
champs_QUADRILABO (chaque colonne du fichier XLS)
Outputs : *datRes* (objet résultat au format QUADRILABO)
codes_QUADRILABO ← conversion *champs/codes_QUADRILABO*
datRes ← data.frame(*dat*, *champs_QUADRILABO, codes_QUADRILABO*)
return *datRes*

Fonction « createXLSfile »

Cette fonction permet de créer et de mettre en forme un nouveau fichier XLS sur le disque. Pour la mise en forme, un code de couleur est utilisé : ROSE = champs EDILABO (obligatoires), BLEU = champs QUADRIGE² (métadonnées), VERT = champs SAISIE (niveau de saisie des résultats).

Fonction 2 : createXLSfile

Inputs : *filename* (nom du fichier XLS)
Outputs : *XLSfile* (fichier créé sur le disque)
datXLS ← *data.frame(champs_QUADRILABO)* : création du data.frame vide
wb ← *createWorkbook(type = "xls")* : création d'un nouveau workbook au format XLS
addDataFrame(wb, datXLS) : ajout du data.frame au workbook
setCellStyle(wb, ...) : format des cellules du workbook
saveWorkbook(wb, filename) : sauvegarde du workbook sur le disque

Fonction « fillXLSfile »

Cette fonction permet de compléter un fichier XLS présent sur le disque.

Fonction 3 : fillXLSfile

Inputs : *filename* (nom du fichier XLS)
dat (objet résultat obtenu après validation : « smpName_valid »)
champs_QUADRILABO (chaque colonne du fichier XLS)
Outputs : *XLSfile* (fichier complété sur le disque)
datRes ← *createResObjXLS(dat, champs_QUADRILABO)*
wb ← *loadWorkbook(filename)* : chargement du workbook
addDataFrame(wb, datRes) : ajout du data.frame au workbook
setCellStyle(wb, ...) : format des cellules du workbook
saveWorkbook(wb, filename) : sauvegarde du workbook sur le disque

Fonction « exportResultsToXLS »

Cette fonction reprend les fonctions précédentes, et permet l'exportation des résultats dans un fichier XLS (existant ou non).

Fonction 4 : exportResultsToXLS

Inputs : *filename* (nom du fichier XLS)
dat (objet résultat obtenu après validation : « smpName_valid »)
champs_QUADRILABO (chaque colonne du fichier XLS)
Outputs : *XLSfile* (fichier créé/complété sur le disque)
if *filename* n'existe pas
| *createXLSfile*(*filename*) : création d'un nouveau workbook sur le disque
| *fillXLSfile*(*filename*, *dat*, *champs_QUADRILABO*) : complétion du workbook
else
| *fillXLSfile*(*filename*, *dat*, *champs_QUADRILABO*) : complétion du workbook

Utilisation (fichier « testQuabdrige.R »)

```
Sys.setenv(JAVA_HOME='C:\\Program Files\\Java\\jre7')
require(rJava)  # Bibliotheque necessaire pour xlsx
require(xlsx)  # Bibliotheque necessaire pour manipuler des fichiers XLS
require(zooimage)  # Chargement de ZooImage
setwd("C:/Users/Guit/Desktop/UMONS – EcoNum/dataQuadrige/")  # Repertoire de travail
# Fichiers necessaires a la creation du fichier QUADRILABO
source("exportResultsToXLS.R")  # Fonctions principales pour l'exportation des resultats
source("sampleInfoNew.R")  # Fonction d'extraction d'informations sur les echantillons

dataValid <- get(loadObjects())  # Chargement des donnees validees : smpName_valid.RData

filename <- "./Results/Quadrilabo_UMONS_2014.xls"  # Nom du fichier XLS
createXLSfile(filename)  # Creation du fichier XLS
fillXLSfile(filename, datValid, reseau = "REPHY", saisisseur = "LER_Boulogne-sur-Mer",
parametre = "FLORTOT")  # Completion du fichier XLS
# OU
exportResultsToXLS(filename, datValid, reseau = "REPHY", saisisseur = "LER_Boulogne-
sur-Mer", parametre = "FLORTOT")  # Creation et completion du fichier XLS
```

Résultats

En ce qui concerne les codes SANDRE des lieux de surveillance (champ « CODE LIEU DE SURVEILLANCE ») et des taxons (champs « TAXON_RESULTAT » et « GROUPE_TAXON_RESULTAT »), une table de correspondance entre noms et codes doit être fournie par la cellule d'administration de Quadrige² (demande envoyée le 13/11/2014 à la cellule d'administration et à Antoine Huguet ; réponse reçue le 10/12/2014).

	A	B	C	D
1	NUMERO	CODE LIEU DE SURVEILLANCE	CODE RESEAU	CODE SANDRE SAISISSEUR
2	1	BL1	0000000014	1394
3	2	BL1	0000000014	1394
4	3	BL1	0000000014	1394
5	4	BL1	0000000014	1394
6	5	BL1	0000000014	1394
7	6	BL1	0000000014	1394
8	7	BL1	0000000014	1394
9	8	BL1	0000000014	1394
10	9	BL1	0000000014	1394
11	10	BL1	0000000014	1394
12	11	BL1	0000000014	1394
13	12	BL1	0000000014	1394
14	13	BL1	0000000014	1394
15	14	BL1	0000000014	1394
16	15	BL1	0000000014	1394
17	16	BL1	0000000014	1394
18	17	BL1	0000000014	1394
19	18	BL1	0000000014	1394
20	19	BL1	0000000014	1394
21	20	BL1	0000000014	1394
22	21	Me2	0000000014	1394
23	22	Me2	0000000014	1394
24	23	Me2	0000000014	1394
25	24	Me2	0000000014	1394
26	25	Me2	0000000014	1394
27	26	Me2	0000000014	1394
28	27	Me2	0000000014	1394
29	28	Me2	0000000014	1394
30	29	Me2	0000000014	1394
31	30	Me2	0000000014	1394
32	31	Me2	0000000014	1394
33	32	Me2	0000000014	1394
34	33	Me2	0000000014	1394

	AX	AY	AZ	BA	BB	BC
1	TAXON_RESULTAT	GROUPE_TAXON_RESULTAT	RESULTAT	RESULTAT_QUALITATIF	CODE SANDRE UNITE	CODE SANDRE ANALYSTE
2	Thalassiosira_cells		4		228	1394
3	Chaetoceros_spp		56		228	1394
4	Ciliophora_small		27		228	1394
5	Gyrodinium_spp		48		228	1394
6	D.brightwelli		97		228	1394
7	Rhizo_Probo_spp		49		228	1394
8	P.micans		7		228	1394
9	Ciliophora_big		2		228	1394
10	C.curvisetum		10		228	1394
11	Pleuro_Gyro_empty		1		228	1394
12	D.fragilissimus		11		228	1394
13	Pleuro_Gyro_small		1		228	1394
14	Pseudo_Nitzschia_spp		28		228	1394
15	P.globosa		98		228	1394
16	C.danicus		1		228	1394
17	Mues		3		228	1394
18	Paralia_spp		1		228	1394
19	G.striata		1		228	1394
20	C.socialis		3		228	1394
21	Thalassiosira_colonies		2		228	1394
22	Pseudo_Nitzschia_spp		1745		228	1394
23	Thalassiosira_colonies		5		228	1394
24	G.delicatula		258		228	1394
25	A.glacialis		18		228	1394
26	Rhizo_Probo_spp		11		228	1394
27	Ciliophora_small		3		228	1394
28	Chaetoceros_spp		1		228	1394
29	Ciliophora_big		4		228	1394
30	P.globosa		2		228	1394
31	N.longissima		2		228	1394
32	Zooplankton_spp		1		228	1394
33	Mues		2		228	1394
34	Thalassiosira_cells		4		228	1394

Conclusion

Pour la mise en forme des données au format Quadrilabo, des fonctions ont été développés grâce à l'utilisation des packages R « XLConnect » et « xlsx ». Aujourd'hui, une réflexion commune est tout de même menée en parallèle afin de prendre en compte toutes les particularités des résultats obtenus avec le système couplé FlowCAM/ZooPhytoImage (bancarisation des données brutes issues du FlowCAM, mesures sur chacune des particules et/ou sur chacun des groupes taxonomiques, résultats dépendants de la méthode et du set d'apprentissage utilisé, etc.). Les conclusions et perspectives dégagées au terme de cette réflexion sont présentées dans le livrable 2 (« Mise en oeuvre opérationnelle de l'outil FlowCAM/ZooPhytoImage dans le cadre de la surveillance REPHY »), ainsi que dans les annexes 1 et 2 (compte-rendus des réunions du comité de pilotage du projet FlowCAM/ZooPhytoImage) de ce livrable.

Annexe 1

Compte-rendu du Comité de Pilotage
du projet FlowCAM/ZooPhytoImage

VisioConférence

28 Mai 2014

COPIL FlowCAM / ZooPhytoImage.

Visio et télé conférence, 28 mai 2014

Participants

Site de Boulogne-sur-Mer : Alain Lefebvre (Ifremer), Felipe Artigas (LOG CNRS ULCO Wimereux), Denis Hamad (ULCO LISIC Calais)

Site de Brest : Florent Colas, Jean-François Rolin, Luis Lampert, Raffaele Siano (Ifremer)

Site de Nantes : Catherine Belin, Nadine Neaud-Masson, Dominique Soudant (Ifremer)

Site d'Arcachon : Danièle Maurer (Ifremer)

Par téléphone : Philippe Grosjean (Univ. Mons)

Liste de diffusion

Participants + Jean François Cadiou (ODE/DIR), René Robert (ODE/UL/DIR), Chantal Compère, Michel Répécaud (REM/RDT), Pascale Hébert, Camille Blondel, Elvire Antajan (LER BL), Claire Méteigner, Myriam Rumèbe-Perrière (LER AR), Martin Plus, Antoine Huguet, Michel Lunven, Marie Madeleine Danielou, Marie Pierre Crassous (DYNECO), tous responsables LERs

Objectifs de la réunion

Mettre à jour la composition du Comité de Pilotage, faire un point sur l'avancement de l'outil opérationnel, examiner les perspectives du FastCam, et prévoir les échéances à venir.

Composition et organisation du COPIL

Le nouveau COPIL est composé de tous les participants + les personnes suivantes : Michel Répécaud, Michel Lunven et Elvire Antajan. En fonction des sujets abordés, des personnes supplémentaires peuvent être invitées. Le COPIL devrait se réunir deux fois par an.

Organisation Ifremer : le projet FlowCAM / ZooPhytoImage est une action au sens de la comptabilité analytique Ifremer, positionnée jusqu'à fin 2014 dans le département REM (projet Nouvelles Technologies). Les membres du COPIL sont d'accord pour que cette action soit scindée en deux actions : l'une restant à REM pour les aspects technologie et matériel, l'autre positionnée à ODE dans le projet REPHY, pour les aspects outil opérationnel et déploiement dans les LERs. Les deux actions devraient porter le même intitulé, plus générique que l'intitulé actuel trop lié au FlowCAM (qui pourrait ne plus être à terme le matériel de référence, voir plus bas) : par exemple « numérisation phytoplancton et traitement d'images par ZooPhytoImage ». L'action positionnée à REM est sous la

responsabilité de JF Rolin / M. Répécaud (à préciser), l'action positionnée à ODE sous la responsabilité de A. Lefebvre.

Alain Lefebvre est le pilote scientifique de l'ensemble du projet, et assure également le pilotage du COPIL.

Pour la gestion des documents, la création d'un site collaboratif est jugée inadéquate. Une DROPBOX partagée pourrait être une solution simplifiée. La création d'une liste mail est envisagée.

Convention ONEMA

Les livrables 2013 ont donné toute satisfaction à l'ONEMA, cf. ci-dessous commentaires de MC Ximenes :

« Globalement OK, synthèses et résumés clairs. Ces documents ont vocation à être diffusés ; il serait utile de faire quelques modifs pour rendre la lecture plus aisée pour des publics non avertis. Ci-dessous des suggestions… »

Pour la convention 2014, deux livrables sont attendus pour fin février 2015 :

Version évolutive de l'outil opérationnel de numérisation et d'analyse semi-automatique d'images de phytoplancton, utilisant le matériel FlowCAM et le logiciel ZooPhytoImage. Nouvelles perspectives	A faire par Univ. Mons + LOG Wimereux
Mise en œuvre opérationnelle de l'outil FlowCAM / ZooPhytoImage dans le cadre de la surveillance REPHY	A faire par Ifremer

Un pré projet de convention a été envoyé à l'ONEMA pour 2015, contenant principalement le budget : la demande de financement est de 120 000 euros au total. La fiche devra être complétée avec le contenu scientifique pour début juillet : C. Belin reviendra vers les principaux interlocuteurs pour finaliser la fiche[1].

Divers

Avancement de la thèse de Nour Ali (co-financée par Univ. Mons et Ifremer, co-encadrée par Mons et ULCO).

D. Hamad : des attributs de texture ont été ajoutés pour la caractérisation des images des particules en plus des attributs classiques fournis par PhytoImage. Pour le moment les résultats de classification obtenus avec les attributs mélangés (attributs classiques et attributs de texture notamment LBP Local Binary Patterns) ne sont pas satisfaisants. Ceci est étonnant car la bibliographie mentionne des améliorations des résultats en classification d'espèces de phytoplancton quand les attributs de texture sont ajoutés. Il faut toutefois mentionner que le matériel utilisé n'est pas le même et ce ne sont pas les mêmes données d'espèces de phytoplancton . La recherche d'explications est en cours.

[1] La dernière version de la fiche 2015 date du 15 juillet 2014

F. Artigas : Nour est revenue à Wimereux depuis mars ; une comparaison est en cours entre échantillons SRN passés au FlowCAM (X4) et au fluorimètre.

Projet JERICO 2 (Horizon 2020). JF Rolin : mettre en avant le projet FlowCAM / ZooPhytoImage pour anticiper de futurs financements européens. A. Lefebvre, en contact avec P. Farcy, signale une réunion prochaine à Boulogne sur JERICO 2 [La réunion a eu lieu le 13/6 en marge du colloque MAREL Carnot – Etaient présents : A. Lefebvre, I. Puillat, P. Riou, G. Charriat, M. Répécaud (Ifremer), F. Artigas, F. Schmitt (UMR LOG). Un « case study » orienté hydrologie-phytoplancton-zooplancton/eutrophisation/D5-DCSMM/Cytométrie/FLowCam/ZooScan/Fluorimétrie spectrale devrait être proposé].

Bancarisation. L'espace disque dédié au projet est actuellement de 4 To. Un doodle va être envoyé bientôt pour une réunion spécifique sur la bancarisation. Une première réunion devrait avoir lieu au plus tard mi-juillet[2].

Perspectives FastCam – Présentation de Florent Colas

Le FlowCAM n'est pas assez performant, en particulier d'un point de vue optique, c'est pourquoi F. Colas a développé un prototype de matériel appelé « FastCam », beaucoup plus rapide en acquisition. L'avancement du projet FastCam, est le suivant :

- optique OK
- informatique : OK pour le calcul des paramètres et des vignettes ; actuellement les fichiers générés sont de type PID (pour Plankton Identifier, logiciel utilisé par le ZooScan pour le zooplancton), mais cela ne demandera qu'une demi-journée de développement pour générer des fichiers ZID utilisables par ZooPhytoImage : ceci sera fait après que la nouvelle interface graphique aura été livrée par Ph. Grosjean (voir ci-dessous). PG confirme qu'il n'y aura pas de problème pour implanter ZooPhytoImage si le FastCam produit des fichiers ZID.
- mécanique : le packaging est actuellement celui d'un prototype de labo. Avant de continuer, il faut faire la validation de l'ensemble du système et surtout de la classification : si cela marche avec Plankton Identifier, cela marchera avec ZooPhytoImage.

Pour Florent, d'ici fin 2014, on aura une idée plus précise des avantages et inconvénients du FastCam par rapport au FlowCAM. Il faudra notamment comparer le temps de calcul nécessaire par l'ordinateur par échantillon. En effet, dans le cas du FastCAM/PhytoImage, tous les calculs seront réalisés après l'acquisition. Dans le cas du FlowCAM/Phytoimage, une partie est réalisée pendant le passage de l'échantillon et une seconde partie après le passage de l'échantillon.

[2] Cette réunion bancarisation a eu lieu le 29 septembre 2014 à Nantes (CR en cours)

Remarque de D. Hamad : l'augmentation de la résolution d'une image, efficace pour le cerveau humain, ne l'est pas forcément pour un logiciel de reconnaissance : il faut en tenir compte dans les spécifications et les tests.

Avancement ZooPhytoImage – Infos Philippe Grosjean

Prochaine version : outil opérationnel

Une première version du logiciel intégrant des améliorations de l'interface graphique, la validation partielle et le module de correction d'erreurs, a été envoyée à Ifremer par Ph. Grosjean en avril dernier et installée sur les FlowCAMs de Boulogne et Nantes, puis testée avec retour de la part de G. Wacquet avant son départ.

Concrètement, la version finale intégrant ces améliorations sera livrée en septembre 2014. PG a prévu de passer un contrat avec G. Wacquet pour l'aider à terminer. Ceci marque le début de la phase opérationnelle de l'outil dans le cadre du REPHY.

L'interface WEB marchera sur tous les systèmes d'information.

Nouvelles perspectives

Ph. Grosjean veut tester le training set adaptatif : au fur et à mesure que la validation est faite sur un échantillon, le training set s'améliore (à la limite, il n'est plus nécessaire d'avoir un training set au départ). L'intérêt est de remplacer des vignettes anciennes par des vignettes récemment validées. Ceci permettra de proposer au niveau national un set d'apprentissage de base qui sera adapté au contexte régional au fur et à mesure des passages d'échantillons. Par ailleurs, cette adaptation permettra d'intégrer des échantillons vivants et lugolés dans le même set.

En termes de perspectives, il serait intéressant de dresser un bilan des utilisations en cours du système et les utilisations potentielles considérant les possibilités qu'offrira un outil opérationnel (en tenant compte des appels d'offre et des sollicitations en cours – Exemples : campagnes IBTS, PELGAS, Camanoc – Projets Interreg, JERICO-NEXT, …).

Comptage des colonies

La prise en compte de la spécificité des taxons en colonie pour leur dénombrement par ZooPhytoImage est une priorité absolue pour tous les participants, dans la mesure où cela concerne de très nombreuses espèces.

Pour Philippe, les outils testés par Pierre Govaerts et basés sur une régression linéaire (cf. livrable 2 de la convention Ifremer / ONEMA 2013) donnaient des résultats assez satisfaisants. Philippe propose donc de continuer à travailler sur ces outils, mais le COPIL juge qu'il s'agit de la priorité 2, la priorité 1 étant la livraison de la version contenant l'interface graphique et le module de correction d'erreurs en septembre. Néanmoins, il serait intéressant dans l'intervalle que les observateurs travaillant sur les FlowCAMs fassent un comptage des particules pour chaque colonie pour améliorer la partie quantitative de l'outil de reconnaissance : au moment de la validation manuelle, indiquer à la

machine quel est le nombre compté manuellement (quand c'est possible, ce qui n'est pas toujours évident au 4X). La question de la conversion des biovolumes en nombre de cellules, ne sera pas réglée de cette façon, mais sera au moins testée.

Conclusion

Les retours sur l'utilisation du FlowCAM par des collègues étrangers nous conduisent à penser qu'il faut effectivement prévoir à moyen terme une alternative au FlowCAM. En attendant dans les prochaines années, les trois FlowCAMs devront être utilisés dans les trois sites de Boulogne, Nantes et Arcachon. Il est proposé que Nadine Neaud-Masson coordonne les utilisateurs des FlowCAMs et suive l'avancement du travail, en relation avec Alain Lefebvre.

Il faudrait d'ores et déjà intégrer dans la fiche ONEMA 2015 un projet de passage en production du FastCam, afin de prévoir l'avenir.

Le prochain CoPil aura lieu en novembre 2014.

Annexe 2

Compte-rendu du Comité de Pilotage
du projet FlowCAM/ZooPhytoImage

Réunion et Formation

2-3 Décembre 2014

COPIL FlowCAM / ZooPhytoImage et formation à la V5 de ZooPhytoImage Boulogne, 2 et 3 décembre 2014

Sommaire

Compte rendu

Coordination : Catherine Belin

Contributeurs : Alain Lefebvre, Luis Lampert, Nadine Neaud-Masson, Guillaume Wacquet, Felipe Artigas, Philippe Grosjean, Florent Colas, Danièle Maurer

Participants

COPIL : Alain Lefebvre, Camille Blondel & Pascale Hébert (Ifremer Boulogne), Philippe Grosjean & Guillaume Wacquet (Université Mons), Felipe Artigas (LOG CNRS ULCO Wimereux), Denis Hamad (ULCO LISIC Calais), Luis Lampert (Ifremer Brest), Catherine Belin & Nadine Neaud-Masson (Ifremer Nantes) - par visio-conférence : Danièle Maurer & Myriam Rumèbe (Ifremer Arcachon) - par téléphone : Florent Colas (Ifremer Brest).

Formation : les participants du COPIL présents à Boulogne + Marie-Madeleine Danielou (Ifremer Brest)

Liste de diffusion

Participants

+ membres du COPIL absents : Jean-François Rolin, Michel Répécaud, Michel Lunven & Raffaele Siano (Ifremer Brest), Dominique Soudant (Ifremer Nantes), Elvire Antajan (Ifremer Boulogne), Nicolas Chomérat (Ifremer Concarneau), Mathilde Schapira (Ifremer Port en Bessin)

+ Jean François Cadiou, René Robert, Chantal Compère, Claire Méteigner, Martin Plus, Antoine Huguet, Marie Pierre Crassous, tous responsables LERs

Objectifs

COPIL

Faire un état d'avancement des travaux (i) en lien avec la fiche ONEMA, (ii) de la thèse de Nour Ali, (iii) du projet FastCAM, (iv) de la bancarisation des données provenant de l'outil FlowCAM/ZooPhytoImage. Faire un bilan de l'utilisation de cet outil : (i) par les utilisateurs REPHY, (ii) lors de la campagne CAMANOC 2014. Faire un bilan des travaux en cours et à venir en lien avec cet outil (thèses, masters, etc)

Formation

Formation des utilisateurs à la dernière version (V5) de ZooPhytoImage, et liste des actions à prévoir.

COPIL – 2 décembre après midi

Bancarisation des données provenant de ZooPhytoImage

Une première réunion sur ce sujet a eu lieu le 29 septembre 2014 : le CR est en cours de finalisation.

Depuis cette date, Antoine Huguet (AH) a transmis à Philippe Grosjean (PG) un modèle de format Quadrilabo, sur lequel travaille Guillaume Wacquet[1] (GW). Il faudrait fournir à GW une liste à jour du référentiel taxinomique Quadrige² pour le phytoplancton : **action Nadine Neaud-Masson** (NNM)[2].

Le contenu de ce qui sera disponible directement dans Q², et de ce qui sera disponible dans Q² au travers d'un lien vers des fichiers hébergés ailleurs, n'est pas encore complètement défini. En particulier, actuellement les données phytoplancton acquises par microscope sont constituées, pour chaque taxon identifié dans un échantillon, d'un seul résultat qui est le nombre de cellules par litre. Avec ZooPhytoImage, en plus du nom du taxon (qui s'appuiera sur le référentiel Q²) et du nombre de cellules par litre, de nouvelles mesures seront disponibles (jusqu'à 26 variables avec la version V5 de ZooPhytoImage), mais parfois à un niveau différent du taxon, c'est à dire soit au niveau de chacune des particules numérisées, soit au niveau de l'échantillon, par exemple :

- la taille de chaque particule est calculée actuellement dans ZooPhytoImage comme le Diamètre Circulaire Equivalent (ECD)
- il en est déduit au niveau de l'échantillon un spectre de tailles, dont les classes sont à définir (par défaut, les classes sont actuellement définies de 10 en 10 µm)
- le biovolume est disponible pour chaque particule

Les réponses à apporter rapidement sont :

- doit-on stocker dans Q² les infos sur les particules, qui incluent de nombreuses particules n'appartenant pas au phytoplancton, ou bien les stocker dans un fichier externe ? (avec un lien par échantillon)
- dans le deuxième cas, quels sont les paramètres qui seront régulièrement utilisés et qui méritent donc d'être gardés dans Q², au niveau de chaque taxon ou au niveau de l'échantillon ?
- pour les paramètres supplémentaires qui seraient retenus, comment agréger au niveau d'un taxon les informations disponibles au niveau des particules qui le composent, par exemple pour le biovolume, cela pourrait être le min et le max des biovolumes de ces particules, mais est-ce satisfaisant et suffisant ?
- pour la taille, doit-on garder une information au niveau de chaque taxon ou au niveau de l'échantillon, ou bien aux deux niveaux ?

[1] GW a été recruté à Mons pour réaliser ce travail, et accueilli au LOG à Wimereux depuis le 17 novembre jusqu'au 31 décembre 2014

[2] Action en cours

- dans le premier cas (taxon), quelle est l'info à garder : l'ECD ou la plus grande longueur, ou bien la longueur et la largeur ? voir aussi la question de l'agrégation de ce résultat
- dans le deuxième cas (échantillon), un graphique de distribution des spectres de taille est actuellement disponible, qui pourrait être attaché à l'échantillon. Il serait intéressant d'avoir en plus le résultat du nombre de particules par classe de taille, ce qui fait autant de paramètres que de classes de taille, et dans ce cas veut-on voir la répartition des spectres de taille de l'ensemble des particules ou bien seulement de celles appartenant au phyto ?

Catherine Belin (CB) fera circuler un questionnaire pour recueillir les avis des membres du COPIL. Lors de ces réflexions, il s'agira de tenir compte des besoins « utilisateurs » mais aussi des conséquences que cela pourra engendrer sur le temps de réponse du serveur lors des extractions, le coût de fonctionnement de l'espace dédié, etc.

Il est suggéré de ne pas négliger ce volet bancarisation dans le montage des projets à venir en lien avec ce genre de système automatisé, voire à définir un work package spécifique.

Par ailleurs, pour un échantillon, les informations doivent être liées à une méthode : soit parce que la méthode a évolué dans le temps, soit parce que l'on doit pouvoir utiliser au choix une ou une autre méthode (par ex une méthode classerait les vignettes en seulement dix groupes, une autre classerait selon des regroupements beaucoup plus précis, etc). Dans l'interface ZooPhytoImage, la possibilité de faire ceci existe, et dans Q² cela ne pose pas de problème car un échantillon est toujours associé à un PSFM[3] qui contient la méthode. Il faudrait cependant ne pas multiplier les méthodes possibles et surtout ne pas faire évoluer les méthodes trop souvent, ce qui poserait un problème de mise à jour du référentiel Q².

Avec le set d'apprentissage actif, il faudra stocker le set en lien avec l'échantillon, car celui-ci évolue à chaque échantillon : cela doublera le volume de données à stocker, et encore seulement si on ne stocke pas les vignettes de l'échantillon. Ce problème est donc à considérer.

***N.B**. le set d'apprentissage actif est une des améliorations en cours de ZooPhytoImage, qui consistera à améliorer le set commun à tous les utilisateurs, en fonction des corrections que chacun d'entre eux aura apporté aux échantillons traités.*

Avancement du projet FastCAM

Les premiers tests de classification ont commencé en octobre 2014 comme prévu. Certaines espèces comme celles du genre *Pseudo-nitzschia* ont montré une mauvaise reconnaissance au 10X. En effet, les images sont apparues de mauvaise qualité suivant le petit axe des cellules. Le problème a été identifié. Il faut améliorer l'optique pour permettre une meilleure reconnaissance qu'au FlowCAM équipé d'un objectif 4X.

[3] PSFM : Paramètre-Support-Fraction-Méthode, entité du référentiel Q² nécessaire pour saisir tout résultat

La difficulté du FlowCAM en termes d'optique est de garantir une image nette pour toutes les cellules passant dans la cuve tout en assurant une bonne résolution. Le compromis alors trouvé s'était avéré bon pour les cellules de type *Alexandrium minutum* (voire compte-rendu de la réunion précédente). Il ne l'est pas pour toutes les cellules.

La société Fluid Imaging a résolu le problème en concevant une optique placée après l'objectif de microscope. Cette solution a fait l'objet d'un brevet abandonné depuis.

FC et Morgan Tardivel (MT) ont depuis travaillé sur une solution alternative. Il paraît en effet préférable de ne pas développer de système entrant dans le périmètre d'un brevet déposé par la société Fluid Imaging en vue d'un transfert industriel. Un produit commercial vendu par la société Keyence a été identifié et devrait être testé début 2015.

Les essais devront commencer le plus tôt possible en 2015 pour ne pas pénaliser la suite du projet. Il est en effet impératif que nous disposions d'une comparaison précise et complète entre le FlowCAM et le FastCAM d'ici la fin du premier semestre 2015.

Concernant les brevets, Fluid Imaging a déposé un premier brevet en 2005, il en a ensuite déposé d'autres, mais ils ne sont pas tous actifs. Florent pense qu'il est possible de contourner ces brevets et que ce ne devrait pas être un problème. PG signale cependant qu'un brevet peut empêcher toute utilisation, y compris en interne, la seule utilisation possible étant en test.

Il est recommandé de tester le FastCam non plus uniquement sur des cultures ou des échantilons mono-spécifiques, mais sur des échantillons ou cultures pluri-spécifiques afin de se mettre dans les conditions d'applications futures.

Le CoPil confirme qu'il n'est pas pertinent d'envisager l'achat d'un nouveau FlowCam en 2015 sans retour de l'état d'avancement du développement du FastCam et des possibilités d'essaimage (notion de brevet). Il faudra par contre être vigilant lors de la phase de préparation de l'EPRD 2016 afin de tenir compte des évolutions du REPHY au regard du déploiement de ces systèmes.

Etat d'avancement des travaux 2014

Deux livrables devront être fournis à l'ONEMA au plus tard fin février 2015.

Le premier est : « Version évolutive de l'outil opérationnel de numérisation et d'analyse semi-automatique d'images de phytoplancton, utilisant le matériel FlowCAM et le logiciel ZooPhytoImage. Nouvelles perspectives. » A faire par Université de Mons + LISIC Calais + LOG Wimereux

Il comprendra donc la livraison de la version 5, les présentations aux Journées REPHY, etc

Le deuxième livrable est « Mise en œuvre opérationnelle de l'outil FlowCAM / ZooPhytoImage dans le cadre de la surveillance REPHY ». A faire par Ifremer. L'état d'avancement est le suivant :

- à Nantes, entre 50 et 100 échantillons (2013 et 2014) ont été numérisés par NNM. Une grande partie des numérisations sont traitées avec ZooPhytoImage et toutes sont en attente

d'analyse avec la nouvelle version de ZooPhytoImage. Grâce à ces analyses, NNM fera aussi une comparaison entre ce qui est obtenu au microscope (résultats déjà saisis dans Q^2) et ce qui est obtenu par l'outil

- à Boulogne, les échantillons sont numérisés systématiquement depuis début 2013, ils sont donc à traiter. Un stage de M2 va contribuer à traiter ces échantillons, et à en faire la comparaison avec les résultats obtenus au microscope (tous saisis dans Q^2)
- à Arcachon, il y a des échantillons numérisés jusqu'en juin 2013, Une grande partie des numérisations sont traitées avec ZooPhytoImage et toutes sont en attente d'analyse avec la nouvelle version de ZooPhytoImage. Il faut recommencer les numérisations le plus vite possible, car il est important que les trois laboratoires possédant des FlowCAMs participent au test de la version 5 et avancent vers une phase opérationnelle, avant le déploiement vers d'autres laboratoires. Danièle Maurer (DM) attend des instructions. NNM rappelle qu'elle a passé deux semaines dont une avec GW en juin-juillet 2013 à Arcachon pour un soutien au laboratoire. GW y a exposé l'état d'avancement du projet et présenté la version 3, NNM a procédé à des numérisations, des comparaisons quantitatives axées sur les abondances de *Dinophysis*, remonté à GW des bugs de la V.4. (50 numérisations au 4x, 32 traitées par ZooPhytoImage V4, 1632 vignettes classées dont 257 *Dinophysis,* taxon qui manquait dans les bibliothèques d'images) et qu'elle est prête à retourner dans les semaines qui viennent (installation de la V5, démonstration, structuration de l'espace disque commun, travaux à faire).

Il y a donc de la matière pour ce livrable, il est important de commencer à rédiger rapidement (pour février 2015).

Pour le fonctionnement opérationnel à venir, DM demande si l'outil et le set d'apprentissage seront communs aux trois laboratoires. La réponse unanime est oui, il faut un set commun Manche – Atlantique. S'il avait été envisagé un outil différent par laboratoire dans un premier temps, ce n'est plus le cas avec la dernière version de ZooPhytoImage, intégrant la correction d'erreurs : ce sera nettement plus performant de travailler sur un même outil.

GW et NNM vont s'occuper de fusionner les trois sets existants dans un espace disque commun qui vient d'être ouvert à Brest pour le projet. NNM s'occupe de finaliser la gestion de cet espace disque avec RIC.

Avancement de la thèse de Nour Ali

Nour a travaillé sur l'acquisition des données et la préparation de la base d'apprentissage, sur trois classifieurs : Bayes, réseaux de neurones, SVM, Random Forest. En particulier, peut-on améliorer les résultats de classification en ajoutant des attributs, par exemple la texture ? Pour le moment, les résultats sont incohérents, et il n'y a pas d'explication. Des pistes d'évaluation/vérification des différentes étapes sont évoquées, ainsi que la suggestion de participation de Nour à l'amélioration de l'outil de classification et à l'alimentation du set d'apprentissage via les données numérisées au printemps 2013 et 2014 (radiale Saint Jean) en 10x et 4x en vivant.

Campagne CAMANOC

Il s'agit d'une campagne test en Manche mise en œuvre dans le cadre de l'approche écosystémique des pêches et de l'optimisation des campagnes halieutiques pour les besoins du Programme de Surveillance de la DCSMM, lors de laquelle de nombreux points ont été échantillonnés, avec plusieurs types d'analyses faites sur chacun de ces points : pocket ferry box couplé à un fluorimètre spectral, cytométrie en flux, numérisation par FlowCAM, etc. Les échantillons numérisés ne pourront pas être traités avant fin février, mais ce travail doit être signalé dans le livrable n°2, même s'il ne concerne pas directement le REPHY.

Projets 2015

La thèse de Laurie Perrot sur les coccolithes et leur détection est co-encadrée par Francis Gohin et Diana Ruiz-Pino du laboratoire LOCEAN de Paris. Des échantillons (env. 100) seront fixés au Lugol lors de la campagne à laquelle elle va participer sur le plateau continental argentin en mars 2015. Laurie et Luis Lampert vont ensuite les numérisera à Nantes avec le soutien de NNM vers le mois de mai au plus tôt. Puis ils seront traités à Nantes ou Brest avec le logiciel ZooPhytoImage.

A ce propos, il est important que les modalités d'utilisation du FlowCAM et de ZooPhytoImage soient claires pour la bonne collaboration entre intervenants REPHY et chercheurs. Les FlowCAMs ne bougent pas de leurs laboratoires respectifs, car leur première fonction est d'être affectés aux échantillons REPHY (c'est pour cela que nous sommes financés par l'ONEMA). Par contre, il serait dommage de ne les utiliser qu'à temps partiel, il est donc évident que toute utilisation supplémentaire sera la bienvenue, mais il faudra que les « laboratoires FlowCAM » fassent un planning des demandes, pouvant venir de chercheurs, étudiants, sans oublier les autres LERs pour lesquels une utilisation sera à prévoir en 2015. Par contre l'utilisation de ZooPhytoImage pourra être déconnectée et faite à distance par d'autres utilisateurs.

A Mons

Les travaux de GW pour 2015 sont en priorité : la préparation de l'intégration des données Q^2, le comptage des colonies dans ZooPhytoImage, l'amélioration de l'outil avec l'apprentissage actif.

La dernière source de financement pour le projet est BELSPO, mais ce n'est pas sûr que cela dure.

Projet microscope holographique : une demande d'accueil jeune chercheur a été déposée par le LOG Wimereux pour Eva-Maria Zetsche, travaillant sur l'analyse d'image pour le phytoplancton via l'utilisation d'un microscope holograpique qui serait rendu disponible par l'équipe VUB (Bruxelles) et NIOZ (Yerseke). Projet déposé fin novembre 2014 pour un démarrage en octobre-novembre 2015, et qui pourrait être complémentaire au projet FlowCAM/ZooPhytoImage sur pas mal d'aspects.

PG se positionne plus sur statistiques et développement d'outils que sur écologie.

A Boulogne

Un stagiaire de M2 FOGEM (ULCO) sera accueilli au LER de Boulogne entre janvier et juin 2015 et va contribuer à traiter les échantillons numérisés, et à en faire la comparaison avec les résultats obtenus au microscope (tous saisis dans Q^2). Un stagiaire de DUT aura éventuellement un lien avec celui de M2.

Les travaux en lien avec la mise en œuvre du Programme de Surveillance de la DCSMM et le développement d'un indice de composition du phytoplancton se poursuivent.

L'arbitrage final quant au démarrage du CPER MARCO (et ses conditions de réalisation) n'a toujours pas eu lieu. Rappelons que le CPER MARCO (Recherches marines et littorales en Côte d'Opale : des milieux aux ressources, aux usages et à la qualité des produits aquatiques, 2015-2018) est un projet structurant multi-laboratoires et multi-organismes, associant la mise en place d'instruments et d'outils (enquêtes, indicateurs) pour une approche globale de l'étude du milieu marin, de la ressource et de la qualité des produits aquatiques. Pour répondre à ces enjeux académiques et sociétaux, le projet s'articule autour de 6 axes :

- Observation et évaluation de l'environnement marin
- Structure, fonctionnement et dynamique des écosystèmes
- Productivité et durabilité des ressources halieutiques et aquacoles
- Qualité et sécurité des Ressources aquatiques
- Vulnérabilité et usages des éco-socio-systèmes marins et littoraux
- Ingénierie marine et littorale

Par ailleurs, l'arbitrage du projet H2020 JERICO-Next est attendu pour le premier semestre 2015. Les work packages concernant le projet FlowCam/ZooPhytoImage sont les :

- WP 3 Technology and methodology developments (JRA) : (Semi)-automated observatory of phytoplankton dynamics in European coastal marine waters
- WP4 Case Study : Application of (semi)-automated technologies for monitoring plankton distribution and dynamics within the North Sea and the Channel

A Wimereux

Fin 2015, Alain Lefebvre (AL) et Felipe Artigas (FA) re-proposeront un sujet de thèse sur FlowCAM / ZooPhytoImage. Ils sont également sur : un projet d'InterReg V Manche, un projet INTERREG V « 2 mers » DYMAPHY bis, H2020 JERICO-NEXT (implication dans l'inter-comparaison de techniques, des cas d'études, etc)

A Calais

Denis Hamad (DH) a participé à la thèse de Kevin Rousseauw, et participe à celle de Nour Ali. DH souligne l'avantage des thèses en co-tutelle.

D'un point de vue opérationnel, pour le REPHY

Le déploiement vers d'autres LERs de l'outil ZooPhytoImage est à discuter dans quelques mois, quand l'outil sera vraiment passé en phase opérationnelle dans les trois labos équipés d'un FlowCAM.

Entretemps, il est nécessaire de faire de la communication en interne sur le projet.

Prochaine réunion du CoPil

Au premier semestre 2015, à préciser en fonction de l'état d'avancement des points évoqués dans le présent compte rendu.

Formation à la V5 de ZooPhytoImage – 3 décembre matin

Conclusions de la formation

Onze personnes sont présentes à cette formation, en plus du formateur PG.

PG signale que cette version 5 est en beta-test, il y a donc une phase de test obligatoire pour faire remonter les bugs. En particulier, il faudrait vérifier que les sorties du logiciel correspondent à ce qui a été calculé par ailleurs avec une autre méthode.

De l'avis de CB, cette phase de test est à faire par les utilisateurs eux-mêmes qui doivent décrire tout ce qui ne correspond pas à un résultat attendu. Ceci est inhérent à tout logiciel. C'est de cette façon que cela fonctionne :

- pour Quadrige²
- pour le produit « alerte, résultats sanitaires en temps réel » avec la remontée des dysfonctionnements par les divers utilisateurs vers VIGIES, qui retransmet aux sociétés de service concernées
- de façon moins formelle, pour tous les produits gérés par VIGIES en R, avec remontée des utilisateurs (CB par exemple) vers Dominique Soudant ou Alice Lamoureux, si des corrections ou des améliorations sont à faire

Dans aucun de ces cas, l'utilisateur n'a en général la compétence ou la possibilité de rentrer dans les programmes. Par contre, c'est l'utilisateur quotidien, et lui seul, qui est capable de voir les petits dysfonctionnements, une fois les bugs bloquants réparés. Il est cependant nécessaire de prévoir une organisation, qui est sûrement dans un premier temps un passage obligé par GW. En particulier, la comparaison avec ce qui a été calculé par une autre méthode devrait être décrite une bonne fois pour toutes (équivalent d'un jeu de test).

Plusieurs types d'utilisateurs pourraient partager ce travail : les trois labos « FlowCAM », mais aussi Luis Lampert (qui est d'accord), Marie Madeleine Daniélou. Le LOG Wimereux est également volontaire pour y participer.

Améliorations en cours ou à prévoir

Une tentative de priorisation est faite sur les actions 2015.

Priorité 1

Comptage des cellules d'une colonie : en cours (GW), avec proposition de participation de Nour Ali.

Déplacer plusieurs vignettes en même temps : Mons doit nous dire si cela demande beaucoup de développement

Réflexion sur l'organisation du consensus pour régler les cas litigieux, du type corrections différentes faites par différents observateurs. Le pilotage de cette action doit être défini, il s'agit d'un travail de collaboration entre les utilisateurs du logiciel (à identifier au sein du CoPIL sur la base du volontariat) avec l'aide de spécialistes de la taxonomie phyto qui pourraient être sollicités.

Avancer sur le cahier des charges de la bancarisation : AH

Réflexion sur le training set actif (en cours de développement) : en mesurer toutes les conséquences en termes de bancarisation

Calcul des biovolumes : il faut calculer et tabuler les relations à utiliser pour ce calcul. Elles existent dans la littérature, mais il reste à les retrouver et les rentrer dans la table *ad hoc*.

Priorité 2

Créer plusieurs catégories « Other », par exemple « Other phytoplankton » et « Other non phyto ». Sur ce sujet, LL demande s'il est possible de pouvoir créer une nouvelle colonne en direct (au moment du tri), dans le cas où on observe des cellules dont on connaît le nom, mais qui ne figurent pas dans la liste.

Pour des particules phytoplancton non reconnues, mais dont la classe n'est pas encore définie : proposer la création d'une classe à partir de la liste Q^2

Traduire une partie du manuel en français

Identification des doublons

Ne pas calculer les biovolumes sur les particules coupées

Priorité non définie, à mettre en priorité 1 si ne demande pas beaucoup de développement

Ajouter une barre d'échelle sur les « full size » des vignettes

Annexe 3

Diapositives
Le FastCAM : une alternative au FlowCAM ?







Le FastCAM: une alternative au FlowCAM ?

M. Tardivel, B. Forest, M.-P. Crassous, M. Lunven, M.-M. Danielou, F. Colas.

Vitesse de numérisation

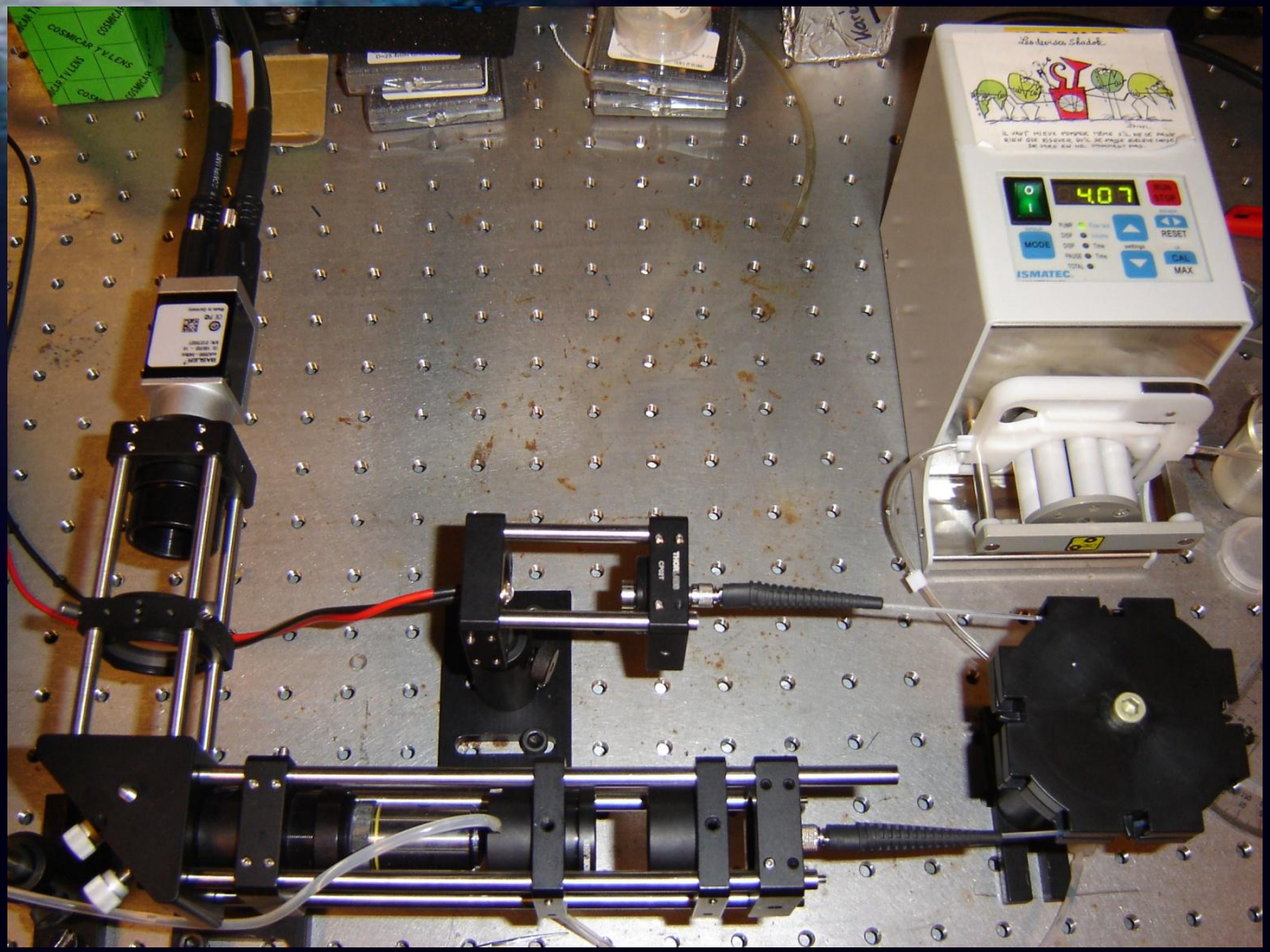
Le débit de numérisation est donné :

$$\Phi_{num} = V_i N \quad (1)$$

avec V_i le volume imagé et N le nombre de trames par minute.

	X4	X10
V_i (µL)	0,9	0,05
Φ_{num} (mL/min)	1,2	0,07
t_{10mL} (min)	8,3	143

Le FastCAM









Vitesse de numérisation

- Caméra rapide (340 i/s) et haute résolution (1024 x 2048 pixels):
  - Un flux brut à sauvegarder de 700Mo/s!
  - Légère compression jpeg + écriture sur la RAM

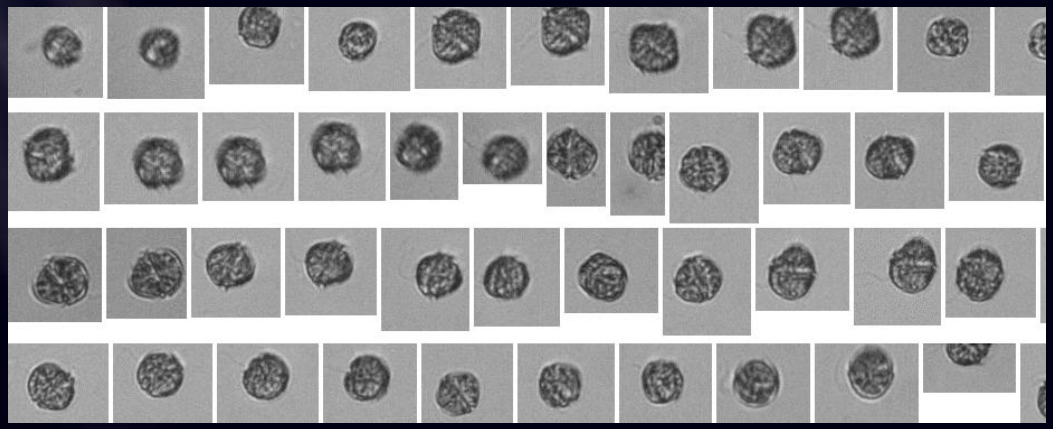
⇒ Numérisation de 10 mL en 13min, soit 11 fois plus rapide !

- Mais post-traitement environ 20-30min.

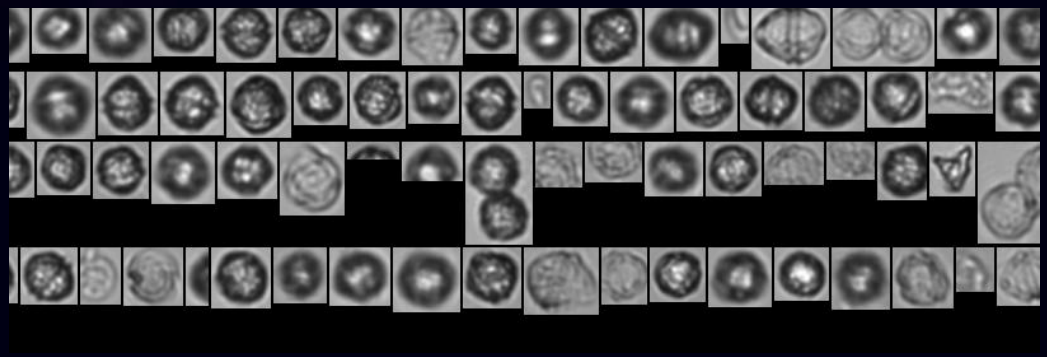
Qualité des images

- Travail sur l'illumination pour avoir une ouverture numérique (résolution) optimale.

FastCAM



FlowCAM



Ifremer







Conclusions et perspectives

- Vitesse d'acquisition ?
  - ⇒ Numérisation au FastCAM X10 aussi rapide qu'au FlowCAM X4.
- Qualité d'image ?
  - ⇒ Nous semble aussi bonne, à confirmer avec d'autres espèces et Set d'apprentissage.

Ifremer

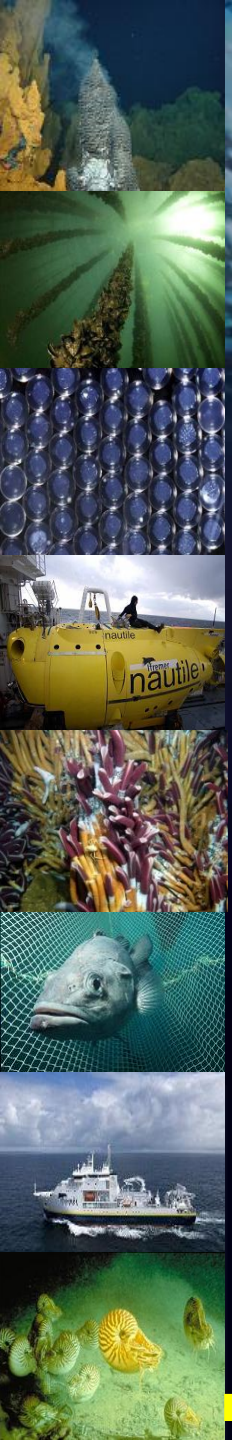






Conclusions et perspectives

- Optique:
  - OK,
- Informatique:
  - Acquisition => OK
  - Calcul des paramètres (environ 60) et des vignettes => OK
  - Génération de PID automatique => OK (bientôt ZID),
- Mécanique:
  - Packaging du prototype de laboratoire (A voir si Go).



Conclusions et perspectives

- Validation du système sur des échantillons naturels:
- Avec des échantillons de la mission Phytec + Irlande,
- Juin-Juillet: constitution d'un set d'apprentissage sur le model de celui du REPHY,
- Juillet-Octobre : Passages des échantillons, prédiction et validation des échantillons comptés au microscope par Pascale Malestroit (DYNECO Pelagos)

Ifremer







Budget pour la suite

- Besoin de budget pour
  - L'intégration mécanique (<5k€),
  - Dédoubler le système (10-12k€)

Annexe 4

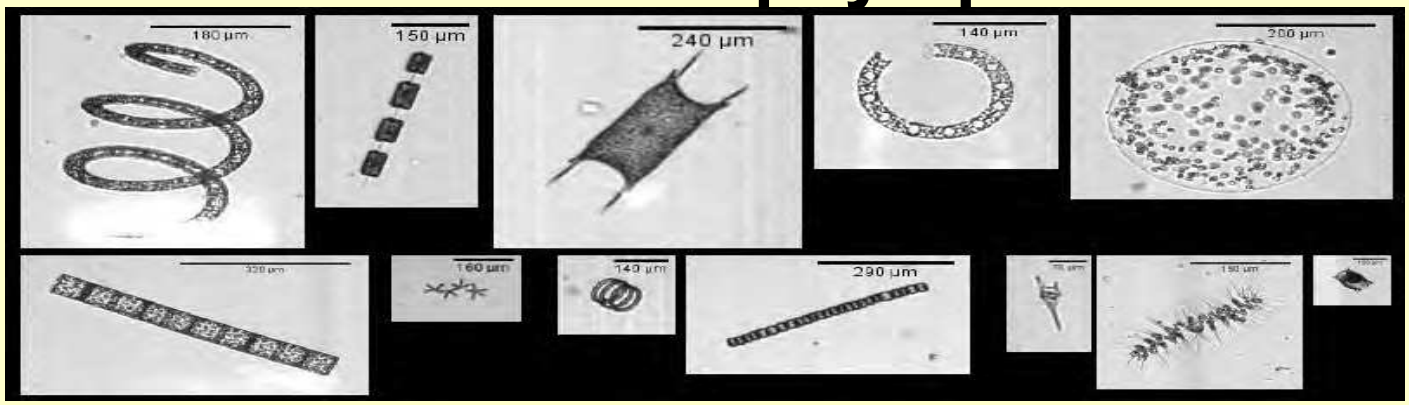
Diapositives

Le système FlowCAM/ZooPhytoImage, évolution de l'observation et de la surveillance du phytoplancton.

Journées REPHY 2014

Nantes

Le système FlowCam/ZooPhytoImage, évolution de l'observation et de la surveillance du phytoplancton.



Alain Lefebvre[1],Guillaume Wacquet[1], Philippe Grosjean[2],
Nadine Neaud-Masson[3], Danièle Maurer[4], Catherine Belin[3]

1 IFREMER,LER Boulogne-sur-mer, France

2 Université de Mons, Mons-Hainaut, Belgique

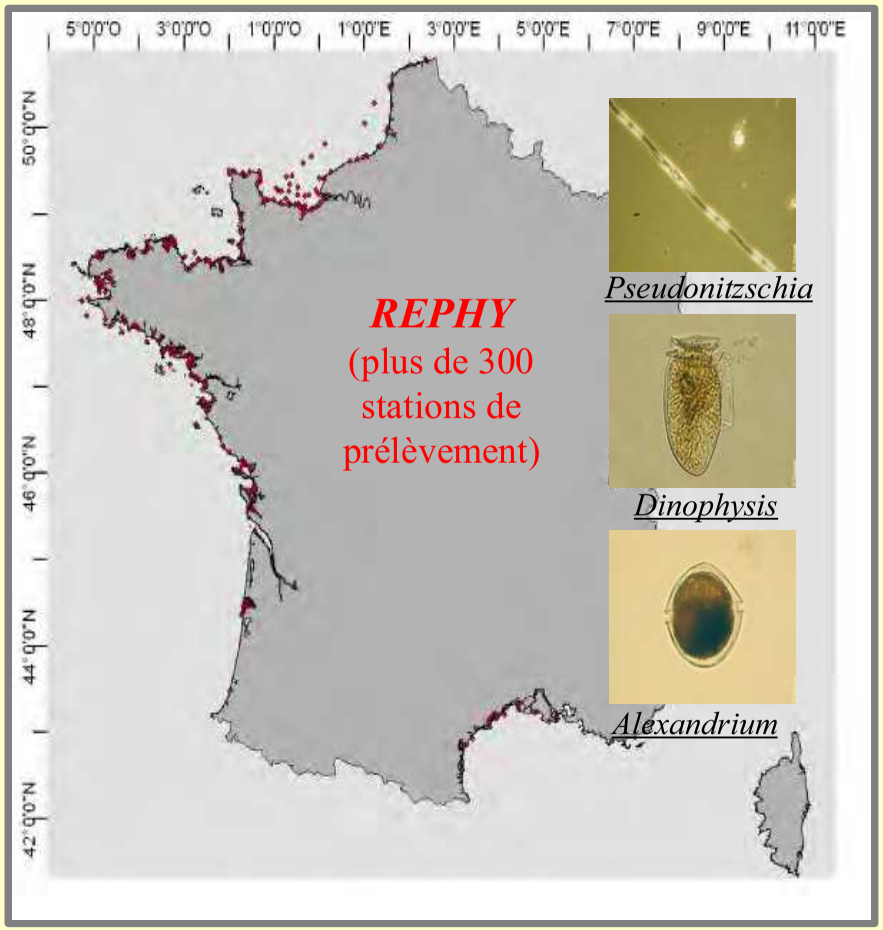
3 IFREMER,DYNECO, Nantes, France

4 IFREMER,LER Arcachon, France

Journées REPHY 2014, Nantes

Contexte et objectif

Surveillance/Observation



Microscope inversé

- Formation approfondie en taxonomie
- Quelques heures pour 10 mL
  - ***Fatigue, déconcentration***
  - ***Identifications erronées***
- Erreur non quantifiable

Nouvelles technologies

(FlowCAM, cytomètre en flux, …)

- Classification automatique
- Nouvelles informations, mesures
- Multiplication des analyses
- Suivi plus régulier
- Erreur quantifiable
- Stockage des données quasi-illimité

Présentation du projet

FlowCAM/PhytoImage

Système couplé de reconnaissance automatique du phytoplancton.

- FlowCAM : - **Flow C**ytometer and **CAM**era
  - Appareil d'acquisition et de numérisation des images
- PhytoImage : - Module de traitement d'images
  - Module de reconnaissance automatique

Objectif du projet

Optimisation de l'identification et du dénombrement du micro-phytoplancton avec le système couplé de numérisation et d'analyse d'images FlowCAM/PhytoImage.

Collaborations

- IFREMER : - Brest : optimisations technologiques et optiques
  - Nantes, Arcachon, Boulogne-sur-mer : utilisateurs du système
- UMONS : développement du logiciel PhytoImage
- ULCO : analyse de la complémentarité des informations obtenues par FlowCAMet par cytomètre en flux.

Bref historique...

- **<u>2008</u> : Première convention avec l'UMONS**

  Développement d'un logiciel pour l'étude du phytoplancton à travers l'analyse d'images.

- **<u>2009</u> : Convention IFREMER/ONEMA**

  Financement obtenu pour 2009.

- **<u>2009-2011</u> : Post-doctorat**
  - Janvier 2009 - Juin 2010 : IFREMER Arcachon
  - Octobre 2010 - Avril 2011 : Université de Bordeaux

<u>Projet stoppé pendant 20 mois</u>

- **<u>2012-2014</u> : CDD de 18 mois à Boulogne-sur-mer (G. Wacquet)**
  - Transfert de la responsabilité scientifique du projet au LER BL

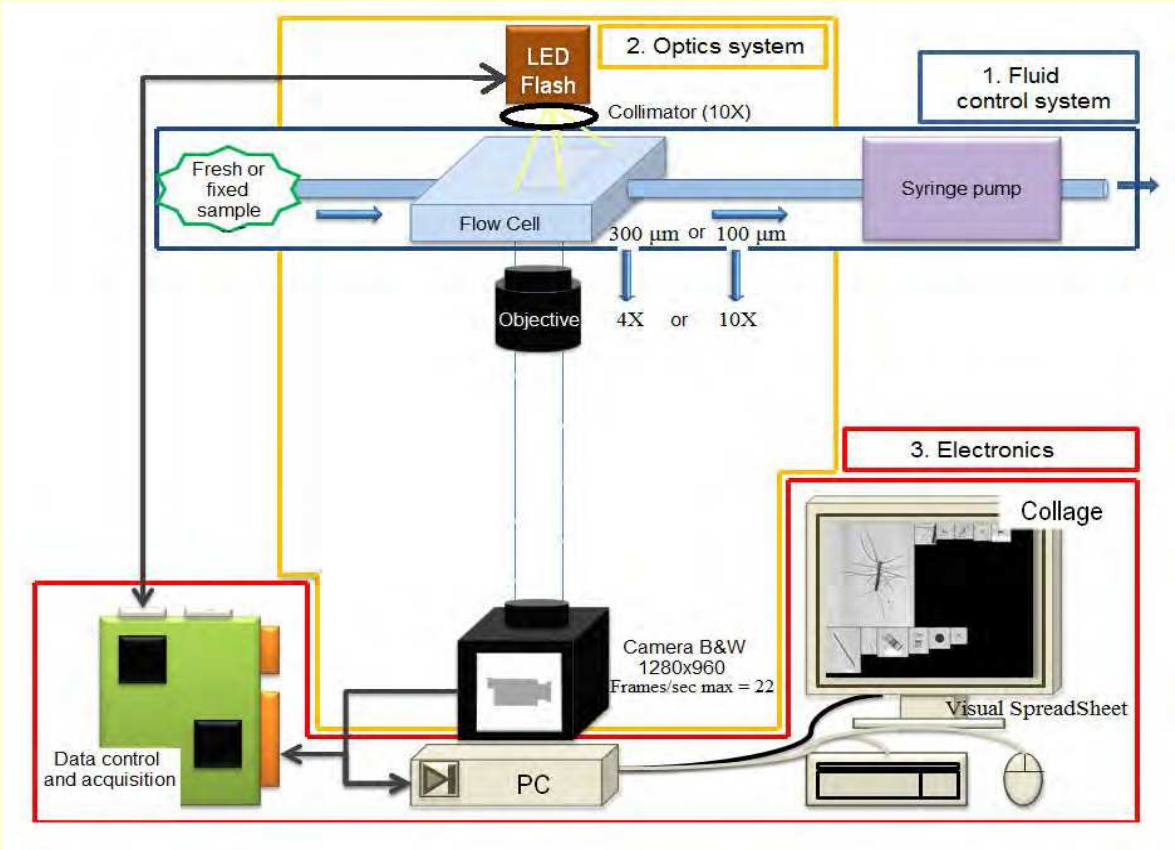
- **<u>2013-2015</u> : Convention IFREMER/ONEMA**
  - Janvier 2013 – Décembre 2015
  - Missions scientifiques : - Finalisation d'un outil opérationnel
    - Participation financière à une thèse de doctorat (N. Ali)

FlowCAM® (Flow Cytometry and CAMera)

Benchtop B2 Series FlowCAM® (Fluid Imaging Technologies)

- Pompe seringue externe C71
- Caméra N&B haute résolution :
  - Résolution : 1280x960
  - Nb images maximal par seconde = 22 fr/sec

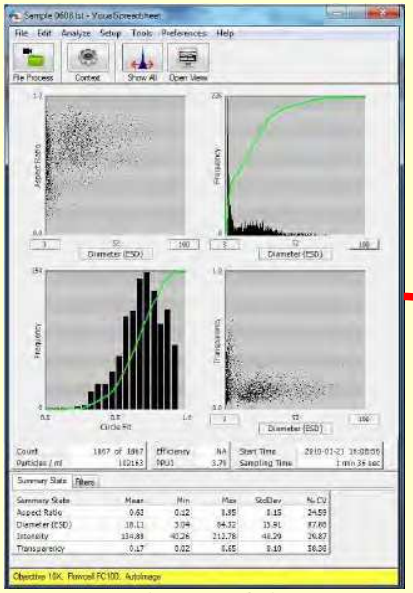
Benchtop B2 Series FlowCAM



Visual SpreadSheet®

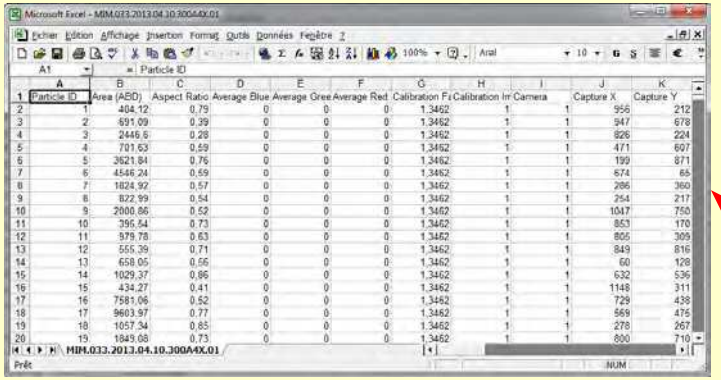
Logiciel couplé au FlowCAM® (Fluid Imaging Technologies)

- Configuration pour l'acquisition de données
- Acquisition et numérisation des données
- Post-traitement des données collectées



Visual SpreadSheet (fenêtre principale)

Collage



Fichier de mesures (format CSV)

21 paramètres par particule
(longueur, largeur, diamètre, périmètre, surface, niveaux de gris, etc.)

Réglages principaux pour l'acquisition

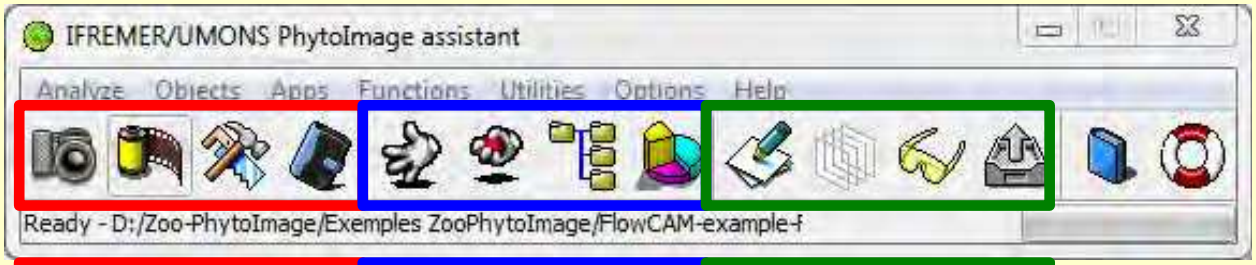
- Caméra : intensité et homogénéité lumineuse, images/sec, durée du flash, etc.
- Fluidique : volume d'échantillon, coefficient de dilution/concentration, vitesse de flux.
- Détection : seuil de segmentation, filtre de taille, etc.

PhytoImage (version 3.0-3)

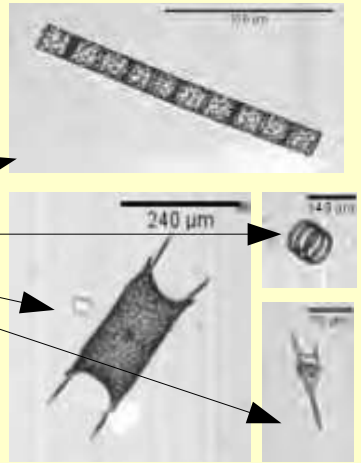
http://www.sciviews.org/zooimage

Logiciel développé par l'UMONS

- Toolbox sous une interface graphique (développé en R)
- Traitement d'images d'origine variée et avec des caractéristiques différentes



Images processing and thumbnails generation | **Training set, recognition tool and performances** | **Spatial and temporal analysis of series**



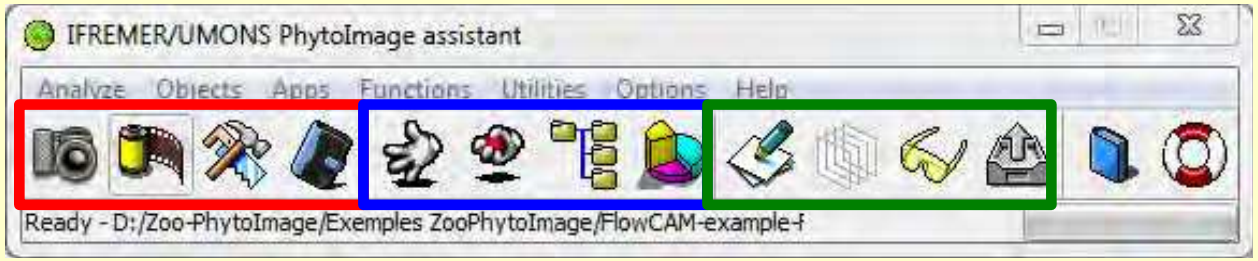
- **Traitement d'images**
  - Méthodes de segmentation et de détection de contours
  - Génération de vignettes (une image par particule)
  - 38 paramètres additionnels
- **Set d'apprentissage et outil de reconnaissance**
  - Classification supervisée basée sur un set d'apprentissage
  - Algorithme Random Forest
- **Analyse spatio-temporelle**
  - Abondances, biomasse, spectres de taille

+ 38 paramètres par particule => *59 paramètres*

PhytoImage (version 3.0-3)

http://www.sciviews.org/zooimage

Principe de fonctionnement



1. Traitement d'images

2. Phase d'apprentissage

Ensemble des vignettes

Tableau de mesures

Set d'apprentissage (sous-ensemble de vignettes représentatives des groupes taxonomiques, identifiées manuellement)

Algorithme Random Forest

Optimisation des performances de l'outil de reconnaissance

Outil de reconnaissance

Échantillon inconnu

Tableau de mesures

Identification basée sur les mesures

Reconnaissance automatique des particules

3. Phase de reconnaissance

Training set

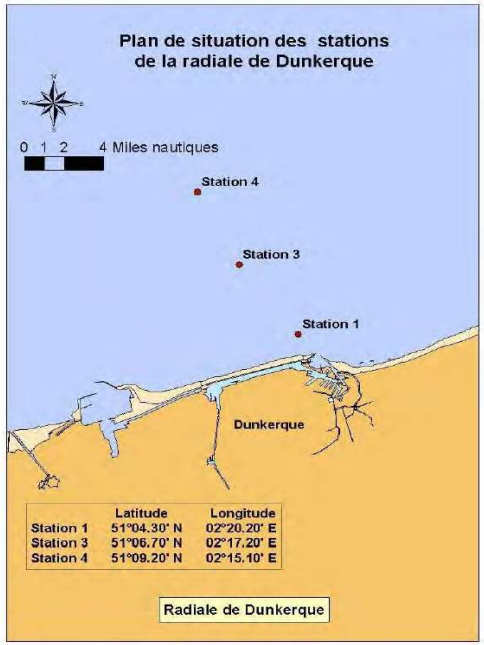
Zones d'étude : - Boulogne

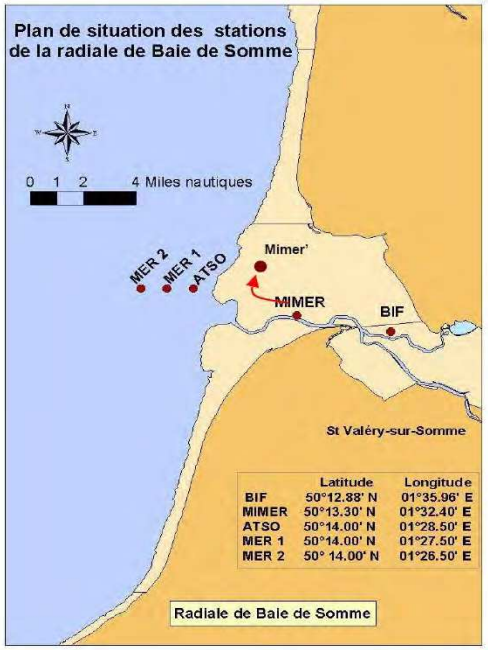
- Dunkerque

- Baie de Somme

Période : mars 2013 à juin 2013

→ 33 échantillons vivants





- **Alter**
  - **Artifact**
    - Bubble
  - **Misc**
    - Debris
      - Black_opaque
      - Clear
      - Dark
      - Granular
      - Menbranous
      - Short_thin
      - Long_thin
    - Fecal_pellets
    - Fiber
- **Phytoplankton**
  - **Bacillariophyceae**
    - Centrales
      - Biddulphiaceae
        - Biddulphia_mobiliensis
        - O.sinensis
        - Odontella_spp
      - Chaetoceraceae
        - Chaetoeros
          - C.socialis
          - Chaetoceros_spp
      - Leptocylindraceae
        - Leptocylindrus
          - L.danicus
      - Lithosdesmiaceae
        - Ditylum
      - Rhizosoleniaceae
        - Dactyliosolen
          - D.fragilissimus
        - Guinardia
          - Guinardia_spp
        - Rhizosolenia _imbricata
        - Proboscia alata
      - Thalassiosiraceae
        - Thalassiosira
          - Thalassiosira_spp
    - Pennales
      - Fragilariaceae
        - Asterionellopsis
          - A.glacialis
        - Thalassionema
          - T.nitzschioides
      - Naviculaceae
        - Pleuro_Gyrosigma
          - Pleuro_Gyrosigma_spp
          - Pleuro_Gyrosigma_empty
          - Pleuro_Gyrosigma_small
      - Nitzschiaceae
        - Nitzschia
          - N.longissima_cylindrotheca
        - Pseudo-nitzschia
  - **Ciliophora**
    - Ciliophora_big
    - Ciliophora_small
  - **Dinophyceae**
    - Gymnodiniales
      - Gyrodinium_spp
    - Peridiniales
      - Protoperidinium
        - Protoperidinium_spp
    - Prorocentrales
      - Prorocentrum
        - P.micans
  - **Euglenophyceae**
  - **Larva**
  - **Prymnesiophyceae**
    - Phaeocystales
      - Phaeocystaceae
        - Phaeocystis
- **Zooplankton**

Optimisation du training set

Groupes	Abondance	Error (%)	Groupes	Abondance	Error (%)
bubble	86	1,163	Short_thin	21	38.095
Gyrodinium_spp	179	3,371	Dark	140	42.143
Phaeocystis	202	3,960	fiber	54	42.593
L.danicus	203	4,433	Chaetoceros_spp	65	50.000
Pleuro_Gyrosigma_small	99	5,051	Protoperidinium_spp	4	50.000
O.sinensis	137	5,109	Fecal_pellets	19	52.632
Clear	126	7,087	larva	13	53.846
P.alata	211	7,109	Ditylum	72	54.930
A.glacialis	199	8,040	Granular	21	57.143
T.nitzschioides	69	8,696	Guinardia_spp	47	57.447
Pleuro_Gyrosigma spp	79	8,861	Odontella_spp	16	62.500
Pseudo-nitzschia	148	10,135	C.socialis	21	71.429
D.fragilissimus	147	12,245	Menbranous	7	71.429
Black_opaque	99	17,172	Long_thin	10	83.333
P.micans	27	18,519	R.imbricata_styli	14	85.714
Thalassiosira_spp	52	21,154	Ciliophora_big	16	93.750
Ciliophora_small	158	22,152	Biddulphia_mobiliensis	1	100.000
N.longissima_Cylindrotheca	11	27,273	Euglenophyceae	1	100.000
Pleuro_Gyrosigma_empty	45	28,889	Zooplankton	1	100.000

OPTIMISATION

Groupes	Abondance	Erreur (%)	Groupes	Abondance	Erreur (%)
Bubble	169	1,76	Short_thin	69	24,64
Gyrodinium_spp	181	2,21	Ciliophora_big	117	24,79
L.danicus	225	4	Dark	197	27,92
Pleuro_Gyrosigma spp	150	4,67	Ciliophora_small	94	28,72
O.sinensis	190	4,74	Chaetoceros_spp	140	30
Phaeocystis	203	4,93	P.micans	29	37,93
Clear	216	6,02	Protoperidinium_spp	22	40,91
Pleuro_Gyrosigma_small	100	6,36	Odontella_spp	20	45
G.flaccida	184	6,52	Fecal_pellets	31	45,16
T.nitzschioides	152	6,58	Zooplankton	8	50
Pseudo-nitzschia	207	7,25	Granular	33	51,52
A.glacialis	199	10,05	Larva	14	64,29
Rhizosolenia_Proboscia	228	10,97	G.delicatula	17	64,71
D.fragilissimus	159	14,47	C.socialis	17	82,35
Black_opaque	133	16,54	Ditylum	27	96,3
N.longissima_cylindrotheca	11	18,18	Biddulphia_mobiliensis	1	100
Pleuro_Gyrosigma_empty	65	18,46	C.fusus	6	100
Menbranous	15	20	Euglenophyceae	1	100
Thalassiosira_spp	54	20,37	G.stiata	1	100
fiber	127	23,62	paralia	12	100

Optimisation du training set

Groupes	Abondance	Error (%)	Groupes	Abondance	Error (%)
bubble	86	1,163	Short_thin	21	38.095
Gyrodinium_spp	179	3,371	Dark	140	42.143
Phaeocystis	202	3,960	fiber	54	42.593
L.danicus	203	4,433	Chaetoceros_spp	65	50.000
Pleuro_Gyrosigma_small	99	5,051	Protoperidinium_spp	4	50.000
O.sinensis	137	5,109	Fecal_pellets	19	52.632
Clear	126	7,087	larva	13	53.846
P.alata	211	7,109	Ditylum	72	54.930
A.glacialis	199	8,040	Granular	21	57.143
T.nitzschioides	69	8,696	Guinardia_spp	47	57.447
Pleuro_Gyrosigma spp	79	8,861	Odontella_spp	16	62.500
Pseudo-nitzschia	148	10,135	C.socialis	21	71.429
D.fragilissimus	147	12,245	Menbranous	7	71.429
Black_opaque	99	17,172	Long_thin	10	83.333
P.micans	27	18,519	R.imbricata_styli	14	85.714
Thalassiosira_spp	52	21,154	Ciliophora_big	16	93.750
Ciliophora_small	158	22,152	Biddulphia_mobiliensis	1	100.000
N.longissima_Cylindrotheca	11	27,273	Euglenophyceae	1	100.000
Pleuro_Gyrosigma_empty	45	28,889	Zooplankton	1	100.000

OPTIMISATION

Groupes	Abondance	Erreur (%)	Groupes	Abondance	Erreur (%)
Bubble	169	1,76	Short_thin	69	24,64
Gyrodinium_spp	181	2,21	Ciliophora_big	117	24,79
L.danicus	225	4	Dark	197	27,92
Pleuro_Gyrosigma spp	150	4,67	Ciliophora_small	94	28,72
O.sinensis	190	4,74	Chaetoceros_spp	140	30
Phaeocystis	203	4,93	P.micans	29	37,93
Clear	216	6,02	Protoperidinium_spp	22	40,91
Pleuro_Gyrosigma_small	100	6,36	Odontella_spp	20	45
G.flaccida	184	6,52	Fecal_pellets	31	45,16
T.nitzschioides	152	6,58	Zooplankton	8	50
Pseudo-nitzschia	207	7,25	Granular	33	51,52
A.glacialis	199	10,05	Larva	14	64,29
Rhizosolenia_Proboscia	228	10,97	G.delicatula	17	64,71
D.fragilissimus	159	14,47	C.socialis	17	82,35
Black_opaque	133	16,54	Ditylum	27	96,3
N.longissima_cylindrotheca	11	18,18	Biddulphia_mobiliensis	1	100
Pleuro_Gyrosigma_empty	65	18,46	C.fusus	6	100
Menbranous	15	20	Euglenophyceae	1	100
Thalassiosira_spp	54	20,37	G.stiata	1	100
fiber	127	23,62	paralia	12	100

Optimisation du training set

Groupes	Abondance	Error (%)	Groupes	Abondance	Error (%)
bubble	86	1,163	Short_thin	21	38.095
Gyrodinium_spp	179	3,371	Dark	140	42.143
Phaeocystis	202	3,960	fiber	54	42.593
L.danicus	203	4,433	Chaetoceros_spp	65	50.000
Pleuro_Gyrosigma_small	99	5,051	Protoperidinium_spp	4	50.000
O.sinensis	137	5,109	Fecal_pellets	19	52.632
Clear	126	7,087	larva	13	53.846
P.alata	211	7,109	Ditylum	72	54.930
A.glacialis	199	8,040	Granular	21	57.143
T.nitzschioides	69	8,696	Guinardia_spp	47	57.447
Pleuro_Gyrosigma spp	79	8,861	Odontella_spp	16	62.500
Pseudo-nitzschia	148	10,135	C.socialis	21	71.429
D.fragilissimus	147	12,245	Menbranous	7	71.429
Black_opaque	99	17,172	Long_thin	10	83.333
P.micans	27	18,519	R.imbricata_styli	14	85.714
Thalassiosira_spp	52	21,154	Ciliophora_big	16	93.750
Ciliophora_small	158	22,152	Biddulphia_mobiliensis	1	100.000
N.longissima_Cylindrotheca	11	27,273	Euglenophyceae	1	100.000
Pleuro_Gyrosigma_empty	45	28,889	Zooplankton	1	100.000

OPTIMISATION

Groupes	Abondance	Erreur (%)	Groupes	Abondance	Erreur (%)
Bubble	169	1,76	Short_thin	69	24,64
Gyrodinium_spp	181	2,21	Ciliophora_big	117	24,79
L.danicus	225	4	Dark	197	27,92
Pleuro_Gyrosigma spp	150	4,67	Ciliophora_small	94	28,72
O.sinensis	190	4,74	Chaetoceros_spp	140	30
Phaeocystis	203	4,93	P.micans	29	37,93
Clear	216	6,02	Protoperidinium_spp	22	40,91
Pleuro_Gyrosigma_small	100	6,36	Odontella_spp	20	45
G.flaccida	184	6,52	Fecal_pellets	31	45,16
T.nitzschioides	152	6,58	Zooplankton	8	50
Pseudo-nitzschia	207	7,25	Granular	33	51,52
A.glacialis	199	10,05	Larva	14	64,29
Rhizosolenia_Proboscia	228	10,97	G.delicatula	17	64,71
D.fragilissimus	159	14,47	C.socialis	17	82,35
Black_opaque	133	16,54	Ditylum	27	96,3
N.longissima_cylindrotheca	11	18,18	Biddulphia_mobiliensis	1	100
Pleuro_Gyrosigma_empty	65	18,46	C.fusus	6	100
Menbranous	15	20	Euglenophyceae	1	100
Thalassiosira_spp	54	20,37	G.stiata	1	100
fiber	127	23,62	paralia	12	100

Optimisation du training set

Groupes	Abondance	Error (%)	Groupes	Abondance	Error (%)
bubble	86	1,163	Short_thin	21	38.095
Gyrodinium_spp	179	3,371	Dark	140	42.143
Phaeocystis	202	3,960	fiber	54	42.593
L.danicus	203	4,433	Chaetoceros_spp	65	50.000
Pleuro_Gyrosigma_small	99	5,051	Protoperidinium_spp	4	50.000
O.sinensis	137	5,109	Fecal_pellets	19	52.632
Clear	126	7,087	larva	13	53.846
P.alata	211	7,109	Ditylum	72	54.930
A.glacialis	199	8,040	Granular	21	57.143
T.nitzschioides	69	8,696	Guinardia_spp	47	57.447
Pleuro_Gyrosigma spp	79	8,861	Odontella_spp	16	62.500
Pseudo-nitzschia	148	10,135	C.socialis	21	71.429
D.fragilissimus	147	12,245	Menbranous	7	71.429
Black_opaque	99	17,172	Long_thin	10	83.333
P.micans	27	18,519	R.imbricata_styli	14	85.714
Thalassiosira_spp	52	21,154	Ciliophora_big	16	93.750
Ciliophora_small	158	22,152	Biddulphia_mobiliensis	1	100.000
N.longissima_Cylindrotheca	11	27,273	Euglenophyceae	1	100.000
Pleuro_Gyrosigma_empty	45	28,889	Zooplankton	1	100.000

Erreur moyenne 17,55%

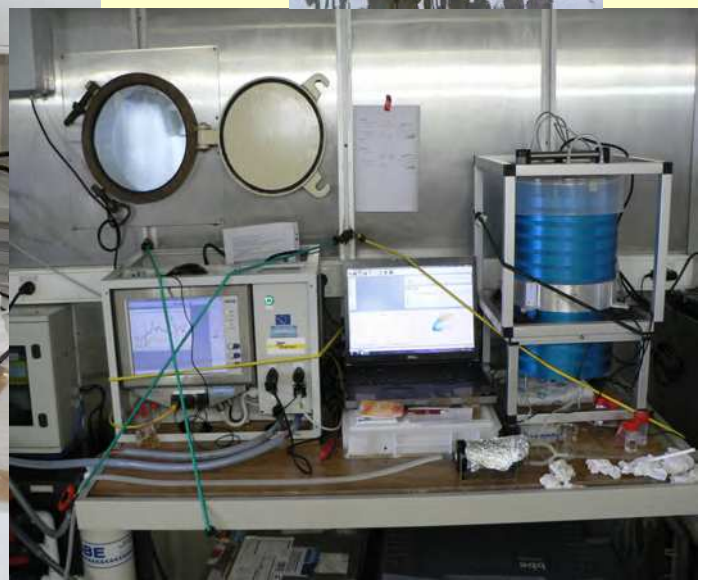
OPTIMISATION

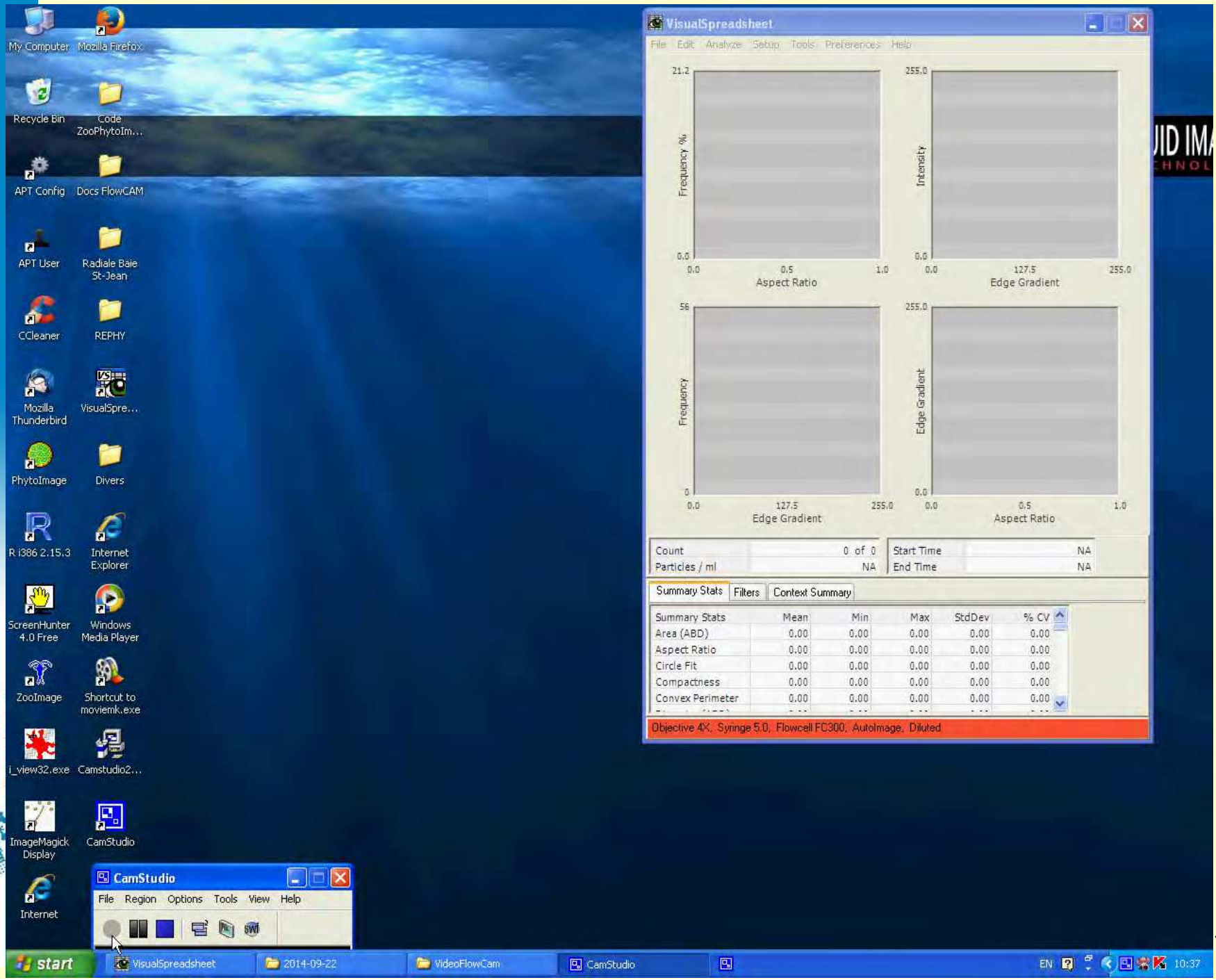
Groupes	Abondance	Erreur (%)	Groupes	Abondance	Erreur (%)
Bubble	169	1,76	Short_thin	69	24,64
Gyrodinium_spp	181	2,21	Ciliophora_big	117	24,79
L.danicus	225	4	Dark	197	27,92
Pleuro_Gyrosigma spp	150	4,67	Ciliophora_small	94	28,72
O.sinensis	190	4,74	Chaetoceros_spp	140	30
Phaeocystis	203	4,93	P.micans	29	37,93
Clear	216	6,02	Protoperidinium_spp	22	40,91
Pleuro_Gyrosigma_small	100	6,36	Odontella_spp	20	45
G.flaccida	184	6,52	Fecal_pellets	31	45,16
T.nitzschioides	152	6,58	Zooplankton	8	50
Pseudo-nitzschia	207	7,25	Granular	33	51,52
A.glacialis	199	10,05	Larva	14	64,29
Rhizosolenia_Proboscia	228	10,97	G.delicatula	17	64,71
D.fragilissimus	159	14,47	C.socialis	17	82,35
Black_opaque	133	16,54	Ditylum	27	96,3
N.longissima_cylindrotheca	11	18,18	Biddulphia_mobiliensis	1	100
Pleuro_Gyrosigma_empty	65	18,46	C.fusus	6	100
Menbranous	15	20	Euglenophyceae	1	100
Thalassiosira_spp	54	20,37	G.stiata	1	100
fiber	127	23,62	paralia	12	100

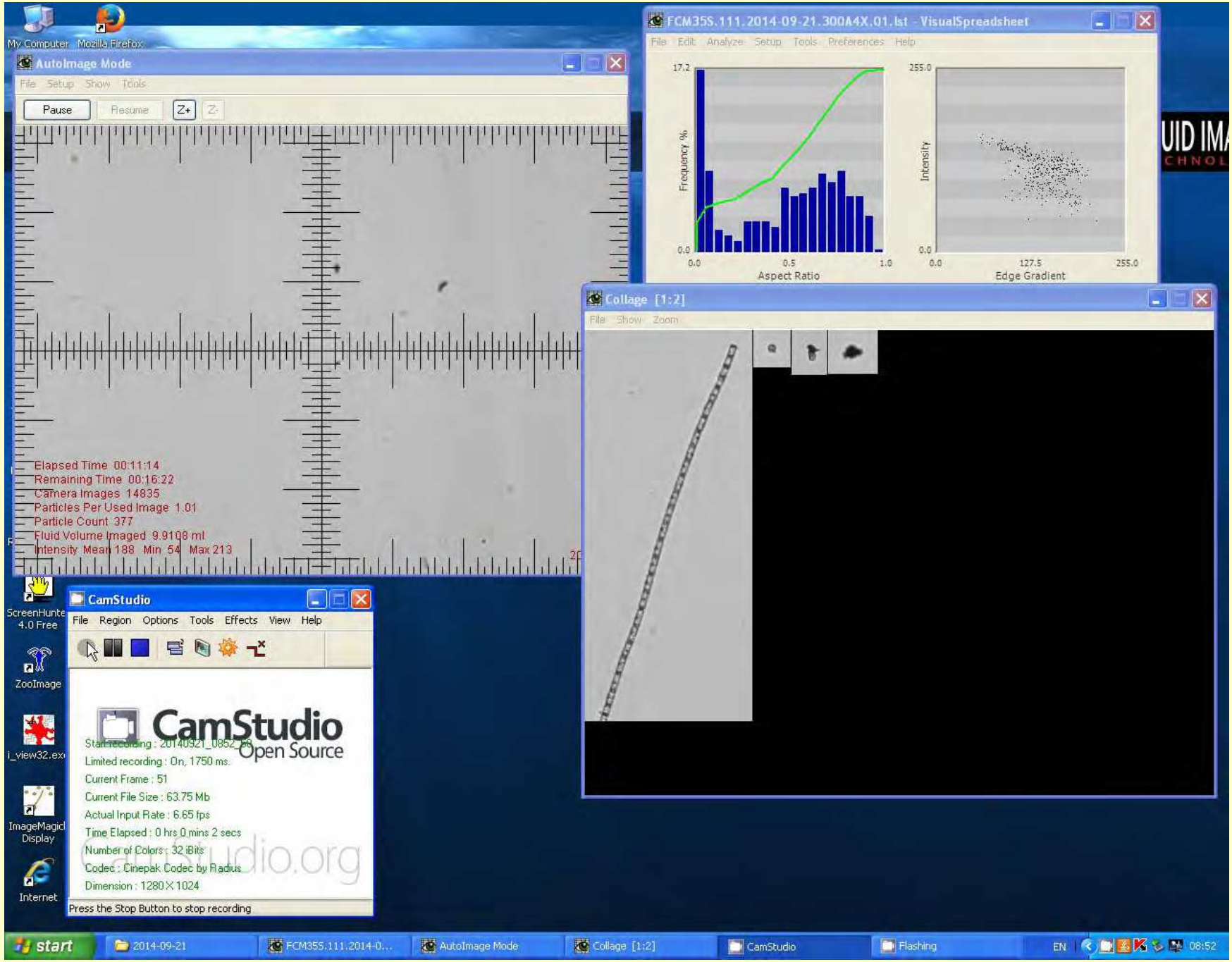
Erreur moyenne 14,66%

Le FlowCam à bord de la Thalassa

(Campagne CAMANOC, 15/9 au 15/10/2014)




My Computer
Mozilla Firefox
Recycle Bin
Code ZooPhytoIm...
APT Config
Docs FlowCAM
APT User
Radiale Baie St-Jean
CCleaner
REPHY
Mozilla Thunderbird
VisualSpre...
PhytoImage
Divers
R i386 2.15.3
Internet Explorer
ScreenHunter 4.0 Free
Windows Media Player
ZooImage
Shortcut to moviemk.exe
i_view32.exe
Camstudio2...
ImageMagick Display
CamStudio
Internet
CamStudio
File Region Options Tools View Help
VisualSpreadsheet
File Edit Analyze Setup Tools Preferences Help
Frequency %
Aspect Ratio
Intensity
Edge Gradient
Frequency
Edge Gradient
Edge Gradient
Aspect Ratio
Count 0 of 0
Particles / ml NA
Start Time NA
End Time NA
Summary Stats
Filters
Context Summary
Summary Stats Mean Min Max StdDev % CV
Area (ABD) 0.00 0.00 0.00 0.00 0.00
Aspect Ratio 0.00 0.00 0.00 0.00 0.00
Circle Fit 0.00 0.00 0.00 0.00 0.00
Compactness 0.00 0.00 0.00 0.00 0.00
Convex Perimeter 0.00 0.00 0.00 0.00 0.00
Objective 4X, Syringe 5.0, Flowcell FC300, AutoImage, Diluted
start
VisualSpreadsheet
2014-09-22
VideoFlowCam
CamStudio
EN
10:37


My Computer
Mozilla Firefox
AutoImage Mode
File Setup Show Tools
Pause
Resume
Z+
Z-
Elapsed Time 00:11:14
Remaining Time 00:16:22
Camera Images 14835
Particles Per Used Image 1.01
Particle Count 377
Fluid Volume Imaged 9.9108 ml
Intensity Mean 188 Min 54 Max 213
FCM35S.111.2014-09-21.300A4X.01.lst - VisualSpreadsheet
File Edit Analyze Setup Tools Preferences Help
Frequency %
Aspect Ratio
Intensity
Edge Gradient
Collage [1:2]
File Show Zoom
CamStudio
File Region Options Tools Effects View Help
CamStudio Open Source
Limited recording : On, 1750 ms
Current Frame : 51
Current File Size : 63.75 Mb
Actual Input Rate : 6.65 fps
Time Elapsed : 0 hrs 0 mins 2 secs
Number of Colors : 32 iBits
Codec : Cinepak Codec by Radius
Dimension : 1280 X 1024
Press the Stop Button to stop recording
ScreenHunter 4.0 Free
ZooImage
i_view32.exe
ImageMagick Display
Internet
start
2014-09-21
AutoImage Mode
Collage [1:2]
CamStudio
Flashing
EN
08:52

Conclusions et perspectives

Système couplé FlowCAM/PhytoImage

- Stockage des données → vérification ou nouvelles analyses ;
- Réduction de la pénibilité et de l'erreur « observateur » ;
- Temps d'analyse réduit → multiplication des analyses → suivi à haute résolution ;
- Nouvelles informations → études écologiques plus poussées ;
- Système opérationnel pour le REPHY optimisé et la DCSMM

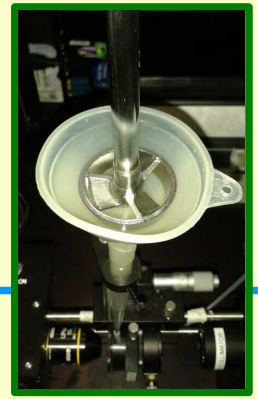
Perspectives

- Optimisation des sets d'apprentissage (pour les espèces sous-représentées) ;
- Dénombrement des cellules en colonies (Bacillariophycées et *Phaeocystis globosa*)
- Étude de la complémentarité des objectifs 4X et 10X ;
- Étude comparative de différents appareils (FlowCAM, cytomètre en flux, …) ;
- Contribution au développement d'un indicateur « indice composition du phytoplancton » DCE et DCSMM compatible.
- Développement d'une interface graphique interactive, ergonomique et intuitive => livraison de la Version 4 (« user friendly ») en octobre 2014.
- Essaimage dans les LERs (Observation, Surveillance & Recherche).

Merci de votre attention

Derniers travaux réalisés

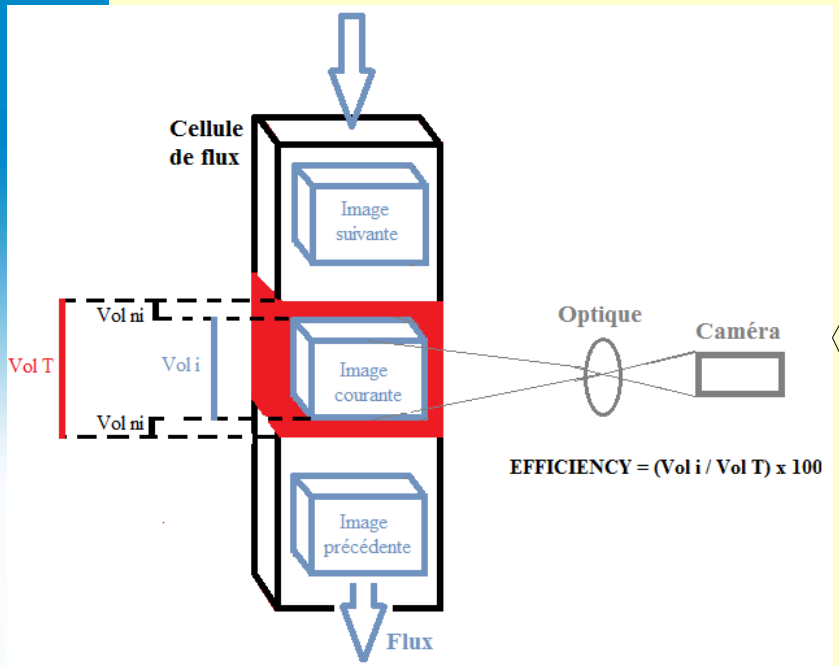
- Brest (LDCM) : améliorations optiques et mécaniques
  - Exemples : développement d'un système de mise au point automatique pour le FlowCAM ; Prototype FASTCAM
- Mons : modules supplémentaires pour ZooPhytoImage
  - Exemple : développement du module de correction statistique de l'erreur → validation des vignettes les plus « suspectes »
- Boulogne : optimisation du protocole de numérisation
  - Collaboration avec Nantes et Arcachon
  - - sédimentation des particules, sens du débit, homogénéisation/agitation, vitesse de flux, dilution/concentration, étude quantitative,
  - Acquisition d'images, constitution de training sets et optimisation des performances de l'outil de reconnaissance

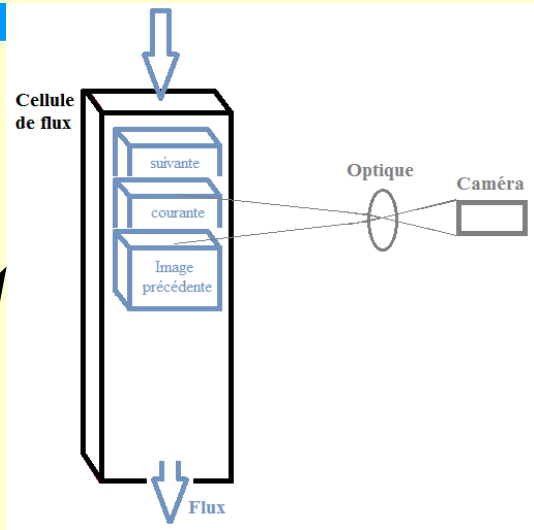


Annexe 1

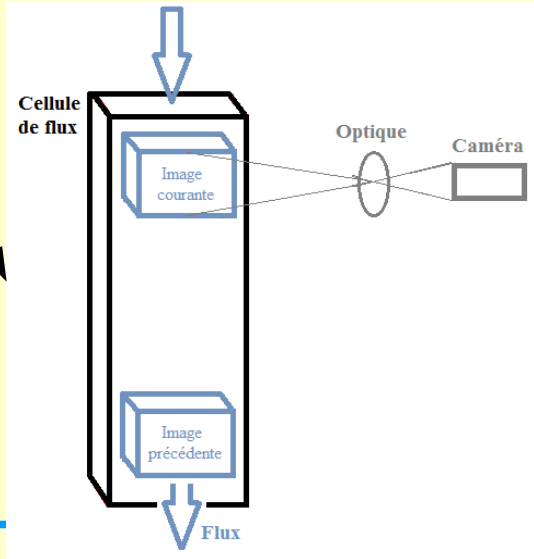
Optimisation du protocole de numérisation

Optimisation de la vitesse





RISQUE DE MULTI-IMAGING



RISQUE DE PERTE DE PARTICULES

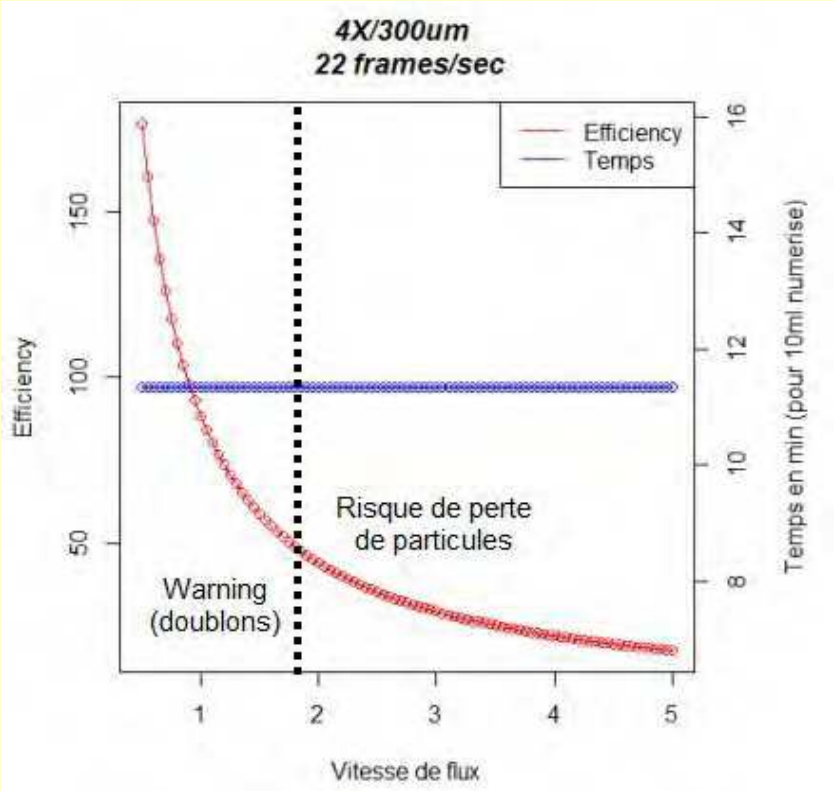
- Vol_T : volume total de la cellule de flux
- Vol_i : volume réellement imagé (selon caméra)

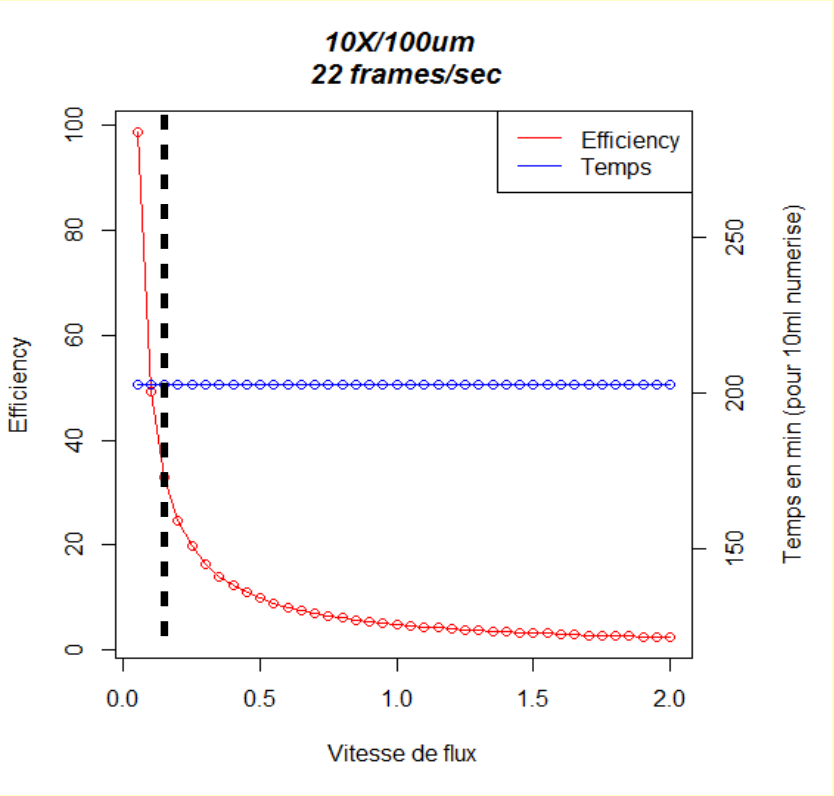
Annexe 1 (suite)

Optimisation du protocole de numérisation

Optimisation de la vitesse de numérisation

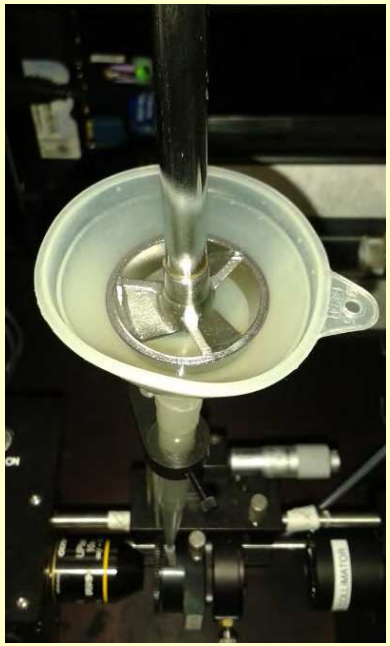
- Efficacité recommandée (par expérimentations et constructeurs) =
  - **50 % pour 4X/300µm**
  - **30% pour 10X/100µm**
- Nombre d'images par seconde prise par la caméra = **22 fr/sec**



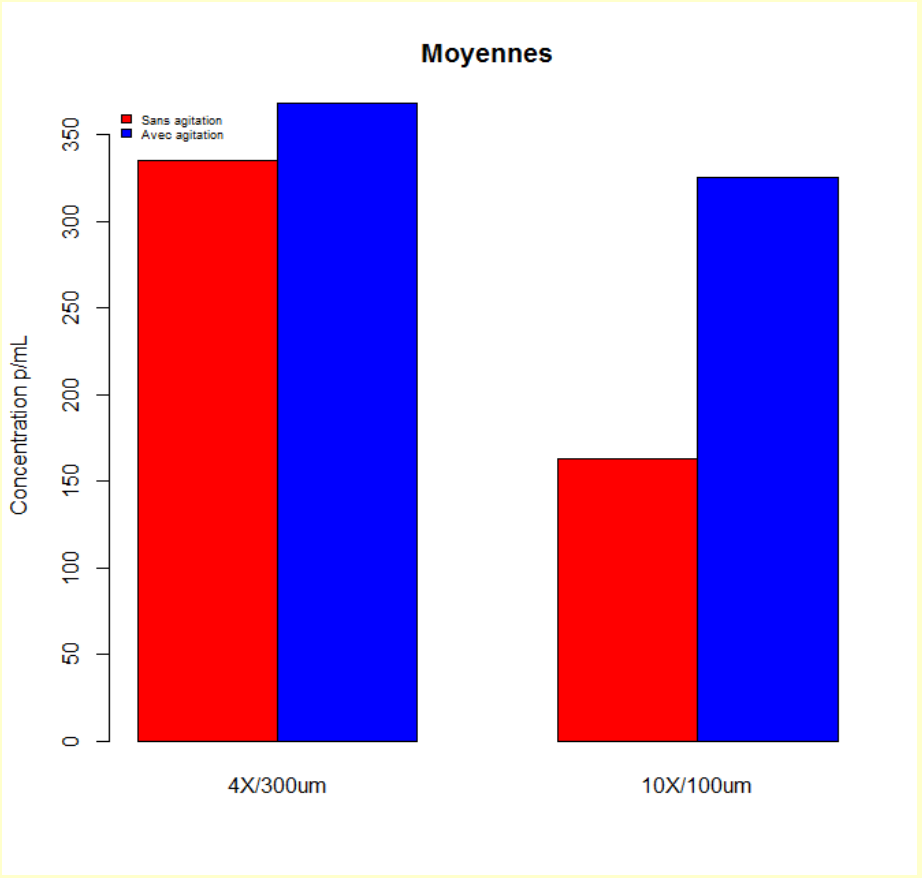


Annexe 2

Homogénéisation/agitation de l'échantillon



Agitateur à hélice

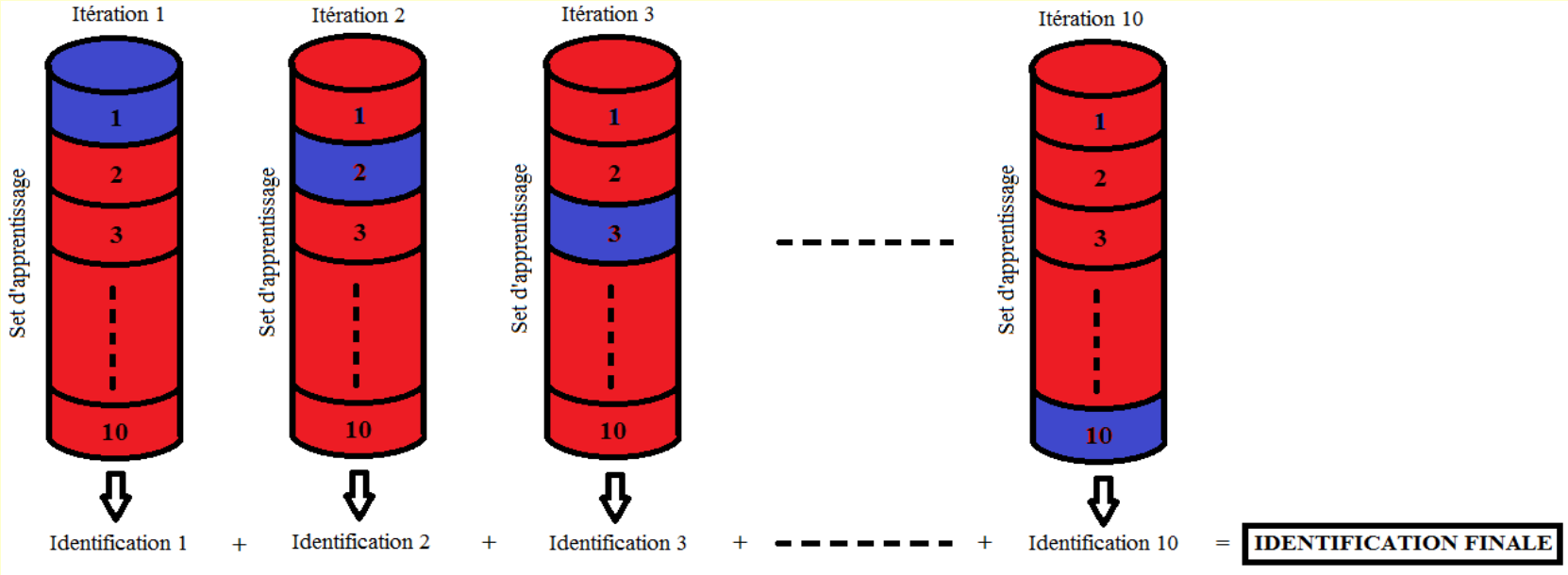


Concentrations moyennes obtenues (sur 3 passages) pour les assemblages 4X/300µm et 10X/100µm

Sans agitation Avec agitation

Annexe 3

Processus de validation croisée (performance de reconnaissance)



- Apprentissage effectué sur 9/10ème du set
- Reconnaissance effectuée sur le 1/10ème restant

→ Opération réalisée 10 fois

(chaque dixième du set est utilisé une fois pour la phase de reconnaissance).

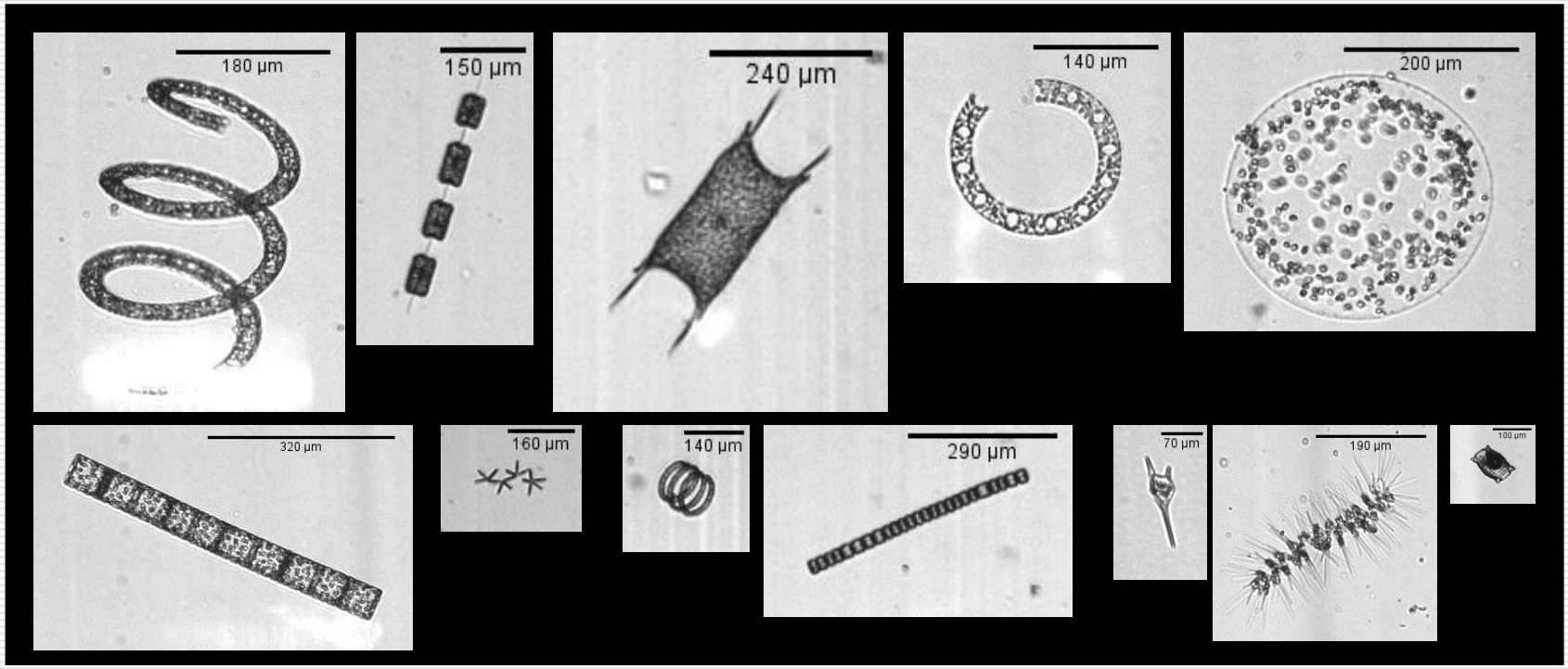
Annexe 5

Diapositives
Zoo/PhytoImage version 5
Simplification de l'interface pour utilisation REPHY en routine

Journées REPHY 2014
Nantes

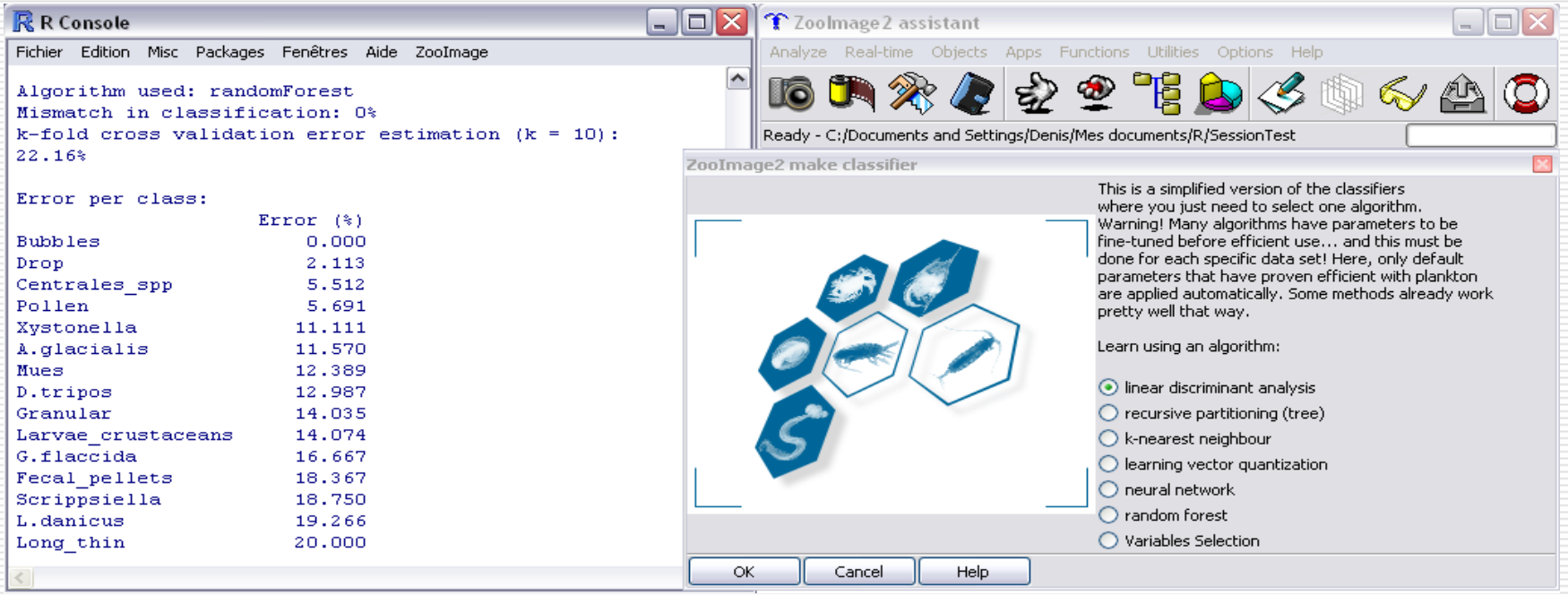
Zoo/PhytoImage version 5

Simplification de l'interface pour utilisation REPHY en routine



Philippe Grosjean, EcoNum lab, UMONS, Belgique

Qu'est-ce que Zoo/PhytoImage ?

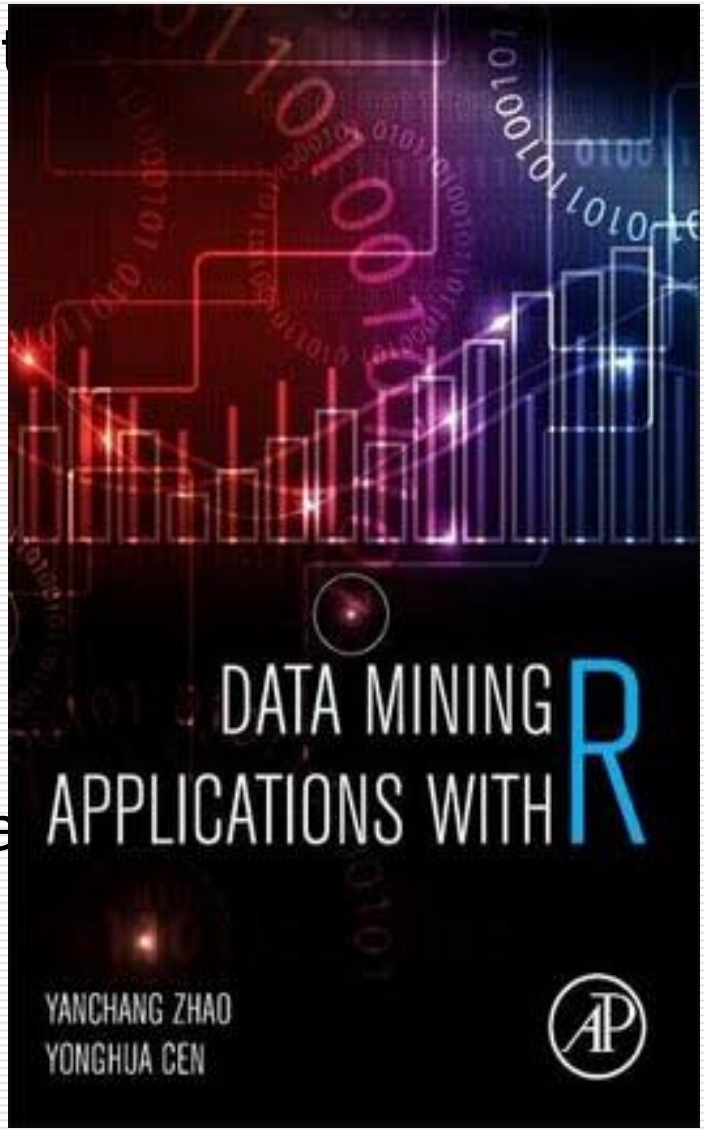


- Logiciel libre (open source), écrit en R et Java et spécialisé dans la classification d'images numériques de zoo- et phytoplancton
- Classification supervisée ("machine learning")
- Adaptable pour l'analyse de *tout type* d'image de plancton, e.g., images provenant du FlowCAM, micro- ou macrophotographies, ...

Version 3 du logiciel

Voir : Data mining application wit
ISBN 978-0124115118,
Décembre 2013.
Academic Press, Elsevier.

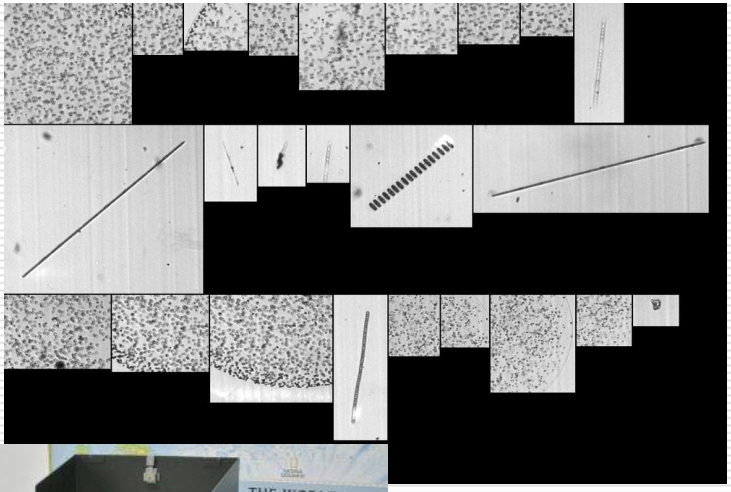
Le chapitre 12 couvre une description complète et un Tutorial de la version 3 de Zoo/PhytoImage

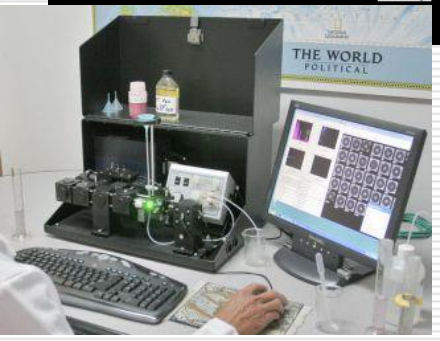


Supervised classification of image applied to plankton samples using R and zooimage.
Ph. Grosjean & K. Denis

Images issues du FlowCAM: le point de départ

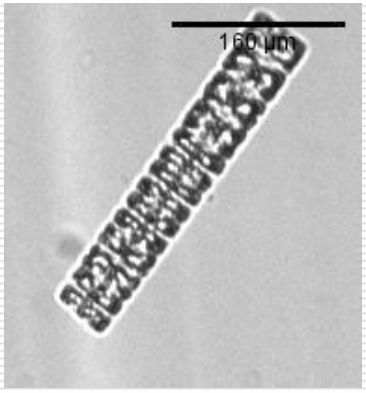
Image du FlowCAM





Analyse d'image

Vignettes (pour l'identification manuelle)



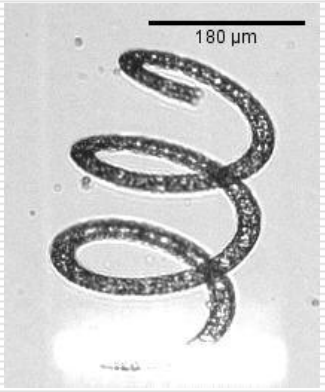


Table de mesures (attributs)

Résultats typiques - Arcachon

Global error : 26 %

Error per class:

	Error (%)		Error (%)
Drop	7	Lauderia_Schroederella	24
Pollen	7	Thalassiosira_spp	25
C.fusus	8	Fecal_pellets	25
L.danicus	9	Long_thin	26
L.undulatum	12	Dark	32
Thalassiosira_spp_cells	13	D.fragilissimus	33
G.flaccida	14	D.brightwelli	35
Black_opaque	15	Aggregates	36
D.tripos	17	P.alata_indica	40
Centrales_spp	17	G.delicatula	46
Mues	18	Membranous	48
G.striata	18	R.imbricata_styli	50
Pseudo-nitzschia	19	Fibers	50
Euglenophyceae	20	Dictyochophyceae	53
T.subtilis	20	Protoperidinum_spp	68
Short_thin	22	Chaetoceros_spp	69
Clear	23	Larvae_crustaceans	71
N.longissima_Cylindrotheca	23	Ceratium_spp	79
A.glacialis	23	P.alata	93
Bubbles	23	C.danicus	95
Granular	24	C.decipiens	97

- **< A. Tuning-Ley & D. Maurer**
- > 40 groupes
- Très bons résultats pour la moitié des groupes
- Plus difficile pour les groupes plus rares

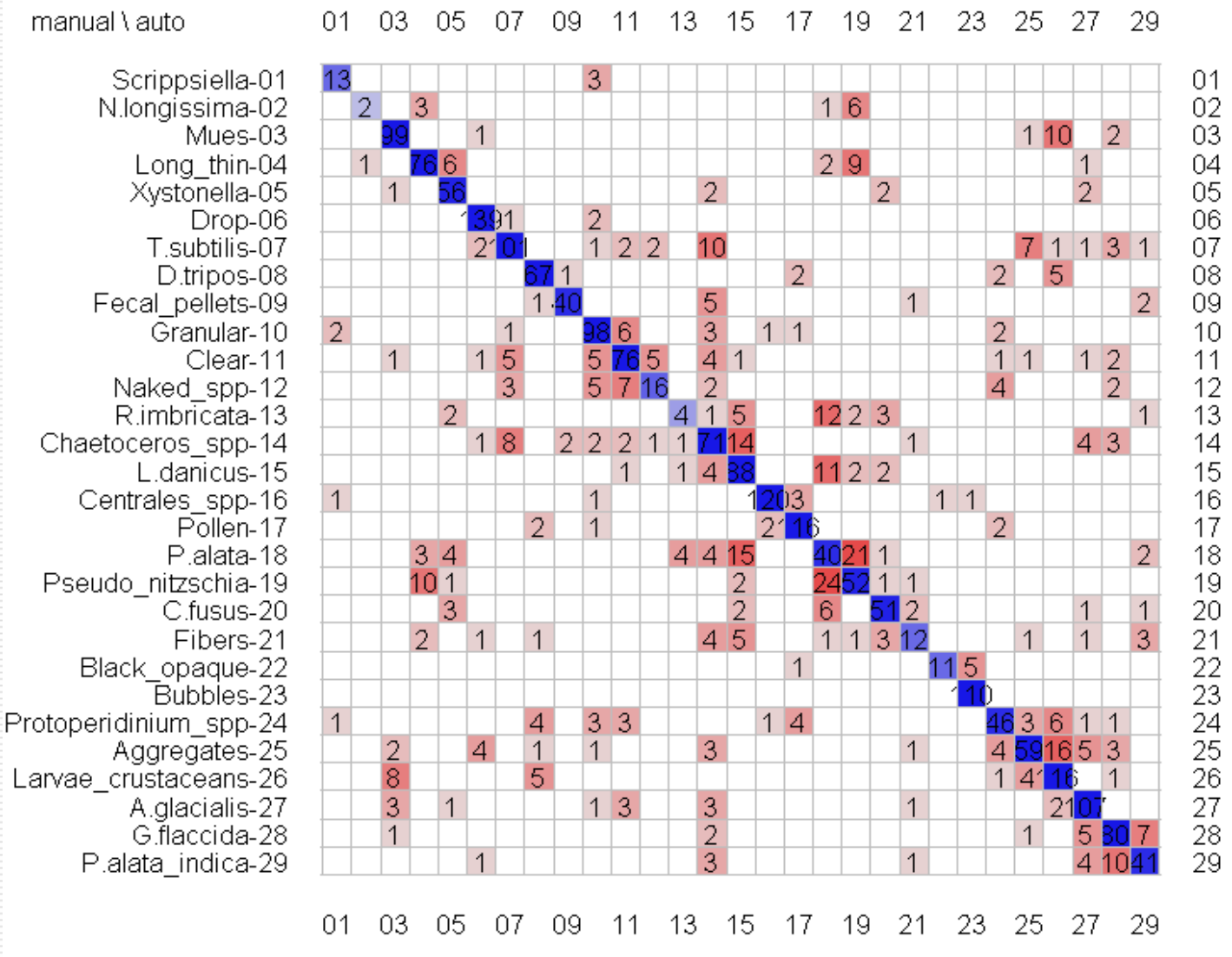
Validation nécessaire pour les groupes rares

Accélération de la validation des échantillons :

validation des suspects et
correction statistique de l'erreur

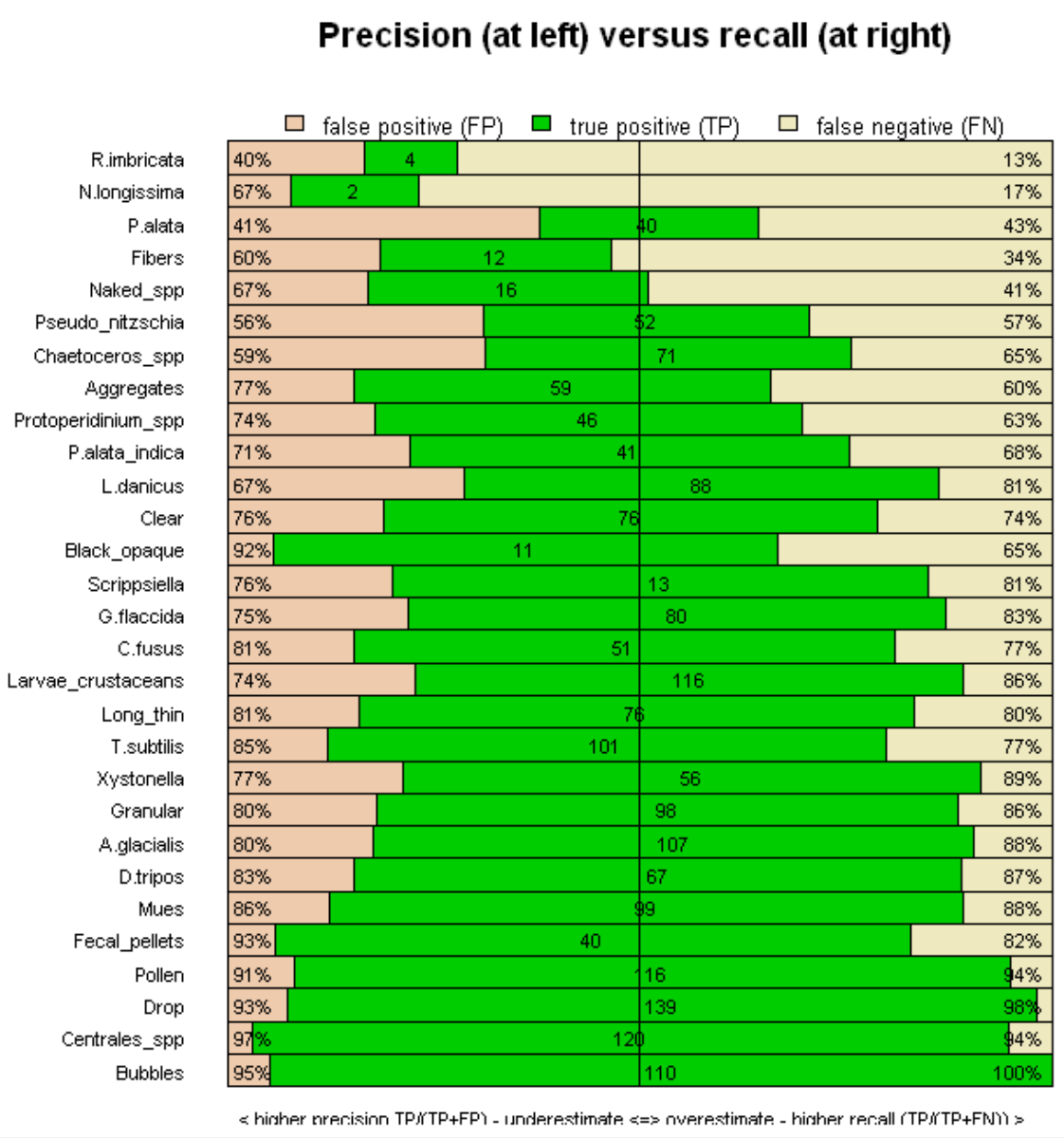
Diagnostique

Matrice de confusion

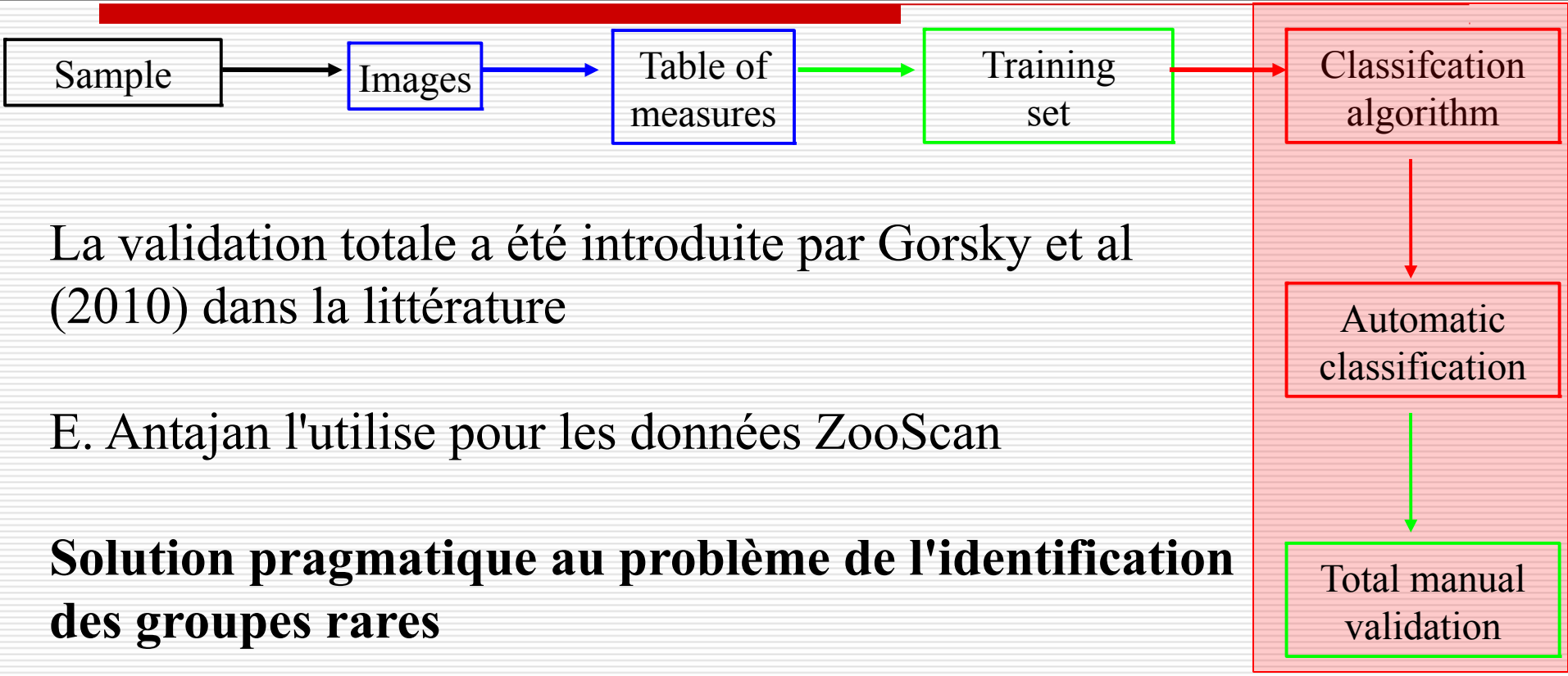


Diagnostique

Graphique « Precision-recall » pour une meilleure visualisation des Performances par classe



Correction de l'erreur (« validation totale »)



La validation totale a été introduite par Gorsky et al (2010) dans la littérature

E. Antajan l'utilise pour les données ZooScan

Solution pragmatique au problème de l'identification des groupes rares

Détection des suspects avec un GLM

$$\text{logit}(\text{Suspect}) = \text{ProbaDiff} + \text{FPDiff} + \text{Bio} + \ldots + \varepsilon$$

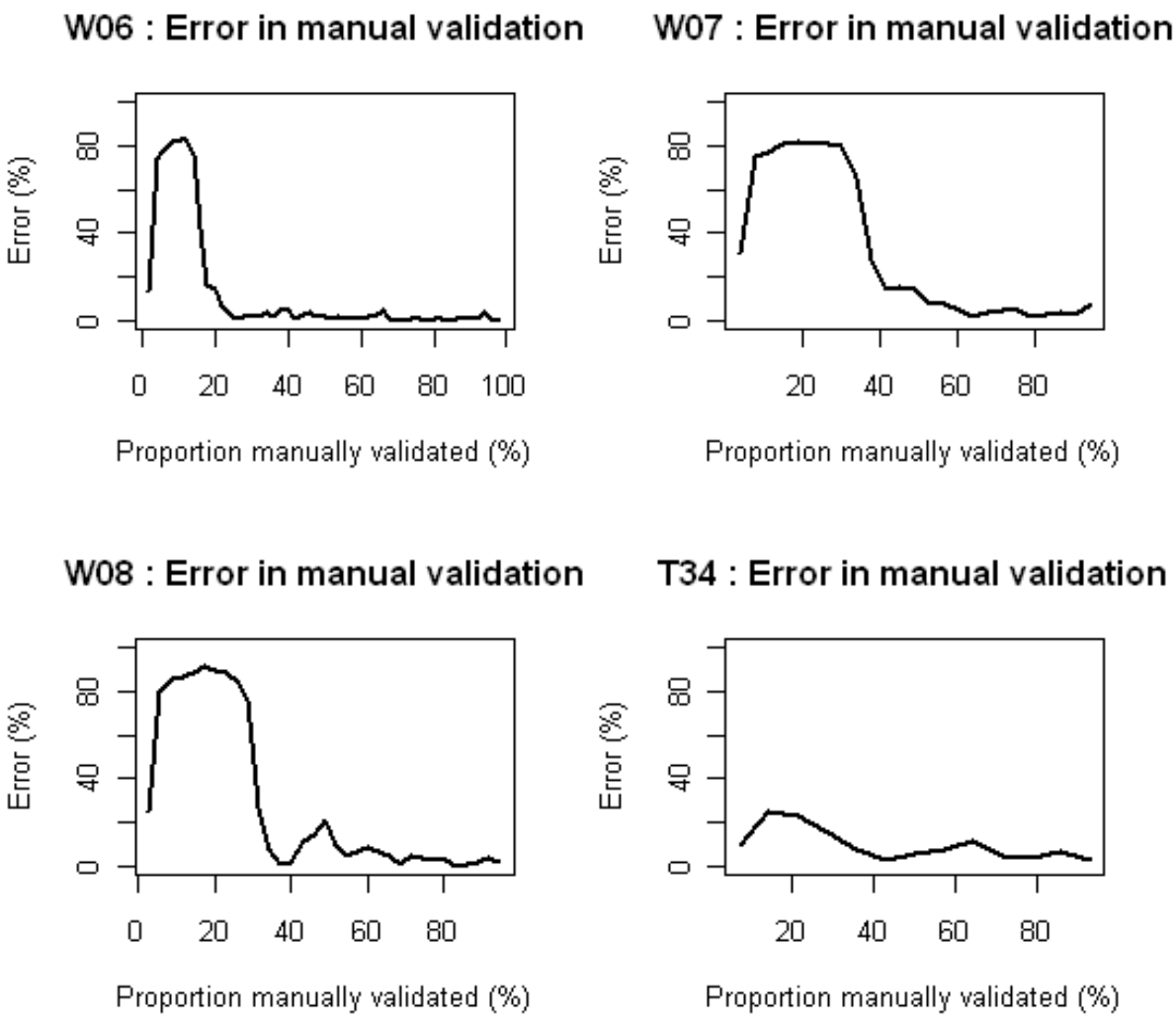
ProbaDiff : probabilités renvoyées par l'outil de classification lyi-même – *critère intrinsèque à la classification supervisée*

FPDiff : proportion des groupes en regard de la probabilité d'avoir des faux positifs – *critère bayésien*

Bio : « probabilité » d'occurence des espèces, information fournie par l'opérateur – *critère basé sur la biologie et l'écologie*

… autres variables

Détection des particules suspectes



Application sur 4 échantillons W06, W07, W08 (BCZ), T34 (Arcachon)

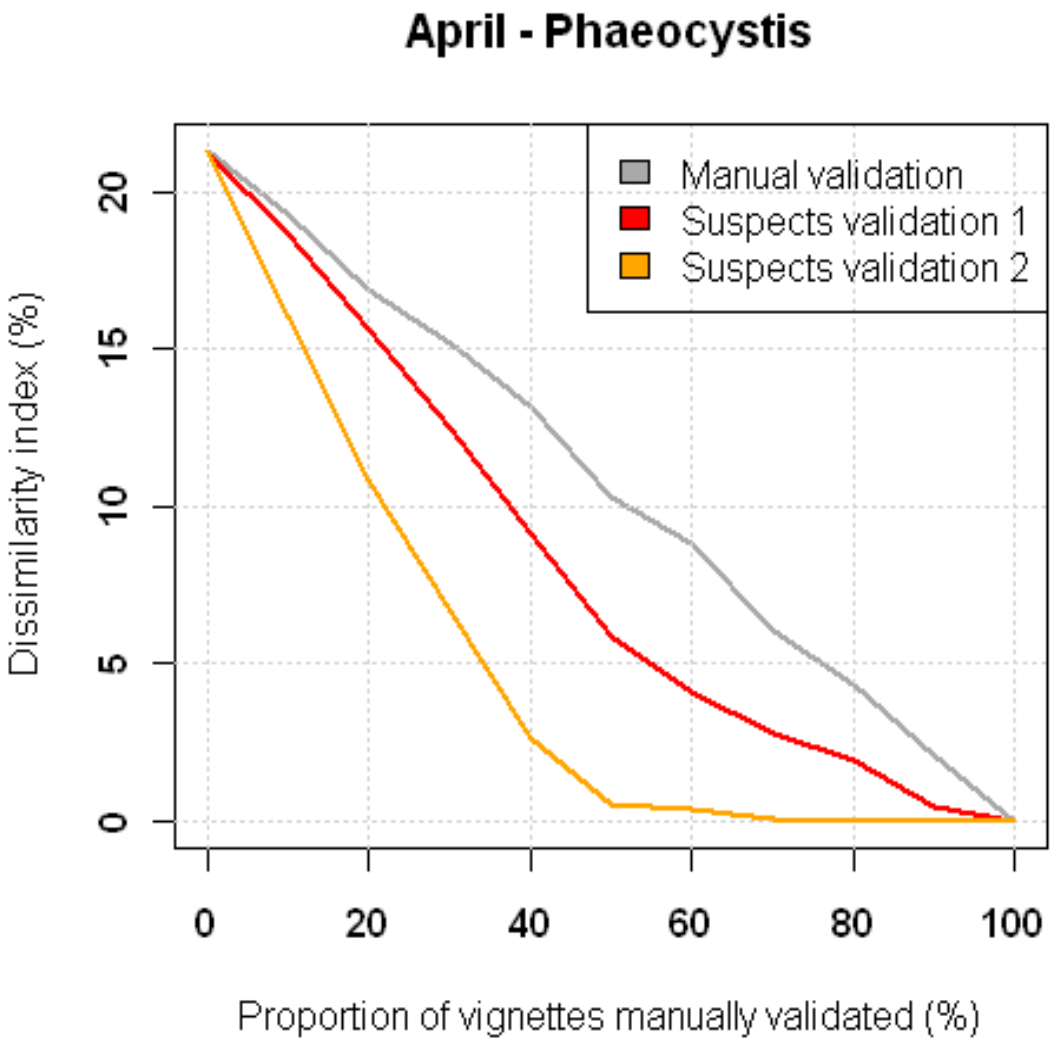
Les particules correspondent à la grande majorité des classements erronés !

Possibilité d'optimiser la validation manuelle en se concentrant sur ces suspects.

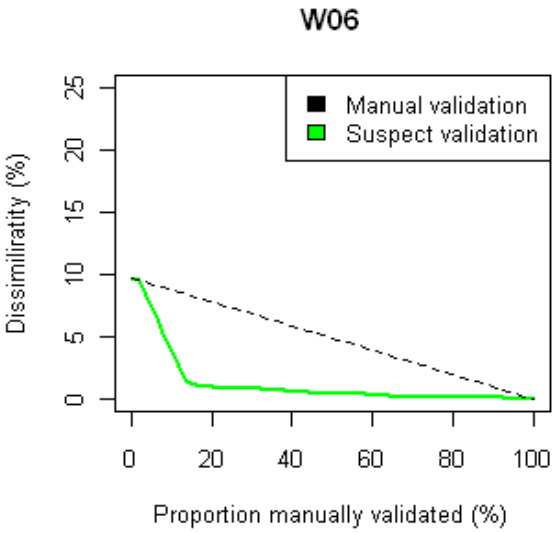
Impact de l'information biologique sur la validation

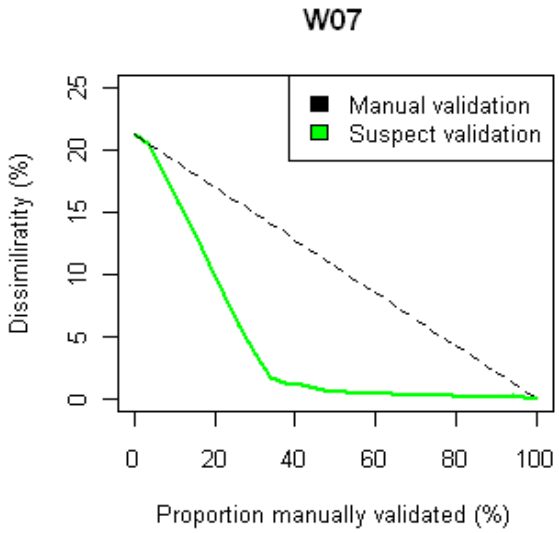
La détection des suspects est grandement améliorée en utilisant des critères biologiques et écologiques !

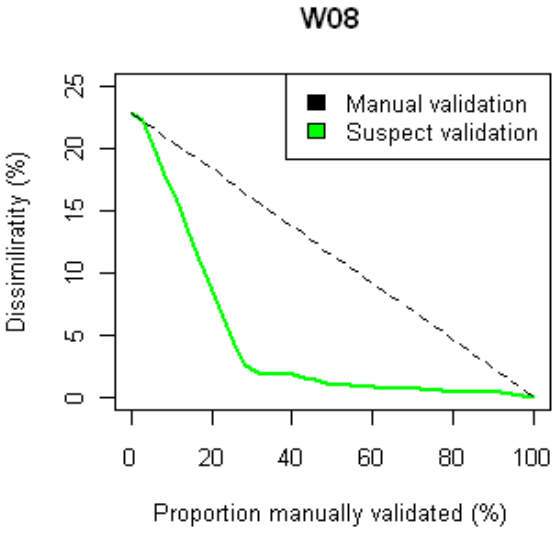
e.g., une espèce est-elle probable à cet endroit ? à ce moment ? A cette température ? ...

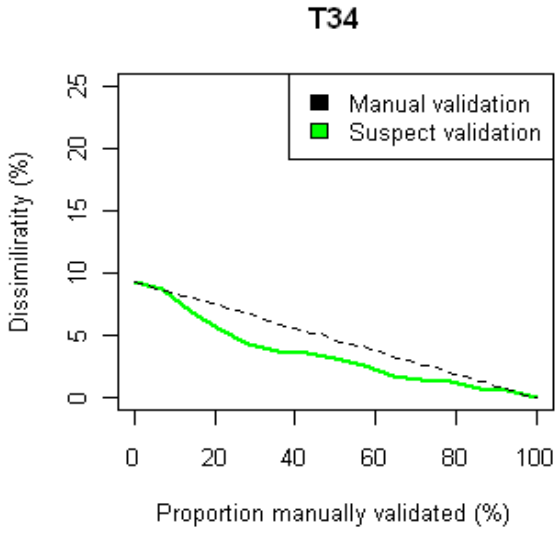


Validation of the suspects





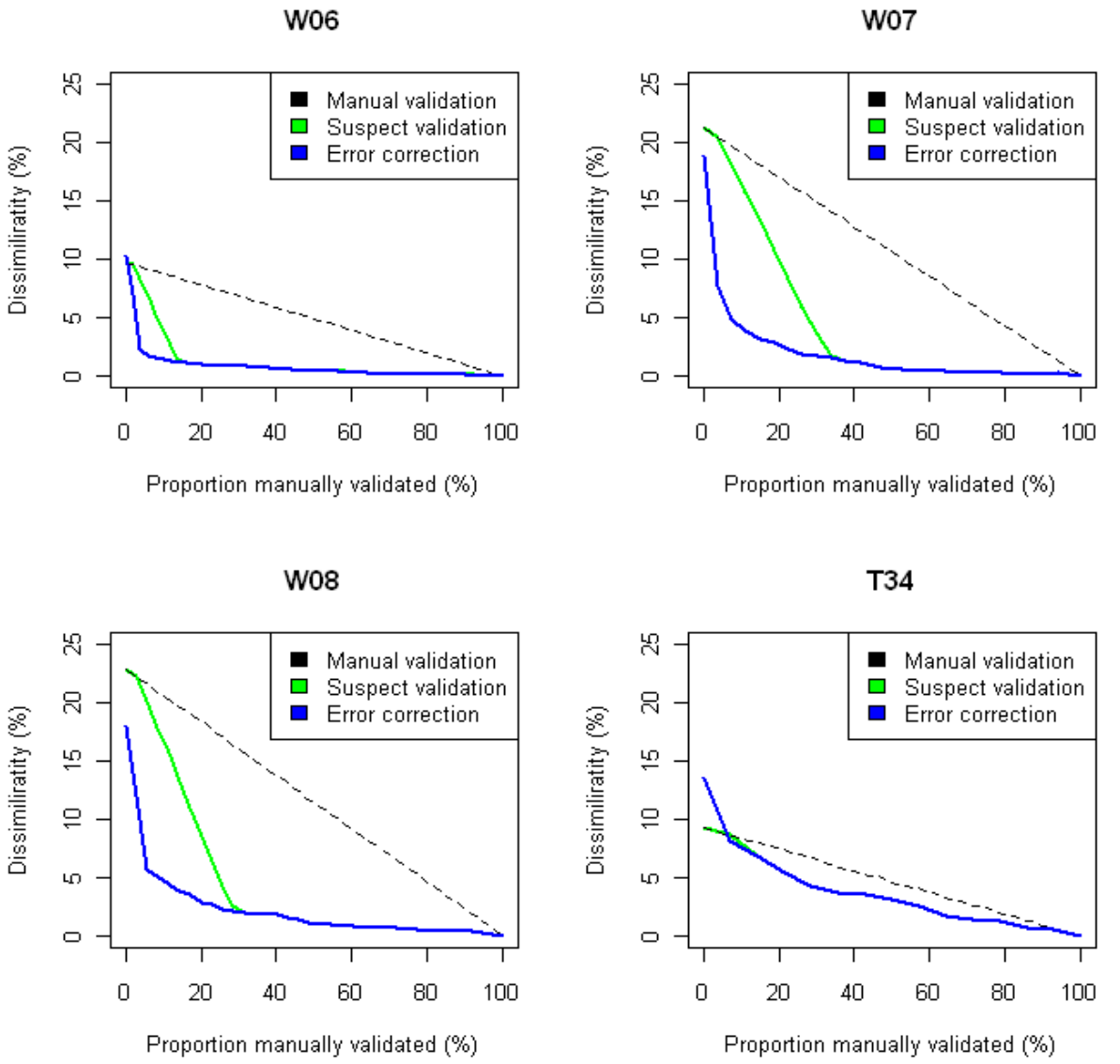




Décroissance linéaire de l'erreur via la validation manuelle

Décroissance plus rapide de l'erreur résiduelle par la validation des suspects

Correction statistique de l'erreur



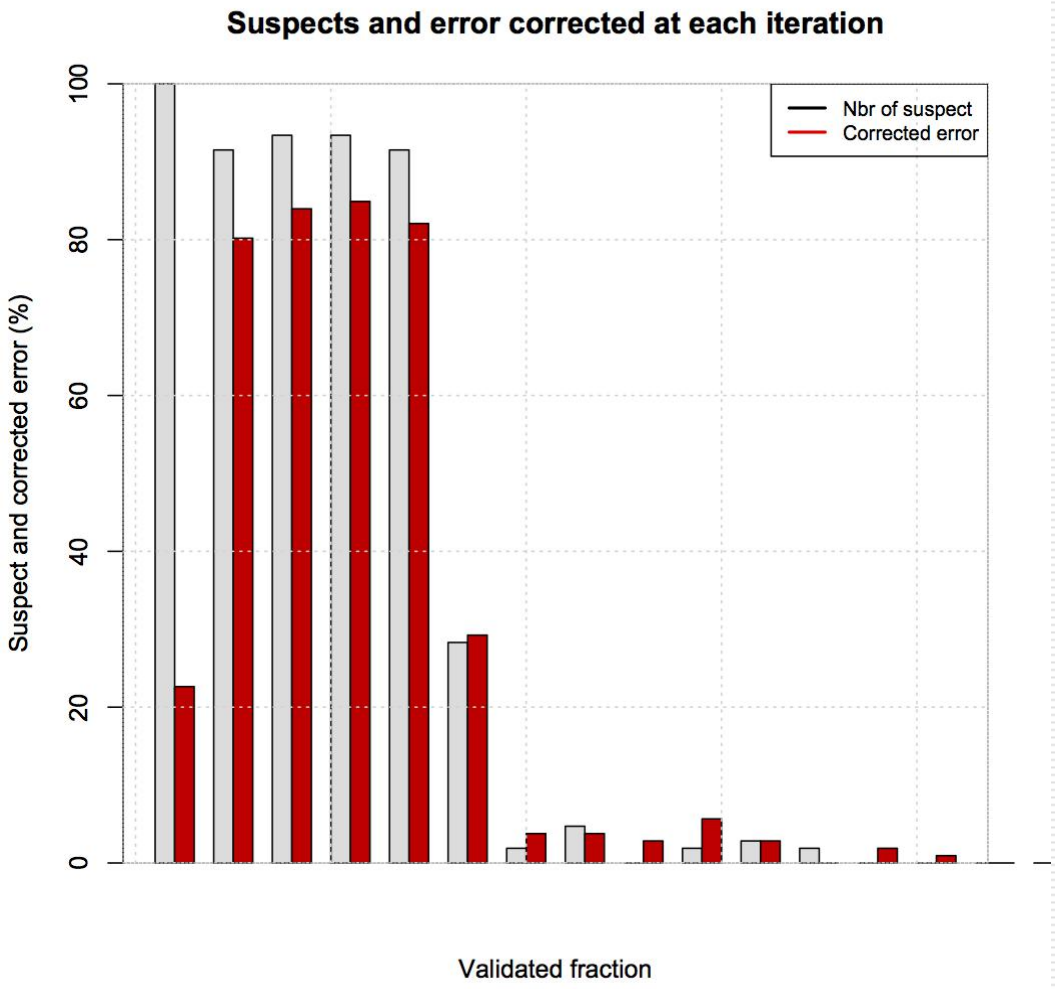
Une étape plus loin : associer à la détection des suspects, un modèle de correction statistique de l'erreur

Moins de 5% d'erreur résiduelle après validation manuelle d'environ seulement 10% de l'échantillon

Zoo/PhytoImage version 5

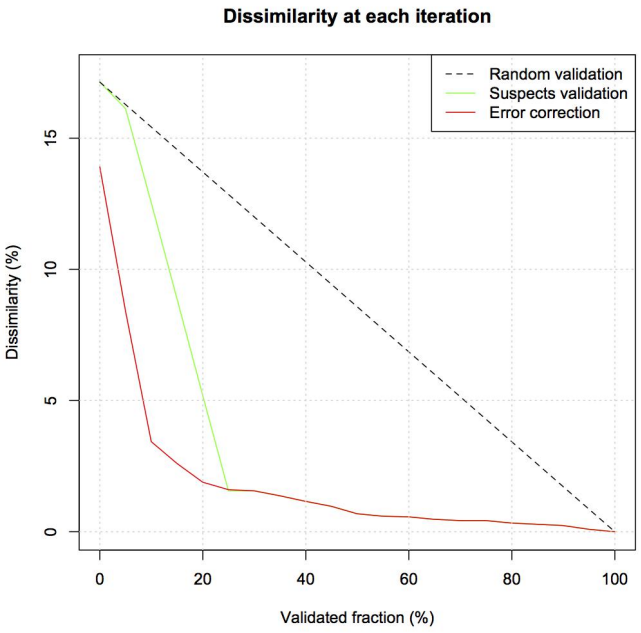
Interface utilisateur simplifiée

Validation étape par étape - démo



Barres en gris = suspects

Barre en rouge error = erreur constatée après validation



Pistes pour le futur ?

Apprentissage actif

Apprentissage actif

- Actuellement, on réalise un gros set d'apprentissage qui mène à un outil de classification figé

- Travail lourd sur au moins un an pour reprendre les variations saisonnières

- A moduler par zone géographique

- L'apprentissage actif utilise les données validées en routine pour enrichir le set d'apprentissage

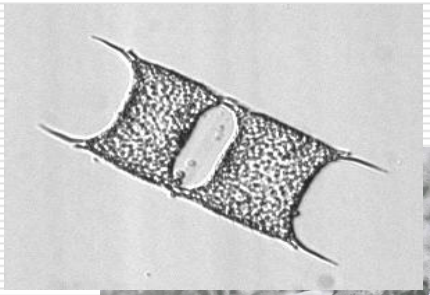
- *Adaptation (géographique, temporelle et saisonnière) du set d'apprentissage, de manière transparente !*

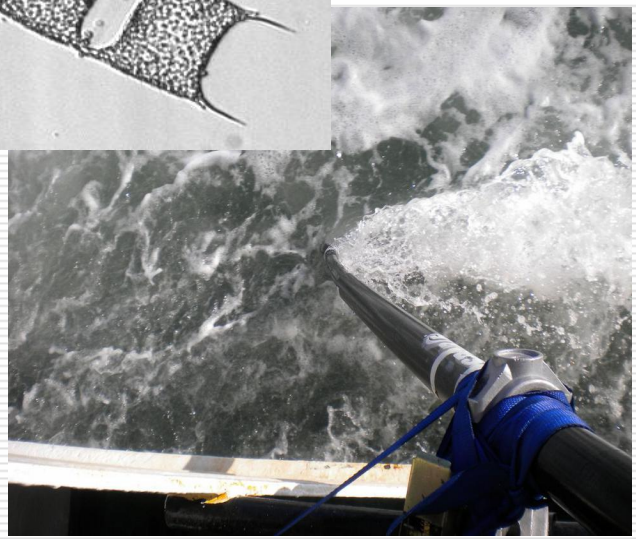
Pistes pour le futur ?

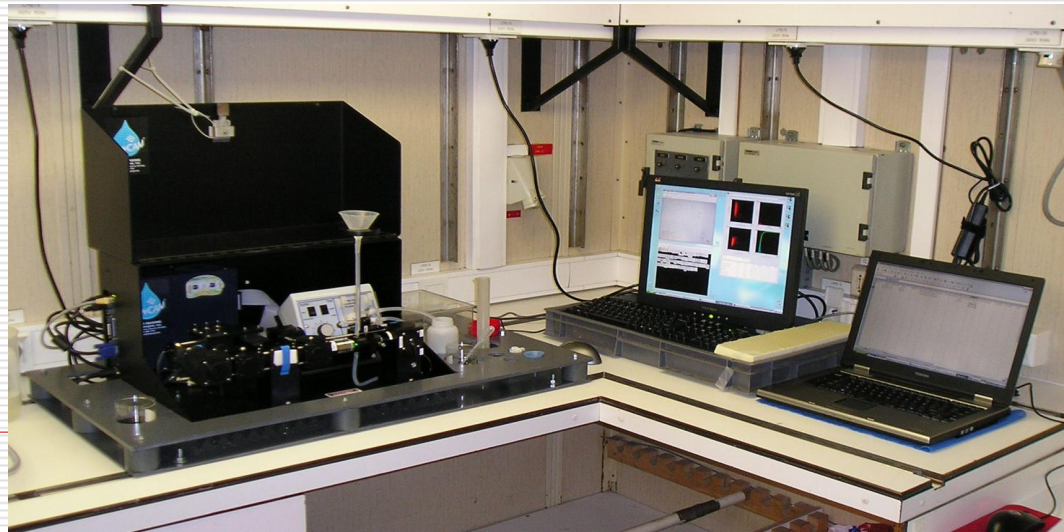
Analyse des données en temps réel sur un navire océanographique

Application en temps réel

- Tests à bord du 'Belgica'
- Pompage via une perche renforcée de fibres de carbone (jusqu'à to 6-7 beaufort)
- 25 groupes discriminés en temps réel (dont 18 groupes de phytoplancton)

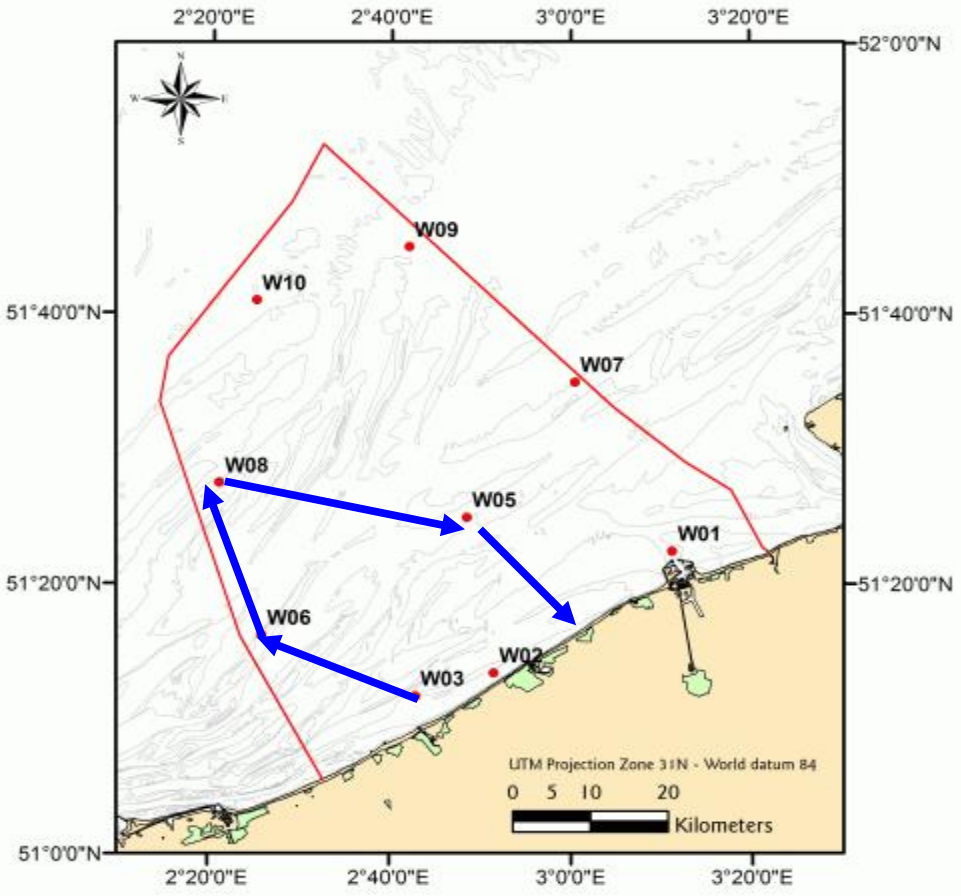






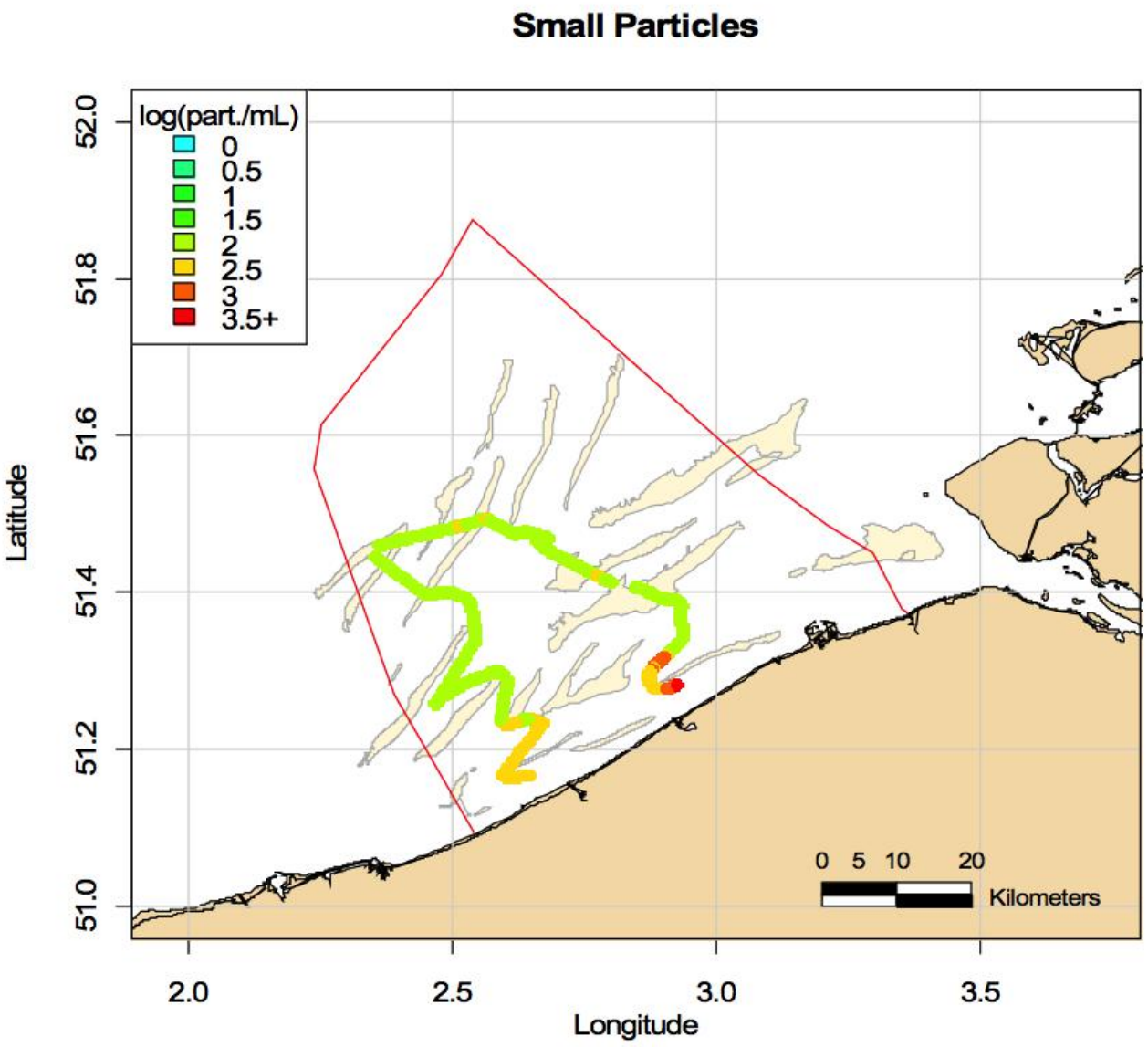
Etude temps réel du phytoplancton de Mer du Nord

Exemple d'application : campagne d'un jour dans la zone côtière belge (BCZ).



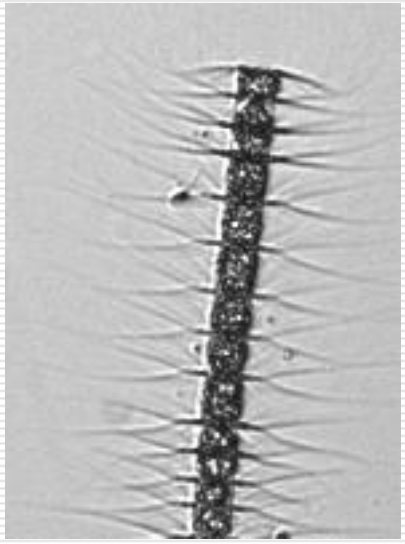
Etude en temps réel - résultats

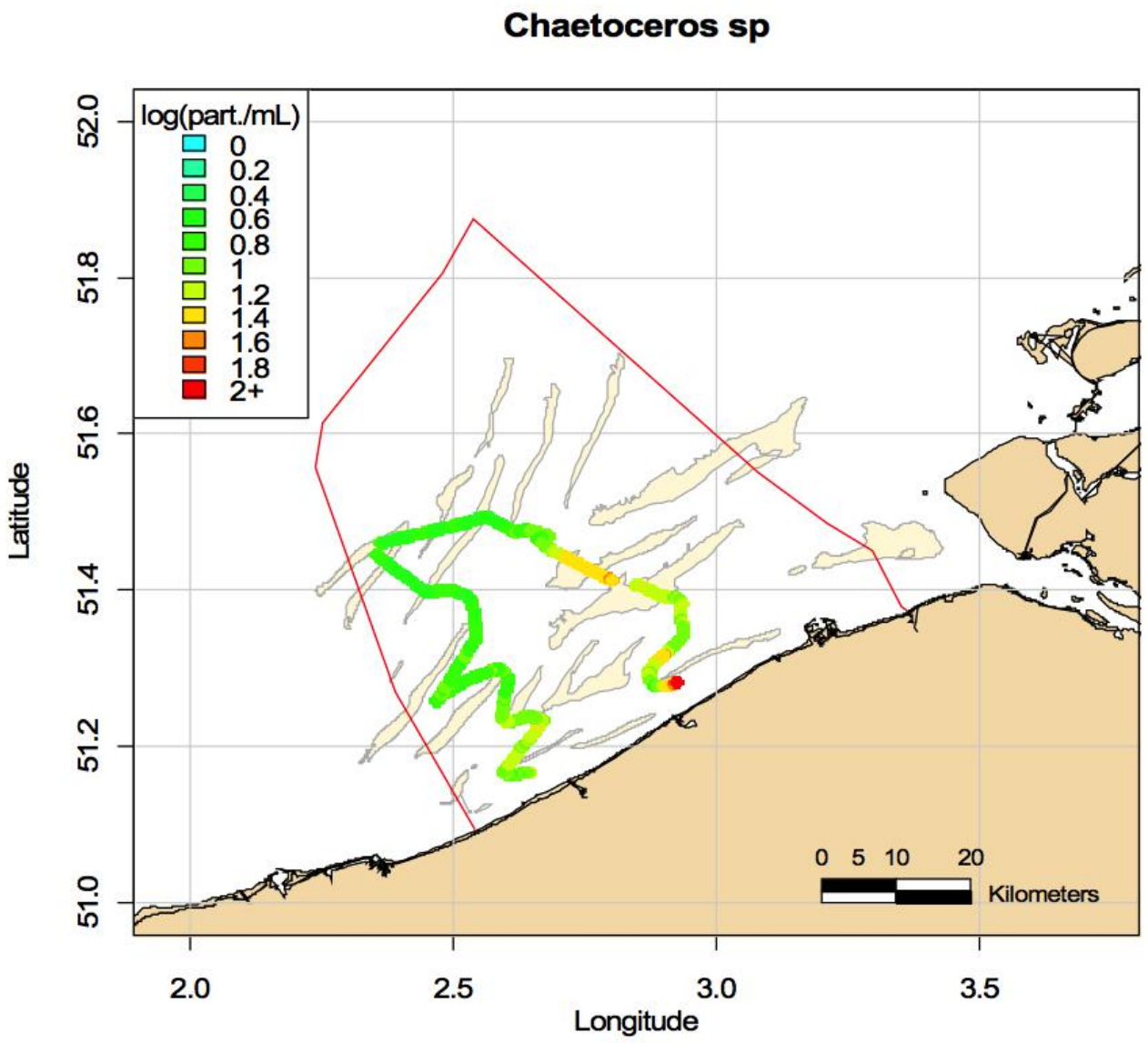
Les particules fines dominent près de l'estuaire de l'Escault.



Etude en temps réel - résultats

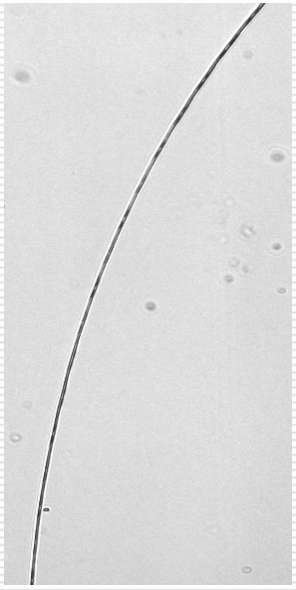
Exemple d'un groupe néritique : *Chaetoceros spp*, répartition semblable aux particules fines.

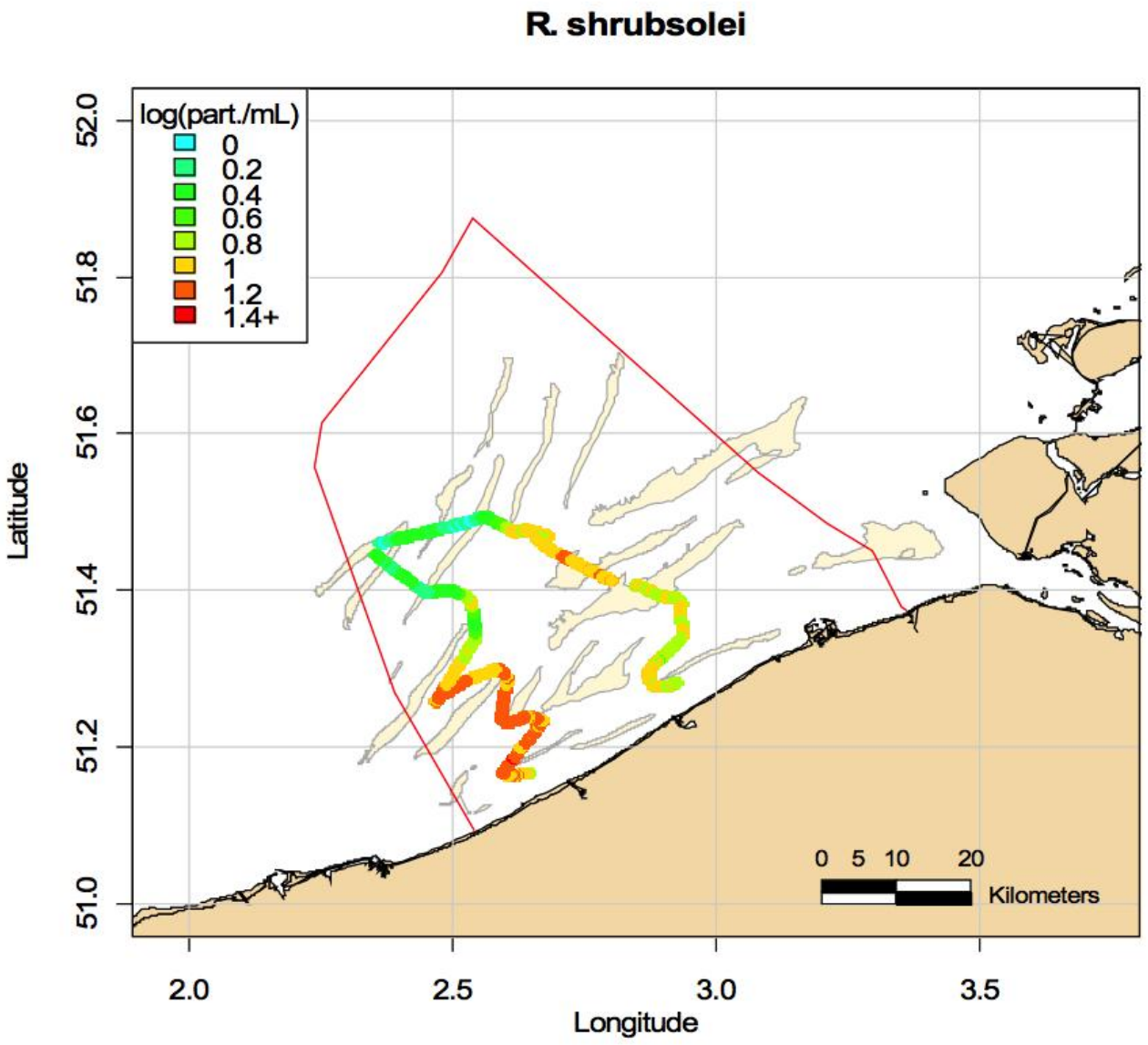




Etude en temps-réel - résultats

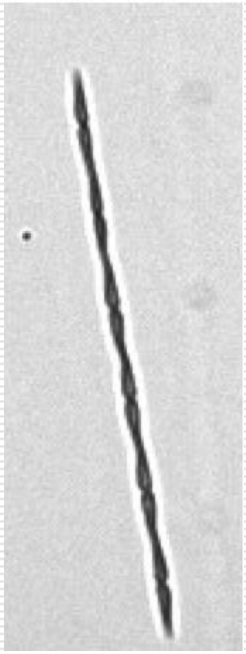
Rhizosolenia shrubsolei a bloomé en cette période, mais avec une distribution spatiale très différente.

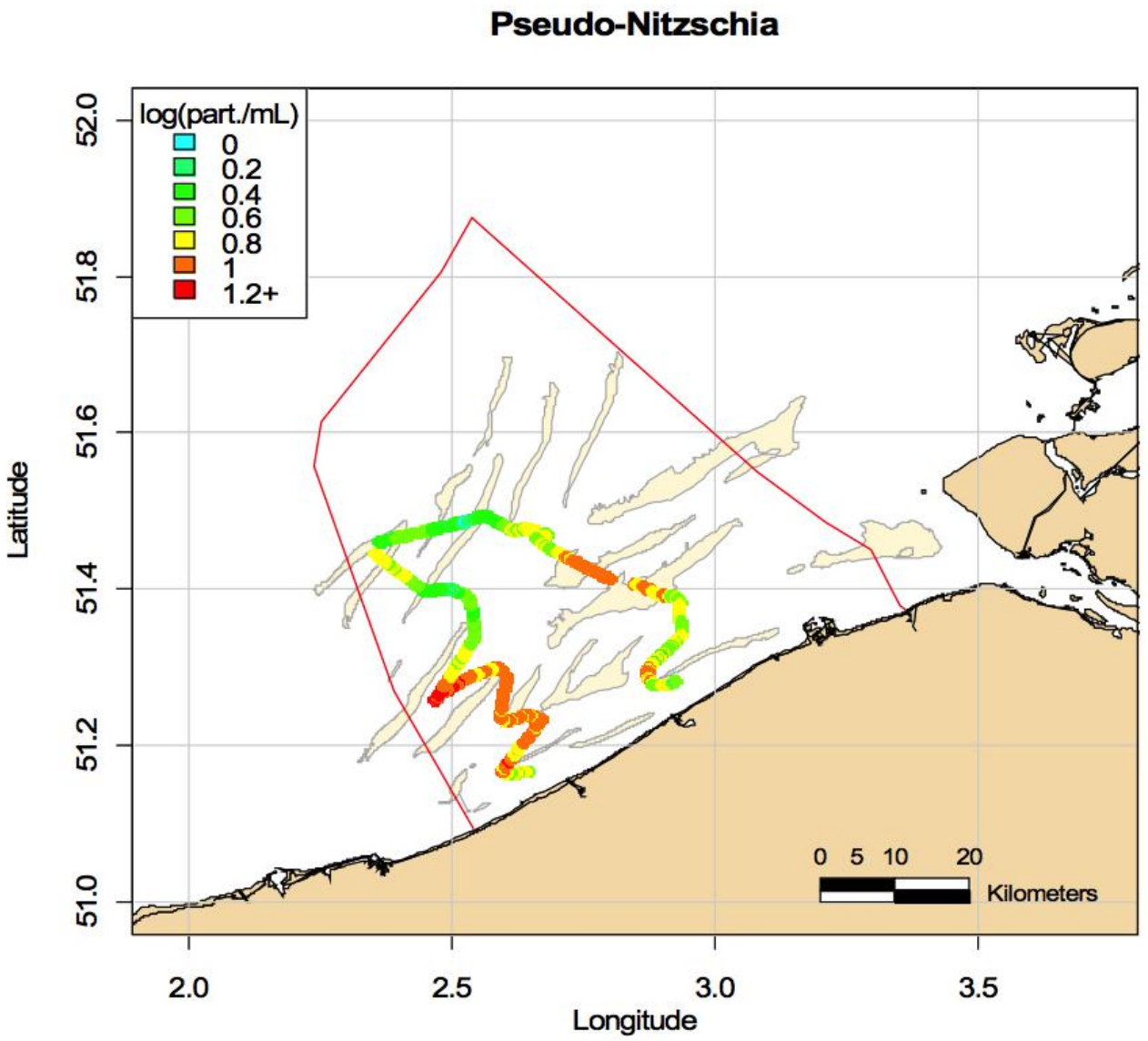




Etude en temps réel - résultats

Un troisième groupe avec une distribution encore différente : *Pseudo-nitzschia*.





Conclusions

- La version 3 de Zoo/PhytoImage offre une boite à outil exhaustive pour la manipulation d'images de plancton

- La version 4 développe une nouvelle interface simplifié ; la version 5 en est l'outil actuel pour une utilisation en routine sur le REPHY

- La détection des suspects et la correction statistique de l'erreur permet de réduire les erreurs de comptage et d'optimiser le traitement

- A l'avenir, un apprentissage actif sera installé progressivement

- Une utilisation en temps réel en mer est étudiée

- Gros travail nécessaire par rapport à Quadrige^2 !

Merci pour votre attention

Travail de collaboration entre ***IFREMER*** et ***UMONS***.

Le projet BelSpo AMORE III a aussi financé les développement pour une utilisation en temps réel.

Zoo/PhytoImage see :

http://www.sciviews.org/zooimage