

2015 - **Domaine** Outils pour la surveillance environnementale

**Action 9** – FlowCAM / ZooPhytoImage

## **Optimisation de l'identification et du dénombrement du microphytoplancton avec le système couplé de numérisation et d'analyse d'images FlowCAM – Zoo/PhytoImage (système innovant)**

**Action 9 – Livrable 1. 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**

**Rapport final, février 2016**

**Philippe GROSJEAN (Université de Mons)**

**Guillaume WACQUET (Université de Mons)**

Février 2016

## AUTEURS

**Philippe GROSJEAN, professeur** (Université de Mons), [Philippe.Grosjean@umons.ac.be](mailto:Philippe.Grosjean@umons.ac.be)

**Guillaume WACQUET, assistant de recherche** (Université de Mons),  
[Guillaume.Wacquet@umons.ac.be](mailto:Guillaume.Wacquet@umons.ac.be)

## CORRESPONDANTS

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

**Ifremer : Catherine BELIN** (Ifremer), [catherine.belin@ifremer.fr](mailto:catherine.belin@ifremer.fr)

## AUTRES CONTRIBUTEURS

**Nadine NAUD-MASSON** (Ifremer), [nadine.masson@ifremer.fr](mailto:nadine.masson@ifremer.fr)

**Alain LEFEBVRE** (Ifremer), [alain.lefebvre@ifremer.fr](mailto:alain.lefebvre@ifremer.fr)

**Rémi CUVELLIEZ** (stagiaire Ifremer)

**Catherine BELIN** (Ifremer), [catherine.belin@ifremer.fr](mailto:catherine.belin@ifremer.fr)

**Danièle MAURER** (Ifremer), [daniele.maurer@ifremer.fr](mailto:daniele.maurer@ifremer.fr)

**Luis Felipe ARTIGAS (ULCO)**, [felipe.artigas@univ-littoral.fr](mailto:felipe.artigas@univ-littoral.fr)

**Denis HAMAD (ULCO)**, [denis.hamad@lisc.univ-littoral.fr](mailto:denis.hamad@lisc.univ-littoral.fr)

**Claire METEIGNER** (Ifremer), [claire.meteigner@ifremer.fr](mailto:claire.meteigner@ifremer.fr)

**Myriam PERRIERE-RUMEBE** (Ifremer), [myriam.rumebe@ifremer.fr](mailto:myriam.rumebe@ifremer.fr)

**Pascale HEBERT** (Ifremer), [pascale.hebert@ifremer.fr](mailto:pascale.hebert@ifremer.fr)

**Camille BLONDEL** (Ifremer), [camille.blondel@ifremer.fr](mailto:camille.blondel@ifremer.fr)

Droits d'usage : **libre accès**

Niveau géographique : **national**

Couverture géographique : **nationale**

Niveau de lecture : **experts**

## RESUME

Ce livrable détaille les travaux réalisés pour l'évolution du logiciel Zoo/PhytoImage afin d'en optimiser l'usage dans le cadre de l'étude d'échantillons de phytoplancton de manière générale, et dans le cadre de son utilisation opérationnelle pour le monitoring des eaux côtières tel que réalisé par le REPHY à l'Ifremer en particulier. Zoo/PhytoImage permet d'analyser des échantillons de plancton fixés numériquement, c'est-à-dire, sur base d'images obtenues à l'aide d'un appareil spécialisé comme le FlowCAM ou le FastCAM (voir livrable n°3). La classification supervisée (machine learning) permet de classer de manière automatique les particules imagées dans les différents groupes taxonomiques, et d'en dériver ensuite des statistiques sur l'échantillon tout entier : dénombrement, biomasse et spectre de tailles par groupe taxonomique.

Deux changements majeurs ont été introduits en 2015 dans les calculs réalisés par Zoo/PhytoImage :

- **le dénombrement des cellules par colonies** qui permet d'exprimer les données par cellules, là où les résultats étaient limités aux colonies dans les versions antérieures,
- **l'apprentissage actif** qui permet un travail optimisé lors de la validation de la classification effectuée par l'ordinateur. En effet, au cours de la validation des résultats, le logiciel va apprendre à corriger l'erreur au fur et à mesure que le taxonomiste effectue manuellement ce travail de correction. Le logiciel va donc accompagner de manière dynamique l'utilisateur dans son travail de vérification et de correction. Il en résulte une diminution importante dans la quantité de données qui doit être vérifiée par l'utilisateur pour atteindre un objectif donné (par exemple, moins de 5 % d'erreur résiduelle dans chaque groupe taxonomique).

**Une optimisation de la procédure de validation** a également été développée. La validation des échantillons fait maintenant intervenir des outils statistiques poussés, puisque : (i) la sélection des particules que l'utilisateur vérifie est déterminée d'une manière probabiliste (détection des suspects), (ii) un second classifieur est utilisé pour seconder l'utilisateur dans la validation (apprentissage actif), (iii) enfin, un algorithme de correction statistique de l'erreur est implémenté sur les résultats finaux. L'ensemble permet de réduire le travail de validation. Cela signifie que le taxonomiste peut arrêter plus tôt de visualiser et corriger manuellement l'attribution des groupes taxonomiques aux particules suspectes. Un tableau de bord présente de manière claire et graphique l'avancement du travail de validation et anticipe le nombre d'étapes nécessaire pour atteindre un objectif donné (par exemple, moins de 5 % d'erreur dans tous les groupes taxinomiques). Toutes ces mesures contribuent à la fois au confort et l'efficacité de la validation.

Le livrable est composé de trois rapports complémentaires :

- Les deux premiers (**rapports d'avancement mars et juin 2015, respectivement pages 10 à 69 et 70 à 88 du document pdf**) détaillent la progression dans le dénombrement des cellules par colonies et dans l'implémentation de l'apprentissage actif. Le premier des deux explore des pistes, et le second concrétise les solutions retenues et les teste sur des échantillons plus larges issus du REPHY.
- Le troisième rapport est une version du **manuel utilisateur de Zoo/PhytoImage en français (pages 89 à 120 du document pdf)** qui se focalise sur l'analyse de données de type FlowCAM (et FastCAM) en y incluant les deux nouvelles fonctionnalités décrites ci-dessus. Bien entendu, associé à ce manuel, le code est disponible et matérialise le travail réalisé sous forme d'une nouvelle version 5.4.0 de Zoo/PhytoImage. Une version 6.0 sera rendue publique à la fin de la phase de beta-test : elle sera très proche de la version 5.4.0 décrite ici.

## **MOTS CLES (THEMATIQUE ET GEOGRAPHIQUE)**

Phytoplancton, REPHY, analyse d'image, classification supervisée, dénombrement de cellules, apprentissage actif, Manche, Atlantique.

## TITLE

Evolutionary 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

This report details the work accomplished to enhance the Zoo/PhytoImage software to optimize its use for the analysis of phytoplankton samples in general, but more particularly, in the framework of an operational survey of coastal seawater (REPHY, IFREMER). Zoo/PhytoImage allows to analyze “numerically recorded” plankton samples, that is, by using digital images gathered with specialized devices such as the FlowCAM, or the FastCAM (see report 3). A machine learning approach allows to automatically classify the digitized particles into various taxonomic groups. Once this is done, global statistics are calculated on each sample, including the number of particles, the biomass, and the size spectrum per taxonomic group.

Two major changes are introduced in the calculations done by Zoo/PhytoImage:

- the enumeration of the cells per colonies allows to express results per cell, where previous versions of the software only expressed results per colonies, and
- an active learning algorithm is implemented in order to optimize the validation step (manual correction of the residual error after automatic classification of the particles by the computer). This way, the software will learn how to perform that correction based on the validations done so far by the taxonomist. Consequently, the computer now assists dynamically the user in the verification and correction procedure. The number of items that the user has to manually check is thus greatly reduced in order to reach a given goal (say, less than 5% of residual error in each taxonomic group).

The report is made of three complementary parts:

- the first two sections (advance reports on March and June 2015) detail the progression in the cells per colonies enumeration and in the implementation of active learning. The former section explores ideas, whereas the latter one finalizes the best solutions and tests them on actual samples from the REPHY.
- the third section is a new French version of the Zoo/PhytoImage user manual that focuses on the analysis of plankton images from the FlowCAM (and the FastCAM), including the new functionalities. Of course, this user manual comes with the code of the new version 5.4.0 of Zoo/PhytoImage that fully implements these new features. A version 6.0 will be made public shortly, at the end of the beta test period: it will be very close to the version 5.4.0 described in the present report.

An optimization of the validation procedure was also developed. The validation of samples is now performed with complex statistical tools: (i) the selection of particles which will be checked by the technician is determined in a probabilistic way (detection of suspects), (ii) a second classification is used to assist the user in validation (active learning), (iii) finally, a statistical correction algorithm of error is implemented on the results. This reduces the validation work. This means that the taxonomist may stop earlier and manually correct the allocation of taxonomic groups to suspicious particles. The progress of the validation work is clearly presented and the number of steps necessary to achieve a given objective (eg, less than 5% error in all taxonomic groups) is anticipated. All these measures contribute to both comfort and efficiency of validation.

## **KEY WORDS (THEMATIC AND GEOGRAPHICAL AREA)**

Phytoplankton, REPHY, image analysis, Machine learning, cells enumeration, active learning, The Channel, Atlantic Ocean

## SYNTHESE POUR L'ACTION OPERATIONNELLE

L'analyse d'échantillons phytoplanctoniques est traditionnellement associée à de longues et fastidieuses séances de comptage des particules fixées de plancton sous microscope. Bien que cette image du planctonologue restera probablement pendant un certain temps, il semble y avoir une autre façon de recueillir des données sur le plancton : l'analyse assistée par ordinateur d'images numériques de ce plancton. Toute une gamme de matériel pour prendre des photos de nos organismes, à la fois *in situ* et/ou à partir d'échantillons fixés, est maintenant disponible : FlowCAM, OPC laser, VPR, Zooscan, ... (plus, à venir, FastCAM, 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 plancton 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 plancton provenant d'origines diverses et les transformer en une table de mesures utilisables. La classification supervisée (machine learning) permet de classer de manière automatique les particules imagées dans les différents groupes taxonomiques, et d'en dériver ensuite des statistiques sur l'échantillon tout entier (c'est-à-dire, les abondances, les spectres de taille totaux et partiels, les biomasses totales et partielles, etc.). 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! Zoo/PhytoImage est basé sur ImageJ et R, et fonctionne sur Linux, mais il peut aussi être exécuté sur Windows ou Mac OS X.

La conception générale de Zoo/PhytoImage est conçue de manière à ce que le logiciel soit capable de traiter efficacement des images de caractéristiques et d'origines diverses. Dans le cadre de cette action, l'accent est mis sur son couplage avec le FlowCAM pour l'analyse d'échantillons phytoplanctoniques issus du REPHY. 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, il restait encore à paramétrer différents modèles utilisés par le logiciel dans le contexte particulier du REPHY. Différents axes ont été proposés par l'UMONS et l'IFREMER : (i) dénombrement des cellules par colonie et (ii) apprentissage. Une optimisation de la procédure de validation a également été développée. Tous ces travaux sont détaillés plus bas.

### Contenu du livrable

Le livrable est composé de trois rapports complémentaires :

- les deux premiers (rapports d'avancement mars et juin 2015) détaillent la progression dans le dénombrement des cellules par colonies et dans l'implémentation de l'apprentissage actif. Le premier des deux explore des pistes, et le second concrétise les solutions retenues et les teste sur des échantillons plus larges issus du REPHY
- le troisième rapport est une version du manuel utilisateur de Zoo/PhytoImage en français qui

se focalise sur l'analyse de données de type FlowCAM (et FastCAM) en y incluant les deux nouvelles fonctionnalités décrites ci-dessus. Bien entendu, associé à ce manuel, le code est disponible et matérialise le travail réalisé sous forme d'une nouvelle version 5.4.0 de Zoo/PhytoImage. Une version 6.0 sera rendue publique à la fin de la phase de beta-test : elle sera très proche de la version 5.4.0 décrite ici. Le manuel utilisateur décrit, outre l'installation et l'exécution du logiciel, l'utilisation des fonctions à partir du menu, l'utilisation de l'interface graphique utilisateur, ainsi que l'utilisation du logiciel en ligne de commande R.

## Dénombrement des cellules par colonies

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. La méthode incluse dans Zoo/PhytoImage consiste à calibrer un modèle prédictif 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 ont été mis en évidence sur douze groupes taxinomiques de phytoplancton colonial grâce à une étude comparative des dénombrements manuels et automatiques effectués sur deux sites (Boulogne-sur-Mer et Nantes).

Dans la présente étude, la prise en compte de la spécificité des taxons en colonie pour leur dénombrement par Zoo/PhytoImage représente une évolution prioritaire dans le cadre du REPHY. Les résultats présentés montrent qu'un tel calcul est possible en utilisant toutes les variables explicatives utilisées par Zoo/PhytoImage (attributs issus de l'analyse des images par Visual Spreadsheet, le logiciel associé au FlowCAM). La classification des particules en groupes taxinomiques, préalablement au dénombrement des cellules par colonies est crucial ici, car le calcul est fortement dépendant de la nature des particules étudiées.

Le module de calcul des cellules par colonies permet d'obtenir des mesures de biovolume, de biomasse et d'équivalent carbone pour chacun des groupes taxinomiques identifiés dans cette étude, moyennant un travail de compilation de la littérature disponible. 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. Pour les besoins du REPHY, et en nous appuyant sur les taxa composant le set d'apprentissage utilisé par les observateurs IFREMER, différentes formules allométriques ont été recensées dans ce travail.

## Apprentissage actif

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'optimiser l'approche de type apprentissage actif. Dans ce rapport, nous quantifions les gains obtenus à l'aide 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.

Dans le contexte de l'apprentissage actif, la classification ne se fait plus *uniquement* sur base d'un set



d'apprentissage initial et d'un « classifieur » relativement figés (approche classique en « machine learning »). Zoo/PhytoImage identifie également de manière probabiliste les particules qui sont sans doute mal classées (les particules suspectes). Lors de la phase de validation par l'expert, l'identification de ces particules suspectes, et les corrections éventuelles sont prises en compte localement au niveau de l'échantillon. Un second algorithme de classification supervisée intervient alors pour apprendre à corriger les autres particules dans l'échantillon. La classification des particules est ainsi extrêmement dynamique et contextuelle : en fonction de l'avancement du taxonomiste dans son travail de correction et validation d'un sous-ensemble de l'échantillon (une centaine de particules à la fois, pour un échantillon constitué de plusieurs milliers de particules), c'est l'ensemble de l'échantillon qui est corrigé par ce biais. Enfin, il est possible également de prendre en compte les corrections déjà réalisées sur des échantillons similaires (par exemples, les échantillons immédiatement antérieurs dans une série temporelle, ou des échantillons prélevés simultanément dans le temps et géographiquement proches). Ainsi, la correction de l'erreur est dopée... en théorie ! Le travail réalisé ici quantifie l'impact de cette approche dans le contexte pratique de l'énumération d'échantillons du REPHY traités à l'aide du FlowCAM.

Les résultats obtenus montrent que cet apprentissage actif permet de réduire de manière significative la fraction de l'échantillon que l'utilisateur doit vérifier manuellement. Par exemple, pour obtenir une erreur résiduelle inférieure ou égale à 25 %, l'introduction de l'apprentissage actif diminue la fraction à valider de 20 à 30 % selon l'échantillon considéré. Le temps nécessaire à la validation des dénombrements en est diminué de manière proportionnelle.

## Optimisation de la procédure de validation

La validation des échantillons fait maintenant intervenir des outils statistiques poussés, puisque : (i) la sélection des particules que l'utilisateur vérifie est déterminée d'une manière probabiliste (détection des suspects), (ii) un second classifieur est utilisé pour seconder l'utilisateur dans la validation (apprentissage actif), (iii) enfin, un algorithme de correction statistique de l'erreur est implémenté sur les résultats finaux. L'ensemble permet de réduire le travail de validation. Cela signifie que le taxonomiste peut arrêter plus tôt de visualiser et corriger manuellement l'attribution des groupes taxonomiques aux particules suspectes.

Cependant, il était indispensable également d'offrir une information claire à l'utilisateur concernant l'erreur résiduelle à chaque étape de la validation, afin de lui permettre de décider quand arrêter. De plus, de nombreux paramètres influent aussi sur la qualité de la validation, tels que le nombre de particules soumises à correction à chaque étape, le nombre d'étapes nécessaires, etc. Tous ces paramètres ont été optimisés ici sur base de l'analyse d'échantillons de la série REPHY elle-même. Un tableau de bord présente de manière claire et graphique l'avancement du travail de validation et anticipe le nombre d'étapes nécessaire pour atteindre un objectif donné (par exemple, moins de 5 % d'erreur dans tous les groupes taxinomiques).

Dans le cas d'échantillons contenant beaucoup de particules numérisées (8.000 voire plus), un sous-échantillon plus limité est validé (limité à 200 particules maximum à chaque étape), profitant ensuite de l'apprentissage actif pour effectuer automatiquement les corrections nécessaires sur l'ensemble de l'échantillon. Toutes ces mesures contribuent à la fois au confort et à l'efficacité de la validation.

Les améliorations décrites ici sont rendues disponibles dans la version 5.4.0 de Zoo/PhytoImage. Cette version est très proche de la future version 6.0 qui sera rendue publique : le logiciel entre en effet dans une phase de beta-test pré-version 6.0 à l'issue de ce travail.

Écologie Numérique des Milieux Aquatiques  
UMONS  
Faculté des Sciences



**Projet FlowCAM/ZooPhytoImage**  
**Rapport d'avancement**  
**– 30/03/2015 -**

Guillaume WACQUET & Philippe GROSJEAN

**UMONS**  
Université de Mons



## **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, il reste encore à paramétrer différents modèles utilisés par le logiciel dans le contexte particulier du Rephy. Différents axes ont été proposés par l'UMONS et l'IFREMER.

Premièrement, 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 calibrer le modèle prédictif déjà inclus dans Zoo/PhytoImage, et permettant d'estimer le nombre de cellules par colonie. La calibration se fait en se basant sur les comptages manuels réalisés sur les particules du set d'apprentissage. Dans cette étude, des scores élevés de performance ont été mis en évidence sur douze groupes taxinomiques de phytoplancton colonial de la Manche Orientale et de la Mer du Nord.

Deuxièmement, 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'optimiser l'approche de type apprentissage actif. Dans ce rapport, nous quantifions les gains obtenus à l'aide 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**

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



*Partie 1*  
**Dénombrement des cellules  
dans les colonies**

## Table des matières

Introduction.....	7
Présentation des données.....	7
Comptages manuels des vignettes.....	7
Distribution du nombre de cellules par colonie.....	9
Sélection des variables explicatives communes à tous les taxa.....	11
Résultats expérimentaux.....	15
Conclusion.....	21
Bibliographie.....	23
Annexe 1 : Outil d'aide au dénombrement des cellules.....	25
Organisation des fichiers.....	25
Dénombrement des colonies.....	26
Arrêt et reprise du dénombrement de cellules.....	29
Modalité de sauvegarde des résultats.....	29
Fonctions R.....	30
Fonction « cellsPerColonie ».....	30
Utilisation (fichier « testColonie.R »).....	30
Modalités de réalisation du test de dénombrement des colonies (proposition du 03/03/2015).....	31
Présentation du test.....	31
Avancement.....	31
Annexe 2 : Formules allométriques.....	33
Liste des taxa du set d'apprentissage REPHY - IFREMER.....	33
Formules allométriques (shape – biovolume).....	34

## 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 numérisés à l'aide du couplage 4X/300µm, provenant de cultures. Ici, les données de cette étude sont reprises et fusionnées avec d'autres taxa provenant d'échantillons naturels. 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 utilisés : outils d'aide au comptage manuel sur ordinateur (cf. Annexe 1), 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 douze groupes taxinomiques de phytoplancton colonial de la Manche Orientale et de la Mer du Nord.

## Présentation des données

### Comptages manuels des vignettes

Les particules constituant le set d'apprentissage utilisé dans cette étude, proviennent d'échantillons naturels (3 points à Boulogne-sur-Mer, 3 points à Dunkerque et 5 points en Baie de Somme) et de culture (*Ditylum brightwellii*, *Chaetoceros compressus* et *Thalassiosira rotula*). Ces derniers ont été numérisés à l'aide du FlowCAM durant les années 2009 et 2013.

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).

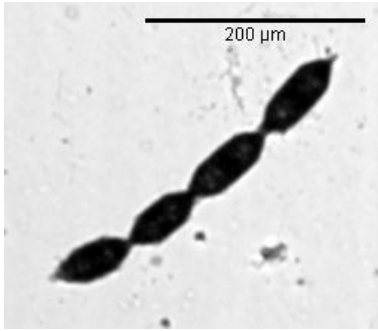
**Table 1** : Table représentant l'ensemble des groupes taxinomiques étudiés, provenant d'échantillons naturels et de culture (300µm/4X).

Groupes taxinomiques (300µm/4X)	Nombre de vignettes	Nombre total de cellules dénombrés
<i>Biddulphia rhombus</i>	101	168
<i>Biddulphia sinensis</i>	213	455
<i>Ditylum brightwellii</i>	453	551
<i>Thalassiosira rotula</i>	414	702
<i>Asterionella glacialis</i>	201	1489
<i>Thalassionema nitzschoides</i>	172	944
<i>Chaetoceros compressus</i>	234	2010
<i>Chaetoceros curvisetum</i>	187	2008
<i>Chaetoceros spp.</i>	127	841
<i>Leptocylindrus danicus</i>	169	1854
<i>Pseudo-Nitzschia spp.</i>	175	532
<i>Phaeocystis globosa</i>	232	???

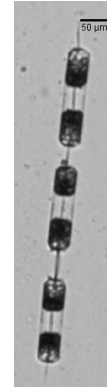
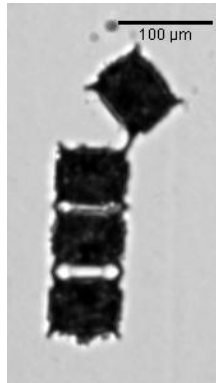


12 groupes taxinomiques contenant des colonies (morphologies et tailles différentes) :

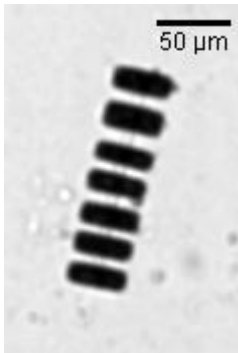
*Biddulphia rhombus* (4 cells)



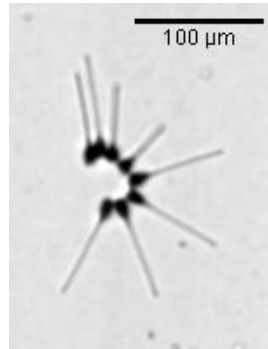
*Biddulphia sinensis* (4 cells)



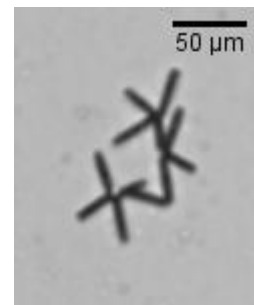
*Ditylum brightwellii* (3 cells)



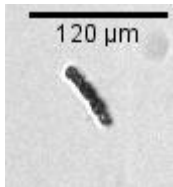
*Thalassiosira rotula* (7 cells)



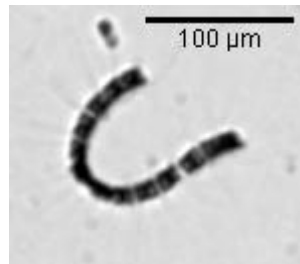
*Asterionella glacialis* (8 cells)



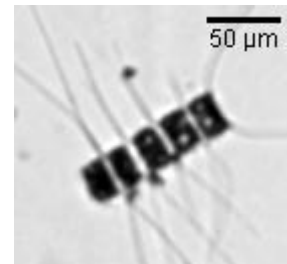
*Thalassionema nitzschooides*  
(12 cells)



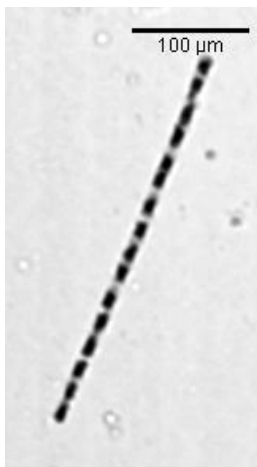
*Chaetoceros compressus*  
(5 cells)



*Chaetoceros curvisetum*  
(13 cells)



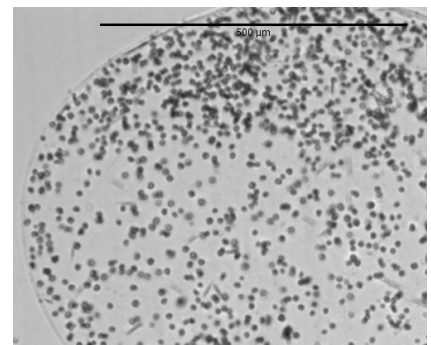
*Chaetoceros* spp. (5 cells)



*Leptocylindrus danicus*  
(16 cells)



*Pseudo-Nitzschia* spp. (3 cells)



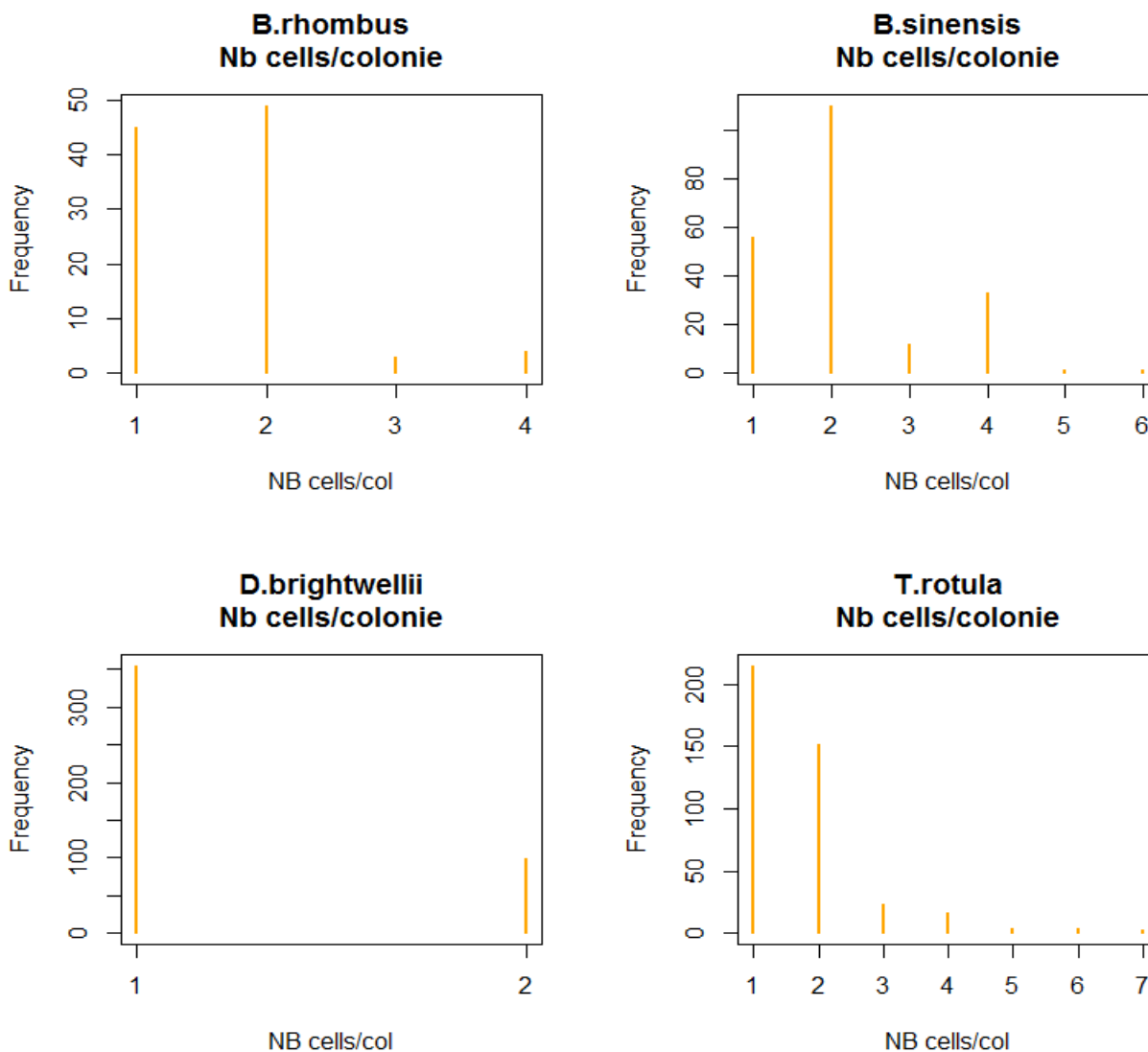
*Phaeocystis globosa* (??? cells)

Dans cette étude, *Phaeocystis globosa* n'est pas pris en compte. En effet, le comptage manuel des cellules dans chacune des colonies s'est avéré être une tâche longue et fastidieuse. C'est pourquoi d'autres pistes doivent être étudiées afin de définir un modèle prédictif pour ce taxon (utilisation d'un abaque en fonction de la taille de la colonie, par exemple).

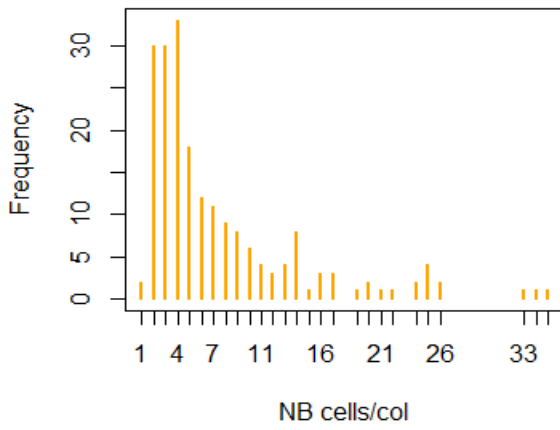
### **Distribution du nombre de cellules par colonie**

Dans cette section, la variabilité du nombre de cellules par colonie pour chacun des taxa est mise en évidence. Les différents graphes présentés ci-dessous permettent alors d'avoir une idée de la distribution du nombre de cellules pour chaque groupe étudié.

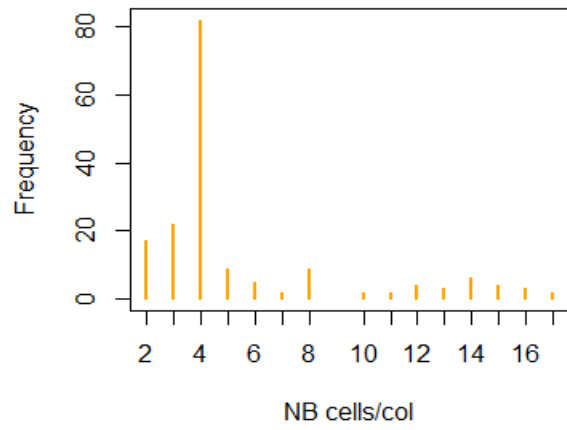
**Table 2** : Distributions des nombres de cellules par colonie pour chaque groupe taxonomique étudié, provenant d'échantillons naturels et de culture (300µm/4X).



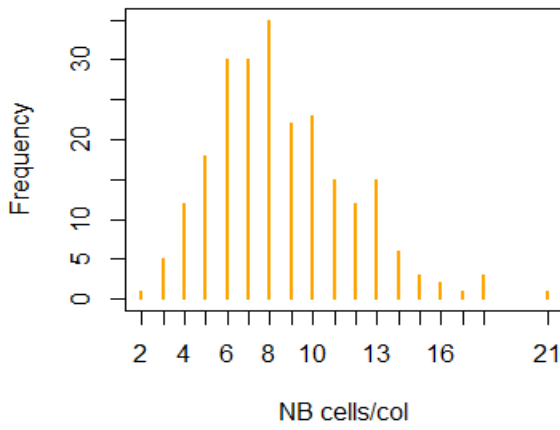
**A.glacialis**  
Nb cells/colonie



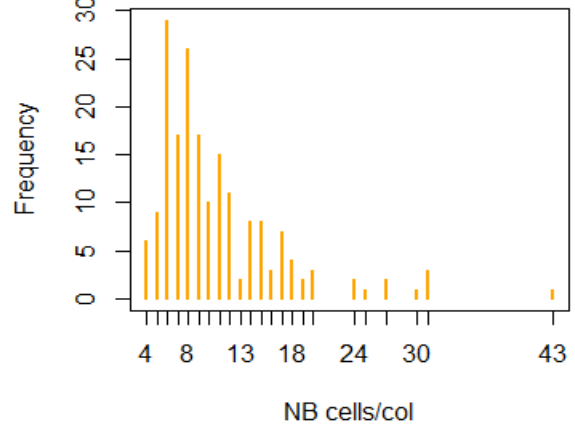
**T.nitzschioides**  
Nb cells/colonie



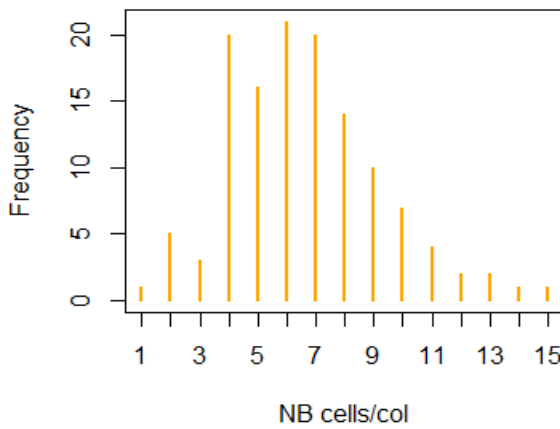
**C.compressus**  
Nb cells/colonie



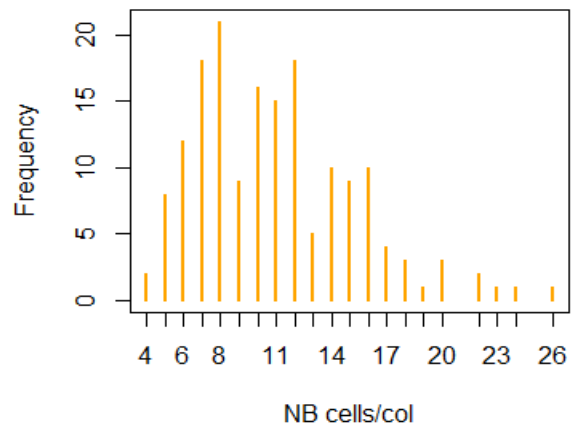
**C.curvisetum**  
Nb cells/colonie

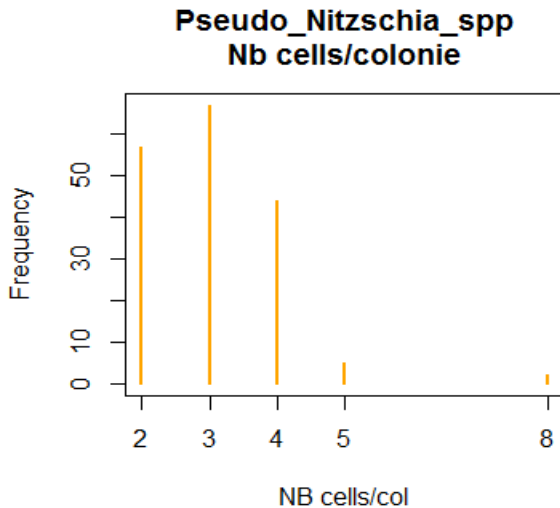


**Chaetoceros\_spp**  
Nb cells/colonie



**L.danicus**  
Nb cells/colonie





A la vue de ces résultats, nous pouvons remarquer que certains groupes coloniaux présentent un faible nombre de cellules. C'est notamment le cas pour : *Biddulphia rhombus*, *Biddulphia sinensis* et *Dytilum brightwellii*. D'autres espèces présentent une variabilité beaucoup plus importante (comme *Asterionellopsis glacialis*, *Chaetoceros compressus*, *Chaetoceros curvisetum* et *Leptocylindrus danicus*). Pour ces dernières, il est donc nécessaire d'obtenir un nombre plus important d'images reflétant la grande variabilité du nombre de cellules par colonie.

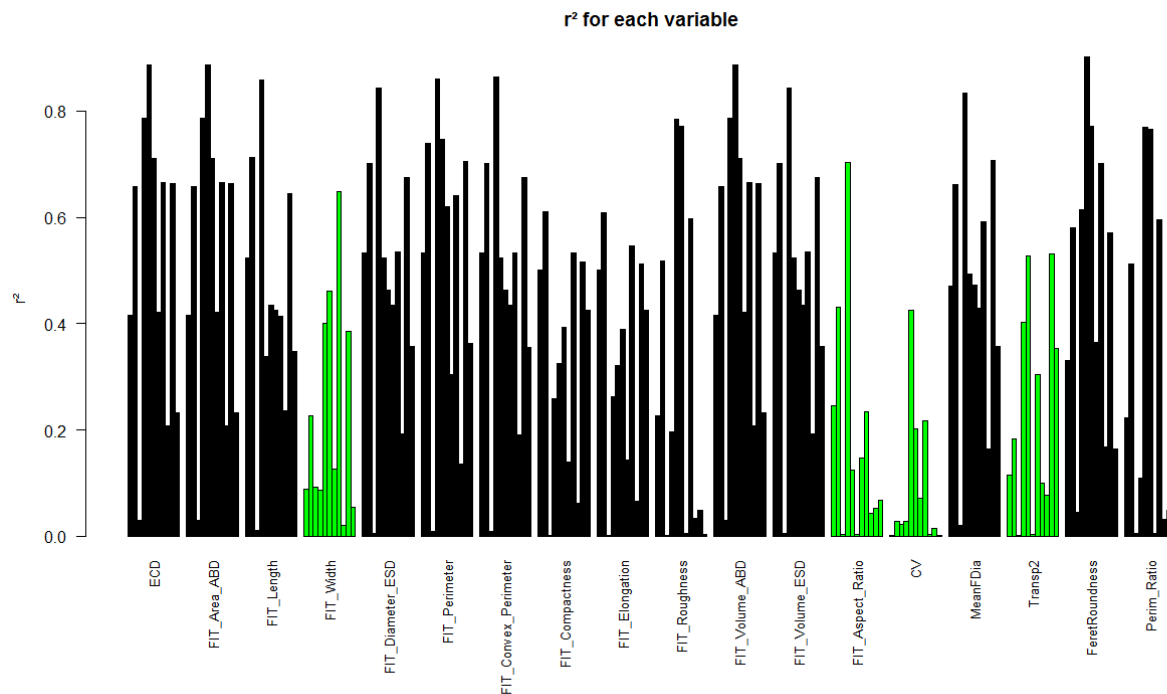
### **Sélection des variables explicatives communes à tous les taxa**

Dans cette section, les variables utilisées pour la construction des modèles prédictifs sont celles issues du FlowCAM (et notamment de Visual SpreadSheet) et de ZooPhytoImage. Ces variables sont au nombre de 18 :

- **FIT\_Area\_ABD, FIT\_Volume\_ABD** : surface en  $\mu\text{m}^2$  et volume en  $\mu\text{m}^3$ , basés sur l'Area Based Diameter.
- **FIT\_Diameter\_ESD, FIT\_Volume\_ESD** : diamètre en  $\mu\text{m}$  et volume en  $\mu\text{m}^3$ , 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 à la limite du masque définissant 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 Feret.
- **FIT\_Roughness** : mesure de la rugosité de la silhouette, grâce au rapport Perimeter/Surface.
- **FIT\_Compactness** : dérivé du périmètre et de l'aire (« ABD »). Cette variable équivaut à 1 pour une particule parfaitement circulaire. Formule :  $\text{Perimeter}^2 / (4 \times \pi \times \text{Area\_ABD})$ .
- **FIT\_Elongation** : rapport longueur/largeur.
- **ECD** : Equivalent Circular Diameter.
- **FeretRoundness** : mesure de la rugosité, à partir des diamètres de Feret.
- **MeanFDia** : Diamètre moyen calculé à partir de tous les diamètres de Feret.
- **Perim\_Ratio** : rapport des périmètres convexe et du périmètre du masque.
- **Transp2** : Transparence calculé en utilisant ECD et MeanFDia.
- **CV** : Coefficient de variation des niveaux de gris.

L'objectif de cette étude est de sélectionner les variables explicatives COMMUNES à TOUS les taxa. En effet, dans ZooPhytoImage, et afin de faciliter le processus, la méthodologie de construction des modèles prédictifs doit être commune à tous les groupes présents dans le set d'apprentissage (mêmes variables, même algorithme, etc.).

Nous choisissons, ici, d'ordonner les différentes variables selon le  $R^2$  (cf. Table 3), et ainsi extraire les « meilleures variables ».



**Figure 1 :**  $r^2$  pour chaque variable et chaque groupe taxonomique étudié, provenant d'échantillons naturels et de culture (300µm/4X).

La Fig. 1 présente les valeurs de  $r^2$  pour chacune des variables (et pour chaque groupe taxonomique) entre la variable et le nombre de cellules. Ici, il est possible de remarquer que certaines variables présentent un  $r^2$  faible pour la quasi-totalité des taxa (FIT\_Width, FIT\_Aspect\_Ratio, CV et Transp2, représentés en vert sur le graphe). Les détails de l'analyse sont repris dans la Table 3.

**Table 3 :** Table résumant les variables non explicatives au sens du  $R^2$ , pour chaque groupe taxonomique étudié, provenant d'échantillons naturels (300µm/4X).

Espèces	Variables (triées selon $R^2$ )	Espèces	Variables (triées selon $R^2$ )
<i>Biddulphia rhombus</i>	0.53	<i>Chaetoceros compressus</i>	0.43
	0.5		0.01
	0.74		0.01

	CV		
<i>Biddulphia sinensis</i>	FIT_Perimeter FIT_Length FIT_Diameter_ESD FIT_Volume_ESD FIT_Convex_Perimeter MeanFDia ECD FIT_Area_ABD FIT_Volume_ABD FIT_Compactness FIT_Elongation FeretRoundness FIT_Roughness Perim_Ratio <hr/> FIT_Aspect_Ratio FIT_Width Transp2 CV	<i>Chaetoceros curvisetum</i>	FeretRoundness FIT_Area_ABD ECD FIT_Volume_ABD FIT_Width FIT_Perimeter FIT_Roughness Perim_Ratio MeanFDia FIT_Elongation FIT_Diameter_ESD FIT_Volume_ESD FIT_Convex_Perimeter FIT_Compactness <hr/> FIT_Length FIT_Aspect_Ratio CV Transp2
<i>Ditylum brightwellii</i>	FIT_Width FeretRoundness FIT_Volume_ABD ECD FIT_Area_ABD CV MeanFDia FIT_Length FIT_Perimeter FIT_Convex_Perimeter FIT_Diameter_ESD FIT_Volume_ESD Perim_Ratio FIT_Aspect_Ratio FIT_Compactness FIT_Elongation Transp2 FIT_Roughness	<i>Chaetoceros spp.</i>	FIT_Length FIT_Volume_ABD FIT_Area_ABD ECD FIT_Diameter_ESD FIT_Volume_ESD FIT_Convex_Perimeter FeretRoundness MeanFDia FIT_Perimeter Transp2 FIT_Elongation FIT_Compactness FIT_Aspect_Ratio FIT_Roughness Perim_Ratio FIT_Width CV
<i>Thalassiosira rotula</i>	FIT_Convex_Perimeter FIT_Perimeter FIT_Length FIT_Diameter_ESD FIT_Volume_ESD MeanFDia FIT_Volume_ABD FIT_Area_ABD ECD FIT_Aspect_Ratio FeretRoundness <hr/> Transp2 FIT_Elongation FIT_Compactness FIT_Roughness Perim_Ratio FIT_Width CV	<i>Leptocylindrus danicus</i>	MeanFDia FIT_Perimeter <b>FIT_Convex_Perimeter</b> FIT_Diameter_ESD FIT_Volume_ESD FIT_Volume_ABD ECD FIT_Area_ABD FIT_Length FeretRoundness Transp2 FIT_Compactness FIT_Elongation <hr/> FIT_Width FIT_Aspect_Ratio FIT_Roughness Perim_Ratio CV
<i>Asterionellopsis glacialis</i>	FeretRoundness ECD FIT_Area_ABD FIT_Volume_ABD	<i>Pseudo-Nitzschia spp.</i>	FIT_Compactness FIT_Elongation FIT_Perimeter FIT_Diameter_ESD

0.70

0.5

0.5

0.00

0.44

0.01

0.01

0.86

0.71

0.5

0.5

0.01

0.03

0.90

0.43

	0.5	FIT_Roughness Perim_Ratio FIT_Perimeter Transp2 FIT_Convex_Perimeter FIT_Diameter_ESD FIT_Volume_ESD ----- MeanFDia CV FIT_Width FIT_Length FIT_Compactness FIT_Elongation FIT_Aspect_Ratio		FIT_Volume_ESD MeanFDia FIT_Convex_Perimeter Transp2 FIT_Length FIT_Volume_ABD ECD FIT_Area_ABD FeretRoundness FIT_Aspect_Ratio FIT_Width FIT_Roughness Perim_Ratio CV	
	0.77	FeretRoundness FIT_Roughness Perim_Ratio ECD FIT_Area_ABD FIT_Volume_ABD FIT_Perimeter ----- MeanFDia FIT_Convex_Perimeter FIT_Diameter_ESD FIT_Volume_ESD FIT_Width FIT_Length FIT_Compactness FIT_Elongation CV Transp2 FIT_Aspect_Ratio			
<i>Thalassionema nitzschoides</i>	0.5		<i>Phaeocystis globosa</i>	???	
	0.01				

0.13  
0.01

Une synthèse des résultats est proposée dans la Table 5. Les variables communes à conserver/supprimer y sont présentées. Ici, dans cette étude, et pour les groupes taxonomiques étudiés, 4 variables semblent peu informatives : CV, FIT\_Aspect\_Ratio, FIT\_Width et Transp2.

**Table 5** : Synthèse des variables explicatives ( $r^2 \geq 0.5$ ) et non explicatives ( $r^2 < 0.5$ ) communes pour les groupes taxonomiques étudiés. --- : Paramètres à conserver ; - - - : Paramètres à supprimer

	Variables explicatives ( $r^2 \geq 0.5$ )	Variables non explicatives ( $r^2 < 0.5$ )
11 espèces		CV
10 espèces		FIT_Aspect_Ratio FIT_Width
9 espèces		Transp2
8 espèces		
7 espèces	FIT_Perimeter	MeanFDia FIT_Roughness Perim_Ratio FIT_Length FIT_Elongation FIT_Compactness
6 espèces	FIT_Convex_Perimeter FIT_Diameter_ESD FIT_Volume_ESD ECD FIT_Area_ABD FIT_Volume_ABD FeretRoundness	

5 espèces		FIT_Volume_ABD ECD FIT_Area_ABD FeretRoundness FIT_Convex_Perimeter FIT_Diameter_ESD FIT_Volume_ESD
4 espèces	FIT_Length FIT_Compactness FIT_Elongation MeanFDia FIT_Roughness Perim_Ratio	FIT_Perimeter
3 espèces		
2 espèces	Transp2	
1 espèces	FIT_Width FIT_Aspect_Ratio	
0 espèce	CV	

### **Résultats expérimentaux**

Dans cette section, nous nous intéressons à l'estimation du nombre de cellules par colonies. Les performances des méthodes prédictives sont comparées par validations croisées (10-fois) sur des modèles de régressions linéaires et non linéaires.

Grâce à la validation croisée, il est possible d'évaluer les taux de reconnaissance des différentes méthodes prédictives. Ici, trois scores sont évalués ([3] Govaerts, 2010) :

- **TL0** : taux de reconnaissance global du nombre de cellules par colonie logarithmique.
- **TL1** : taux de reconnaissance à 1 classe près du nombre de cellule par colonie logarithmique.
- **Estimation totale** : estimation totale du nombre de cellules, en somme.

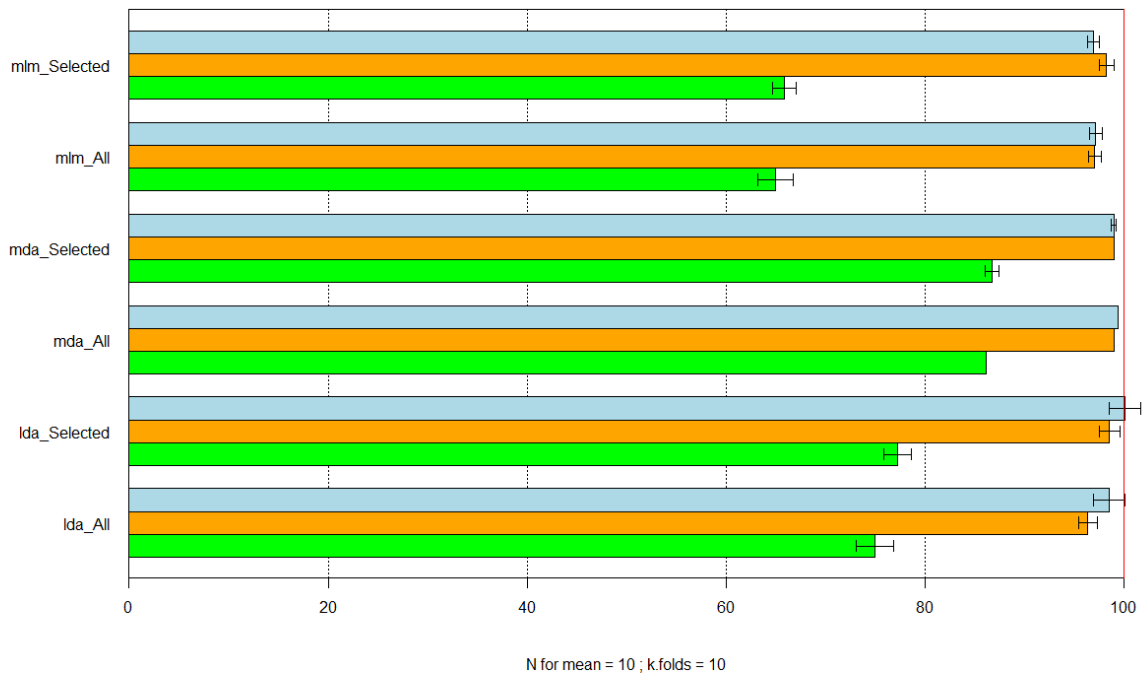
Nous choisissons ici d'utiliser 3 méthodes de régression :

- LM : Linear Model (Modèle Linéaire Multivarié),
- LDA : Linear Discriminant Analysis (Analyse Discriminante Linéaire),
- MDA : Mixture Discriminant Analysis (extension de LDA). 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.



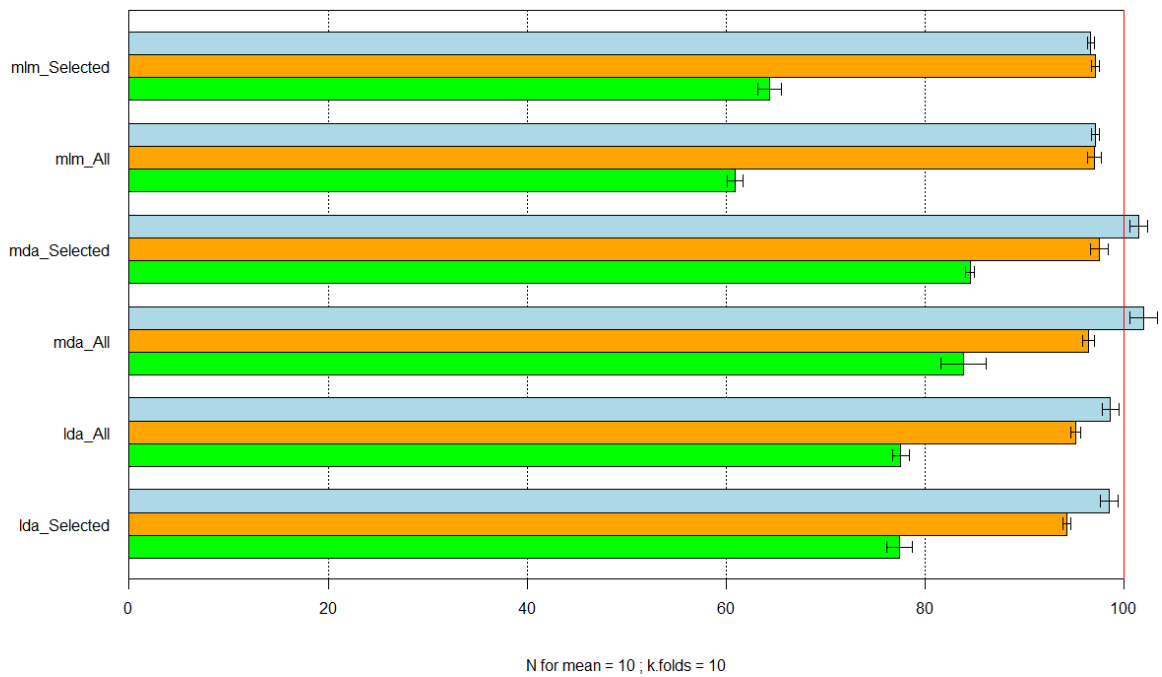
*Biddulphia rhombus*

Barplot of TL0 & TL1 & Total. estimation scoring means of 6 methodologies in *B.rhombus* species



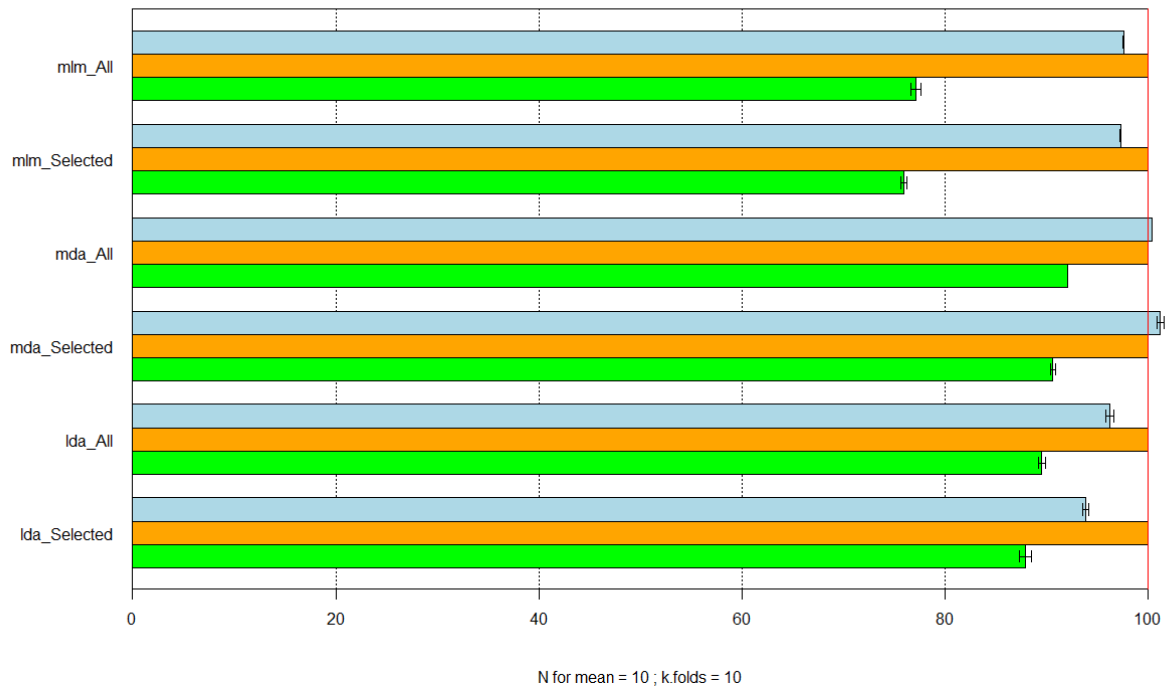
*Biddulphia sinensis*

Barplot of TL0 & TL1 & Total. estimation scoring means of 6 methodologies in *B.sinensis* species



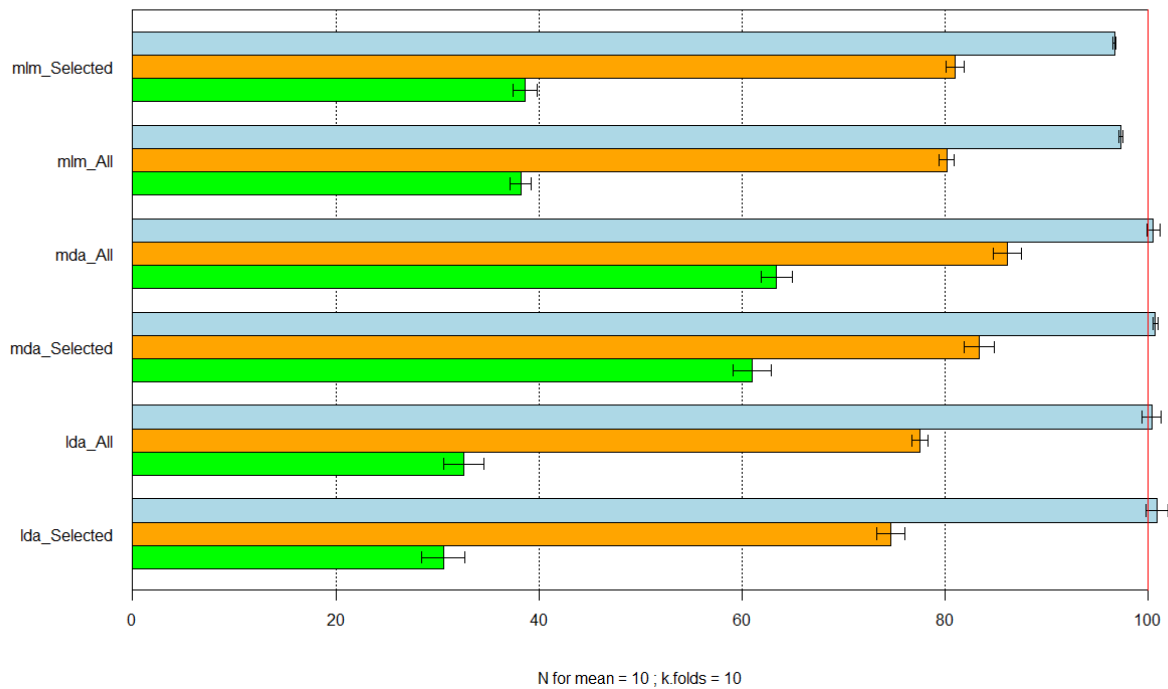
*Dytilum brightwellii*

Barplot of TL0 & TL1 & Total. estimation scoring means of 6 methodologies in *D.brightwellii* species



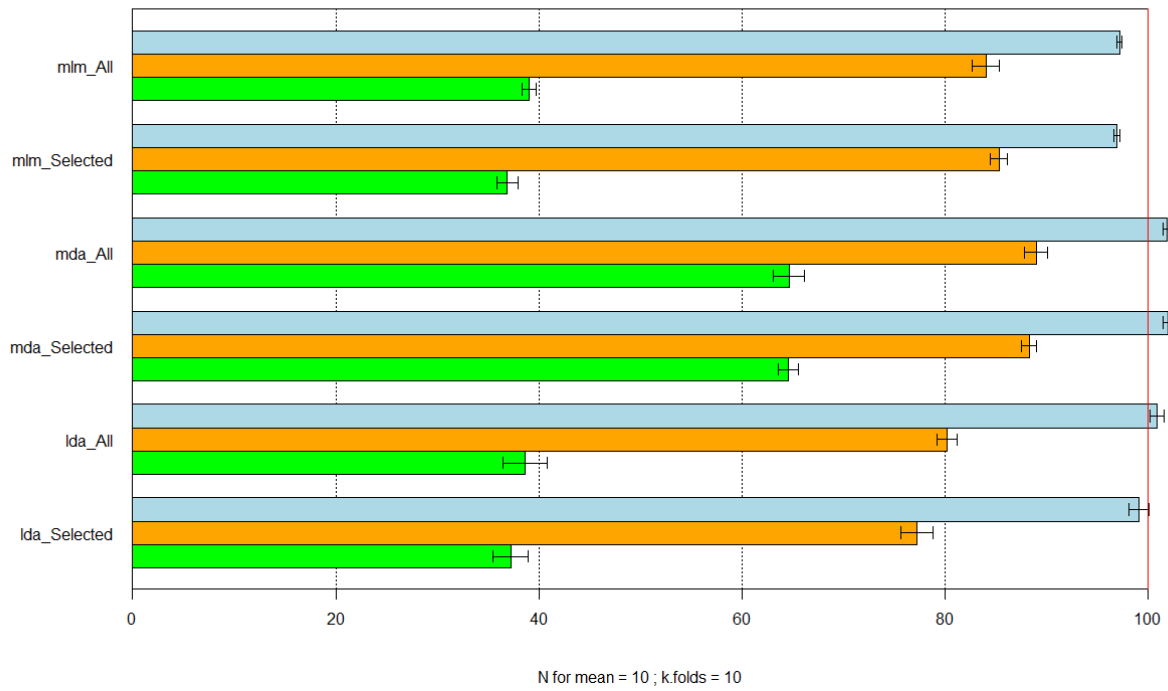
*Chaetoceros compressus*

Barplot of TL0 & TL1 & Total. estimation scoring means of 6 methodologies in *C.compressus* species



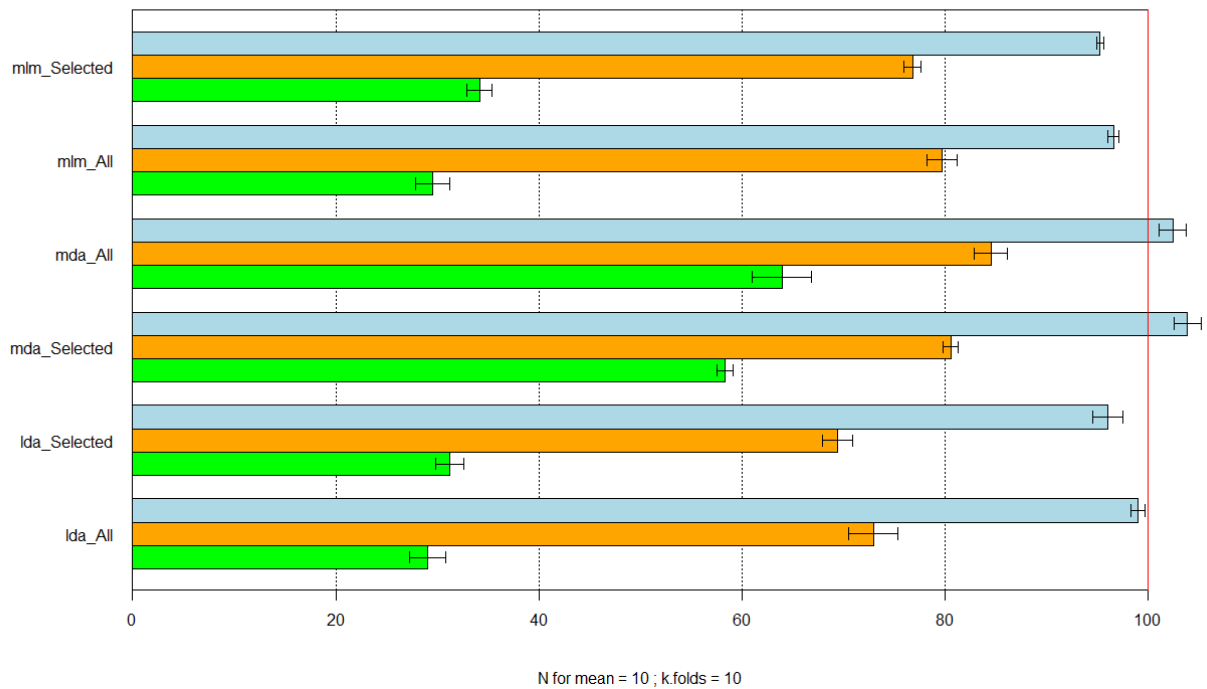
*Chaetoceros curvisetum*

Barplot of TL0 & TL1 & Total. estimation scoring means of 6 methodologies in *C. curvisetum* species



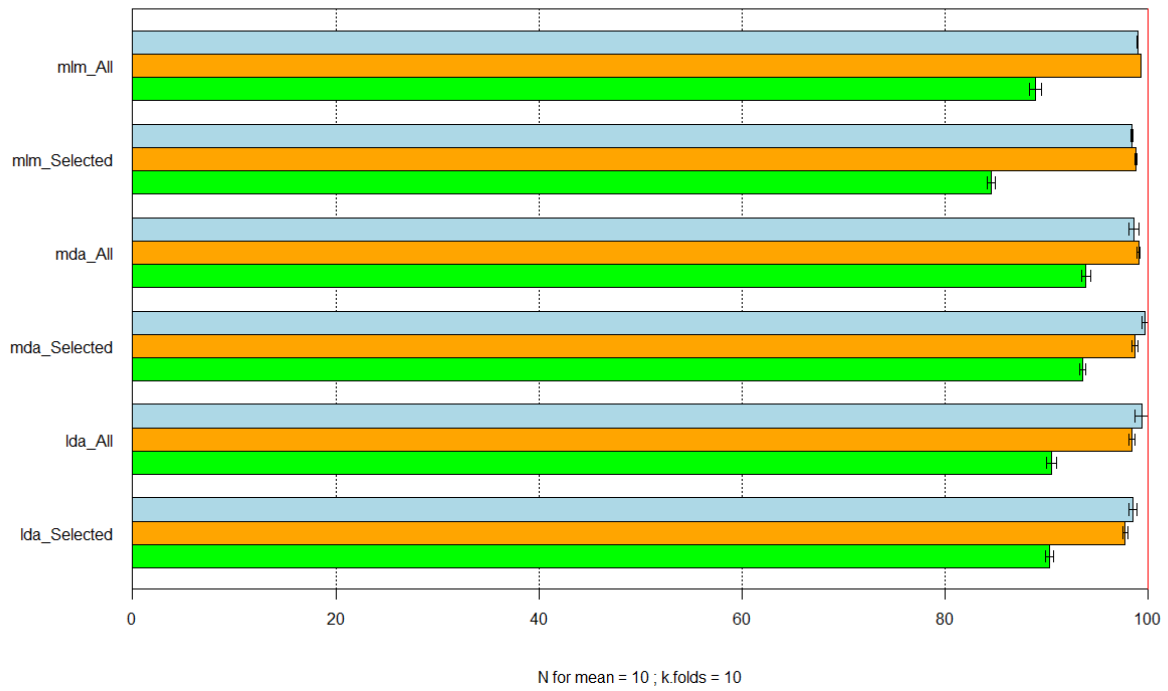
*Chaetoceros\_spp*

Barplot of TL0 & TL1 & Total. estimation scoring means of 6 methodologies in *Chaetoceros\_spp* species



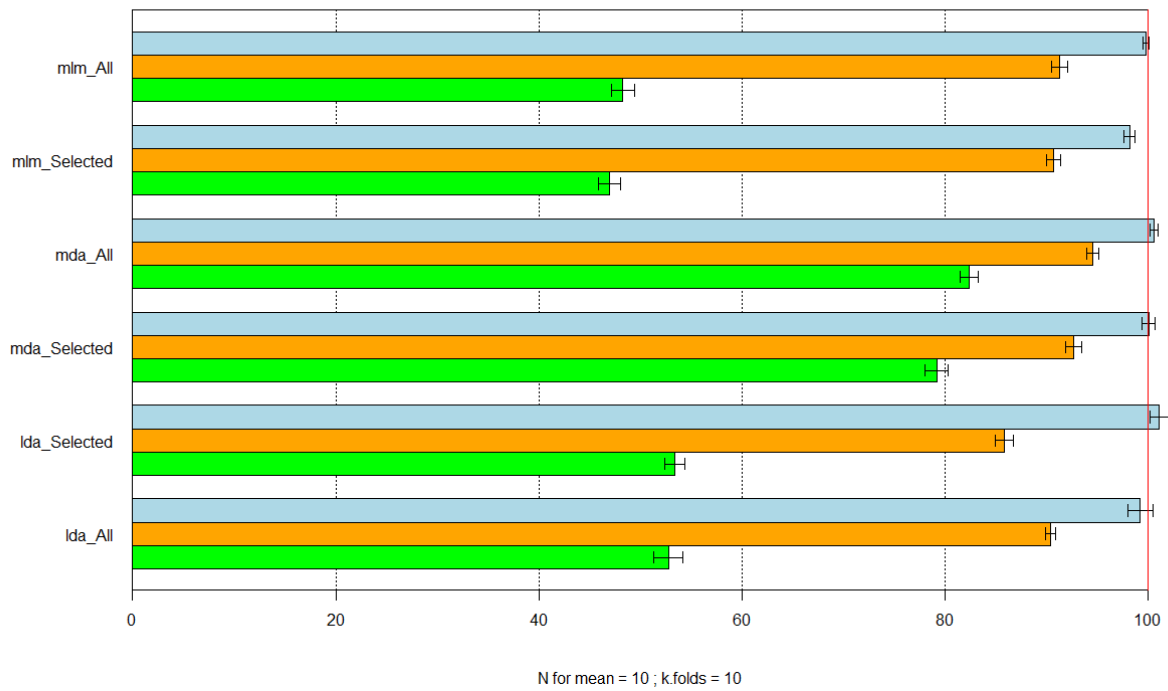
*Thalassiosira rotula*

Barplot of TL0 & TL1 & Total. estimation scoring means of 6 methodologies in *T. rotula* species



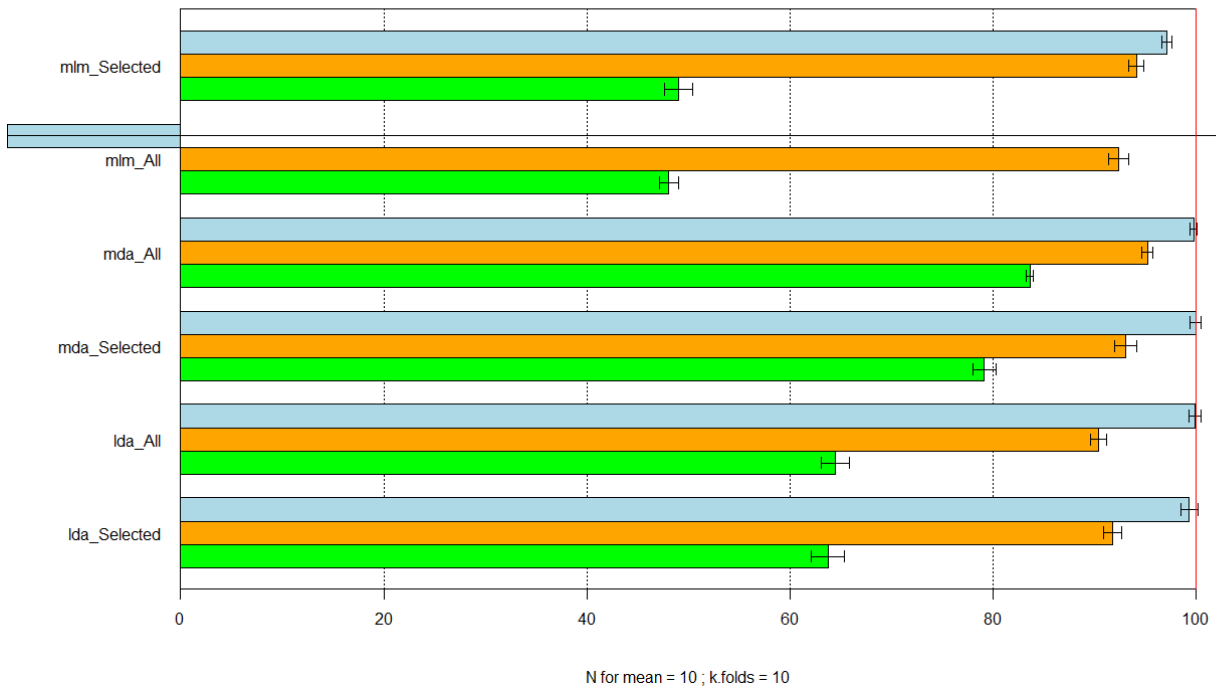
*Asterionellopsis glacialis*

Barplot of TL0 & TL1 & Total. estimation scoring means of 6 methodologies in *A. glacialis* species



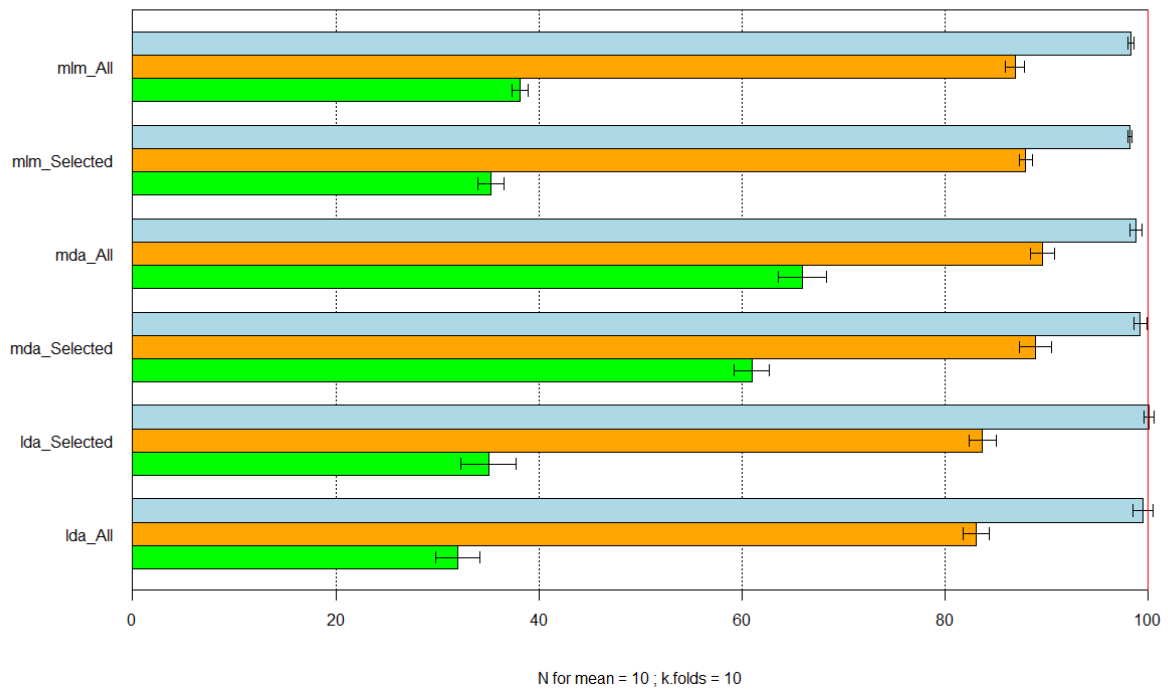
*Thalassionema nitzschioides*

Barplot of TL0 & TL1 & Total. estimation scoring means of 6 methodologies in *T.nitzschioides* species



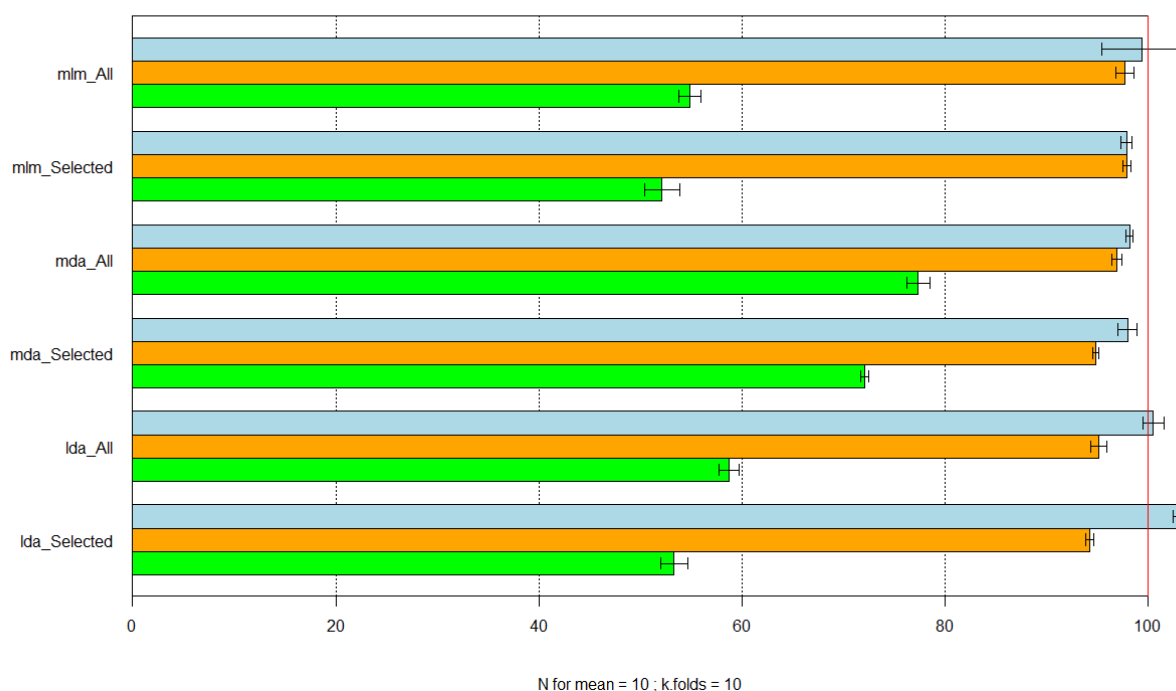
*Leptocylindrus danicus*

Barplot of TL0 & TL1 & Total. estimation scoring means of 6 methodologies in *L.danicus* species



## Pseudo\_Nitzschia\_spp

Barplot of TL0 & TL1 & Total. estimation scoring means of 6 methodologies in Pseudo\_Nitzschia\_spp species



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

- le score **TL0** (représenté en **vert**),
- le score **TL1** (représenté en **orange**),
- le score d'**estimation totale** (représenté en **bleu**).

Globalement, pour toutes les espèces étudiées, un modèle multivarié basé sur le MDA permet d'obtenir les meilleurs scores TL0 et TL1. Cependant, même si le gain pour l'estimation totale est faible (ou dans certains cas, inexistant), il est important que la prédiction soit la plus fiable possible. C'est pourquoi, au vu de la combinaison des résultats TL0, TL1 et Estimation.Totale, nous pouvons remarquer une nette supériorité de la méthode MDA.

De plus, il est important de noter que la sélection de variables explicatives ne permet pas d'améliorer la prédiction. Dans certains cas, cette sélection tend même à dégrader les performances globales de la méthode.

## Conclusion

La prise en compte de la spécificité des taxons en colonie pour leur dénombrement par Zoo/PhytoImage représente une évolution prioritaire dans le cadre du Rephy. Les résultats préliminaires présentés ici montrent qu'un tel calcul est possible en utilisant toutes les variables explicatives, couplées à un algorithme MDA.

Les outils présentés dans ce rapport et en particulier 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 classification. Dans ce sens, un travail collaboratif avec les centres IFREMER de Boulogne-sur-Mer, Nantes et Arcachon a débuté début mars 2015. L'outil d'aide au dénombrement manuel des cellules a alors été mis à leur disposition (cf. Annexe 1).

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. Pour les besoins du REPHY, et en nous appuyant sur les taxa composant le set d'apprentissage utilisé par les observateurs IFREMER, différentes formules allométriques seront recensées dans la suite de ce travail.

## Bibliographie

- [1] Alcaraz M., Saiz E. et al. (2003). **Estimating zooplankton biomass through image analysis**, *Marine Biology*, 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*.





# Annexe 1 : Aide au dénombrement des cellules

## Organisation des fichiers

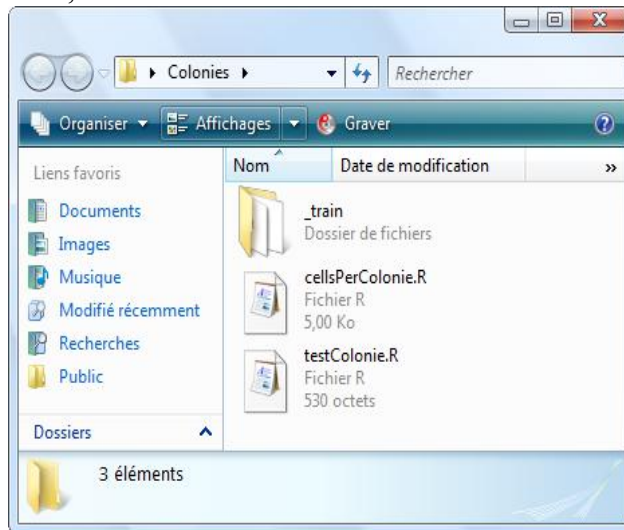
Un outil d'aide au dénombrement des cellules dans les vignettes est disponible dans Zoo/PhytoImage. Il consiste en :

- 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,
- la correction manuelle (par clics souris) ou la validation du nombre de cellules.

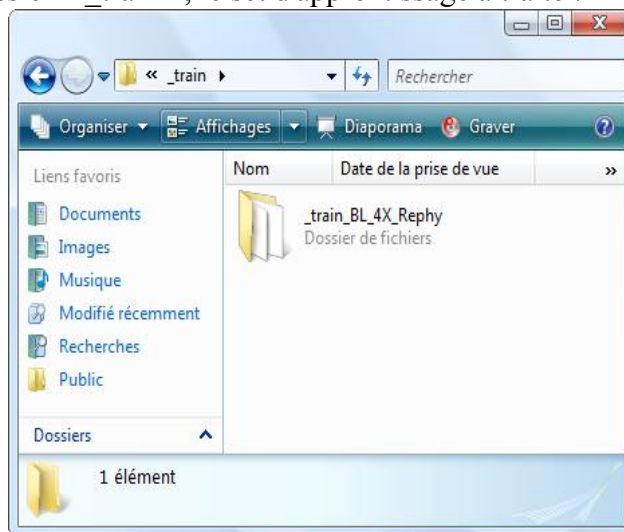
Une fois validé, une nouvelle entrée est créée directement et automatiquement pour chacune des particules dans le set d'apprentissage.

Pour l'utilisation du module de dénombrement, il est nécessaire d'adopter une organisation spécifique des fichiers dans le répertoire de travail. Ce dernier doit donc contenir :

- **les fichiers R** correspondant aux fonctions développées pour le dénombrement des cellules par colonie, ainsi qu'**un sous-dossier « \_train »** contenant le set d'apprentissage avec les vignettes à dénombrer,



- Dans ce sous-dossier « \_train », le set d'apprentissage à traiter.



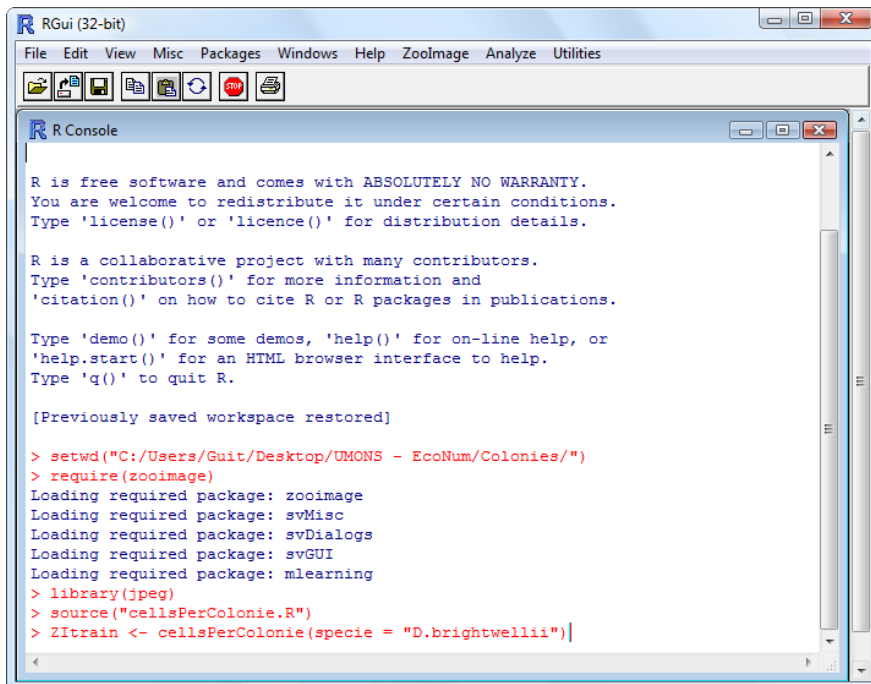
Une fois cette structure de dossiers et de fichiers établie, vous pouvez commencer à dénombrer les cellules dans les vignettes du set d'apprentissage.

**Remarque :** Seules les vignettes contenues dans le répertoire « **Phytoplankton** » du set d'apprentissage, devront être dénombrées par l'utilisateur.  
En effet, les particules détritiques ne nécessitent pas de dénombrement.

## Dénombrement des colonies

Afin de démarrer le processus de dénombrement des colonies, entrez dans la console R, les commandes suivantes :

```
setwd("C:/Users/Desktop/Colonies/") # modifier le chemin d'accès au répertoire de travail
require(zooimage)
library(jpeg)
source("cellsPerColonie.R")
Zltrain <- cellsPerColonie(specie = "D.brightwellii")
```



```
RGui (32-bit)
File Edit View Misc Packages Windows Help ZooImage Analyze Utilities

R Console

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

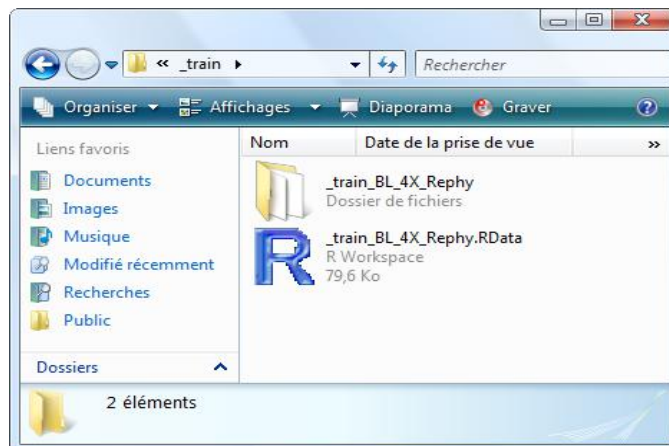
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

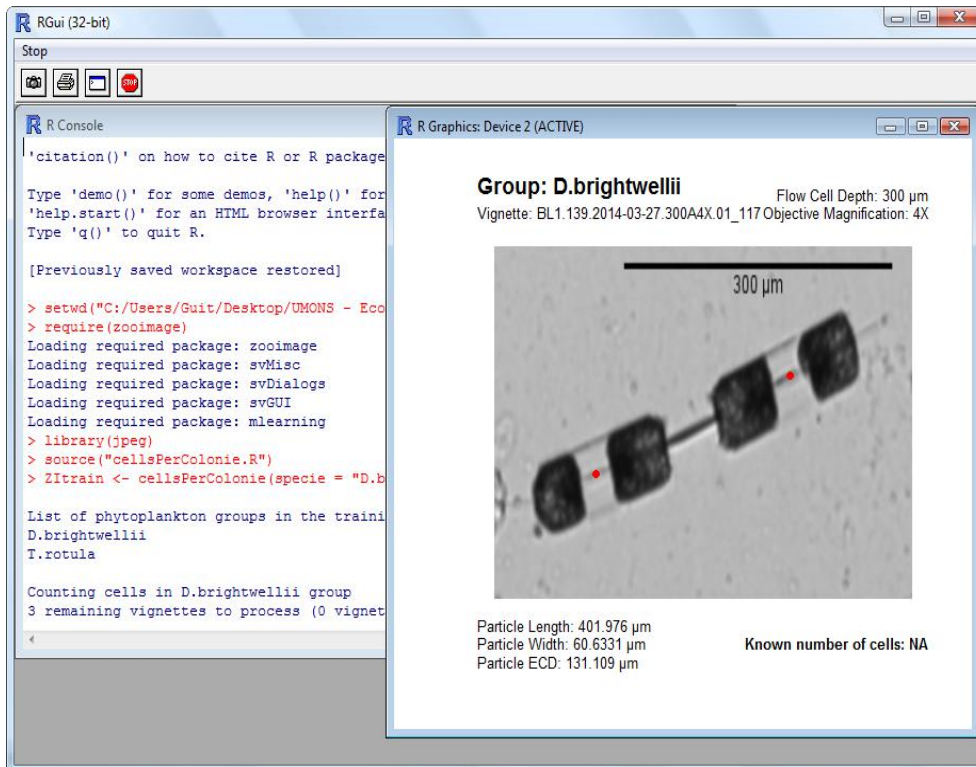
[Previously saved workspace restored]

> setwd("C:/Users/Guit/Desktop/UMONS - EcoNum/Colonies/")
> require(zooimage)
Loading required package: zooimage
Loading required package: svMisc
Loading required package: svDialogs
Loading required package: svGUI
Loading required package: mlearning
> library(jpeg)
> source("cellsPerColonie.R")
> Zltrain <- cellsPerColonie(specie = "D.brightwellii")
```

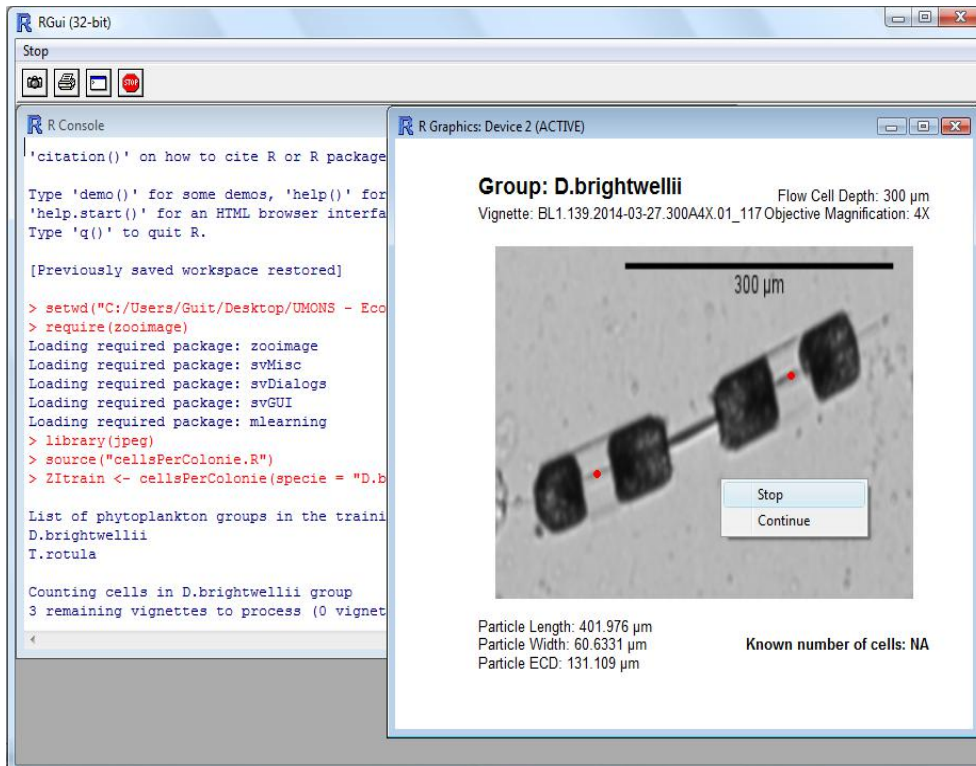
Lors d'une première analyse du set d'apprentissage, **un objet R (de format RData) est créé dans le sous-répertoire « \_train »**. Cet objet correspond au résultat de la lecture de ce set d'apprentissage dans R. Pour utiliser un nouvel ensemble d'apprentissage, il est impératif de supprimer cet objet R afin que le programme puisse recréer un nouvel objet.



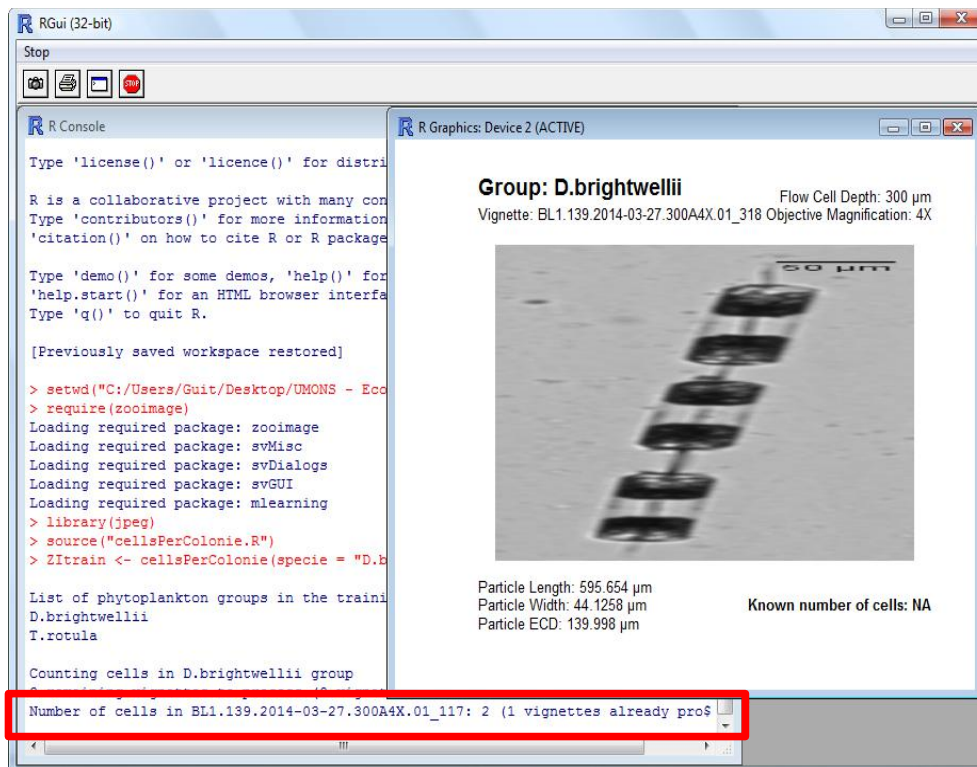
De plus, sous R, une fenêtre contenant une vignette est apparue. Sur cette dernière, le nombre de cellules connu, mais également la longueur, la largeur et l'ECD (Equivalent Circular Diameter) de la particule sont affichés. Pour dénombrer les cellules de cette vignette, il suffit de **cliquer à la souris sur chacune des cellules identifiées** (chaque cellule est alors marquée d'un point rouge afin d'aider l'utilisateur dans le comptage) :



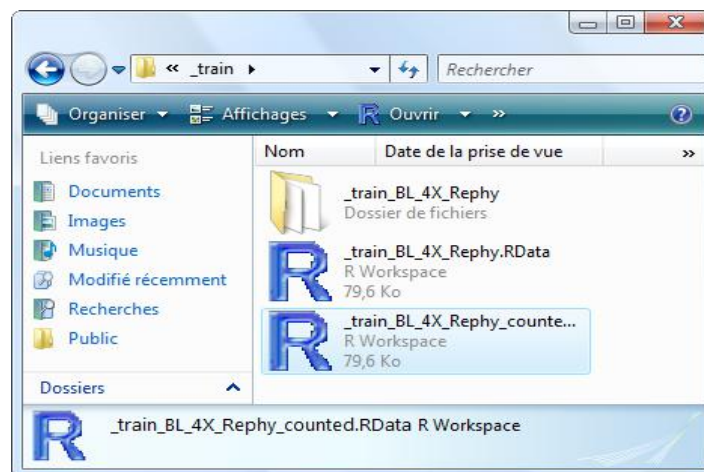
Lorsque le dénombrement des cellules est terminé, et afin de passer à la vignette suivante, il suffit de **cliquer droit avec la souris, et de sélectionner « STOP »**.



La vignette suivante à traiter est alors proposée à l'utilisateur et le nombre de cellules dénombrée dans la vignette précédente, ainsi que le nombre total de vignettes dénombrées dans le groupe sont affichés dans la console R.



Pour chaque vignette traitée, le nombre de cellules correspondant est directement et automatiquement sauvegardé dans le set d'apprentissage. **Un nouvel objet R (au format Rdata) et nommé « trainingSetName\_counted.RData » est alors créé dans le sous-dossier « \_train ».**



**Remarque :** Pour le dénombrement, plusieurs possibilités s'offrent à l'utilisateur :

- si **le nombre de cellules est dénombrable**, l'utilisateur clique sur chacune des cellules identifiées, puis passe à la vignette suivante en cliquant droit et en sélectionnant « STOP ». **La valeur est alors enregistrée dans le set d'apprentissage.**
- si **le nombre de cellules n'est pas dénombrable** (forme complexe, mauvaise qualité d'image, etc.), l'utilisateur passe directement à la vignette suivante en cliquant droit puis en sélectionnant « STOP ». **Aucune valeur (NA) n'est alors enregistrée dans le set.**
- pour stopper le processus de dénombrement, il suffit de fermer la fenêtre contenant la vignette. Le processus peut alors être repris plus tard en retapant les commandes R.

## ***Arrêt et reprise du dénombrement de cellules***

### ***Reprendre le dénombrement***

Si l'utilisateur souhaite stopper le processus de dénombrement, et le reprendre plus tard, il est possible de relancer le processus en tapant, dans la console R, les commandes précédemment présentées :

```
setwd("C:/Users/Desktop/Colonies/")      # modifier le chemin d'accès au répertoire de travail
require(zooimage)
library(jpeg)
source("cellsPerColonie.R")
Zltrain <- cellsPerColonie(specie = "D.brightwellii")
```

Par défaut, le processus reprend alors à partir de la dernière vignette traitée par l'utilisateur, dans la session précédente. Le nombre de vignettes restantes à traiter ainsi que le nombre de vignettes déjà traitées dans le groupe sont affichés dans la console R.

### ***Recommencer le dénombrement***

Cependant, il est également possible de stopper le processus afin de recommencer complètement les comptages sur toutes les vignettes d'un groupe donné. En effet, un argument supplémentaire (« continue ») peut être ajouté à l'appel de la fonction afin de réinitialiser le nombre de cellules pour chacune des particules. Les commandes à taper dans la console R, deviennent alors :

```
setwd("C:/Users/Desktop/Colonies/")      # modifier le chemin d'accès au répertoire de travail
require(zooimage)
library(jpeg)
source("cellsPerColonie.R")
Zltrain <- cellsPerColonie(specie = "D.brightwellii", continue = FALSE)
```

Le processus reprend alors à partir de la première vignette du groupe défini.

### ***Modalité de sauvegarde des résultats***

Par défaut, une sauvegarde du nombre de cellules par colonie est effectuée à chaque traitement de vignette. Cependant, dans le cas d'un set d'apprentissage volumineux, cette sauvegarde à chaque itération peut alors prendre du temps. L'utilisateur peut alors choisir de ne sauvegarder les résultats qu'à la fin du traitement de TOUTES les vignettes d'un groupe donné. Les commandes à taper dans la console R, sont alors :

```
setwd("C:/Users/Desktop/Colonies/")      # modifier le chemin d'accès au répertoire de travail
require(zooimage)
library(jpeg)
source("cellsPerColonie.R")
Zltrain <- cellsPerColonie(specie = "D.brightwellii", saves = FALSE)
```

Cependant, pour le moment, en choisissant cette modalité de sauvegarde, si l'utilisateur stoppe le traitement avant d'avoir traité la totalité des vignettes du groupe, alors aucune sauvegarde n'est effectuée. Lorsque le processus est relancé, le traitement reprend donc depuis la première vignette (même si l'argument « continue = TRUE » - cf. section précédente) .

Une fois la totalité des vignettes traitées, le processus est automatiquement stoppé et la fenêtre active est fermée.



## Fonctions R

### Fonction « cellsPerColonie »

Cette fonction permet de lire le set d'apprentissage, de créer un objet RData correspondant à celui-ci, de proposer à l'utilisateur un outil d'aide au dénombrement des cellules par colonie, et de sauvegarder automatiquement les comptages pour chaque vignette dans le set d'apprentissage.

#### Fonction 1 : cellsPerColonie

```
Input : specie, mode ("manual" ou "auto"), continue (TRUE ou FALSE)
Output : ZItrain (set d'apprentissage contenant les comptages sur chaque vignette)

if Pas de training set dans sous-dossier
| affichage d'un message d'erreur (« Pas de training set dans le sous-dossier !!! »)
else
| vigs ← liste des vignettes présentes dans le groupe specie du set d'apprentissage
| if Training set non lu dans R (objet RData absent du sous-dossier)
| | ZItrain ← lecture du set d'apprentissage
| | ZItrain[specie]$NbCells ← NA : initialisation des comptages dans le set
| | sauvegarde de l'objet RData dans sous-dossier
| else if Dénombrements pas encore commencés (objet _counted.RData absent)
| | ZItrain ← chargement de l'objet RData
| else
| | ZItrain ← chargement de l'objet _counted.RData
| | if !continue
| | | ZItrain[specie]$NbCells ← NA : ré-initialisation des comptages dans le set

if length(vigs avec NbCells) > 0
| for vigs avec NbCells == NA
| | affichage de la vignette dans une fenêtre interactive + quelques mesures sur la particule
| | nbManual ← comptage manuel des cellules dans la vignette
| | ZItrain[specie]$NbCells ← nbManual
| | affichage du nombre de cellules dans la console R
| | sauvegarde de l'objet _counted.RData dans sous-dossier
else
| affichage d'un message d'erreur (« toutes les vignettes ont été traitées !!! »)
return(ZItrain)
```

### Utilisation (fichier « testColonie.R »)

```
setwd("C:/Users/Desktop/Colonies/") # Répertoire de travail (à modifier)
require(zoomimage) # ZoomImage version 5
library(jpeg) # Bibliothèque nécessaire pour manipuler les images au format JPEG
# Fichier nécessaire pour le comptage des cellules par colonie
source("cellsPerColonie.R") # Fonction principale pour le dénombrement des cellules
# Comptage manuel des cellules dans les colonies
ZItrain <- cellsPerColonie(specie = "D.brightwellii") # Mode "auto" pas encore créé
```

## **Modalités de réalisation du test de dénombrement des colonies (proposition du 03/03/2015)**

### **Présentation du test**

L'objectif principal du travail est la construction de modèles prédictifs pour le dénombrement des cellules en colonies. Dans le livrable ONEMA n°1, quelques résultats préliminaires ont été présentés et ont montré des scores intéressants pour quelques taxa. Cependant, afin d'améliorer et valider les modèles construits, il est nécessaire d'avoir un nombre plus important de vignettes dénombrées manuellement pour les espèces recensées dans le cadre du REPHY.

C'est pourquoi, il est proposé de mettre à disposition des principaux utilisateurs de l'IFREMER, un outil de dénombrement manuel (sous forme de script R) permettant de compter manuellement et simplement le nombre de cellules par colonie (par simple clic souris sur chacune des cellules identifiées sur chaque image), et d'enregistrer automatiquement ces valeurs dans le set d'apprentissage (un tutoriel sera fourni aux partenaires). Cependant, pour certaines espèces et/ou vignettes (de formes complexes et/ou de mauvaise qualité), cela reste quasi-impossible. Il suffira alors simplement de les passer.

Il serait intéressant de travailler sur le set d'apprentissage commun (résultant de la fusion des 3 sets provenant de Boulogne-sur-Mer, Nantes et Arcachon) que Nadine Neaud-Masson (NNM) a créé (décembre 2014 - janvier 2015) car il reflète la diversité des groupes rencontrés habituellement. Cependant, il est vrai que ce set comporte un nombre très important de vignettes et que cela risque de prendre du temps, mais le script R permet de réaliser les comptages en plusieurs fois (il est possible de stopper le comptage n'importe quand, et de le reprendre plus tard). De plus, il serait souhaitable que ces dénombrements manuels soient effectués dans un premier temps, par Boulogne-sur-Mer mais également par Nantes et Arcachon de manière complètement indépendante (chaque site devra donc traiter entièrement le set d'apprentissage), afin de confronter les résultats des dénombrements manuels effectués par chaque site et ainsi mettre en évidence les possibles biais de l'observateur (d'où l'intérêt de travailler sur le même set d'apprentissage).

### **Avancement**

Demande d'implication des laboratoires :

- LER Boulogne-sur-Mer :
  - Alain Lefebvre (AL), Pascale Hébert ?, Camille Blondel ?
- LER Arcachon :
  - Danièle Maurer (DM), Myriam Rumebe ?, Claire Meteigner?
- VIGIES et LER Nantes :
  - Nadine Neaud-Masson (NNM), Mireille Fortune ?

**03/03/2015 :** N N M propose de se charger de dupliquer le set « Rephy\_MancheAtlantique\_4X\_V2 » construit et optimisé durant la période décembre 2014 – janvier 2015 (cf. livrable n°2 ONEMA 2014), dans chaque répertoire dédié à chaque site sur le disque réseau, soit :

- zoophytoimage(\iota1)(Z:) BOULOGNE /comptage colonie/  
\_train\_Rephy\_MancheAtlantique\_4X-V2
- zoophytoimage (\iota1) (Z:) NANTES/comptage colonie/  
\_train\_Rephy\_MancheAtlantique\_4X-V2
- zoophytoimage (\iota1) (Z:) ARCACHON/comptage colonie/  
\_train\_Rephy\_MancheAtlantique\_4X-V2

**04/03/2015 :** AL propose d'organiser une audio-conférence ou visio-conférence. L'ordre du jour est le suivant :



- Livrables ONEMA : suites à donner (contenu, deadlines),
- Dénombrement des colonies :
  - définition des modalités du test,
  - niveau d'avancement de la méthode,
  - valorisation prévue,
- Pilotage scientifique IFREMER :
  - gestion des échanges,
  - demandes IFREMER ↔ UMONS.

**10/03/2015** : Compte-rendu de l'audio-conférence du 10/03/2015 (14h-16h)

- Participants : C. Belin, N-N. Masson, Ph. Grosjean, A. Lefebvre et G. Wacquet.
- Feu vert à NNM pour la copie des sets d'apprentissage dans chacun des répertoires dédiés à chaque site (copies pouvant prendre plusieurs heures).
- AL a fait remonter les problèmes de connexion au serveur dédié aux données du FlowCAM. Ce problème doit être résolu prochainement.
- PhG donne quelques indications sur la façon d'aborder le problème :
  - suppression des taxa non coloniaux (1 cellule par vignette),
  - suppression des taxa peu abondants (pas suffisamment de vignettes pour construire des modèles prédictifs pertinents),
  - suppression des taxa indénombrables (tels que *P. globosa*, *C. socialis*, etc.),
  - dénombrement d'une cinquantaine de vignettes par groupe.
  - autant que possible, une variabilité du nombre de cellules par colonie.
- NNM et GW proposent de fournir un cahier des charges, décrivant :
  - la liste des taxa à traiter dans le training set « Manche-Atlantique » v2,
  - l'utilisation du module de dénombrement des colonies.

***Partie 2***  
**Apprentissage actif**  
**Set d'apprentissage adaptatif**

## Table des matières

Introduction.....	3
Présentation des données.....	3
Deux méthodologies d'apprentissage actif.....	4
Complétion du set d'apprentissage APRES validation.....	5
Complétion du set d'apprentissage PENDANT validation.....	6
Résultats expérimentaux.....	7
Comparaison des méthodologies.....	7
Résultats détaillés .....	14
Utilisation en routine dans ZooPhytoImage.....	20
Conclusion.....	24
Bibliographie.....	26

## Introduction

Actuellement, la reconnaissance semi-automatisée des particules de phytoplancton dans un échantillon d'eau doit passer par la construction d'un set d'apprentissage reflétant la variabilité des espèces rencontrés dans le milieu naturel, mais également la variabilité morphologique des particules. Pour cela, il est nécessaire de réaliser un set d'apprentissage volumineux sur au moins un an pour reprendre les variations saisonnières et de le moduler par zone géographique.

Cependant, un tel set d'apprentissage mène à un outil de classification figé, et ne permet donc pas une adaptation temporelle aux échantillons analysés. L'apprentissage actif utilise les données validées en routine sur les échantillons pour enrichir le set d'apprentissage, et ainsi de l'adapter géographiquement, temporellement et saisonnièrement, de manière totalement transparente.

Afin de paramétrer au mieux le module de correction d'erreur qui avait été implémenté précédemment dans Zoo/PhytoImage version 5, et pour l'adapter au mieux à la réalité du travail en routine RePHY, nous étudions deux méthodes d'apprentissage actif : (i) complétion du set avec les données validées d'autres échantillons, (ii) complétion du set d'apprentissage avec les données validées à chaque étape de correction de l'erreur. Les scores de performance obtenus sont mis en évidence sur 19 échantillons prélevés en Manche Orientale.

## Présentation des données

Le couple objectif/cellule de flux choisi pour l'analyse des échantillons au FlowCAM est la combinaison 4X/300µm. Dans cette étude, 19 échantillons prélevés entre janvier et octobre 2014 en Manche Orientale sont utilisés. Voici la liste des échantillons triés par ordre chronologique (du plus ancien au plus récent) :

- 11 échantillons de Baie de Somme,
- 8 échantillons de Boulogne-sur-Mer.

[1]	"BL1.130.2014-01-14.300A4X.01"	→ Rhizosolenia + Proboscia
[2]	"ME2.132.2014-01-21.300A4X.01"	
[3]	"BL1.135.2014-03-07.300A4X.01"	
[4]	"ME2.137.2014-03-17.300A4X.01"	
[5]	"BL1.139.2014-03-27.300A4X.01"	→ Début bloom Phaeocystis
[6]	"ME2.142.2014-03-31.300A4X.01"	
[7]	"ME2.147.2014-04-28.300A4X.01"	→ Plein bloom Phaeocystis
[8]	"ME2.148.2014-05-14.300A4X.01"	→ Plein bloom Phaeocystis
[9]	"BL1.152.2014-05-16.300A4X.01"	
[10]	"Me2.159.2014-05-26.300A4X.01"	→ Fin bloom Phaeocystis + Début bloom PSN
[11]	"B11.160.2014-05-28.300A4X.01"	→ Plein bloom PSN
[12]	"Me2.166.2014-06-11.300A4X.01"	→ Fin bloom PSN + Bloom Chaetoceros
[13]	"Me2.169.2014-06-25.300A4X.01"	→ Rhizosolenia + Proboscia
[14]	"Me2.173.2014-07-16.300A4X.01"	
[15]	"BL1.174.2014-07-22.300A4X.01"	
[16]	"Me2.176.2014-08-25.300A4X.01"	→ Rhizosolenia + Proboscia
[17]	"Me2.177.2014-09-19.300A4X.01"	
[18]	"B11.178.2014-09-29.300A4X.01"	
[19]	"BL1.179.2014-10-27.300A4X.01"	

Afin d'avoir une idée sur la similarité du contenu des échantillons, nous calculons un coefficient de corrélation entre les abondances obtenues après validation totale des particules (Table 1).

Table 1 : Matrice de corrélation entre les abondances obtenues après validation totale des vignettes.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
1	1																			
2	0.99	1																		
3	0.98	0.96	1																	
4	0.98	0.98	0.96	1																
5	0.97	0.96	0.98	0.94	1															
6	0.91	0.87	0.96	0.87	0.96	1														
7	0.49	0.48	0.51	0.47	0.6	0.62	1													
8	0.84	0.84	0.82	0.82	0.87	0.82	0.85	1												
9	0.95	0.95	0.94	0.93	0.95	0.9	0.68	0.95	1											
10	0.23	0.23	0.24	0.22	0.25	0.25	0.09	0.19	0.24	1										
11	0.98	0.98	0.97	0.96	0.96	0.89	0.48	0.83	0.94	0.39	1									
12	0.51	0.5	0.53	0.48	0.54	0.55	0.31	0.4	0.47	0.29	0.52	1								
13	0.97	0.95	0.98	0.94	0.97	0.93	0.5	0.81	0.91	0.28	0.97	0.54	1							
14	0.93	0.91	0.96	0.89	0.95	0.94	0.5	0.78	0.88	0.24	0.93	0.55	0.98	1						
15	0.99	0.99	0.97	0.97	0.97	0.9	0.49	0.84	0.94	0.32	0.99	0.55	0.97	0.94	1					
16	0.6	0.53	0.72	0.53	0.68	0.83	0.42	0.46	0.55	0.17	0.59	0.51	0.72	0.77	0.6	1				
17	0.73	0.7	0.78	0.69	0.78	0.8	0.47	0.59	0.68	0.18	0.72	0.77	0.78	0.8	0.74	0.74	1			
18	0.65	0.58	0.76	0.58	0.73	0.86	0.44	0.5	0.6	0.18	0.64	0.5	0.76	0.81	0.65	0.99	0.79	1		
19	0.62	0.55	0.74	0.56	0.7	0.85	0.43	0.48	0.57	0.17	0.62	0.45	0.74	0.79	0.62	0.99	0.74	0.99	1	

Les résultats présentés dans ce tableau peuvent être facilement interprétés. Ici, nous remarquons une faible corrélation entre l'échantillon 7 et les autres. Ceci s'explique principalement par la dominance importante de l'espèce *Phaeocystis globosa* (plein bloom) dans cet échantillon, et l'absence de cette micro-algue dans les autres. Nous notons, cependant, qu'à partir de l'échantillon 5 (qui correspond au début du bloom), le coefficient de corrélation augmente. Ces mêmes remarques peuvent être formulées pour les échantillons 10 (bloom de *Pseudo-Nitzschia*) et 12 (bloom de *Chaetoceros*).

## Deux méthodologies d'apprentissage actif

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.

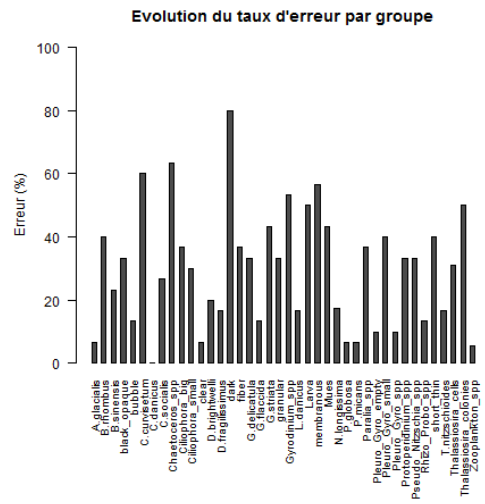
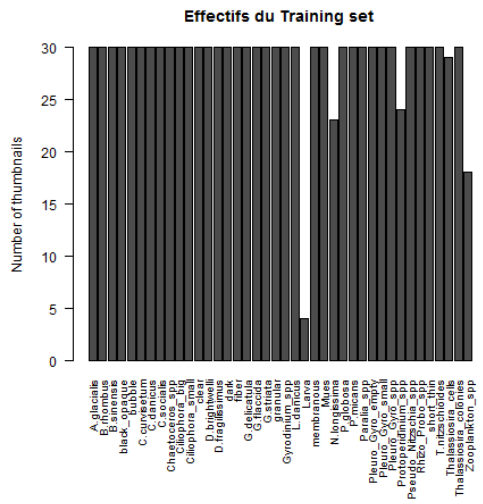
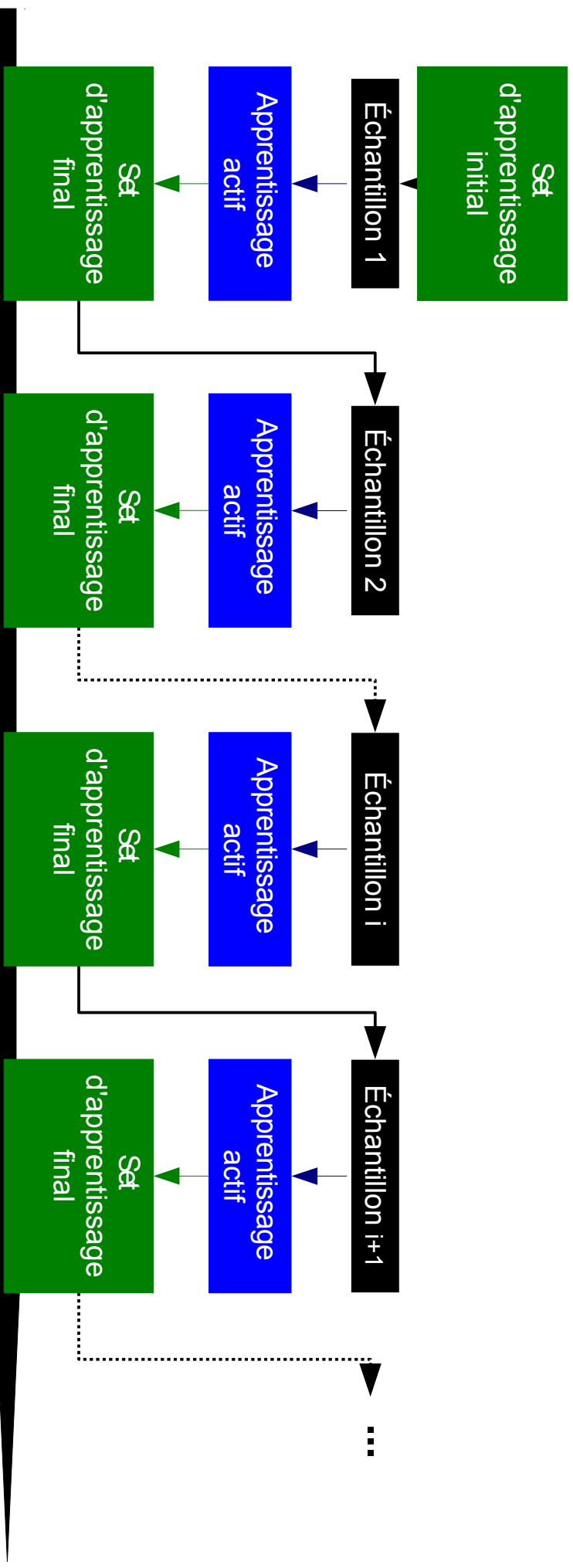


Figure 1 - Effectifs du set d'apprentissage « initial » et taux d'erreur par groupe (Random Forest).

## Complétion du set d'apprentissage APRES validation



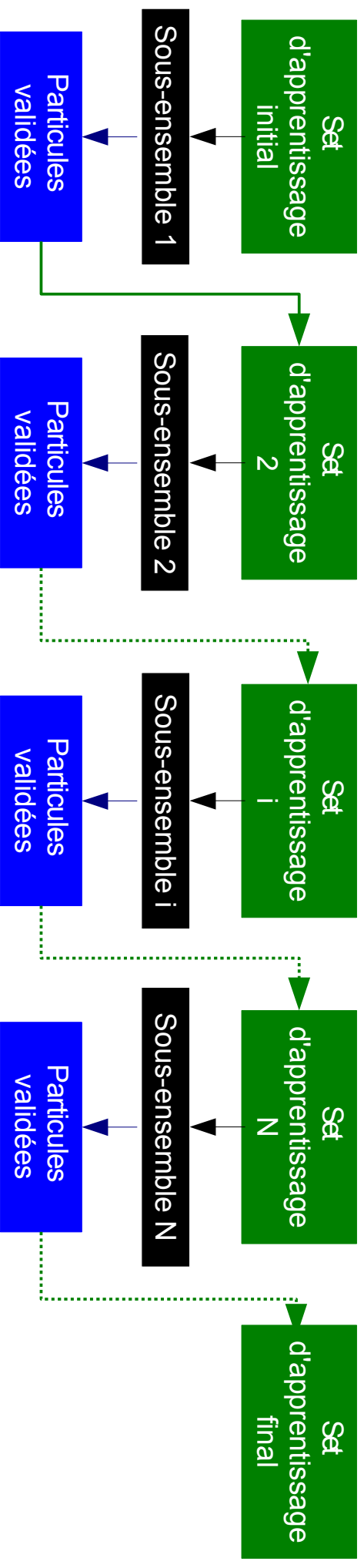
### Traitement chronologique des échantillons

Figure 2 – Complétion du set d'apprentissage APRES validation.

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 (Figure 2), 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 (ici, 5%) puis on ajoute la même quantité de nouveaux items.

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

## Complétion du set d'apprentissage PENDANT validation



## Succession des étapes de validation des données

Figure 3 – Complétion du set d'apprentissage PENDANT validation.

L'objectif est de compléter le set d'apprentissage de manière automatisée au fur et à mesure des étapes de validation (Figure 3). Pour cela, les règles mises en place et citées précédemment sont reprises. L'idée est donc la suivante : après chaque étape de validation, nous complétons le training set avec les données validées à l'étape précédente ; puis nous régénérons l'outil de reconnaissance sur base du training set modifié ; pour enfin prédire de nouveaux les classes d'appartenance des particules.

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



# Résultats expérimentaux

## Comparaison des méthodologies

### Complétion du set d'apprentissage PENDANT la validation (training set initial et modifié)

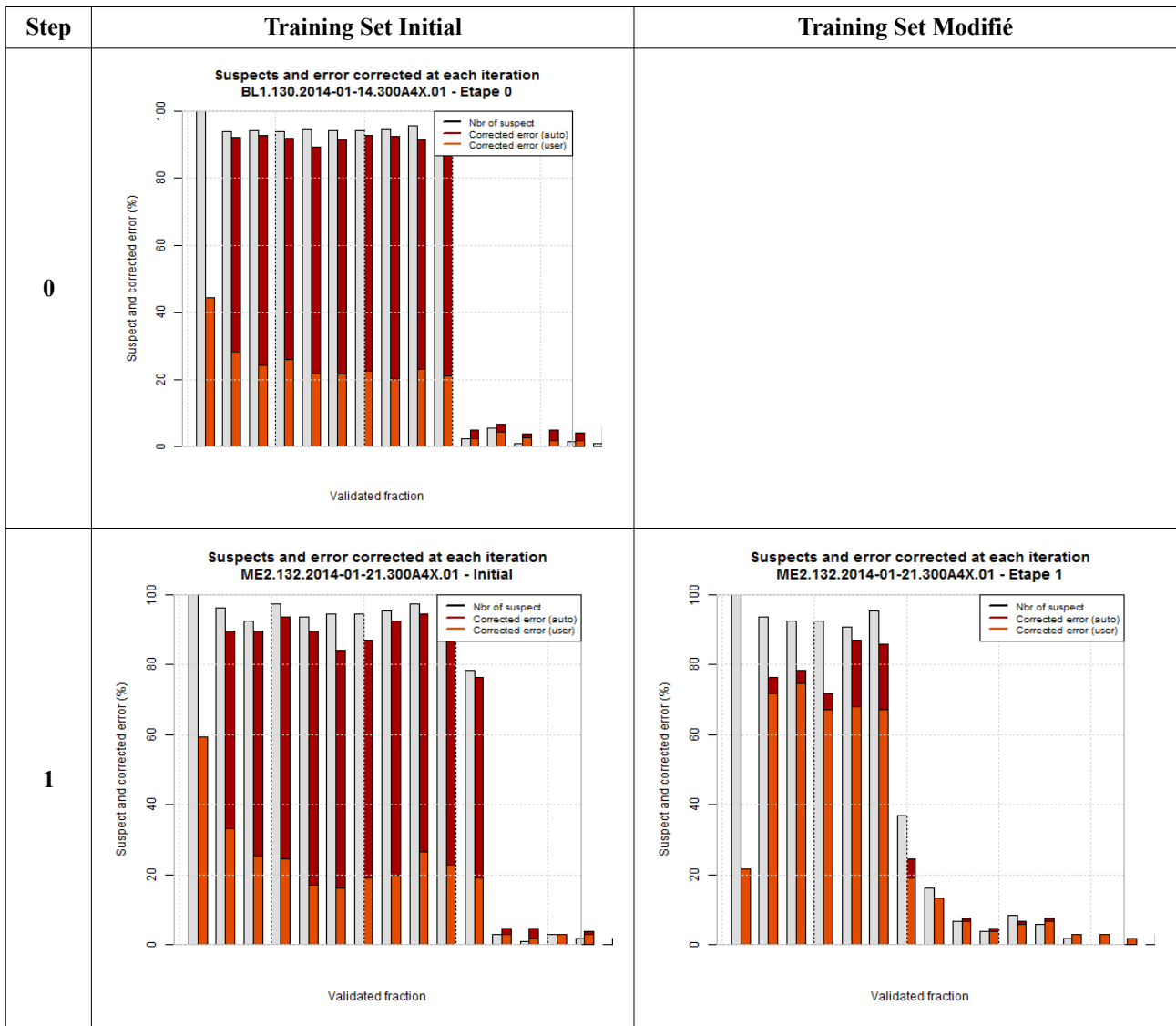
Après chaque étape de validation :

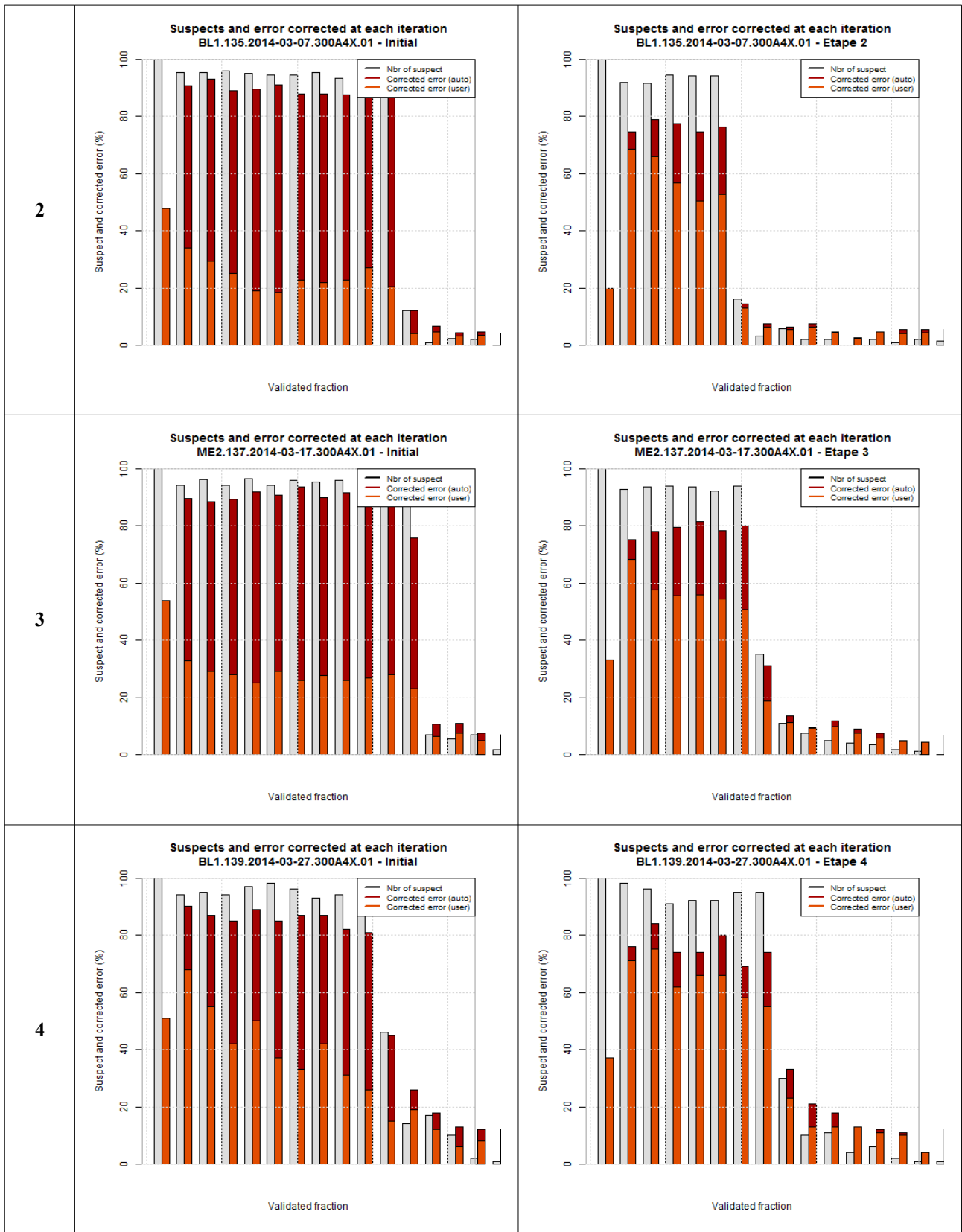
- complétion du training set avec les données validées à l'étape précédente,
- régénération de l'outil de reconnaissance sur base du training set modifié,
- affichage de l'erreur corrigée automatiquement par l'outil de reconnaissance créé (représentée en **rouge** sur les graphes) et de l'erreur corrigée manuellement par l'utilisateur (représentée en **orange** sur les graphes).

### Complétion du set d'apprentissage APRES validation (2nde colonne)

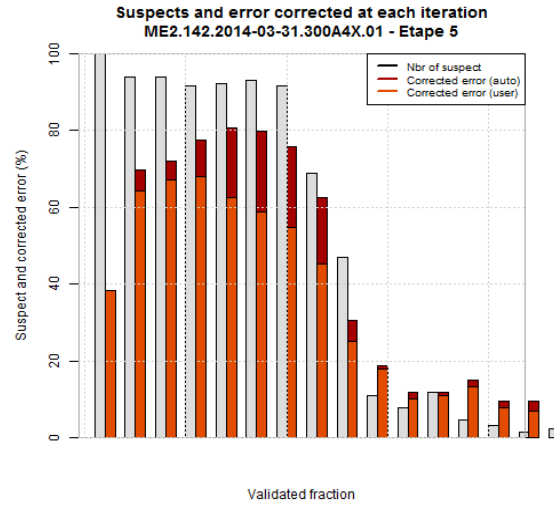
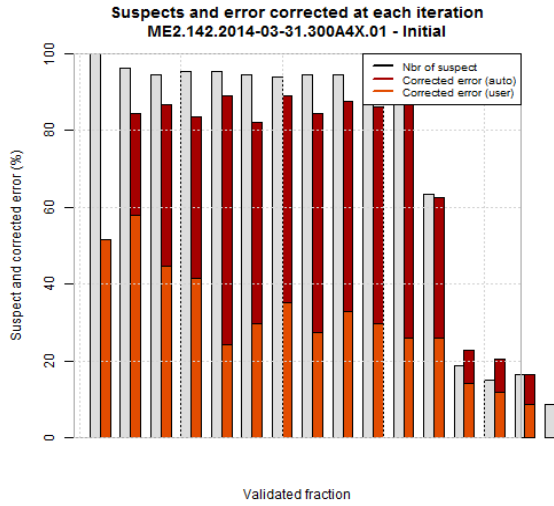
Après validation d'un échantillon :

- complétion du training set initial avec les données validées,
- ré-génération de l'outil de reconnaissance sur base du training set modifié,
- application sur un nouvel échantillon,
- affichage de l'erreur corrigée automatiquement par l'outil de reconnaissance créé (représentée en **rouge** sur les graphes) et de l'erreur corrigée manuellement par l'utilisateur (représentée en **orange** sur les graphes).

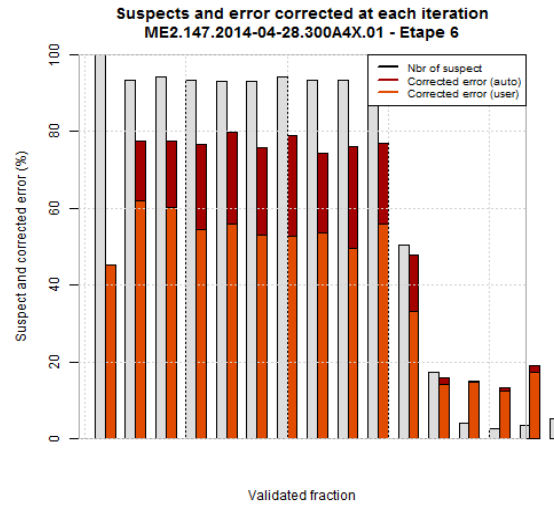
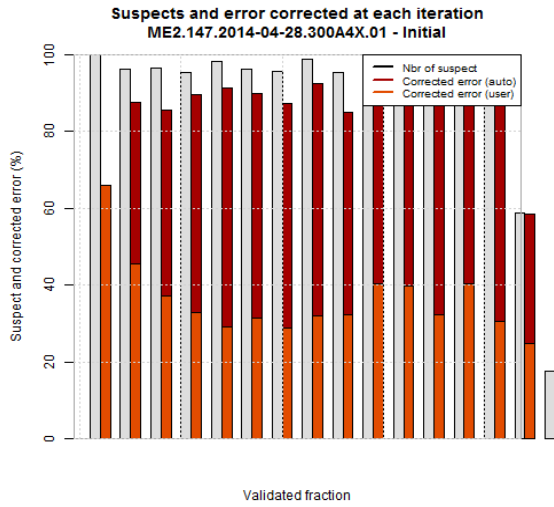




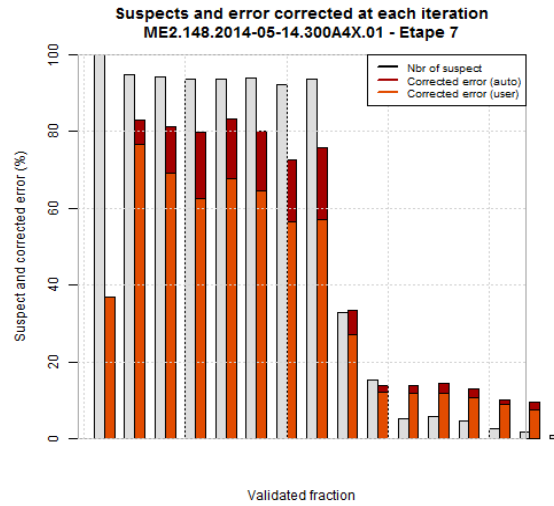
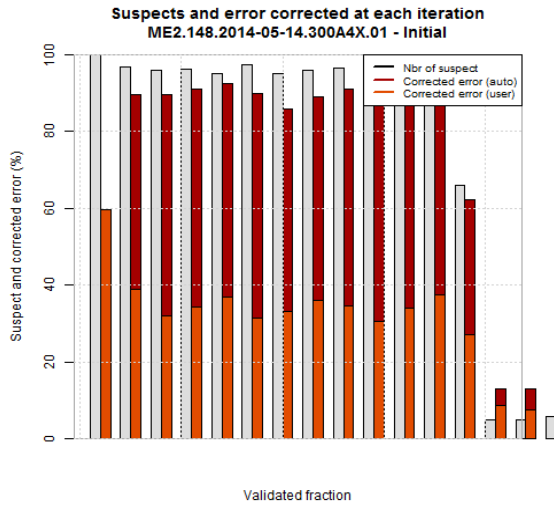
5



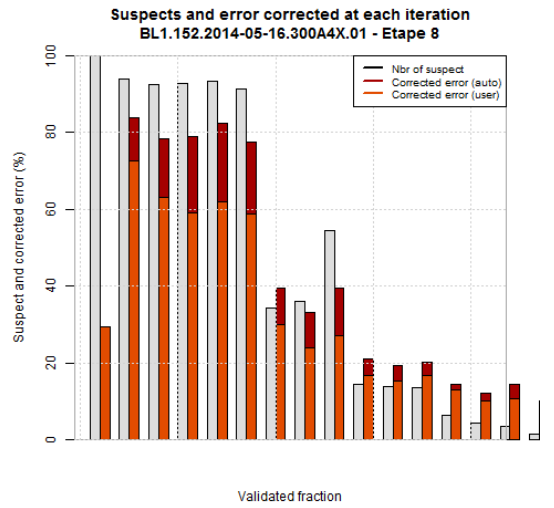
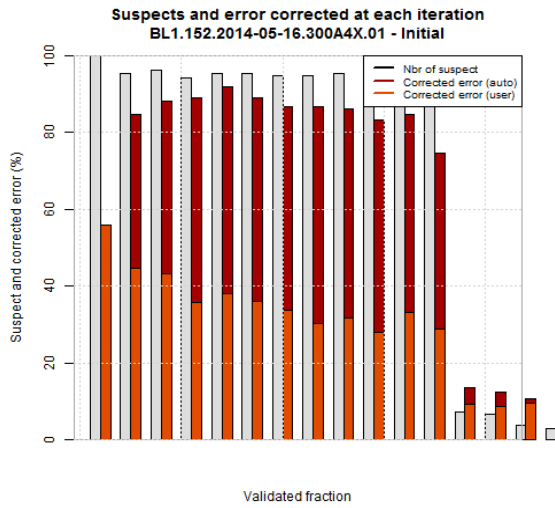
6



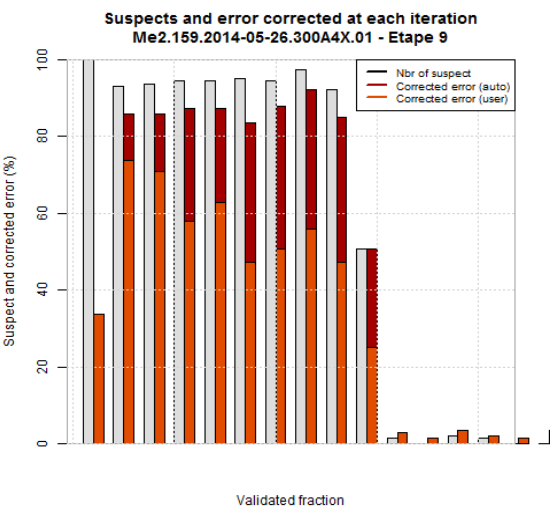
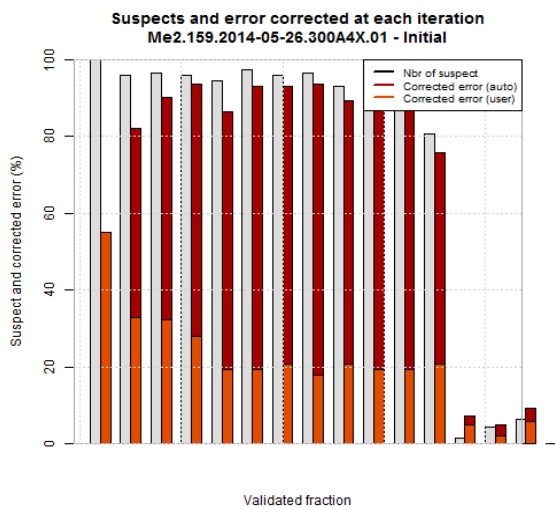
7



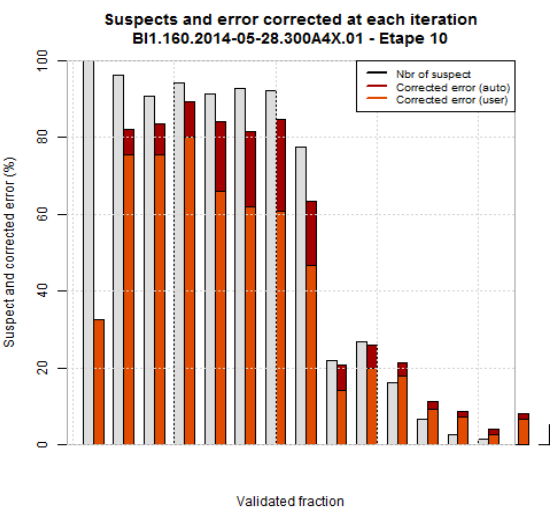
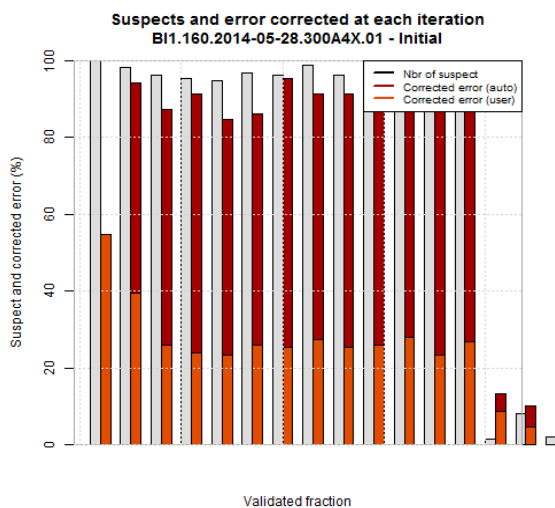
8



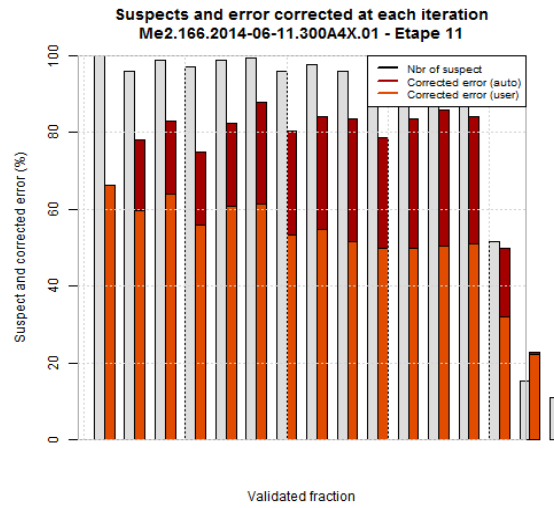
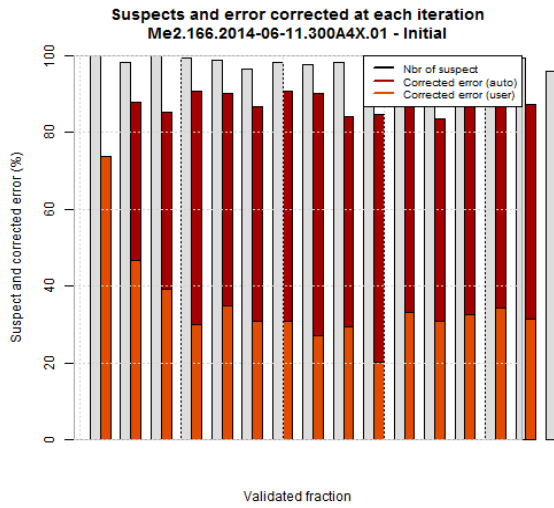
9



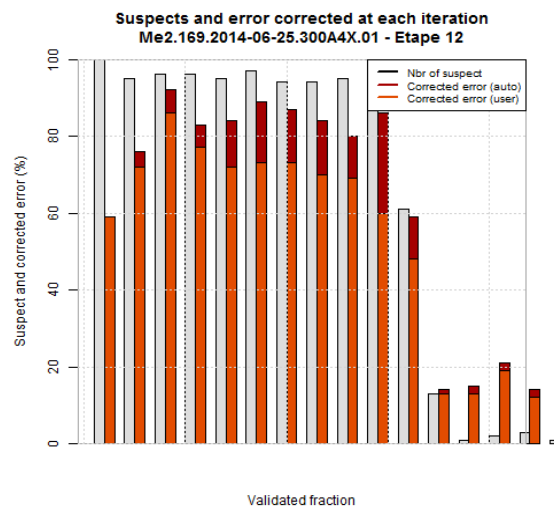
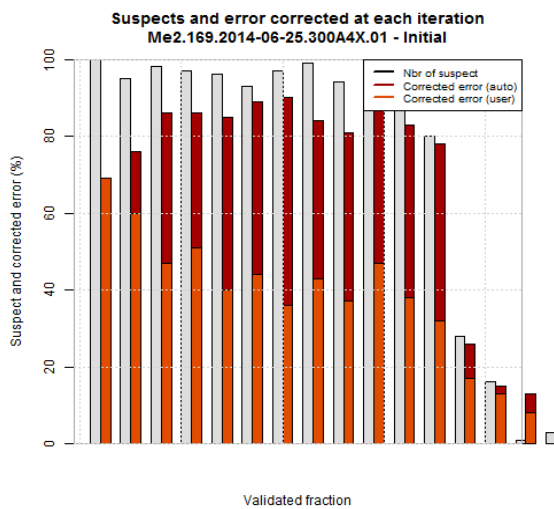
10



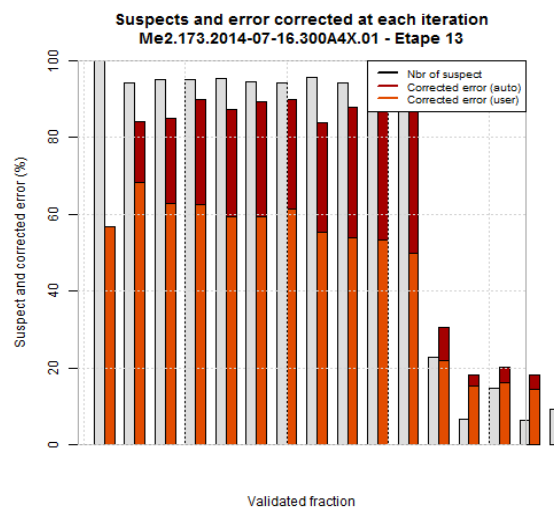
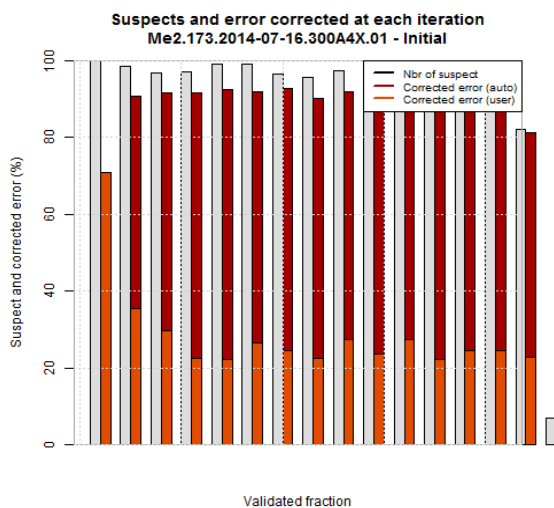
11

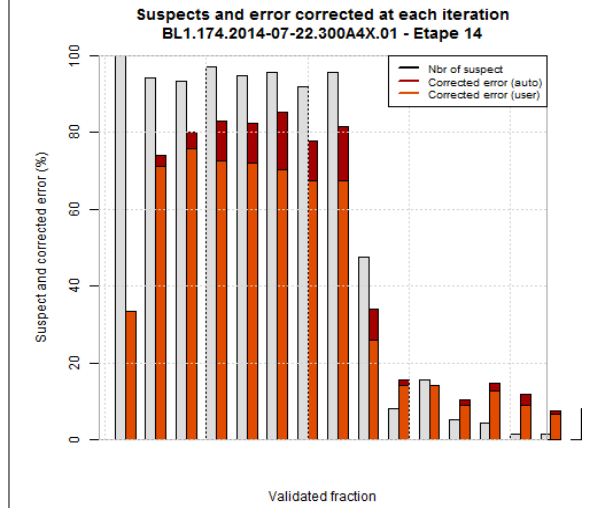
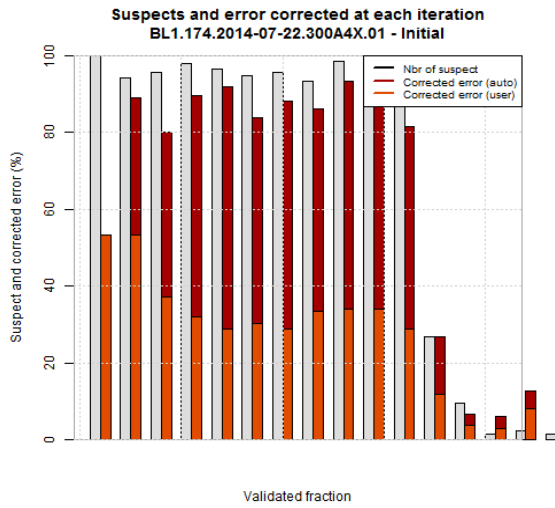


12

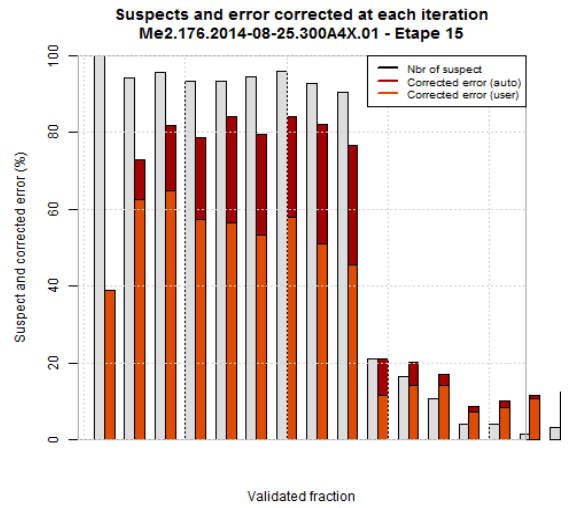
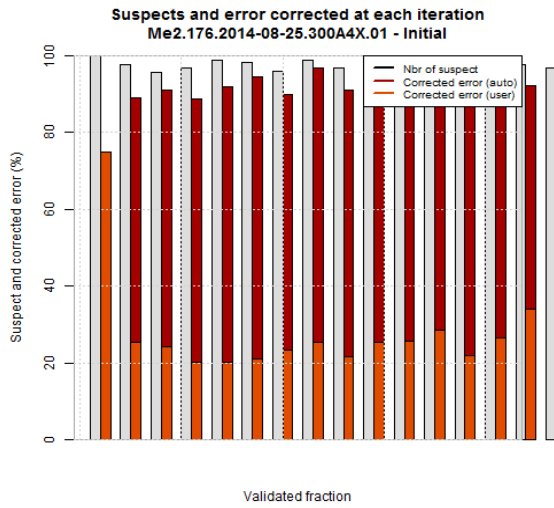


13

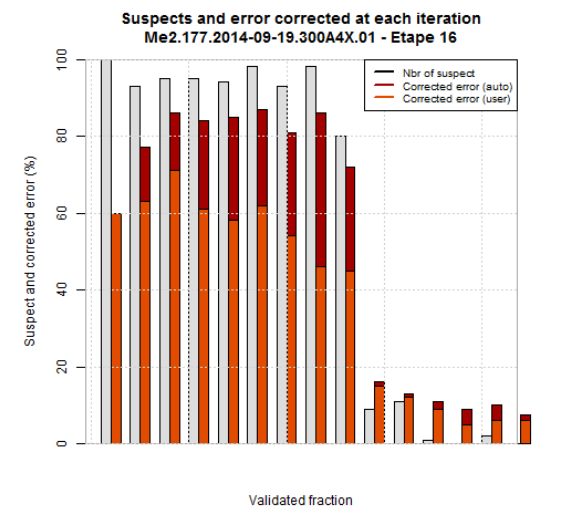
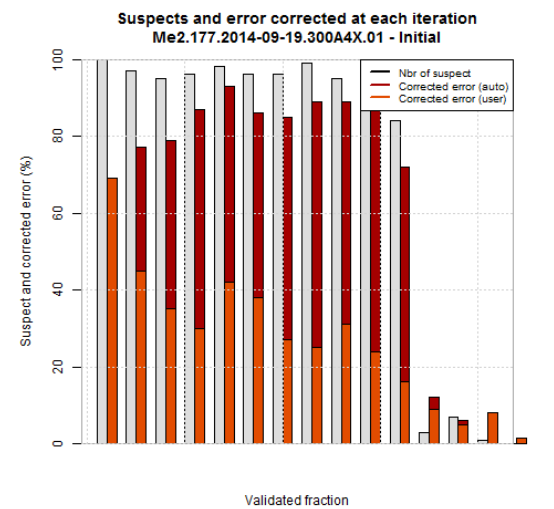




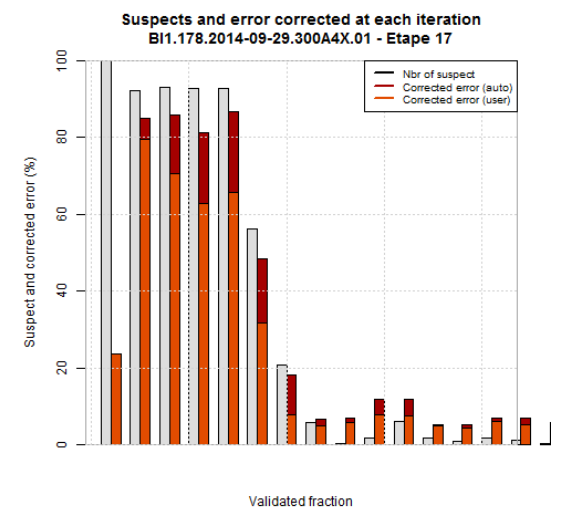
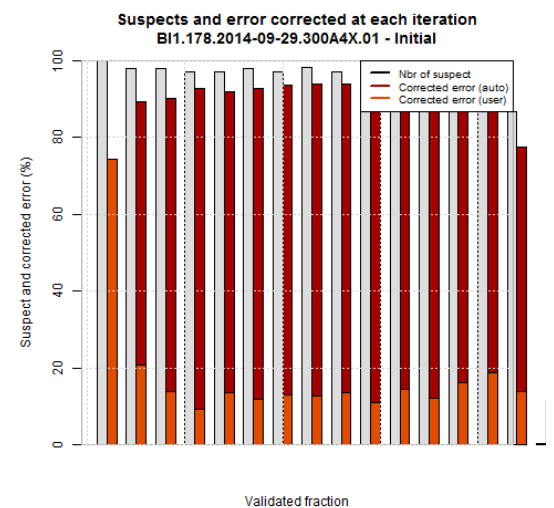
15

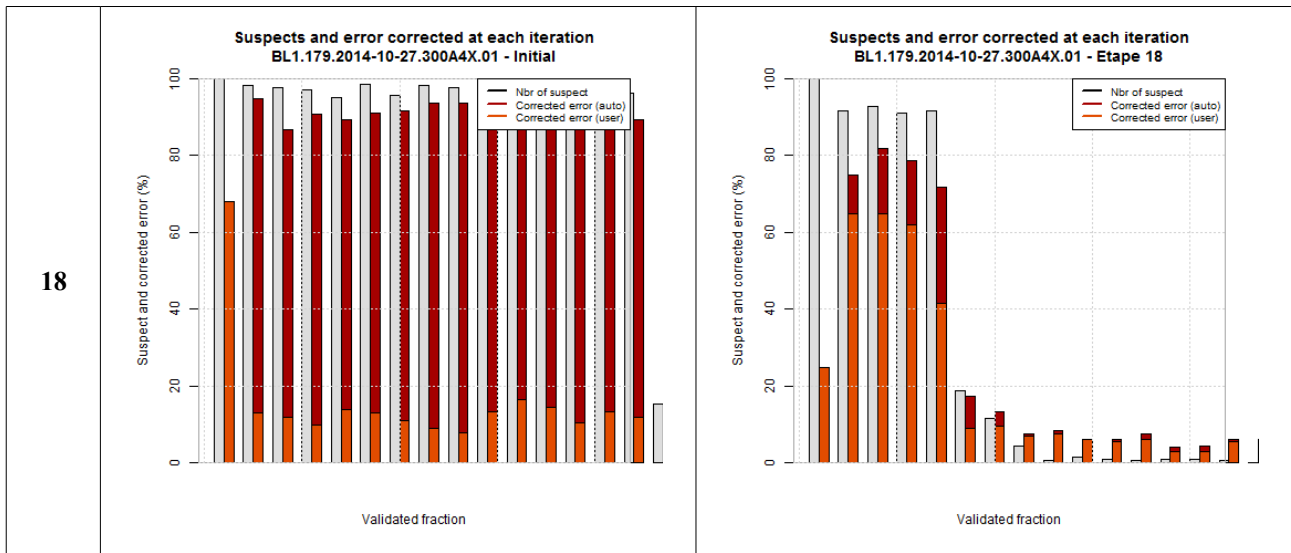


16



17





### Complétion du set d'apprentissage APRES validation (2nde colonne)

Globalement, les résultats présentés montrent une nette réduction du nombre de vignettes suspectes à valider après complétion du set d'apprentissage avec les données validées des échantillons précédents.

### Complétion du set d'apprentissage PENDANT la validation

- Avec le set d'apprentissage initial (**1ère colonne**)
  1. 1ère étape de validation : l'utilisateur doit corriger la totalité du sous-ensemble de vignettes proposé,
  2. étapes suivantes de validation : il est possible d'observer une diminution significative (dans certains cas, + de 50%) des erreurs de prédiction des vignettes considérées comme « suspectes ». En effet, avec l'apprentissage actif, ces dernières sont corrigées automatiquement par les outils de reconnaissance recalculés à chaque étape.
- Avec le set d'apprentissage modifié (initial + échantillons précédents) (**2nde colonne**)
  1. 1ère étape de validation : l'utilisateur doit corriger la totalité du sous-ensemble de vignettes proposé,
  2. étapes suivantes de validation : **perte de l'intérêt de la méthode** puisque l'utilisateur doit corriger la quasi-totalité des vignettes suspectes proposées.

#### En résumé...

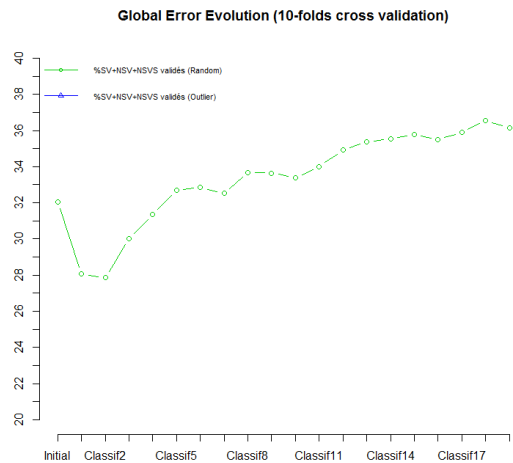
- Complétion du set d'apprentissage APRES validation :  
**Peu de vignettes** sont proposées à la validation, mais l'utilisateur doit en **corriger une grande partie**.
- Complétion du set d'apprentissage actif PENDANT la validation :  
**Beaucoup de vignettes** sont proposées à la validation, mais l'utilisateur ne doit en **corriger qu'une petite partie**.

Dans la suite de l'étude, nous choisissons donc de retenir la méthodologie de complétion du set d'apprentissage APRES validation. En effet, cette méthode nécessite une durée de traitement beaucoup moins longue et permet de ne pas noyer l'utilisateur dans un volume important de vignettes suspectes à valider.

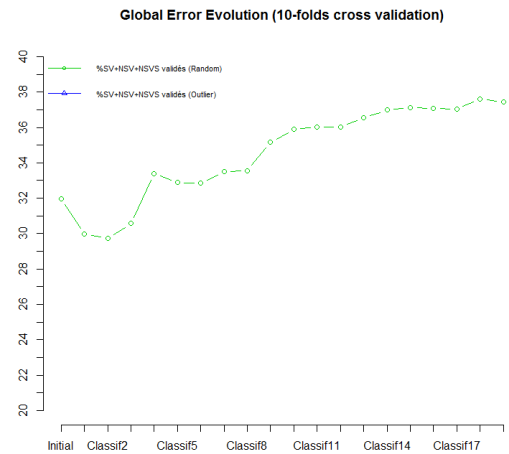
### **Résultats détaillés**

En observant les performances de l'apprentissage actif en détail, une tendance peut être extraite des scores de prédiction par validation croisée (10-fois).





(a)



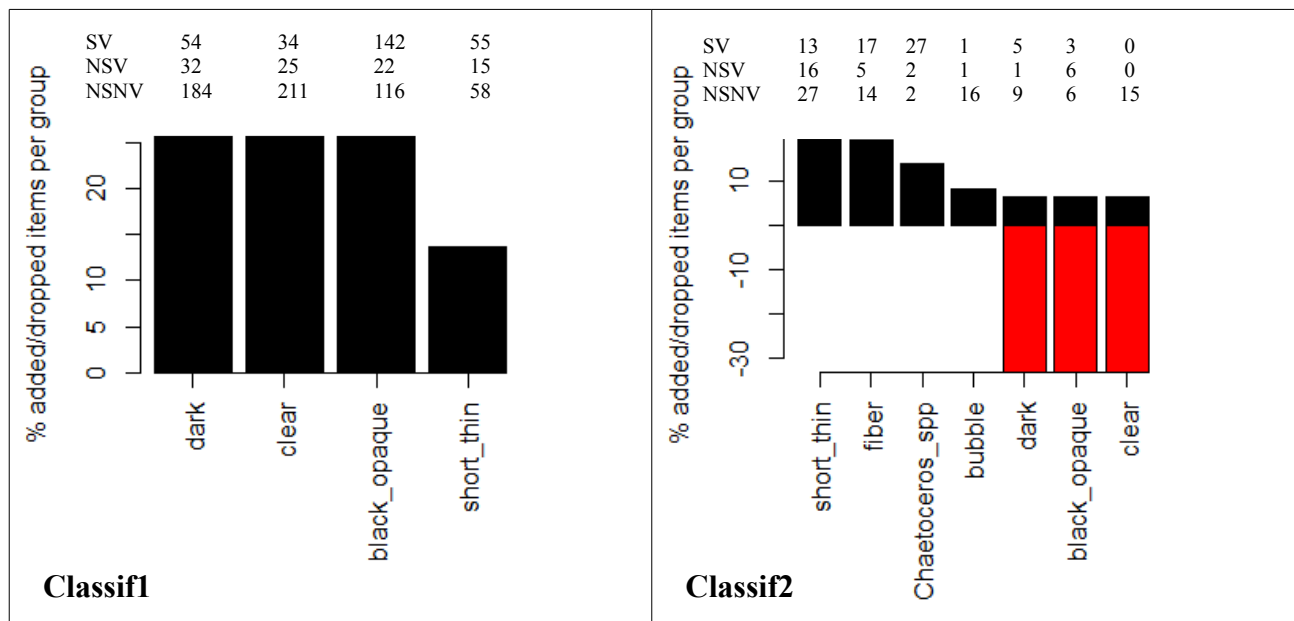
(b)

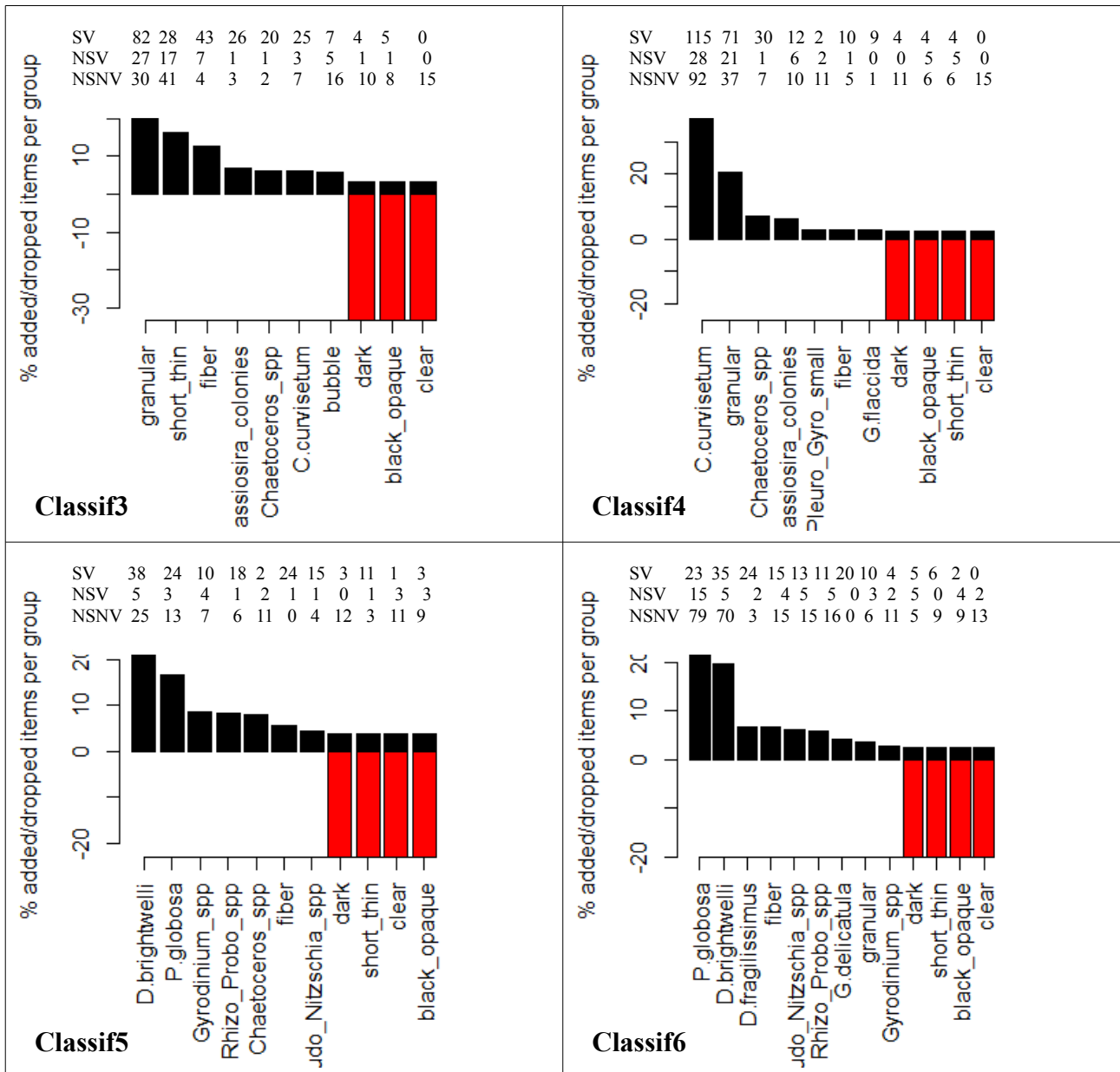
Figure 4 – Pourcentage d'erreur de classification par validation croisée (10-fois) au fur et à mesure de la complétion du set d'apprentissage initial. (a) : traitement des échantillons par ordre chronologique (échantillon 1 à 19) ; (b) : traitement décalé des échantillons (échantillon 5 à 19, puis 1 à 4).

Les résultats présentés sur les graphes (a) et (b) de la Fig. 4 montrent que la complétion du set d'apprentissage avec les données validées d'un nombre important d'échantillons tend à dégrader les performances de reconnaissance automatique par validation croisée.

Pour comprendre cette évolution, nous nous intéressons aux particules ajoutées dans le set d'apprentissage à chaque étape. Nous choisissons donc de représenter l'évolution des effectifs (par groupe) selon le protocole suivant :

- espèces retenues : celles représentant 80% de l'ajout/suppression des vignettes.
- représentation des pourcentages de vignettes ajoutées/supprimées.





Grâce à ces graphes et aux matrices de confusion ci-dessous (Figs. 5-11), nous pouvons formuler les observations suivantes :

1. **Classif1** : ajout de particules détritiques uniquement (qui représentent la majorité de la composition des échantillons) et qui sont à l'origine des confusions les plus importantes (surtout entre les catégories « black\_opaque » et « dark »).
2. **Classif2** : même que Classif1, mais ajout de « Chaetoceros\_spp » également. Apparition de confusions importantes entre « Chaetoceros\_spp » et les particules détritiques (particulièrement « short\_thin ») qui n'existaient pas initialement.
3. **Classif3** : mêmes observations que Classif2, mais avec « C.curvisetum » et « Thalassiosira\_colonies ».
4. **Classif4, Classif5, Classif6, etc.** : mêmes observations.

Actual // Predicted	01	04	07	10	13	16	19	22	25	28	31	34	37	40
P.micans 01	28	1	1											
dark 02	2	12	5	1	1	1	1							
Thalassiosira_cells 03	1	3	22	1										
Ciliophora_small 04	3	1	18	1	4									
Gyrodinium_spp 05	1		15	2	3	1	1	1	3	1				
Protoperdinium_spp 06			1	5	14	2				1	1			
black_opaque 07	1			24	2									
Zooplankton_spp 08				1	17									
G.flaccida 09				26	2	1					1			
halassiosira_colonies 10	3	1	2		1	12	2	2					1	1
granular 11				1	2	19	2	1						
B.rhombus 12	1		2	1	1	1	18	4						
B.sinensis 13			1	1		2	2	24						
membranous 14				2		13	6			2	2	1	2	1
Mues 15						8	17			4	1			
N.longissima 16								15	7					
'seudo_Nitzschia_spp 17							4	22						
Pleuro_Gyro_spp 18							28	2						
Paralia_spp 19	1	2		1			3	20	2					
Pleuro_Gyro_small 20							4	3	18					
C.socialis 21						2	4		11	3	2	6		
fiber 22						3	1		4	19	1			
G.striata 23				2					1	1	14	2	1	1
P.globosa 24									1	3	19		3	1
Ciliophora_big 25		1		1					3		21	3		
D.brightwelli 26							1				26			
clear 27	1			1							27	1		
Pleuro_Gyro_empty 28					1						1	25		
C.danicus 29											30			
A.glacialis 30						1					27	2		
short_thin 31	1					1	4	1	1	2		11	18	
bubble 32			1	1	1								26	
Larva 33									3		1			
C.curvisetum 34									2		1			
T.nitzschioides 35									1	1				
L.danicus 36						1			2					
Rhizo_Probo_spp 37									1					
D.fragilissimus 38													22	1
Chaetoceros_spp 39	1			1	1			1	1	1	2	1	1	9
G.delicatula 40	1							1	3	2			2	1

Figure 5 – Matrice de confusion pour le classifieur Classif\_initial.

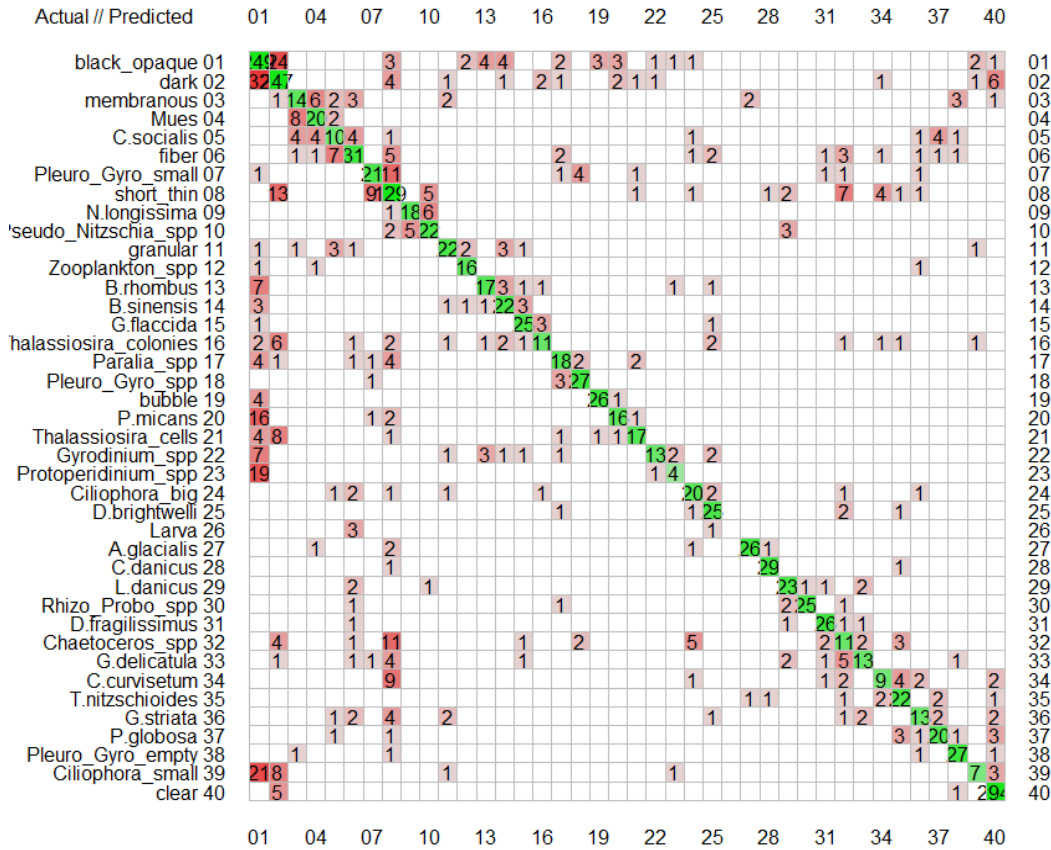


Figure 6 – Matrice de confusion pour le classifieur Classif1.

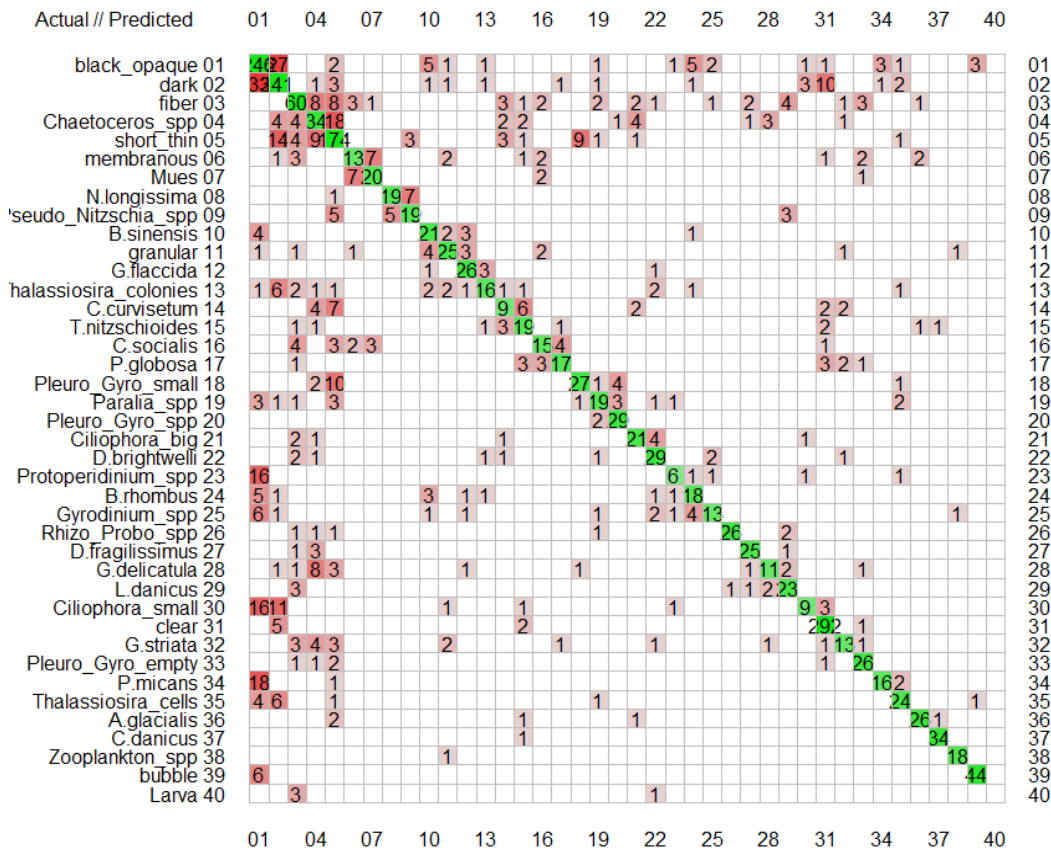


Figure 7 – Matrice de confusion pour le classifieur Classif2.



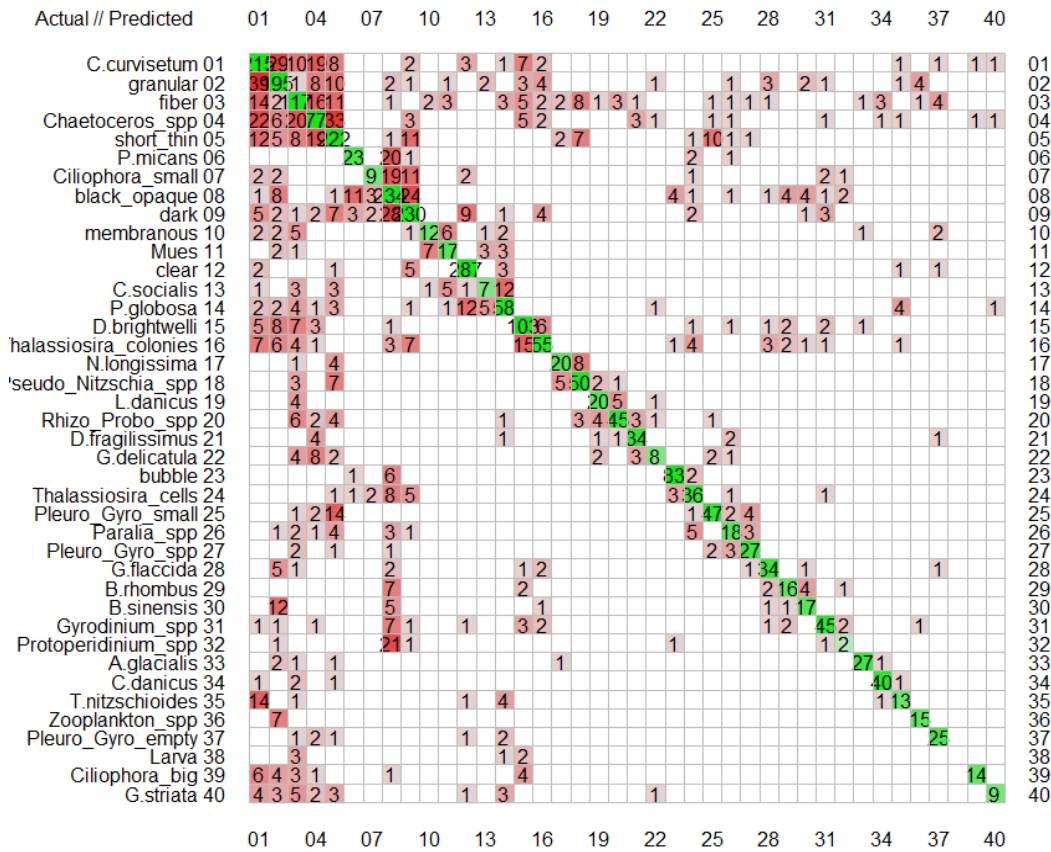


Figure 10 – Matrice de confusion pour le classifieur Classif5.

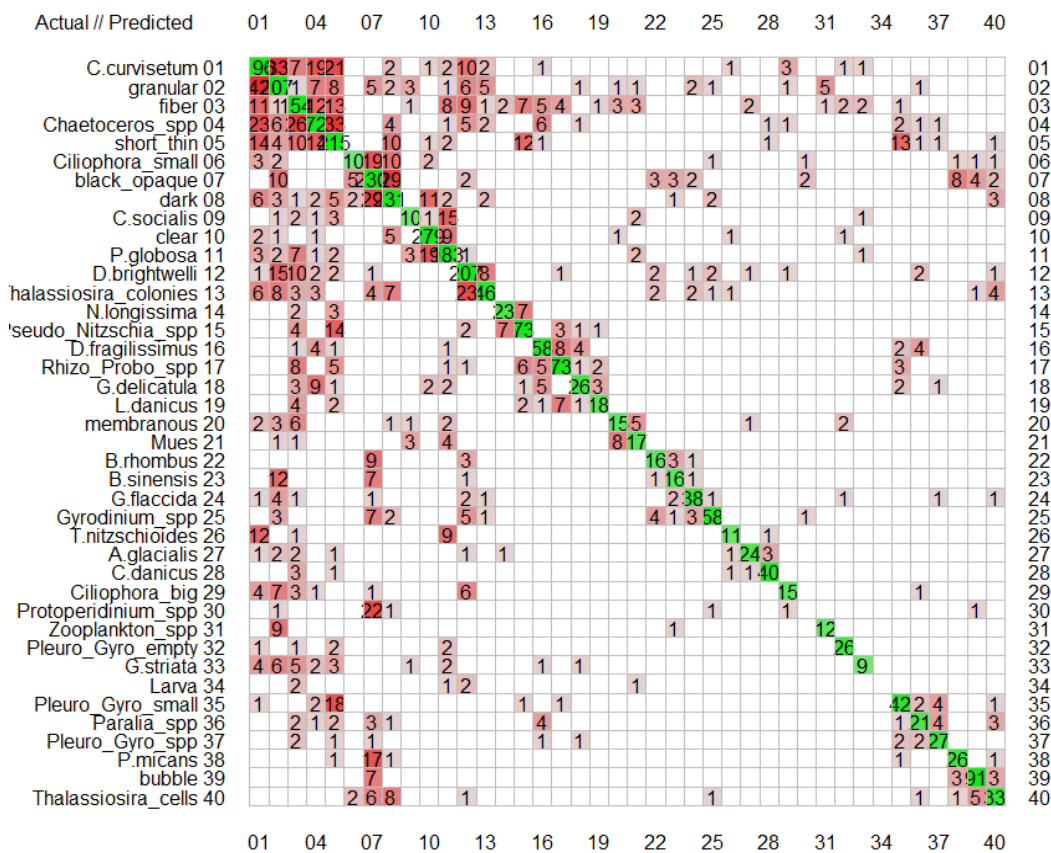


Figure 11 – Matrice de confusion pour le classifieur Classif6.

### Évolution de l'erreur globale du training set (par validation croisée)

- Diminution de l'erreur globale avec l'ajout de vignettes provenant du(es) premier(s) échantillon(s) : la quantité plus importante des particules détritiques par rapport aux restes des particules phytoplanctoniques, entraîne une augmentation des confusions entre elles mais une stabilité relative des confusions avec les particules phytoplanctoniques.
- Lorsque l'on ajoute un nombre important de particules phytoplanctoniques provenant d'autres échantillons, les confusions entre particules détritiques et celles-ci augmentent. Cela peut s'expliquer, en partie, par l'ajout de vignettes problématiques. Bien que cette proportion soit faible (ici, 5%), ce sont ces vignettes qui sont en grande partie responsables des confusions observées.

### **Utilisation en routine dans ZooPhytoImage**

A la vue des résultats précédents, il semble préférable d'utiliser la méthodologie d'apprentissage actif APRES validation. Cependant, en routine, il paraît judicieux que l'utilisateur sélectionne lui-même les échantillons qu'il souhaite intégrer au set (cf. Fig. 12). Ce processus présente plusieurs avantages :

- pas d'obligation de traitement chronologique des échantillons,
- apport d'informations contextuelles (quelles espèces ? à ce moment de l'année ? dans cette zone géographique...?),
- si utilisation d'un nombre faible d'échantillons contextuels, pas (ou peu) de dégradation des scores de performances par validation croisée,
- si utilisation d'un set d'apprentissage contenant un nombre quasi-identique de vignettes par groupe, pondération des groupes présents dans les échantillons contextuels (donc, censés être présents dans l'échantillon à traiter),
- répétabilité de l'analyse : set d'apprentissage initial figé, et contexte (échantillons contextuels) sauvegardé.

#### **Vignettes disparates**

- différents taxa,
- différents points,
- différentes dates,

#### **Abondances quasi-similaires.**

Set  
d'apprentissage  
initial

Échantillon 1  
Échantillon 2  
...  
Échantillon i

Échantillons contextuels  
- même saison,  
- même zone géographique,  
- etc.  
que l'échantillon à traiter.

Apprentissage  
actif

#### **Ajout des particules validées**

- dans le set d'apprentissage initial,
- pour chaque échantillon contextuel.

Set  
d'apprentissage  
final

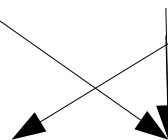
#### **Reflète le contenu probable de l'échantillon à traiter**

- composition phytoplanctonique,
- poids plus important pour certains taxa.

Échantillon  
à traiter

*Figure 12 – Méthodologie utilisée dans ZooPhytoImage pour l'utilisation de l'apprentissage actif en routine.*

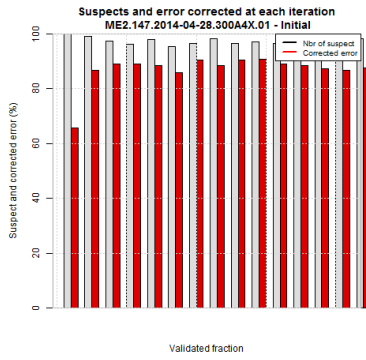
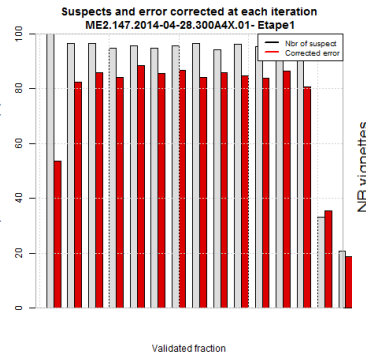
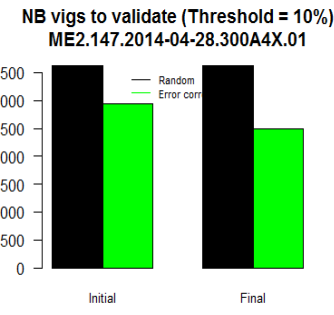
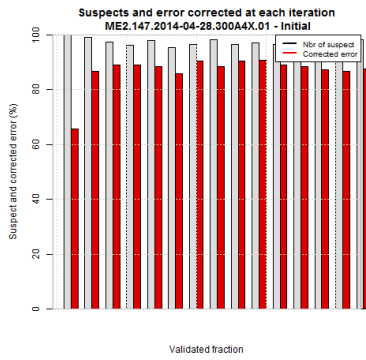
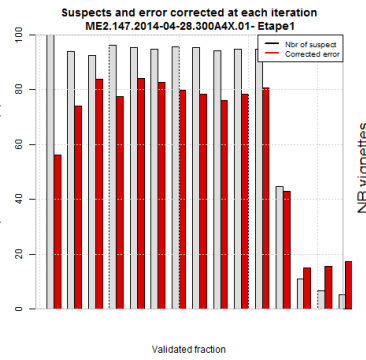
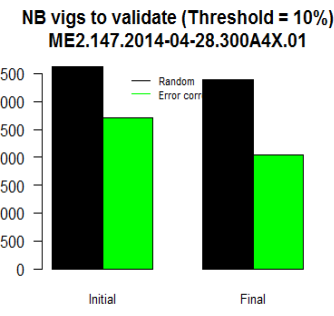
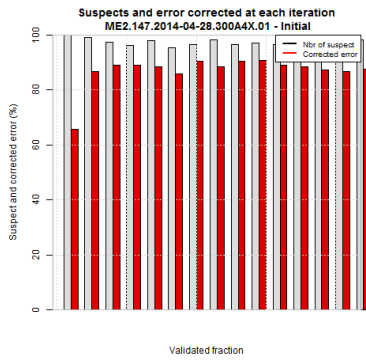
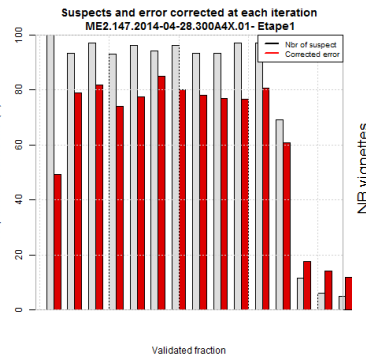
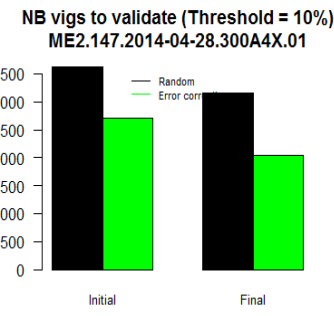
### **Résultats expérimentaux préliminaires**





Afin de mettre en évidence les performances liés à l'apprentissage actif en routine, nous choisissons d'appliquer la méthode sur un échantillon d'avril 2014 correspondant à un bloom de *Phaeocystis globosa* (en rouge dans la liste des échantillons). Les données contextuelles sélectionnées par l'utilisateur (en bleu dans la liste) sont les suivantes : un premier échantillon hors bloom, un second échantillon correspondant au début de bloom et un troisième échantillon en plein bloom.

- [4] "ME2.137.2014-03-17.300A4X.01" | Hors Bloom Phaeocystis
- [6] "ME2.142.2014-03-31.300A4X.01" | Début du bloom Phaeocystis
- [7] "ME2.147.2014-04-28.300A4X.01" → **BLOOM PHAEOCYSTIS**
- [8] "ME2.148.2014-05-14.300A4X.01" | Bloom Phaeocystis

Échantillons contextuels	Résultats initiaux Sans apprentissage actif	Résultats finaux Après apprentissage actif	Nombre de vignettes à valider
ME2.137.2014-03-17.300A4X.01			
ME2.142.2014-03-31.300A4X.01			
ME2.147.2014-04-28.300A4X.01			



ME2.137.2014-03-17.300A4X.01

ME2.142.2014-03-31.300A4X.01

ME2.148.2014-05-14.300A4X.01

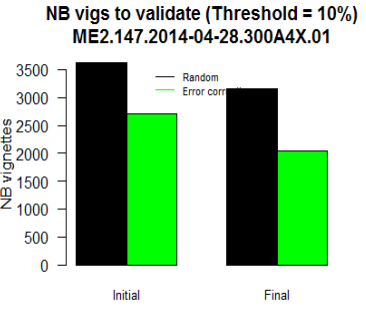
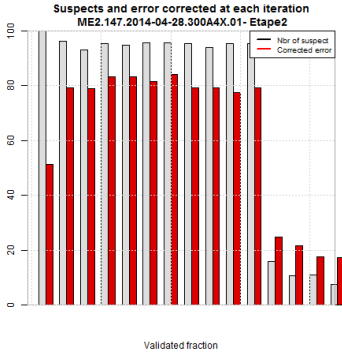
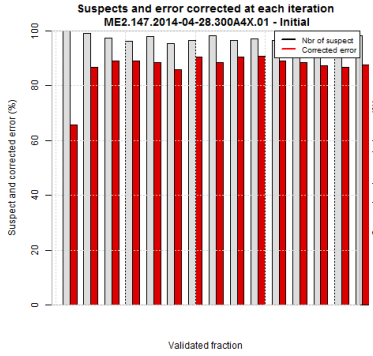
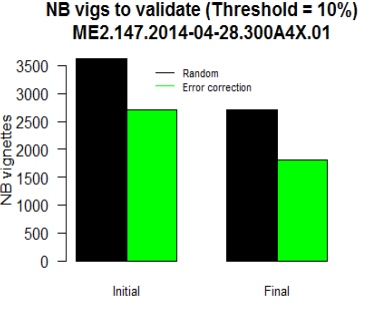
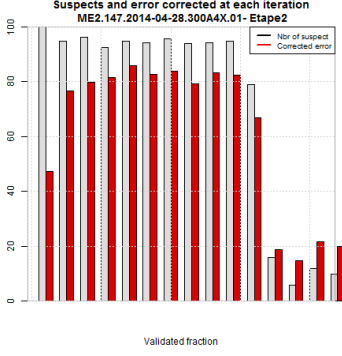
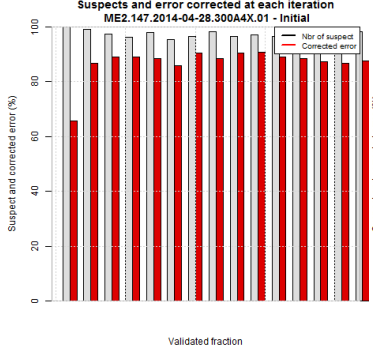
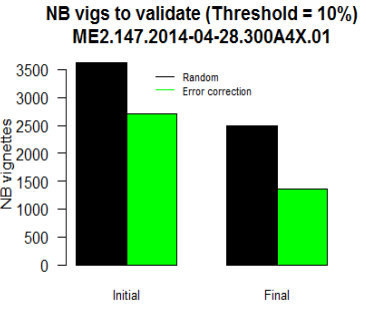
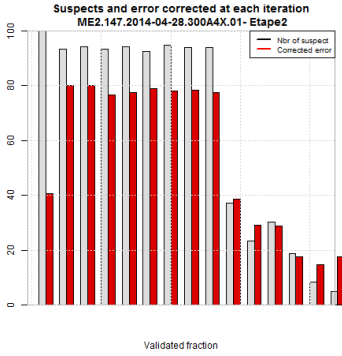
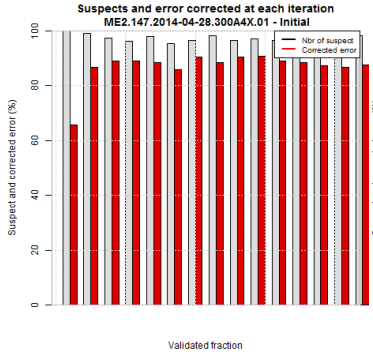
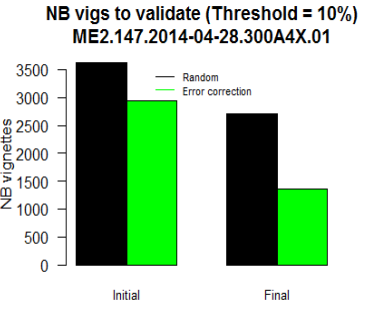
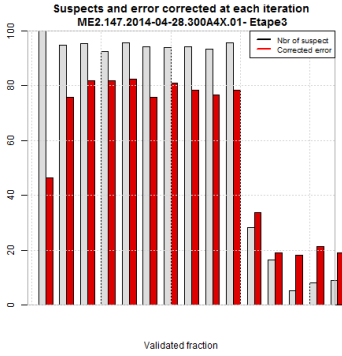
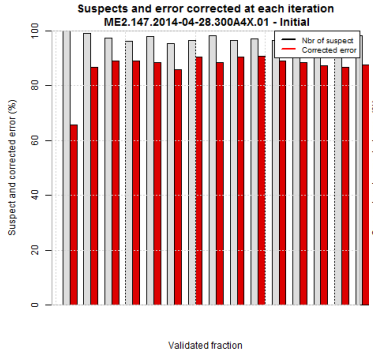
ME2.137.2014-03-17.300A4X.01

ME2.142.2014-03-31.300A4X.01

ME2.148.2014-05-14.300A4X.01

ME2.137.2014-03-17.300A4X.01

ME2.142.2014-03-31.300A4X.01



Les résultats présentés sur les graphes représentant le nombre de vignettes à valider peuvent être interprétés de la manière suivante :

- En noir : nombre de vignettes à valider si le tirage est aléatoire, avant (« *Initial* ») et après (« *Final* ») complétion avec les données validées de(s) échantillon(s) contextuel(s).
- En vert : nombre de vignettes à valider si le tirage est basé sur les vignettes « suspectes », avant (« *Initial* ») et après (« *Final* ») complétion avec les données validées de(s) échantillon(s) contextuel(s).

Ici, nous pouvons noter que le nombre de vignettes à valider après apprentissage actif, est toujours plus faible que le celui avant complétion du set. De plus, lorsque les échantillons contextuels sont sélectionnés de manière pertinente, cette réduction est alors beaucoup plus significative.

Nous présentons maintenant les différentes combinaisons d'échantillons contextuels pour l'apprentissage actif, triées par ordre croissant de vignettes à valider :

1. ME2.142 + ME2.148
2. ME2.137 + ME2.142 + ME2.148
3. ME2.137 + ME2.148
4. ME2.137 + ME2.142
5. ME2.148
6. ME2.142
7. ME2.137

Le contexte offrant les plus mauvais résultats (en terme de nombre de vignettes à valider) correspond à l'échantillon hors bloom. Ici, le contexte permettant d'obtenir le plus faible nombre de vignettes à valider correspond à la combinaison des deux échantillons prélevés pendant le bloom de *Phaeocystis globosa*. Ces résultats montrent l'intérêt de la pertinence du choix de l'utilisateur concernant les échantillons contextuels. Une synthèse de ces conclusions est représentée sur la Fig. 13.

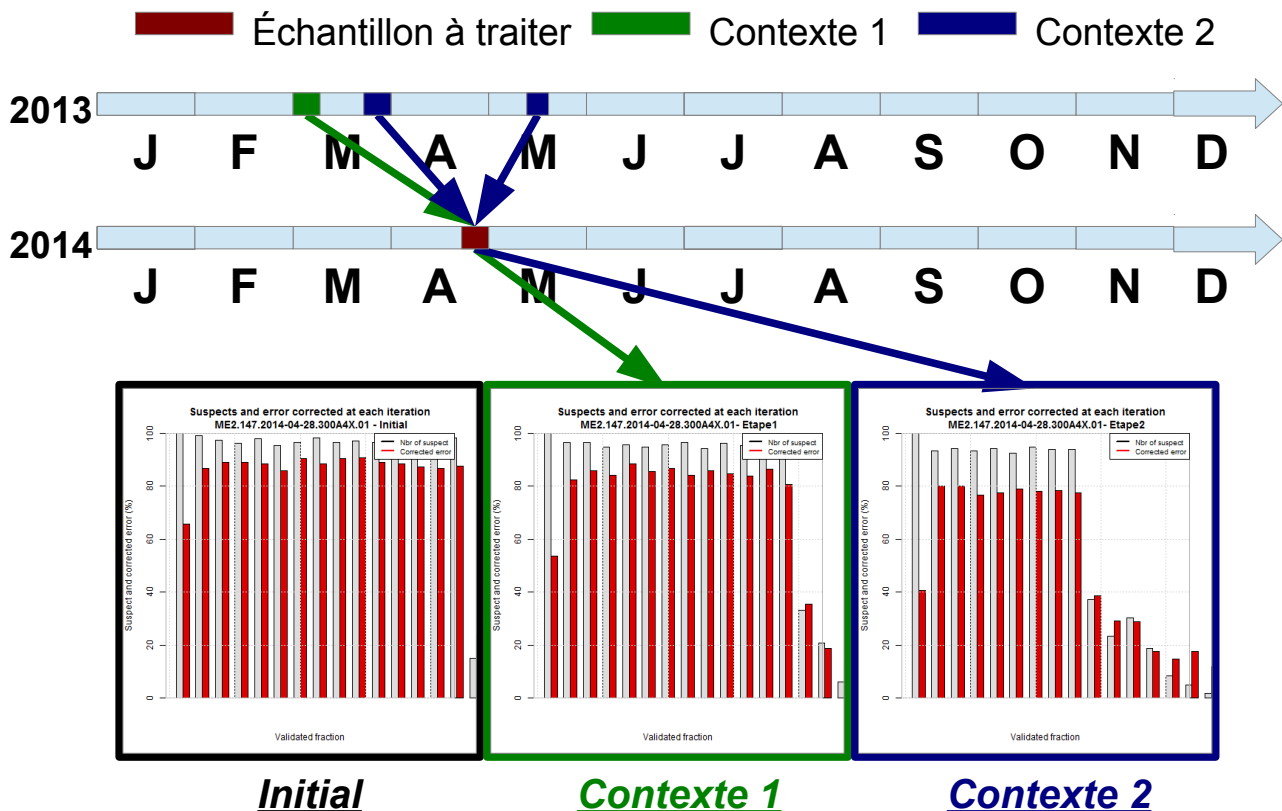
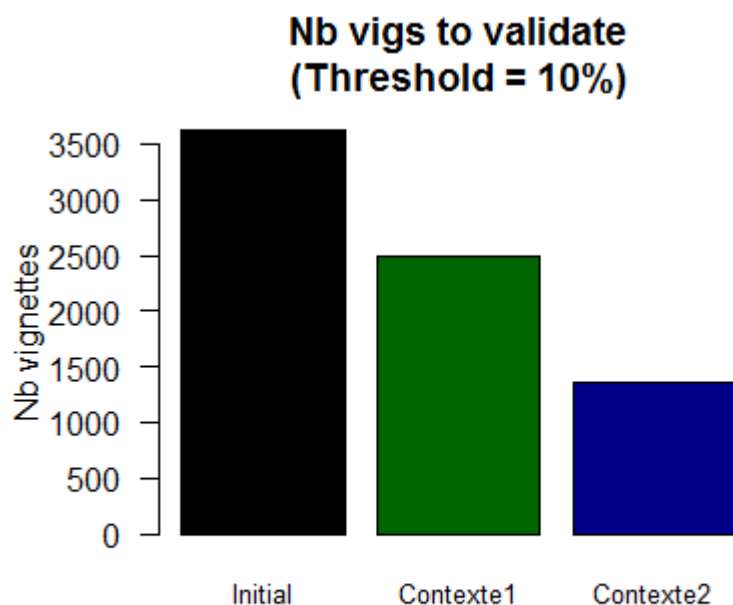


Figure 13 – Synthèse des résultats obtenus par apprentissage actif selon le contexte sélectionné.

La Fig. 14 présentent une comparaison des résultats obtenus avant et après apprentissage actif, selon le contexte sélectionné par l'utilisateur :

- **Contexte 1** : plus mauvaise combinaison d'échantillons contextuels (ici, ME2.137 → hors bloom),
- **Contexte 2** : meilleure combinaison d'échantillons contextuels (ici, ME2.142 + ME2.148 → période de bloom).



*Figure 14 – Nombre de vignettes à valider selon le contexte, pour obtenir moins de 10% d'erreur par groupe.*

Comme montré sur la Fig. 14, le nombre de vignettes à valider afin d'obtenir un taux d'erreur inférieur à 10% par groupe, est beaucoup plus faible pour le contexte 2. Ces résultats tendent à montrer l'importance du choix de contexte par l'utilisateur. Cependant, même s'il devra valider une plus grande proportion de vignettes, un choix moins pertinent permet tout de même de réduire ce nombre par rapport à celui obtenu avec le set d'apprentissage initial.

## Conclusion

Une des grandes évolutions de ZooPhytoImage version 5 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 (géographique, temporelle et saisonnière) 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. Dans cette étude, les expérimentations réalisées montrent qu'une solution possible pour remédier à ce problème, tient dans le set d'apprentissage adaptatif. En

d'autres termes, il paraît 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) afin d'améliorer 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 est également considérablement réduit, ce qui implique un gain de temps pour l'utilisateur.

## 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



**Projet FlowCAM/ZooPhytoImage**  
**Rapport d'avancement**  
**- 30/06/2015 -**

Guillaume WACQUET & Philippe GROSJEAN

**UMONS**  
Université de Mons



## **INTRODUCTION**

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, il reste encore à paramétrer les différentes fonctionnalités du logiciel dans le contexte particulier du REPHY. Différents axes ont donc été proposés par l'UMONS et l'IFREMER.

Premièrement, 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. La méthode incluse dans Zoo/PhytoImage consiste à calibrer un modèle prédictif 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, une étude comparative des dénombrements manuels effectués sur deux sites (Boulogne-sur-Mer et Nantes) est présentée.

Deuxièmement, 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'optimiser l'approche de type apprentissage actif. Dans ce rapport, nous quantifions les gains obtenus à l'aide de ce processus pour la reconnaissance semi-automatisée du phytoplancton, à savoir : l'évolution du taux global d'erreur et le gain de temps lors de la validation des prédictions automatiques dans le cadre de la correction de l'erreur.





*Partie 1*  
**Dénombrement des cellules  
dans les colonies**

L'ensemble d'apprentissage utilisé pour les expérimentations résulte de la fusion des 3 ensembles d'apprentissage construit à Boulogne-sur-Mer, Nantes et Arcachon. Dans ce cadre, seuls les groupes phytoplanctoniques sont considérés :

*\_train\_Rephy\_MancheAtlantique\_4X* → 3787 vignettes réparties en 22 espèces.

**Table 1 :** Table représentant l'ensemble des groupes taxinomiques étudiés dans le set d'apprentissage, provenant d'échantillons naturels (300µm/4X). Total vignettes : nombre total de vignettes dans le groupe ; BOULOGNE : nombre de vignettes dénombrées par Boulogne (et ratio en pourcentage) ; NANTES : nombre de vignettes dénombrées par Nantes (et ratio en pourcentage) ; ARCACHON : nombre de vignettes dénombrées par Arcachon (et ratio en pourcentage). En gris : vignettes dénombrées peu nombreuses ou absentes dues à la difficulté ou l'impossibilité de comptage manuel.

<b>Group</b>	<b>Total vignettes</b>	<b>BOULOGNE</b>	<b>NANTES</b>
<i>Odontella</i>	160	160 (100%)	116 (72.50%)
<i>Guinardia flaccida</i>	260	260 (100%)	101 (38.85%)
<i>Dinophysis tripos</i>	36	36 (100%)	36 (100%)
<i>Proboscia Rhizosolenia</i>	265	265 (100%)	...
<i>Dactyliosolen fragilissimus</i>	238	238 (100%)	146 (61.34%)
<i>Guinardia striata</i>	84	76 (90.48%)	58 (69.05%)
<i>Thalassionema</i>	169	169 (100%)	83 (49.11%)
<i>Leptocylindrus</i>	256	256 (100%)	118 (46.10%)
<i>Pseudo-nitzschia</i>	260	260 (100%)	177 (68.08%)
<i>Phaeocystis</i>	264	...	...
<i>Chaetoceros socialis</i>	60	...	1 (1.67%)
<i>Chaetoceros</i>	260	135 (51.92%)	62 (23.85%)
<i>Asterionellopsis glacialis</i>	227	227 (100%)	60 (26.43%)
<i>Ditylum brightwellii</i>	261	261 (100%)	157 (60.15%)
<i>Paralia</i>	93	8 (8.60%)	...
<i>Thalassiosira big chaines</i>	99	99 (100%)	92 (92.93%)
<i>Guinardia delicatula</i>	56	56 (100%)	52 (92.86%)
<i>Chaetoceros curvisetus</i>	200	195 (97.50%)	77 (38.50%)
<i>Alexandrium affine</i>	93	93 (100%)	93 (100%)
<i>Bacteriastrum</i>	180	164 (91.11%)	61 (33.89%)
<i>Skeletonema</i>	162	162 (100%)	84 (51.85%)
<i>Thalassiosira small chaines</i>	104	104 (100%)	94 (90.38%)

Dans cette étude, nous nous intéressons à l'estimation du nombre de cellules par colonies pour chacun des 2 sites. Les performances des méthodes prédictives sont comparées par validations croisées (10-folds) sur des modèles de régressions linéaires et non linéaires. Grâce à la validation

croisée, il est possible d'évaluer les taux de reconnaissance des différentes méthodes prédictives. Ici, deux scores sont évalués (Govaerts, 2010) :

- **%TL** : taux de reconnaissance global du nombre de cellules par colonie logarithmique.
- **%EstTot** : estimation totale du nombre de cellules, en somme.

Trois méthodes sont utilisés pour la construction des modèles prédictifs :

- **LM** : Linear Model (Modèle Linéaire Multivarié),
- **LDA** : Linear Discriminant Analysis (Analyse Discriminante Linéaire),
- **MDA** : Mixture Discriminant Analysis (Analyse Discriminante par Mélange Gaussien).

## **RESULTATS - SITE DE BOULOGNE**

*Dénombrements manuels réalisés par R.Cuvelliez, A.Lefebvre, C.Blondel, P.Hébert.*

**Table 2** : Table représentant les % d'estimation du nombre de cellules par colonie pour l'ensemble des groupes taxinomiques étudiés.

<b>Group</b>	<b>LM</b>	<b>LDA</b>	<b>MDA</b>
	<b>%TL - %EstTot</b>	<b>%TL - %EstTot</b>	<b>%TL - %EstTot</b>
<i>Odontella</i>	52.69 – 138.43	75.87 – 101.12	<b>84.94 – 100.77</b>
<i>Guinardia flaccida</i>	84.15 – 98.09	89.04 – 98.33	<b>91.23 – 98.99</b>
<i>Dinophysis tripos</i>	60.28 – 153.93	86.67 – 105.65	<b>100.00 – 100.00</b>
<i>Proboscia Rhizosolenia</i>	<b>86.75 – 97.55</b>	88.53 – 95.99	<b>89.89 – 97.12</b>
<i>Dactyliosolen fragilissimus</i>	68.24 – 97.63	<b>77.35 – 101.23</b>	<b>81.47 – 97.40</b>
<i>Guinardia striata</i>	37.76 – 135.75	42.50 – 105.63	<b>74.61 – 96.26</b>
<i>Thalassionema</i>	41.60 – 95.65	58.22 – 97.49	<b>79.70 – 98.47</b>
<i>Leptocylindrus</i>	32.97 – 95.91	35.08 – 98.98	<b>57.50 – 99.35</b>
<i>Pseudo-nitzschia</i>	56.15 – 97.86	<b>56.81 – 98.54</b>	<b>68.54 – 96.81</b>
<i>Phaeocystis</i>	... - ...	... - ...	... - ...
<i>Chaetoceros socialis</i>	... - ...	... - ...	... - ...
<i>Chaetoceros</i>	29.41 – 97.17	<b>30.74 – 99.88</b>	<b>57.56 – 100.84</b>
<i>Asterionellopsis glacialis</i>	36.04 – 98.36	45.24 – 100.57	<b>71.54 – 99.60</b>
<i>Ditylum brightwellii</i>	88.70 – 98.39	<b>93.72 – 98.49</b>	<b>95.82 – 98.24</b>
<i>Paralia</i>	... - ...	... - ...	... - ...
<i>Thalassiosira big chaines</i>	<b>62.83 – 99.66</b>	65.66 – 97.96	<b>84.34 – 99.58</b>
<i>Guinardia delicatula</i>	37.32 – 95.26	<b>48.57 – 98.41</b>	<b>70.00 – 95.54</b>
<i>Chaetoceros curvisetus</i>	41.33 – 97.47	38.87 – 98.62	<b>64.36 – 99.99</b>
<i>Alexandrium affine</i>	82.04 – 99.34	87.31 – 99.59	<b>97.85 – 100.34</b>
<i>Bacteriastrum</i>	43.90 – 98.93	40.91 – 101.37	<b>73.78 – 99.61</b>
<i>Skeletonema</i>	39.44 – 97.86	34.81 – 102.05	<b>54.63 – 100.43</b>
<i>Thalassiosira small chaines</i>	46.73 – 97.33	49.90 – 100.88	<b>69.52 – 100.12</b>

## RESULTATS - SITE DE NANTES

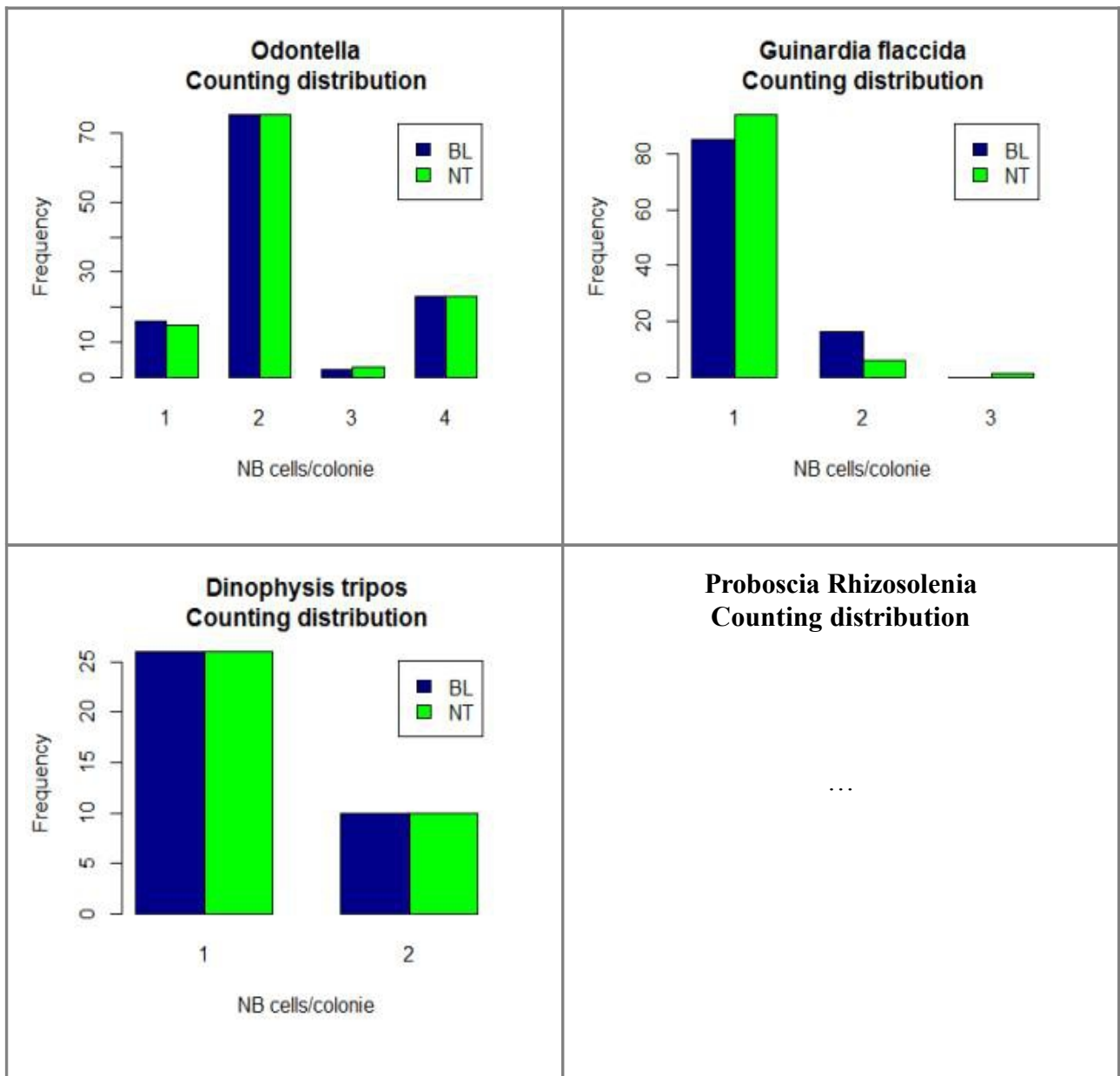
### Dénombrements manuels réalisés par N.Neaud-Masson.

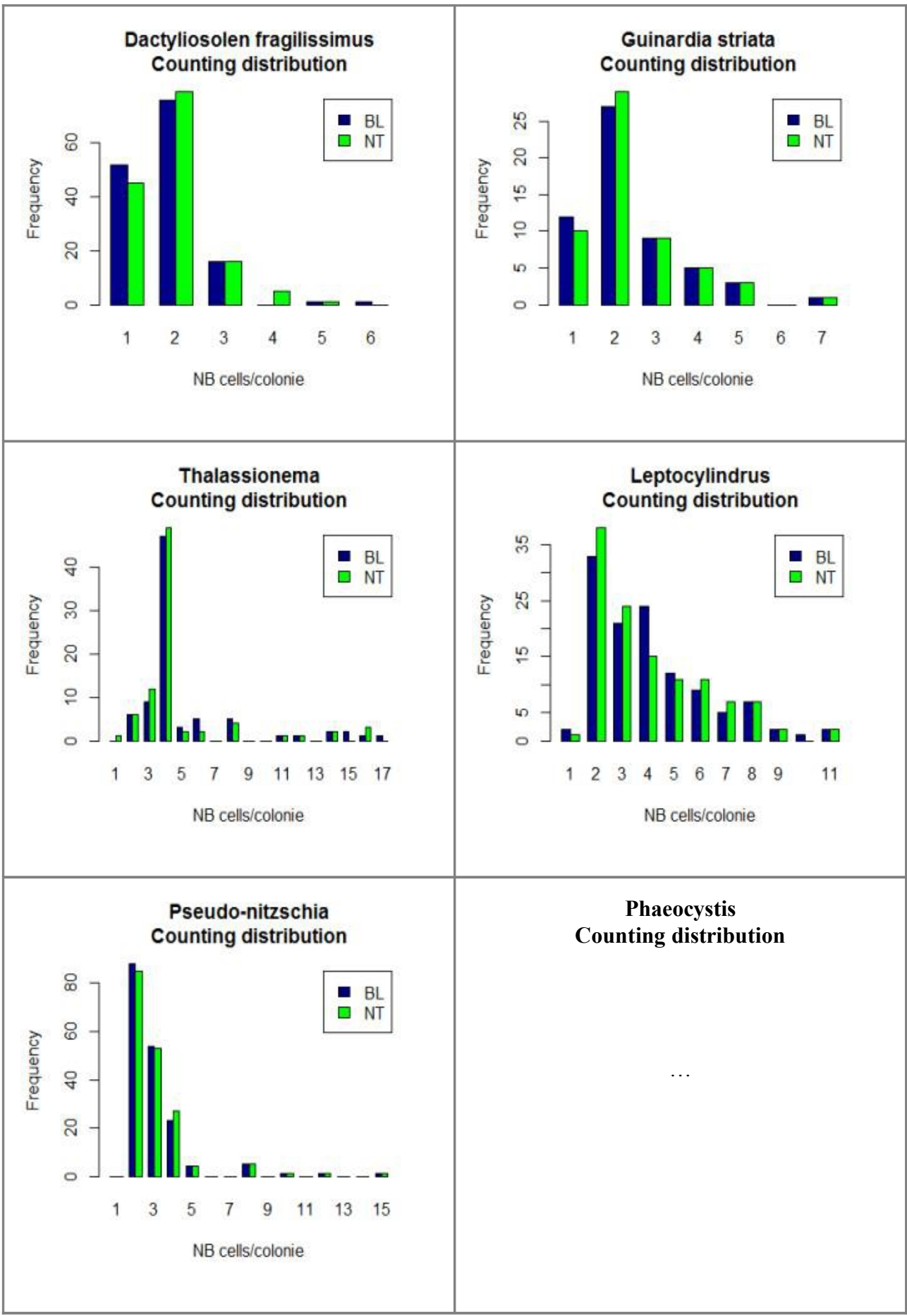
**Table 3 :** Représentation des distributions des nombres de cellules par colonie pour chaque groupe taxonomique étudié dans le set d'apprentissage et dénombré par Nantes, provenant d'échantillons naturels (300µm/4X).

Group	LM	LDA	MDA
	%TL - %EstTot	%TL - %EstTot	%TL - %EstTot
<i>Odontella</i>	63.02 – 97.24	<b>81.12 – 102.18</b>	<b>87.07 – 104.14</b>
<i>Guinardia flaccida</i>	85.15 – 103.04	92.97 – 102.57	<b>97.72 – 98.44</b>
<i>Dinophysis tripos</i>	59.44 – 107.89	84.72 – 104.57	<b>100.00 – 100.00</b>
<i>Proboscia Rhizosolenia</i>	... - ...	... - ...	... - ...
<i>Dactyliosolen fragilissimus</i>	66.37 – 96.78	<b>79.79 – 97.49</b>	<b>85.96 – 94.82</b>
<i>Guinardia striata</i>	48.45 – 99.13	42.41 – 107.76	<b>81.03 – 99.67</b>
<i>Thalassionema</i>	48.43 – 97.15	65.18 – 98.07	<b>82.77 – 99.27</b>
<i>Leptocylindrus</i>	37.63 – 97.19	<b>42.88 – 100.88</b>	<b>66.53 – 97.39</b>
<i>Pseudo-nitzschia</i>	53.95 – 98.44	<b>54.24 – 99.29</b>	<b>69.66 – 96.44</b>
<i>Phaeocystis</i>	... - ...	... - ...	... - ...
<i>Chaetoceros socialis</i>	... - ...	... - ...	... - ...
<i>Chaetoceros</i>	24.68 – 98.68	30.00 – 102.03	<b>73.06 – 101.24</b>
<i>Asterionellopsis glacialis</i>	36.50 – 98.42	36.33 – 105.03	<b>69.45 – 98.72</b>
<i>Ditylum brightwellii</i>	<b>85.86 – 99.80</b>	90.45 – 99.77	<b>95.35 – 98.57</b>
<i>Paralia</i>	... - ...	... - ...	... - ...
<i>Thalassiosira big chaines</i>	60.98 – 96.45	65.76 – 95.24	<b>86.74 – 97.33</b>
<i>Guinardia delicatula</i>	28.85 – 94.76	39.23 – 98.98	<b>76.92 – 100.56</b>
<i>Chaetoceros curvisetus</i>	36.88 – 96.89	30.00 – 104.31	<b>70.91 – 102.38</b>
<i>Alexandrium affine</i>	82.58 – 99.55	87.96 – 98.30	<b>97.85 – 100.34</b>
<i>Bacteriastrum</i>	42.30 – 154.34	38.52 – 98.77	<b>89.84 – 98.97</b>
<i>Skeletonema</i>	33.45 – 84.71	26.79 – 101.08	<b>61.79 – 100.91</b>
<i>Thalassiosira small chaines</i>	49.57 – 97.87	<b>55.21 – 100.32</b>	<b>72.66 – 98.32</b>

Dans les deux cas, la méthode MDA semble fournir (presque chaque fois) les meilleurs scores de prédiction. Cette méthode non linéaire peut donc être retenue pour la construction des modèles prédictifs dans Zoo/PhytoImage.

**COMPARAISON DES COMPTAGES**  
**(pour les mêmes vignettes validées sur les 2 sites)**

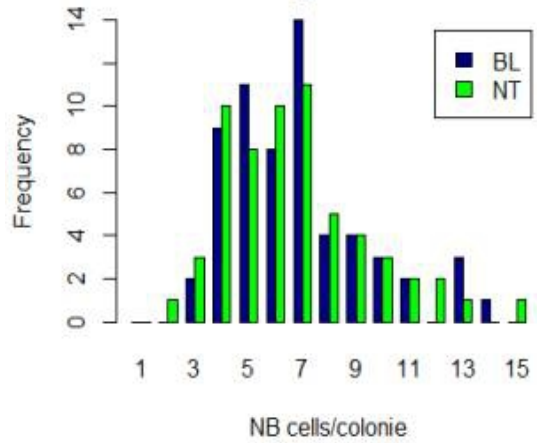




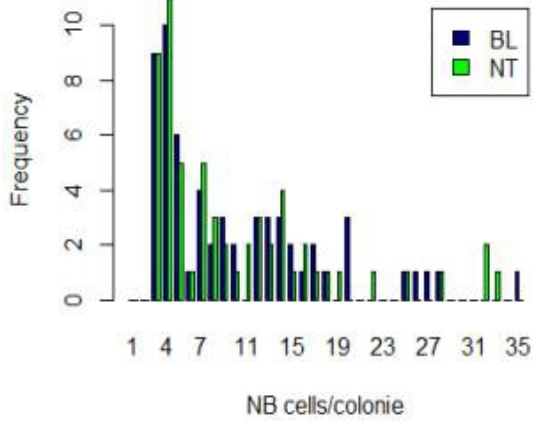
**Chaetoceros socialis**  
Counting distribution

...

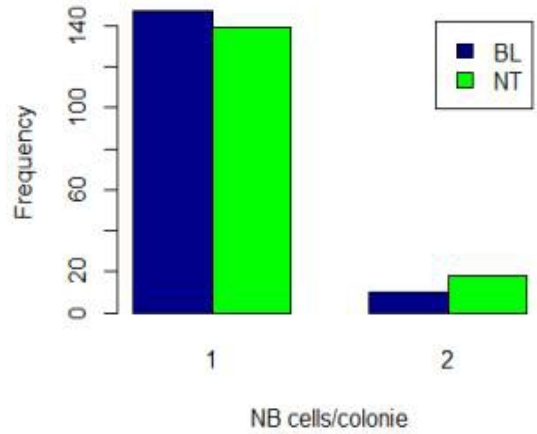
**Chaetoceros**  
Counting distribution



**Asterionellopsis glacialis**  
Counting distribution



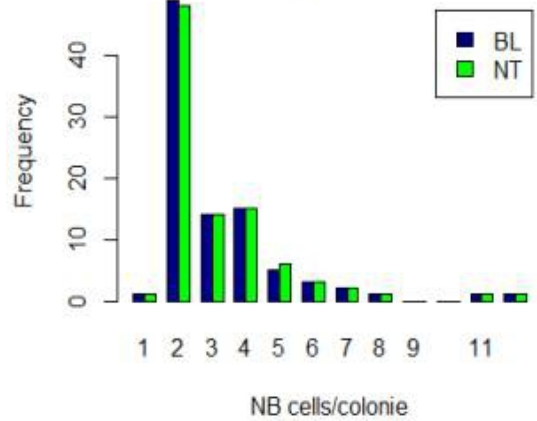
**Ditylum brightwellii**  
Counting distribution



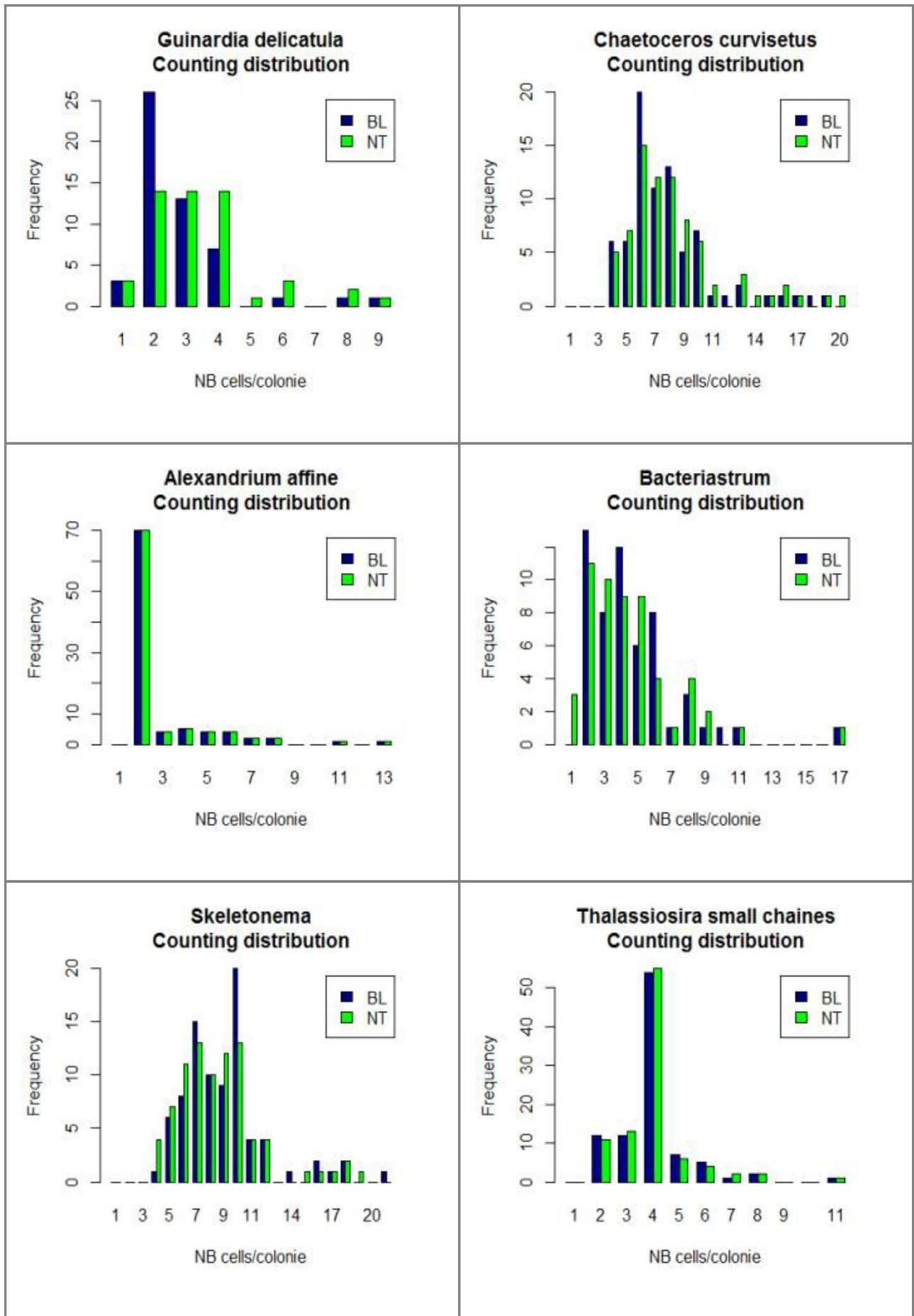
**Paralia**  
Counting distribution

...

**Thalassiosira big chaines**  
Counting distribution







**Table 4** : Comparaison des estimations totales des nombres de cellules comptées par Nantes et Boulogne. En rouge : sous-estimation de Nantes par rapport à Boulogne (ou sur-estimation de Boulogne par rapport à Nantes) ; en vert : sur-estimation de Nantes par rapport à Boulogne (ou sous-estimation de Boulogne par rapport à Nantes).

<b>Group</b>	<b>Comp. NT vs BL %EstTot</b>
<i>Odontella</i>	100.76
<i>Guinardia flaccida</i>	93.16
<i>Dinophysis tripos</i>	100.00
<i>Proboscia Rhizosolenia</i>	... - ...
<i>Dactyliosolen fragilissimus</i>	104.94
<i>Guinardia striata</i>	101.48
<i>Thalassionema</i>	93.43
<i>Leptocylindrus</i>	98.55
<i>Pseudo-nitzschia</i>	101.33
<i>Phaeocystis</i>	... - ...
<i>Chaetoceros socialis</i>	... - ...
<i>Chaetoceros</i>	98.08
<i>Asterionellopsis glacialis</i>	98.69
<i>Ditylum brightwellii</i>	104.79
<i>Paralia</i>	... - ...
<i>Thalassiosira big chaines</i>	101.04
<i>Guinardia delicatula</i>	122.07
<i>Chaetoceros curvisetus</i>	105.13
<i>Alexandrium affine</i>	100.00
<i>Bacteriastrum</i>	96.46
<i>Skeletonema</i>	94.87
<i>Thalassiosira small chaines</i>	100.27

Grâce à ce tableau, des différences de dénombrements entre sites peuvent être observées pour certains groupes phytoplanctoniques. Voici les groupes sur-ou sous-estimés (+ 5%) par un site ou l'autre :

- *Guinardia delicatula*
- *Guinardia flaccida*
- *Thalassionema*
- *Chaetoceros curvisetus*
- *Skeletonema*

Le critères de dénombrements manuels concernant ces groupes devront donc être révisés afin d'en extraire des règles de décisions communes permettant d'obtenir des comptages pertinents et comparables.

## **ADAPTATIONS NECESSAIRES DANS ZOO/PHYTOIMAGE**

Plusieurs difficultés ont été rencontrées lors de l'intégration du module de dénombrement des cellules en colonie dans Zoo/PhytoImage. Des adaptations/modifications du code R doivent donc être envisagées pour certaines fonctions. En particulier :

- la compatibilité avec le module de correction statistique de l'erreur. Pour le moment, la correction statistique des abondances est réalisée en terme de nombre de particules, et non en terme de nombre de cellules. Pour y parvenir, des modifications du code R permettant de calculer des coefficients de correction pour le nombre de cellules, le biovolume, etc., sont nécessaires.
- l'exportation des résultats et la présentation des graphes dans l'interface graphique de Zoo/PhytoImage, en prenant en compte le nombre de cellules par colonie.

***Partie 2***  
**Apprentissage actif**  
**Adaptation du set d'apprentissage**

L'objectif est d'adapter un ensemble d'apprentissage initial à un échantillon à analyser, de manière automatisée. Pour cela, des données « contextuelles » contenues dans des échantillons validés préalablement et similaires à celui à analyser, peuvent être utilisées. Concrètement, ces données contextuelles correspondent aux items validés en routine lors du processus de correction de l'erreur. L'objectif est alors d'inclure ces items validés dans l'ensemble d'apprentissage afin de garantir des performances de reconnaissance supérieures ou égales à celles obtenues initialement.

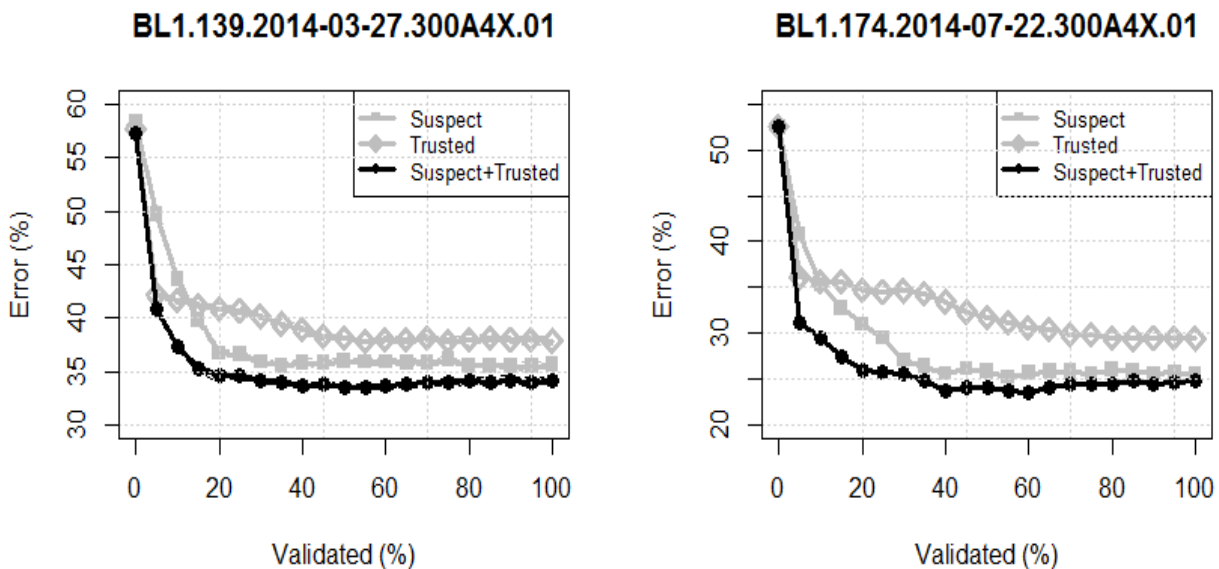
Le paramétrage de cette méthode consiste en le choix du type d'items à ajouter au set d'apprentissage :

- uniquement les items suspects validés (nommé « Suspect »),
- uniquement les items non suspects validés (nommé « Trusted »),
- un mélange d'items suspects et non suspects validés (nommé « Suspect+Trusted »).

Dans cette étude, nous choisissons d'appliquer les critères d'ajout d'items ci-dessus, sur 2 échantillons, et d'utiliser, pour chacun d'eux, 1 échantillon « contextuel » prélevé dans la même zone géographique et à une même période de l'année (dans un intervalle d'un mois, cf. Table 1). L'évolution de l'erreur globale en fonction de la proportion d'items validés ajoutés dans le set d'apprentissage, est présentée sur la Figure 1.

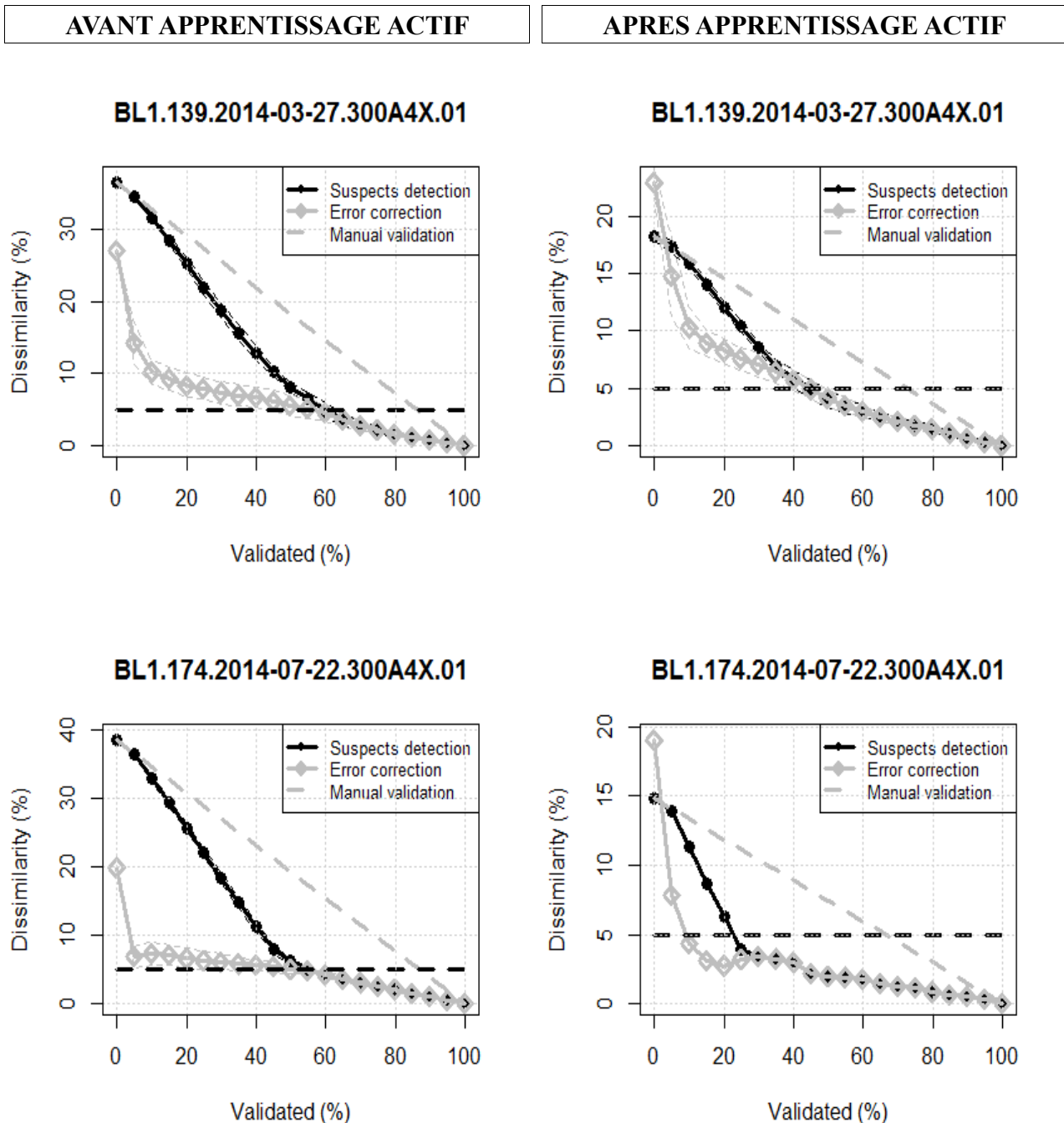
**Table 2:** Contexte des échantillons utilisés pour l'apprentissage actif. Ce contexte est choisi dans le but de représenter au mieux la composition phytoplanctonique de l'échantillon à analyser.

<i>Ech. à analyser</i>	<i>Ech. contextuel</i>	<i>Date</i>	<i>Zone géographique</i>
BL1.139	BL1.135	Mars, 2014	Boulogne-sur-Mer
BL1.174	BL1.160	Juillet, 2014	Boulogne-sur-Mer



**Figure 1:** Erreur observée après ajout d'un pourcentage d'items dans la fraction des suspects, des non suspects et d'un mélange des deux, provenant de l'échantillon contextuel.

Comme observé sur la Figure 1 et dans chacun des cas, l'erreur observée après ajout d'items validés dans le mélange des fractions suspects et non suspects provenant des échantillons contextuels, décroît rapidement dès les premiers 10%. Le processus d'apprentissage actif permet alors de diminuer l'erreur de 20-30%. De plus, ces expérimentations montrent que le critère proposé (Suspect+Trusted) permet d'obtenir les meilleures performances que les autres critères qui n'utilisent que les items suspects ou non suspects validés. La combinaison des items suspects et non suspects semble permettre l'élargissement des frontières des classes pour englober les items les plus suspects de l'échantillon à analyser, tout en maintenant une cohésion globale importante intra-classe.



**Figure 2:** Coefficient de dissimilarité entre les “abondances réelles” (obtenues par validation manuelle de la totalité de l'échantillon) et la validation manuelle des items suspects, la correction de l'erreur ou la validation manuelle usuelle. La dissimilarité-cible de 5% est représentée par la

*ligne noire horizontale pointillée.*

La Figure 2 permet de comparer les résultats obtenus en terme de dissimilarité entre les abondances réelles et celles obtenues après validation manuelle des suspects ou correction de l'erreur, avant et après apprentissage actif. La dissimilarité-cible est ici fixée à 5% (ligne noire en pointillé). Pour les deux échantillons (BL1.139 et BL1.174), nous pouvons constater que la dissimilarité initiale entre les abondances réelles et celles obtenues par prédiction automatique (c'est-à-dire, à l'abscisse 0%), diminue significativement après apprentissage actif. De plus, nous notons que la proportion de vignettes à valider après apprentissage actif pour atteindre une valeur de dissimilarité de 5%, est toujours plus faible que celui avant complétion du set (-20% pour BL1.139 et -30% pour BL1.174). Néanmoins, il est important de noter que cette réduction peut fluctuer de manière significative selon la pertinence de la sélection des échantillons contextuels par l'utilisateur.

### **ADAPTATIONS NECESSAIRES DANS ZOO/PHYTOIMAGE**

Pour pouvoir utiliser en routine le module d'apprentissage actif, une adaptation des routines de traitement des échantillons dans Zoo/PhytoImage est nécessaire. En particulier :

- pour chacun des items présents dans les échantillons contextuels (échantillons validés), il est important de connaître le statut (validé ou non) et la fraction d'appartenance (suspect ou non suspect). Il est donc nécessaire de modifier la routine d'exportation des résultats après correction de l'erreur.
- les changements opérés dans le set d'apprentissage, ainsi que les différences entre les classes des échantillons contextuels et celles du set d'apprentissage utilisé, doivent être considérées lors de l'apprentissage actif. Pour cela, une adaptation du code R est nécessaire.

## **CONCLUSION**

Afin que toutes les fonctionnalités du système couplé FlowCAM/ZooPhytoImage puissent être utilisées pleinement, une mise à jour des menus du logiciel a été effectuée. Cependant, le paramétrage et l'adaptation de certaines routines sont nécessaires. Les modules présentés dans ce rapport permettent, à terme, de fournir des résultats pertinents dans un laps de temps acceptable.

De plus, dans un but d'optimisation de l'interactivité et de l'ergonomie, une réflexion concernant d'autres aspects du logiciel a été menée :

- Limitation du nombre maximal de vignettes à valider à chaque étape du processus de correction de l'erreur (fixé à 200),
- Affichage d'un graphique de dissimilarité (après correction de l'erreur) entre chaque étape de validation, comme critère d'aide à la décision,
- Visualisation du résultat de classification, sous forme de dossiers contenant les vignettes de chaque classe. Ceci permet alors à l'utilisateur de pouvoir reclasser manuellement les particules classées dans la colonne « Other »,
- Création de raccourcis clavier/souris pour accélérer le reclassement via le module de correction de l'erreur.



Zoo/PhytoImage, logiciel gratuit (Open Source) pour l'analyse d'images numériques de plancton  
<http://www.sciviews.org/zooimage>

# Zoo/PHYTOIMAGE VERSION 5.4-0

Analyse d'Images de Plancton Assistée par Ordinateur

## MANUEL UTILISATEUR

L'équipe de développement de ZooImage  
Septembre 2015

*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 planctonologue 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<sup>1</sup>. 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 (5.4-0), qui sera 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 3.0-5, qui est publique et téléchargeable à partir du site du CRAN (<http://cran.r-project.org>).*

---

<sup>1</sup> 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, 4 ET 5

La version 3.0 de Zoo/PhytoImage est la dernière version publique distribuée sur <http://www.sciviews.org/zooimage> jusqu'à présent. Les versions 4 et 5 du logiciel n'étaient pas publique et contenaient 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 5.4-0 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 cette version 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,
- Optimisation du module de correction d'erreur, d'un point de vue technique et ergonomique,
- Intégration d'un module de dénombrement des colonies, et adaptation du code pour l'exportation des résultats,
- Développement de routines pour l'apprentissage actif, utilisant des échantillons validés comme données contextuelles,
- Corrections liées à l'utilisation des fonctionnalités de Zoo/PhytoImage à partir des menus.

Parmi ces changements, un des 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 les versions antérieures. 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 5.4-0, 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.*

Dans les versions précédentes de Zoo/PhytoImage, le critère d'arrêt pour la validation manuelle des vignettes dans le cadre de la correction de l'erreur, était étroitement lié aux nombres de suspects et à l'erreur résiduelle correspondante. Cette méthode nécessitait alors une validation de la quasi-totalité des vignettes considérées comme suspectes. Suite aux besoins des principaux partenaires en terme de temps de réaction lié aux risques environnementaux, il est apparu nécessaire de revoir la stratégie de mise en œuvre du module de correction de l'erreur afin de réduire la durée de validation manuelle des vignettes.

Jusqu'à présent, Zoo/PhytoImage permettait d'obtenir des identifications semi-automatiques pertinentes du plancton 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 contexte, un module de dénombrement des colonies, répondant aux besoins des principaux partenaires, a été intégré au logiciel. Cette intégration a cependant nécessité des modifications importantes dans le code pour les calculs des biovolumes, biomasses, etc.

Généralement, la reconnaissance semi-automatisée des particules de plancton dans un échantillon d'eau, doit passer par la construction d'un set d'apprentissage reflétant la variabilité des espèces rencontrés dans le milieu naturel, mais également la variabilité morphologique des particules. Pour cela, il est nécessaire de réaliser un set d'apprentissage volumineux sur au moins un an pour reprendre les variations saisonnières et de le moduler par zone géographique. Cependant, un tel set d'apprentissage mène à un outil de classification figé, et ne permet donc pas une adaptation temporelle aux échantillons analysés. L'apprentissage actif utilise les données validées en routine sur les échantillons pour enrichir le set d'apprentissage, et ainsi l'adapter géographiquement, temporellement et saisonnièrement, de manière totalement transparente. Dans cette nouvelle version de Zoo/PhytoImage, des routines permettant l'intégration de données contextuelles ont été développées et intégrées au code, afin de réduire le nombre d'erreurs de prédiction et par conséquent, la proportion de vignettes à valider.

## 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 décentement. 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.

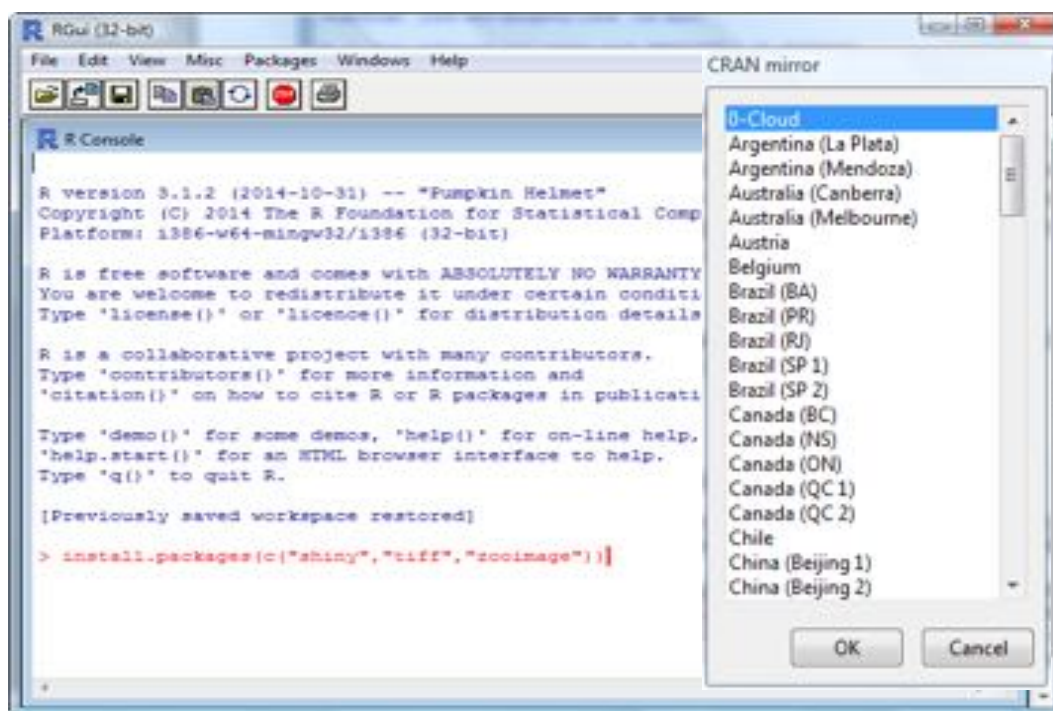
### 3.2. Installation de Zoo/PhytoImage sous Windows

La version 5.4-0 de Zoo/PhytoImage nécessite une version récente de R<sup>2</sup> (version 3.0.x ou plus). Elle peut être téléchargée directement sur le site du CRAN (<http://cran.r-project.org>).

Pour le lancement et l'utilisation 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.

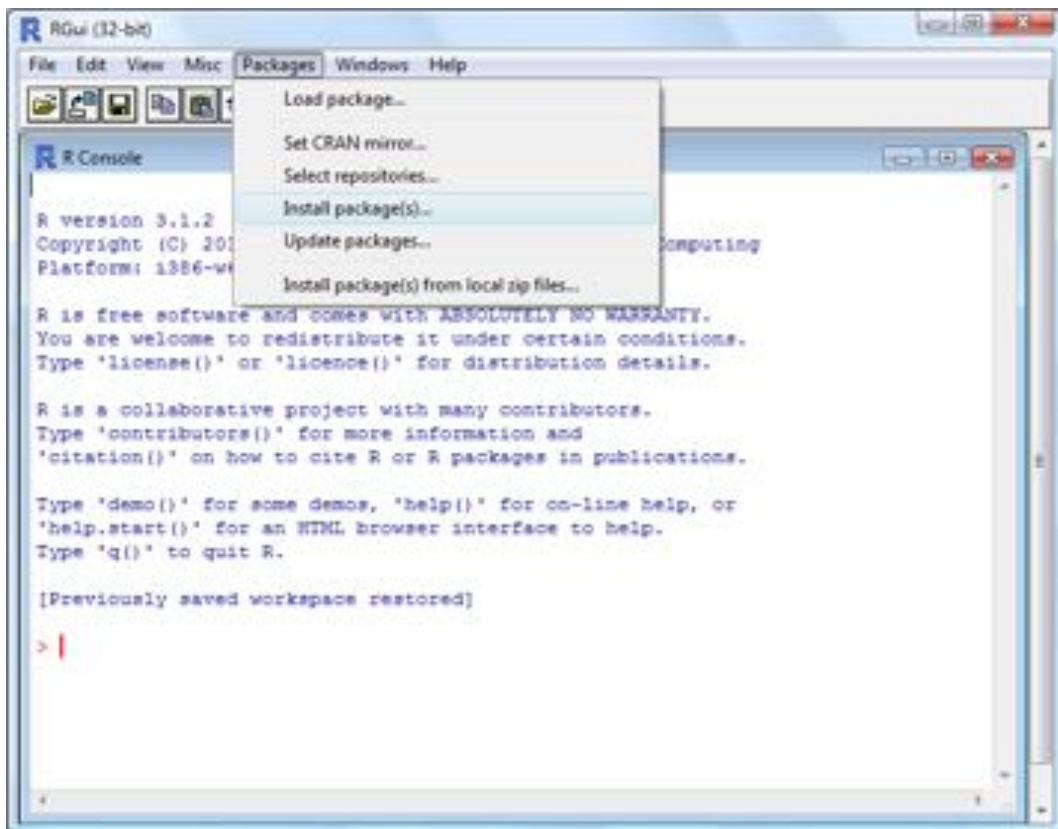
Les packages nécessaires (« shiny », « tiff » et « zooimage ») peuvent être installés directement à partir de la console R, en tapant : `install.packages(c("shiny", "tiff", "DT", "zooimage"))`. Choisissez ensuite un miroir (par défaut : 0-cloud) pour démarrer les téléchargements et les installations.



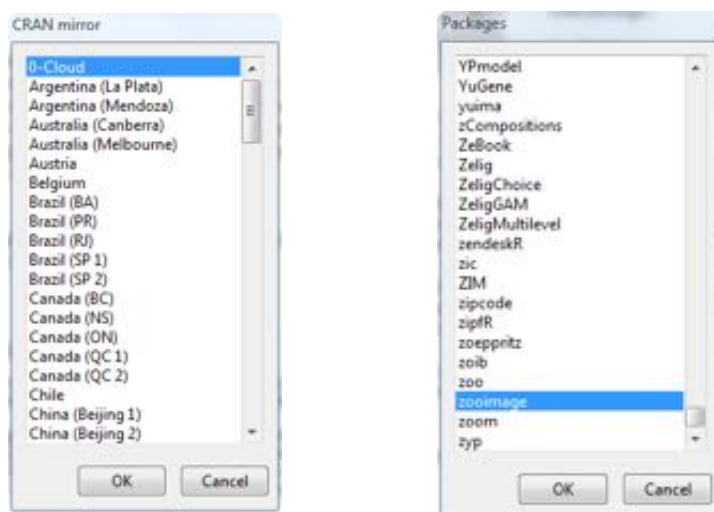
Capture d'écran de la console R pour l'installation de Zoo/PhytoImage version 5.

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".

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



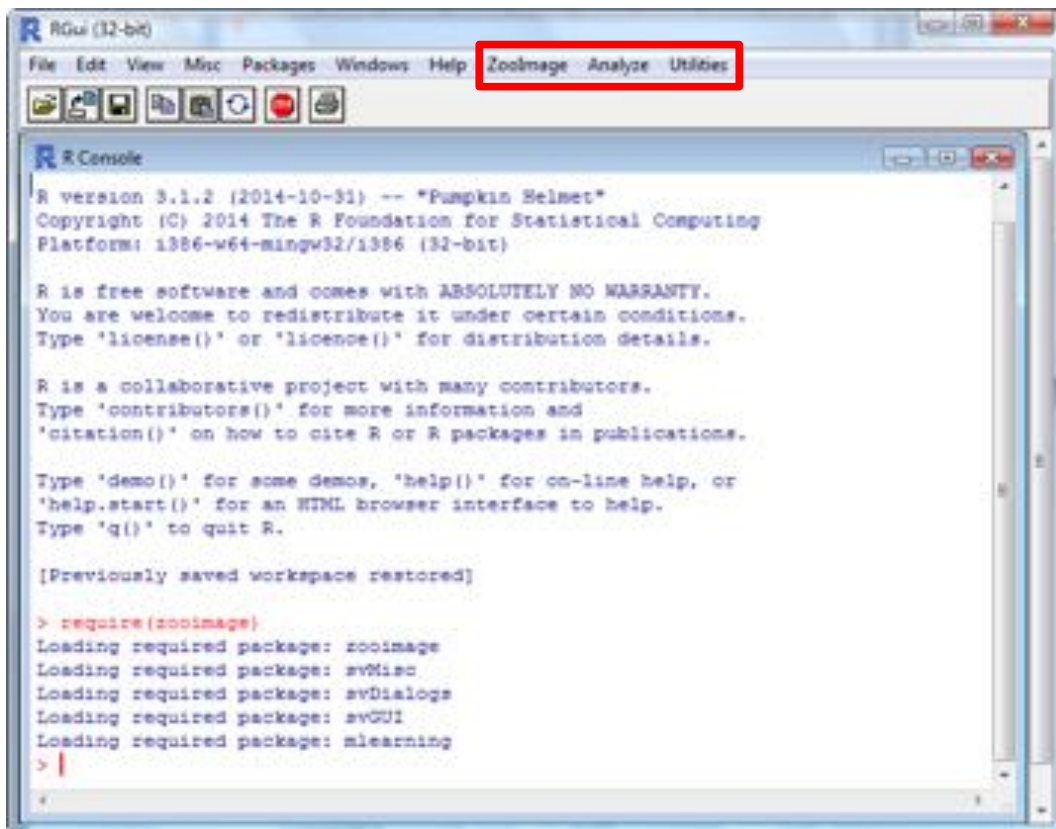
Capture d'écran pour l'installation manuelle de Zoo/PhytoImage version 5.



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

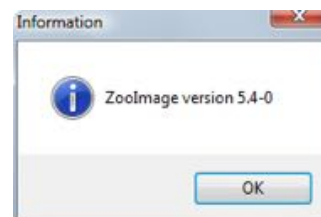
Une fois l'installation des packages terminée, il est possible de s'assurer du bon déroulement des étapes précédentes en vérifiant que la version installée est bien 5.4-0. Pour cela, dans un premier temps, tapez dans la console R : **require(zooimage)**, pour lancer Zoo/PhytoImage. Trois nouvelles entrées 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.

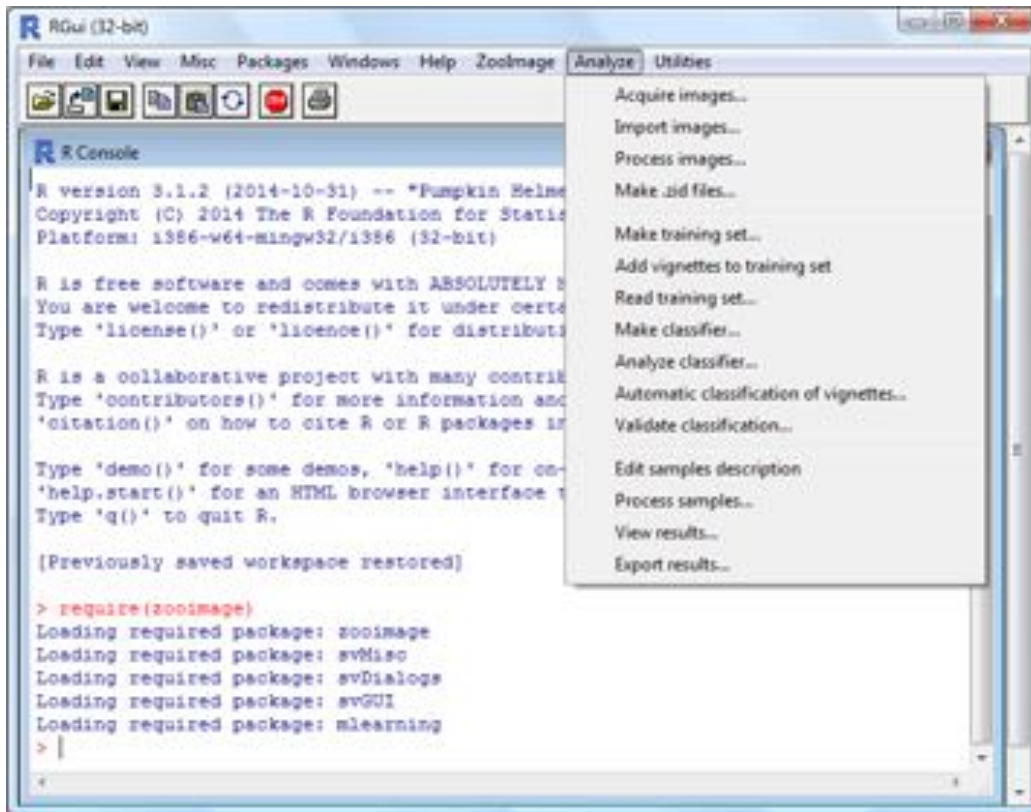


#### 4. UTILISATION DES FONCTIONS A PARTIR DU MENU DE ZOO/PHYTOIMAGE

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.

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.



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

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<sup>3</sup>, leurs mesures et les métadonnées. Toutes ces informations sont alors compressées<sup>4</sup>.

- 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.

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 du set.

3. **Count cell(s) in colonie.** Cette fonction permet de compter le nombre de cellules présentes sur chacune des vignettes des groupes de l'ensemble d'apprentissage. Ces dénombrements sont alors sauves (fichier « RData », à la racine de l'ensemble d'apprentissage) puis utilisés pour la construction des modèles prédictifs.

4. **Compare training sets...** Cette fonction permet de comparer deux sets d'apprentissage. Elle peut être notamment utilisée pour recenser les modifications effectuées au cours de la construction, de l'optimisation (ajout/suppression de vignettes) ou de la validation (changement du nom des classes, reclassement de vignettes) d'un set

<sup>3</sup> Ces images particulières sont appelées « vignettes » dans la terminologie ZooImage.

<sup>4</sup> 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.



d'apprentissage. Un fichier texte listant les différences est alors créé sur le disque, à la racine des sets d'apprentissage.

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

6. **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).

7. **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<sup>5</sup> sont calculés.

8. **Automatic classification of vignettes...** Cette fonction permet de sélectionner un échantillon (et éventuellement, des échantillons contextuels préalablement validés pour l'apprentissage actif) 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.

9. **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 abondances par groupes avec typiquement moins de 10% d'erreur pour tous les groupes.

10. **Active learning...** Cette fonction permet d'utiliser l'apprentissage actif. Après sélection manuelle des données contextuelles (correspondant à des échantillons préalablement validés et similaires à celui à traiter), l'ensemble d'apprentissage est automatiquement complété et adapté à l'échantillon étudié. L'outil de reconnaissance associé est ensuite appliqué et l'interface utilisateur pour la validation manuelle des prédictions est exécutée. Le reste du traitement est alors similaire à celui effectué par l'outil « Validate classification... ».

- 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.

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

---

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

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<sup>6</sup> et la teneur en carbone, le poids sec, etc. Les données sont converties par m<sup>3</sup>, 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<sup>7</sup> d'abondances ou de biomasses, soit des spectres de taille d'échantillons données.

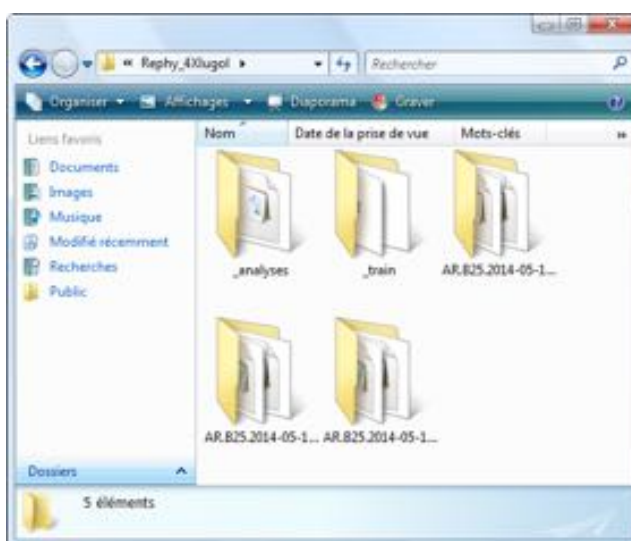
4. **Export results...** Les résultats sont écrits sur le disque dur dans un format ASCII. Ce format est lisible par d'autres logiciels (Excel, Matlab, etc.).

*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.*

## 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 :

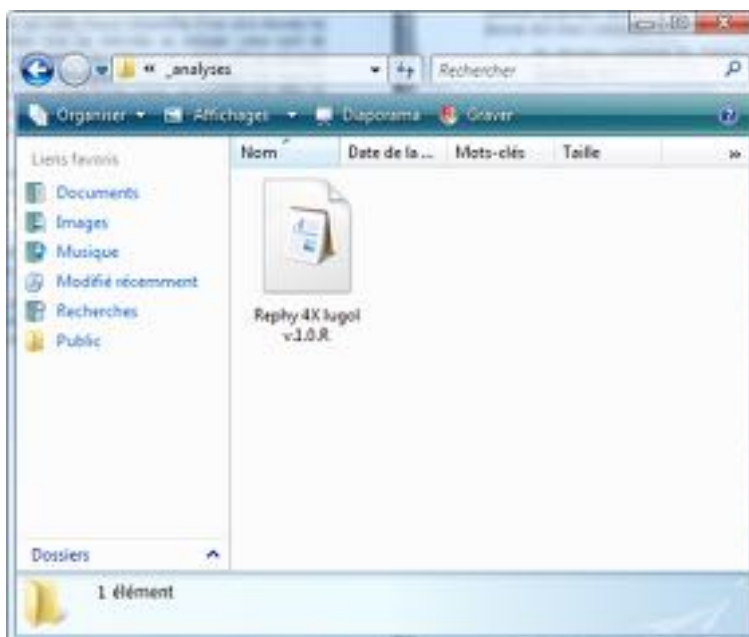
- les dossiers contenant les fichiers bruts en sortie de l'appareil d'acquisition (FlowCAM, ZooScan, etc.),



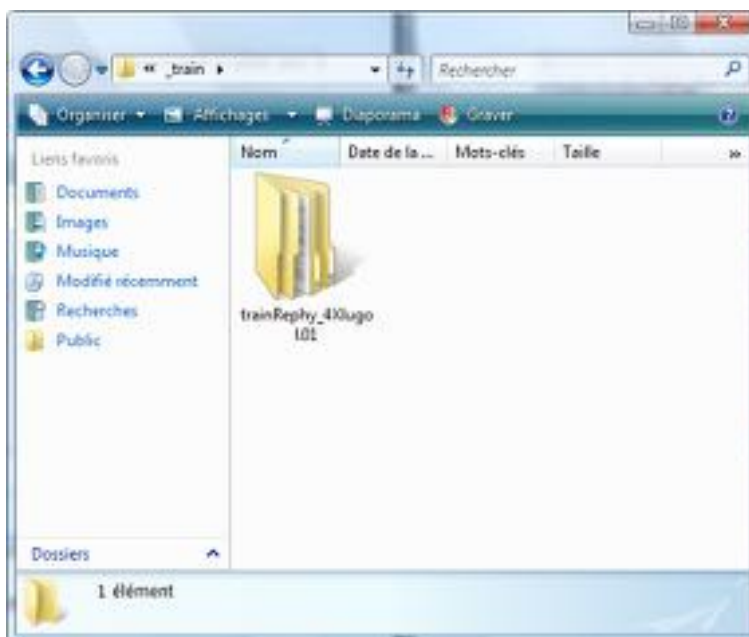
6 ECD = Equivalent Circular Diameter. (Diamètre Équivalent Circulaire)

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

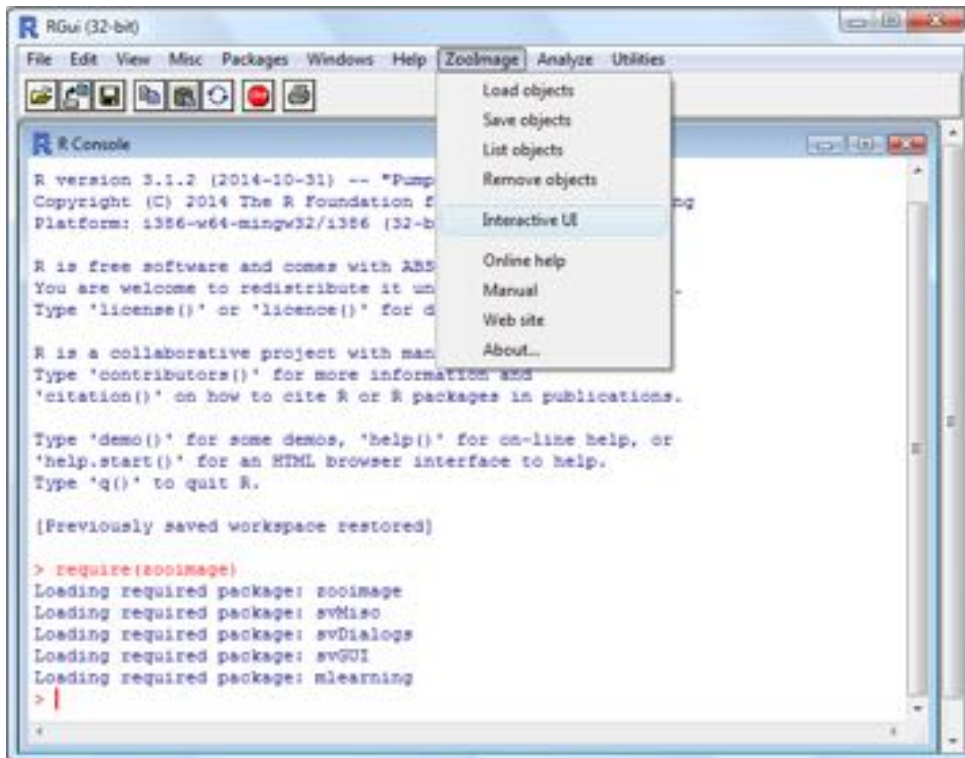
- un sous-répertoire « \_analyses » contenant les fichiers méthodes R définissant les règles d'analyse des échantillons, et qui contiendra les résultats des analyses pour chacun des échantillons,



- 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 :



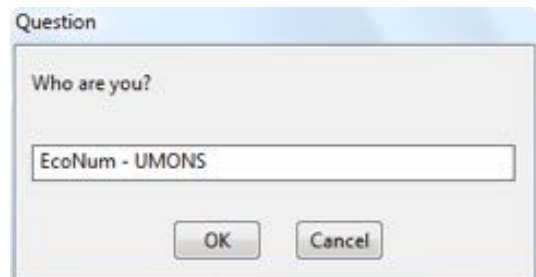
**Note :** le fichier méthode doit être édité selon les besoins et les objectifs de l'analyse. Les différents champs éditables sont présentés ci-dessous (en caractères verts et en gras dans le texte) :

- **.ZI** → nom de la méthode (method),
- **.ZI\$train** → nom du set d'apprentissage à utiliser,
- **.ZI\$classif** → nom de l'outil de reconnaissance à utiliser,
- **.ZI\$classifcmd** → algorithme de classification (method) et nombre de folds pour la validation croisée (cv.k),
- **.ZI\$biovolume** → paramètres allométriques (P1, P2, P3) pour chaque groupe défini dans Class.
- **.ZI\$breaks** → définition des classes de taille des particules pour l'affichage de la distribution.

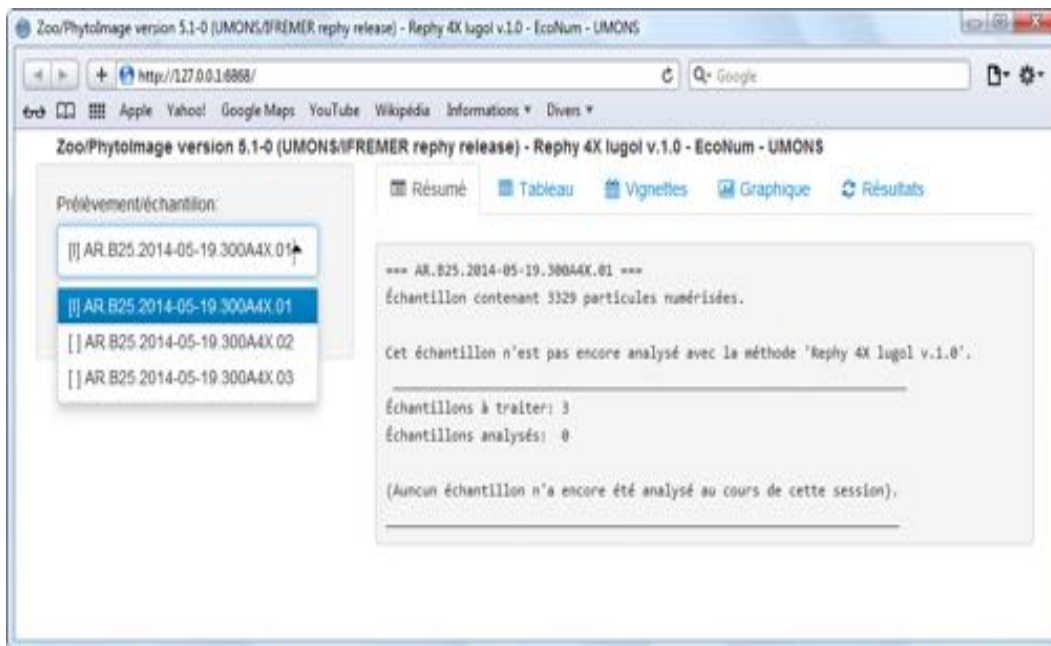
```
#####
#####
#### Parameters for this method
## This is the name of this method
.ZI <- list(user = "", date = Sys.time(), method = "Rephy 4X lugol
v.1.0", wdir = getwd(), system = "")
.ZI$scriptfile <- paste(.ZI$method, "R", sep = ".")
## This is the training set to use
.ZI$train <- "trainRephy_4Xlugol.01"
.ZI$traindir <- file.path("../", "_train", .ZI$train)
.ZI$trainfile <- paste(.ZI$traindir, "RData", sep = ".")
.ZI$classif <- "classrfRephy_4Xlugol.01"
.ZI$classifile <- file.path("../", "_train", paste(.ZI$classif,
"RData", sep="."))
.ZI$classifcmd <- paste0('ZIClass(Class ~ ., data = ', .ZI$train,
', method = "mlRforest", calc.vars = calcVarsVIS, cv.k = 10)')

## Conversion factors for biovolume
## Biovolume calculation is P1 * ECD^P3 + P2
## TODO: fill this table, or use read.delim() on a text file
## TODO: also use number of cells per colony here...
.ZI$biovolume <- data.frame(
  Class = c("Chaetoceros_spp", "[other]"),
  P1 = c(1, 1),
  P2 = c(0, 0),
  P3 = c(1, 1)
)
.ZI$breaks <- seq(0, 200, by = 10) # In um
```

Une fois le fichier méthode modifié et sauvegardé, sélectionnez ce 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.**

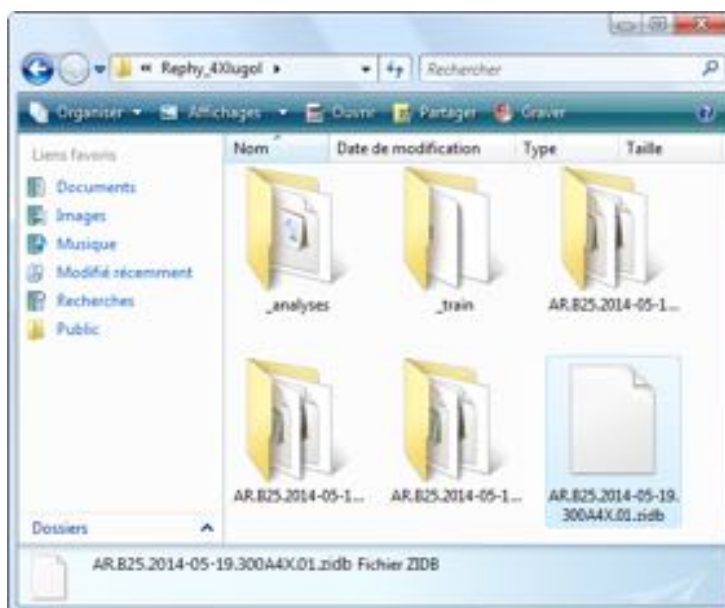
**Prélèvements/Échantillons.** Les échantillons présents dans le répertoire de travail sont listés ici. Chacun d'entre eux est précédé d'un code entre crochets.

[ ] Échantillon non importé et donc non analysé,

[ I] Échantillon importé mais non analysé,

[ A] Échantillon importé et analysé.

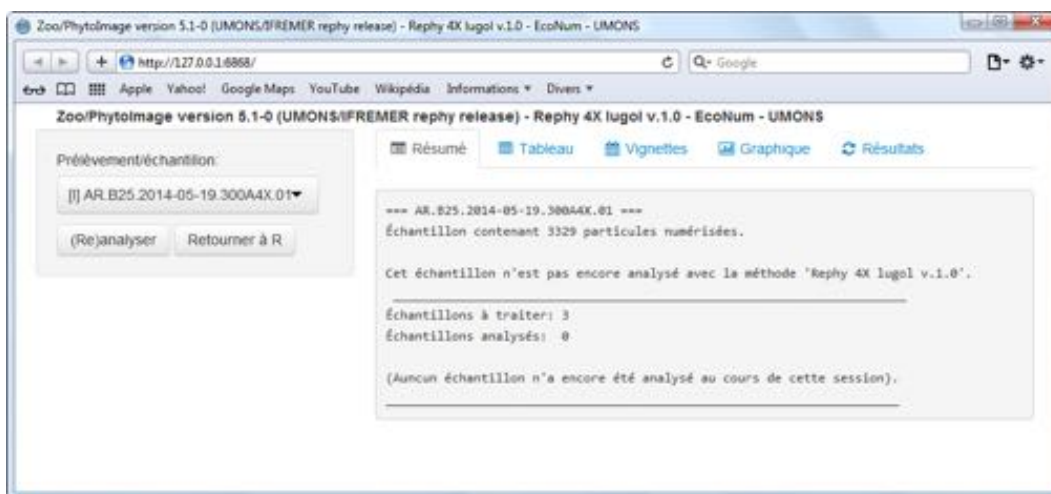
Dans cette nouvelle version de Zoo/PhytoImage, l'importation des données dans le logiciel est effectué de manière complètement automatique. Il vous suffit donc de cliquer sur l'échantillon que vous souhaitez importer afin de créer le fichier .zidb associé sur le disque (dans votre répertoire de travail).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



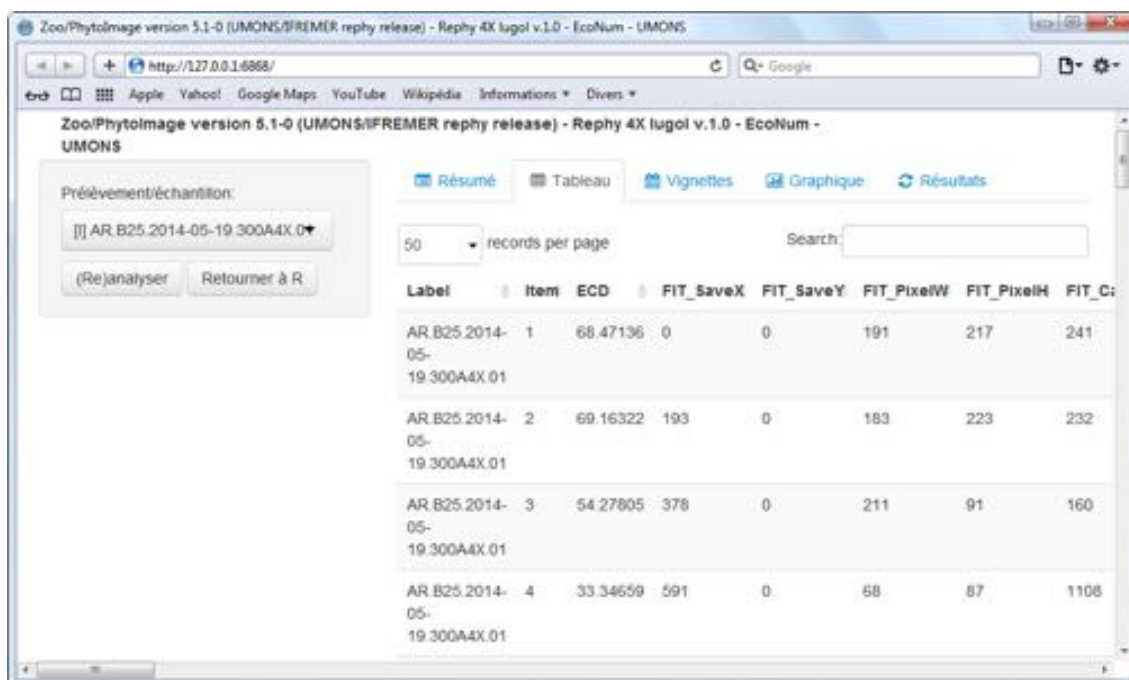


- **Visualisation des données importées.**

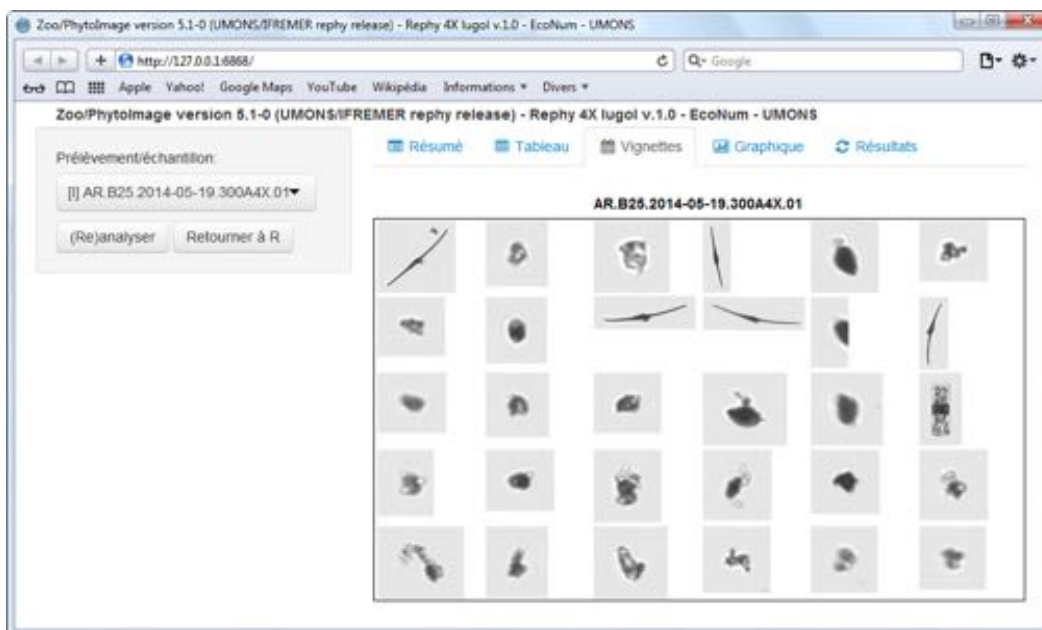
**Résumé.** Dans cet onglet, vous pouvez retrouver des informations générales sur l'échantillon sélectionné et importé : nombre de particules numérisés, état de l'analyse avec la méthode sélectionnée, nombre d'échantillons à traiter, nombre d'échantillons analysés.



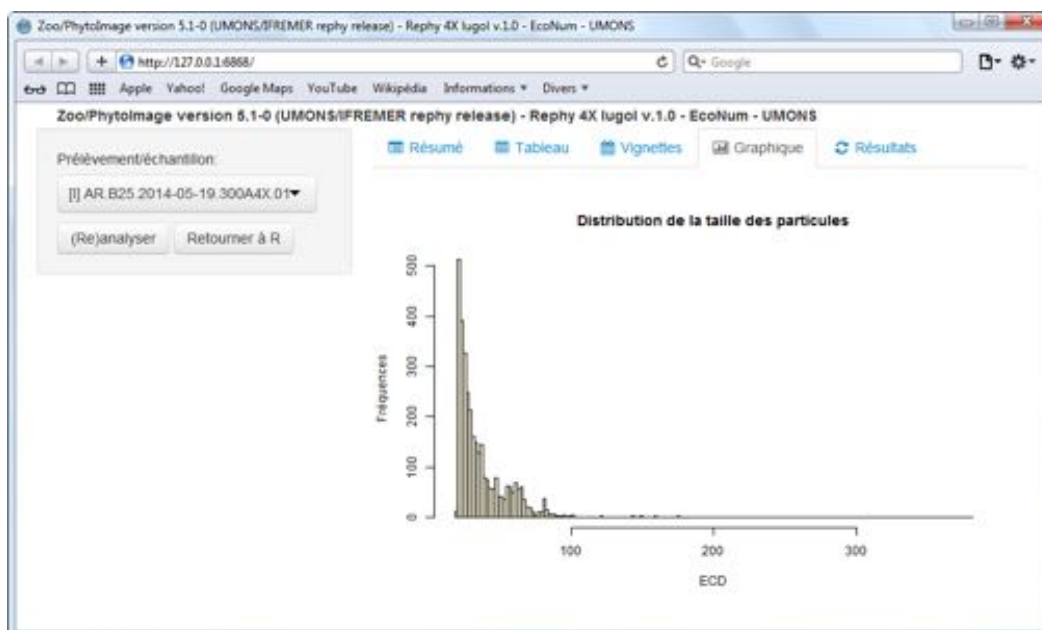
**Tableau.** Cet onglet permet une visualisation rapide et simple des mesures pour chacun des individus de l'échantillon. Il est également possible de trier par valeurs croissantes (ou décroissantes) les différentes colonnes du tableau.



**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 :

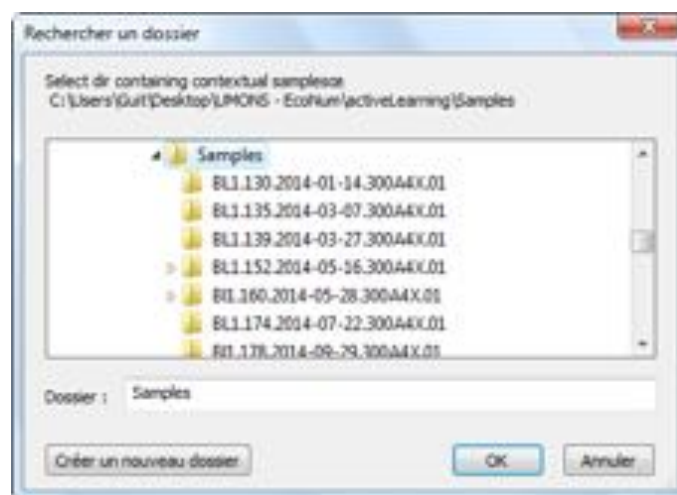
- Il est possible de ré-analyser un échantillon déjà analysé avec une méthode différente. Ceci permet une comparaison des performances de reconnaissance entre différents algorithmes.



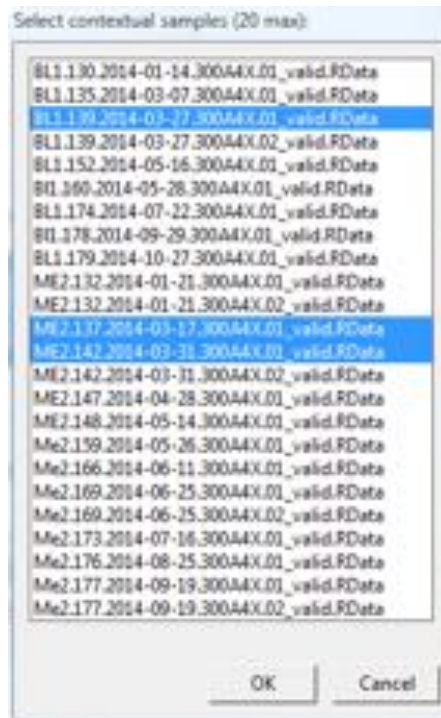
- Lors d'une première analyse avec l'interface de Zoo/PhytoImage, deux objets R sont créés dans le sous-répertoire « \_train » de votre répertoire de travail : le premier correspondant à l'ensemble d'apprentissage et le second correspondant à l'outil de reconnaissance automatique généré à partir de l'ensemble 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.



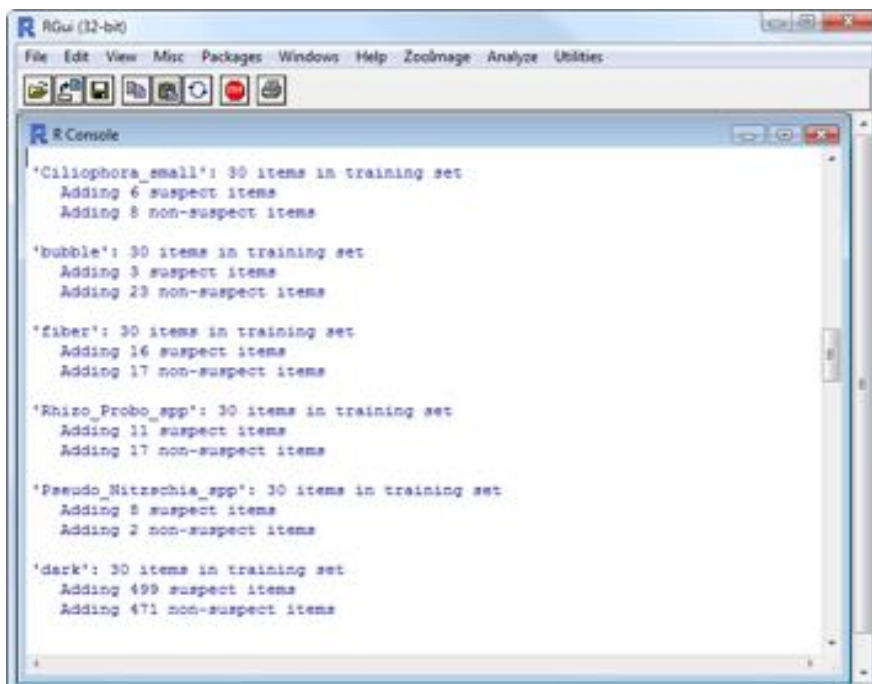
**Apprentissage actif.** Lors du traitement d'un nouvel échantillon, il est possible d'utiliser des données « contextuels » afin d'adapter l'ensemble d'apprentissage, et par conséquent, l'outil de reconnaissance associé, à l'échantillon. Selon les ensembles de données contextuels sélectionnés, l'apprentissage actif peut alors amener à une réduction significative de l'erreur de prédiction ainsi qu'à un gain de temps considérable lors du processus de validation manuelle et de correction de l'erreur. Pour cela, une boîte de dialogue est affichée et propose à l'utilisateur de sélectionner le répertoire dans lequel sont stockés des échantillons préalablement validés. **Ces échantillons correspondent, en réalité, aux fichiers « <échantillon>\_valid.RData »** générés à la fin du processus de correction de l'erreur.



En cliquant sur « **OK** », et afin d'aider l'utilisateur dans le choix des échantillons contextuels, une liste contenant les échantillons validés suivis du nombre d'items validés dans chacun d'eux est affiché. La sélection peut alors être multiple.



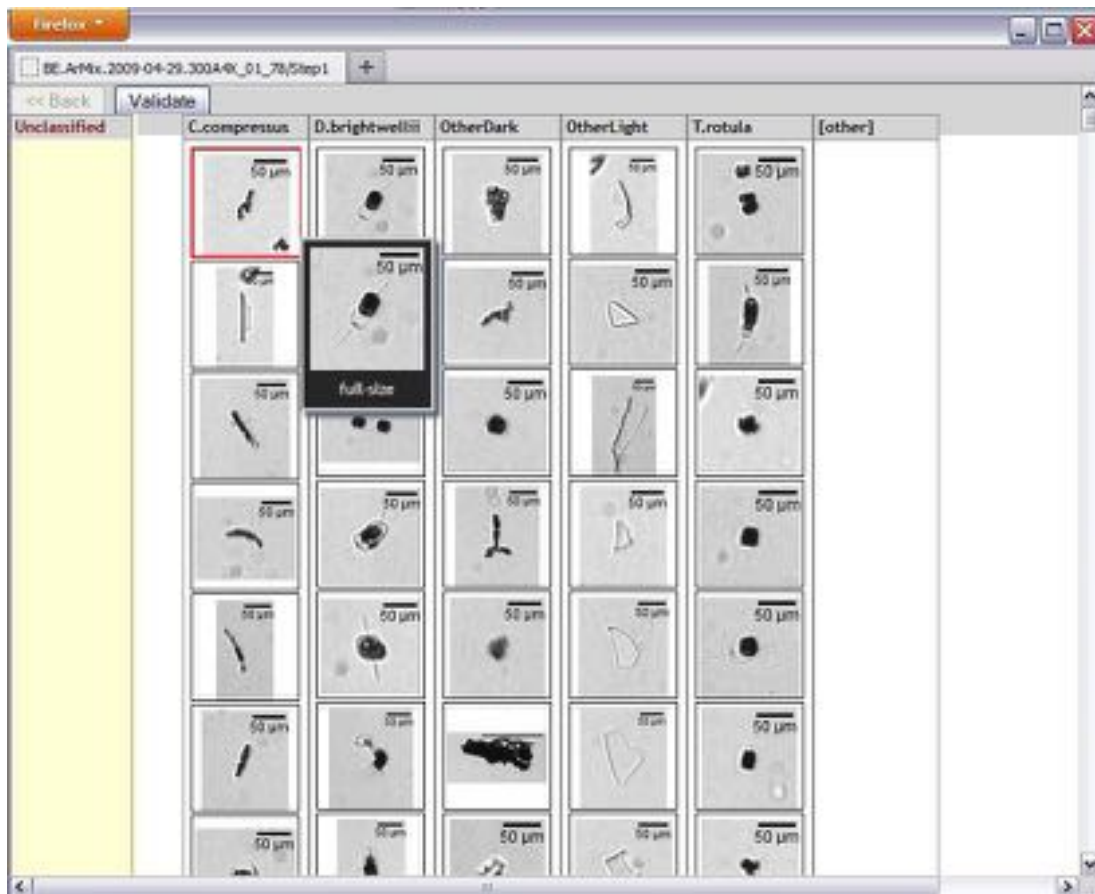
Intuitivement, la sélection de ces fichiers peut être basée sur différents critères : même zone géographique, même période (+/- 1 semaine, +/- 1 mois, ...), même contexte de numérisation, etc. Une fois la sélection effectuée, vous pouvez choisir d'autres répertoires et ainsi compléter la liste d'échantillons contextuels. Lorsque vous avez fini ces sélections, le processus d'apprentissage actif est lancé, et les groupes composant l'ensemble d'apprentissage initial sont complétés à l'aide des items validés provenant des échantillons contextuels. Cet ensemble d'apprentissage modifié est ensuite utilisé pour l'analyse de l'échantillon.



**Notes:** Les performances associées à l'apprentissage actif sont étroitement liées aux choix des échantillons contextuels.

- **Si les échantillons contextuels sont différents de l'échantillon à analyser** (en termes de zone géographique, de période, de contexte de numérisation, de qualité des images, etc.) : le gain en performance est faible (voire nul).
- **Si les échantillons contextuels sont similaires à l'échantillon à analyser** (en termes de zone géographique, de période, de contexte de numérisation, de qualité des images, etc.) : le gain en performance est important.

**Correction erreur.** Ici est décrit le fonctionnement de l'outil de validation de la classification. Une fois l'adaptation de l'ensemble d'apprentissage terminée, Zoo/PhytoImage crée une page web qui vous présente un premier ensemble de (par défaut) 1/20ème des vignettes dans l'échantillon (avec un nombre minimum fixé à 100 et un nombre maximum fixé à 200 vignettes).

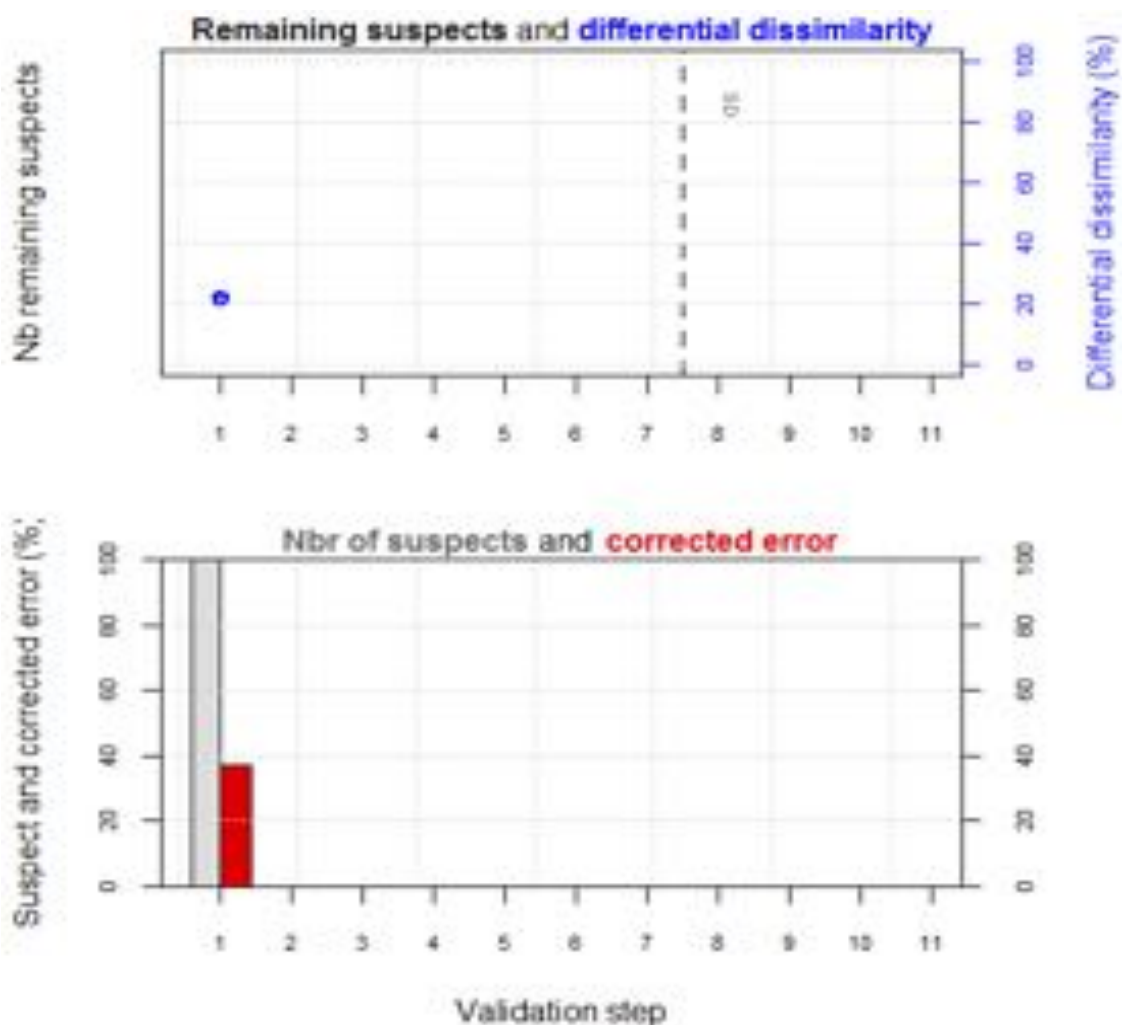


Cette page présente une première série de particules, sélectionnées aléatoirement dans l'échantillon, triées automatiquement par l'outil de reconnaissance choisi. Chaque classe est représentée par une colonne dans la page (e.g., *Ceratium\_furca*, *Ceratium\_fusus*, etc. dans l'exemple). Toutes les vignettes 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.



Ce rapport présente deux graphes :

- un diagramme en bâtons représentant le nombre total d'objets suspects restant à valider dans l'échantillon, ainsi qu'un premier point bleu indiquant la valeur de dissimilarité différentielle globale entre les abondances corrigées obtenues à l'étape précédente et celles obtenues à l'issue de l'étape courante. Pendant cette première étape, aucun modèle n'est calculé... donc, tous les objets sont considérés comme suspects, et la dissimilarité différentielle est calculée en utilisant les prédictions automatiques et celles obtenues après cette première phase de validation.

En plus de ces courbes, un premier indicateur (ligne pointillée grise étiquetée SD) est affichée. Ce dernier permet de donner une bonne indication sur le nombre d'étape restantes (et donc sur le travail restant à fournir) afin de valider la totalité des suspects dans l'échantillon et donc de corriger (quasi-)complètement l'erreur.

- un diagramme en bâtons avec des bâtons gris clairs 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 30%. *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 30% 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 4/5, ce qui représente environ 16/20ème 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, probablement classée à tort) ou non.

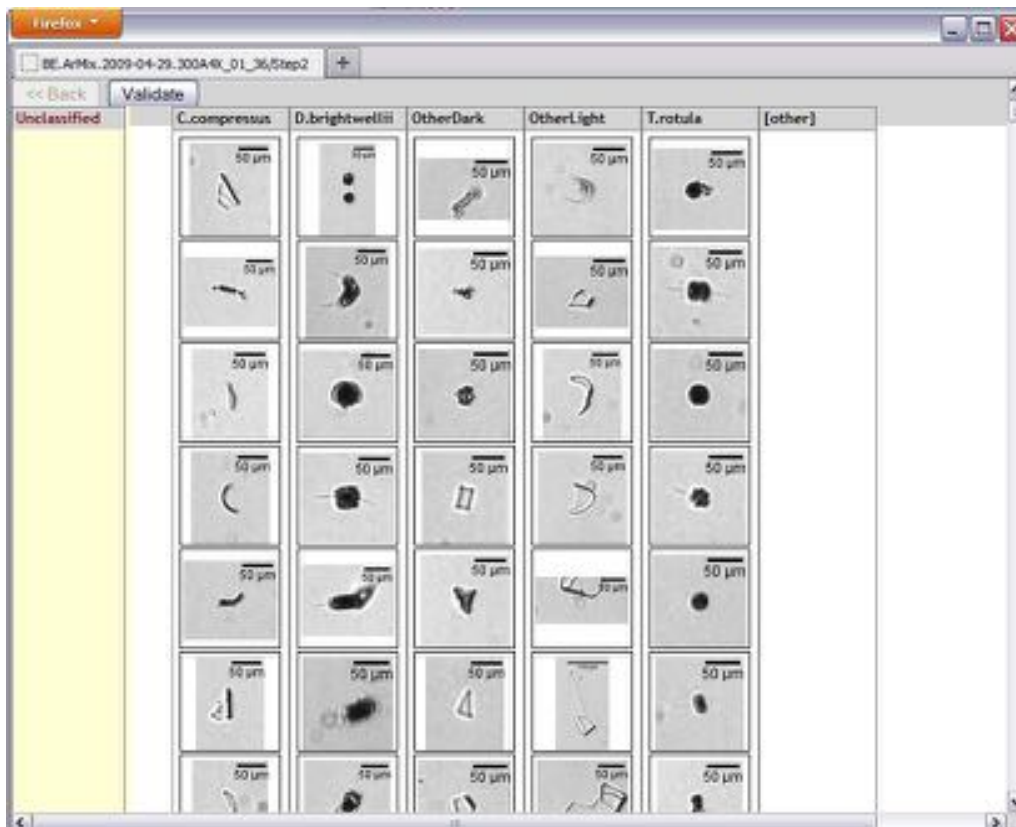
Le modèle considère également plusieurs autres aspects :

- 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.

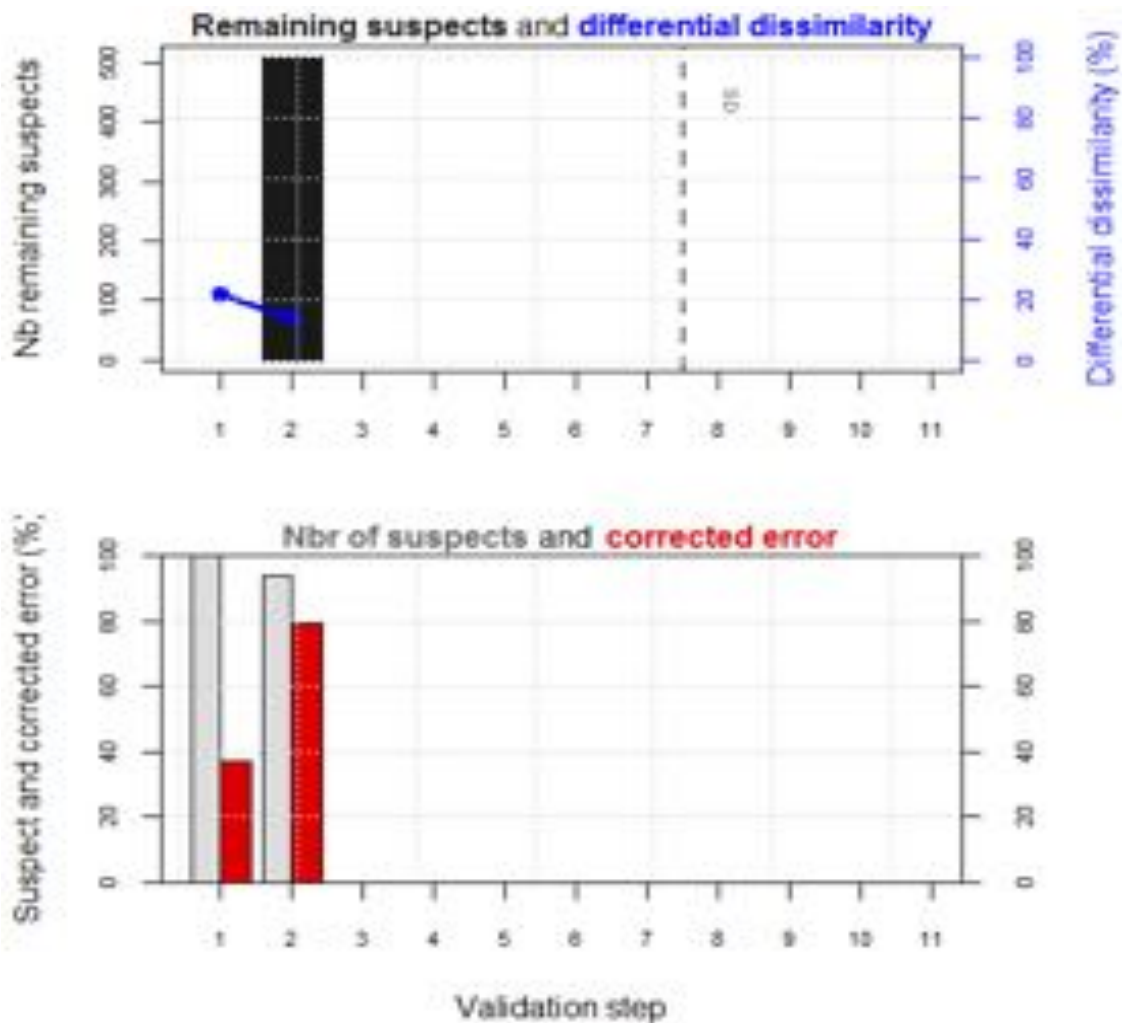
Zoo/PhytoImage utilise le premier ensemble de particules comme un ensemble d'apprentissage pour détecter les objets suspects, en utilisant tous les attributs mesurés sur ces particules, ainsi que les variables additionnelles décrites ci-dessus. Plusieurs algorithmes peuvent être utilisés, mais Random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



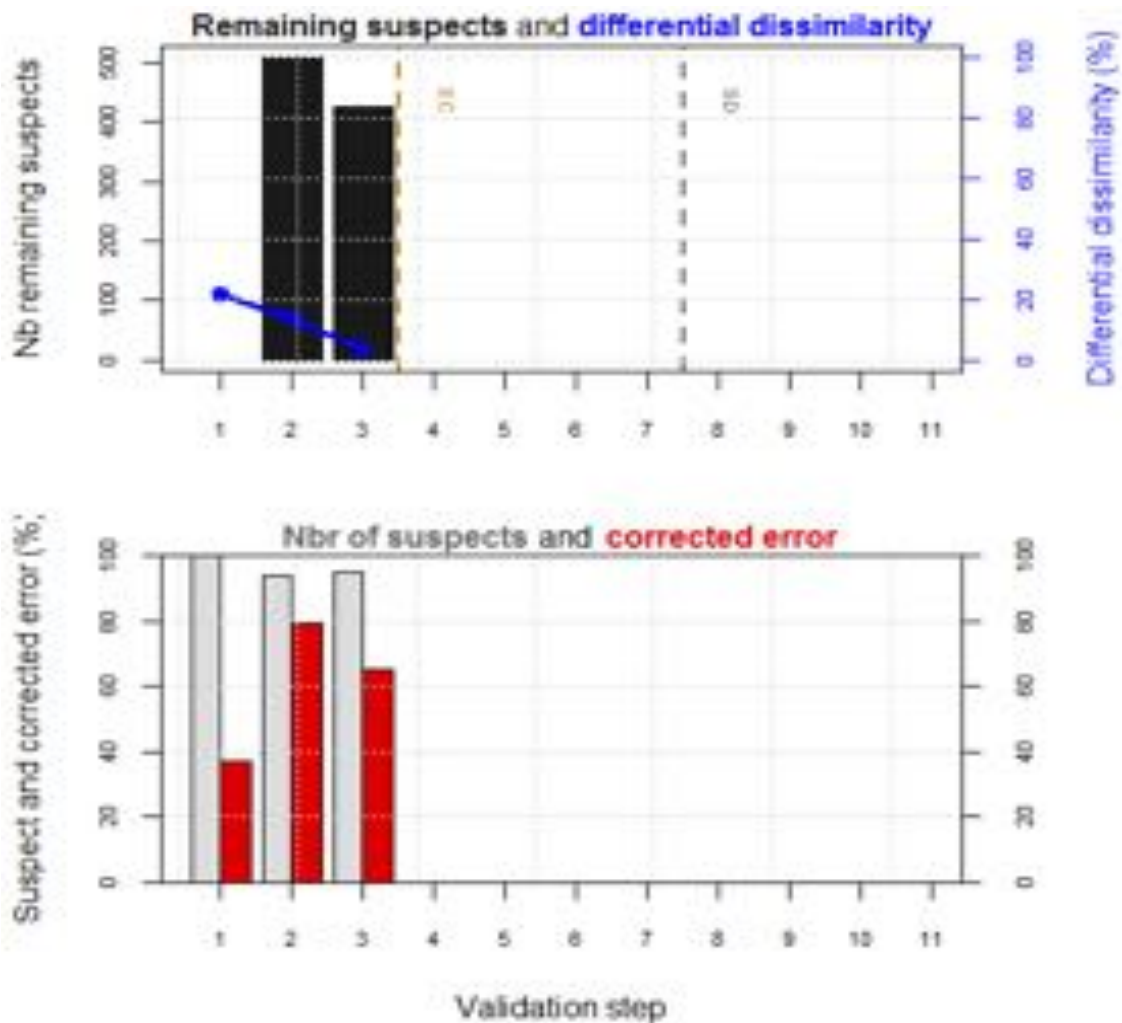
Généralement, il est assez clair que ce second ensemble présente beaucoup plus d'erreur que le précédent... et vous remarquerez également que, en effet, vous avez également beaucoup plus de particules « problématiques » (difficulté à reconnaître les particules, objets coupés, blobs avec une forme étrange, 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 faites ici). Cliquez sur **Validate** lorsque vous en avez fini avec cette deuxième étape.





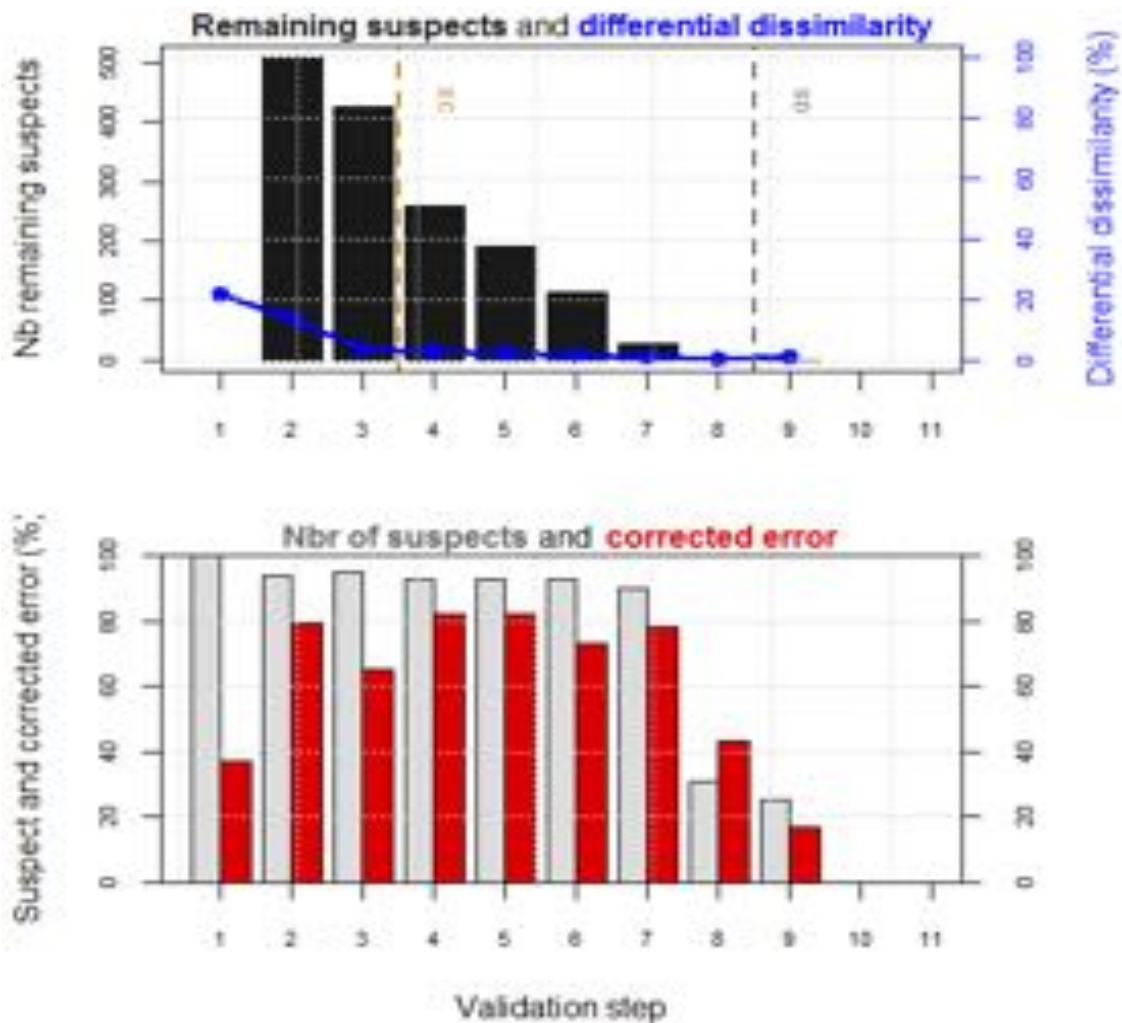
Dans le rapport, le premier graphe possède maintenant une barre grise foncée et un second point bleu. Cette barre grise permet de se faire une bonne idée du travail qu'il reste à fournir afin de valider l'ensemble des objets suspects restants. Vous remarquez également que la dissimilarité différentielle a diminué. Ceci montre que la différence entre les abondances corrigées obtenues après l'étape 1 et l'étape 2 sont moins importantes que celle entre les abondances obtenues automatiquement et après l'étape 1.

Pour le second diagramme en bâtons, une seconde série de barres grises/rouges est maintenant affichée. 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 triplé la fraction de particules erronées à cette étape. Vous avez maintenant concentré l'erreur plus efficacement. Lancez le une troisième fois.



Pour cet échantillon, à la troisième étape, un nouvel indicateur apparaît sur le premier graphe (ligne pointillée brune étiquetée EC). Cette ligne indique que la dissimilarité différentielle a atteint un certain seuil (par défaut, 5%). Cela signifie que les abondances obtenues à l'étape courante sont fortement similaires à celles obtenues à l'étape précédente, même après validation d'une nouvelle fraction de suspects. Si vous le souhaitez, vous pouvez alors stopper le processus de validation manuelle et faire confiance à la correction introduite par cette validation partielle, et par la correction statistique grâce au modèle de détection des suspects, afin d'obtenir des abondances par classe corrigées qui seront pertinentes. Cependant, si vous cherchez à obtenir des prédictions fiables pour chaque classe (et en particulier pour les groupes rares ou peu abondants), vous pouvez continuer la validation jusqu'à atteindre le premier indicateur gris SD. Continuez avec quelques autres sous-ensembles :





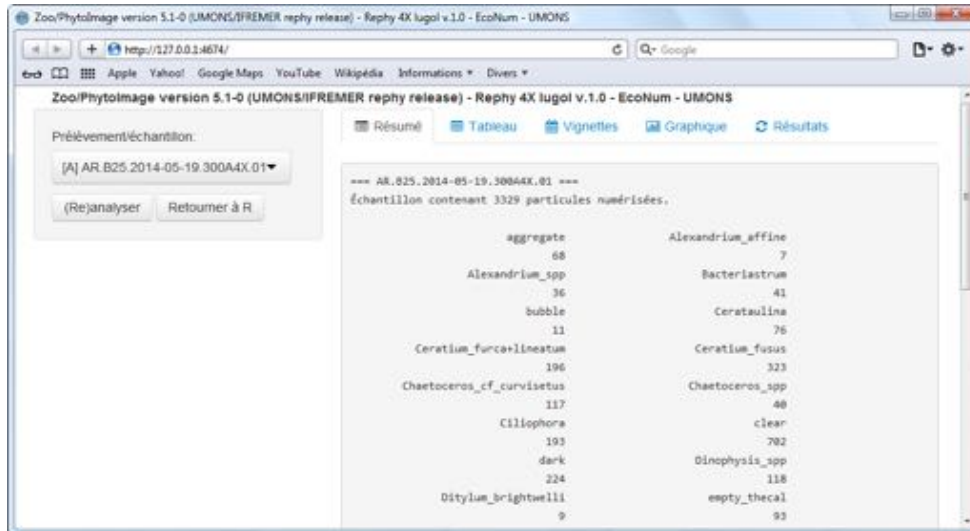
Ici, après l'étape 7, vous remarquez deux choses importantes. Premièrement, la détection des suspects correspond maintenant étroitement avec l'erreur réelle. La détection est améliorée avec la fraction de l'échantillon déjà validée qui peut être utilisée pour entraîner l'algorithme de détection. Deuxièmement,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

A partir de ce moment, vous savez que vous avez validé manuellement toutes les particules erronées jusqu'à environ 5%. 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. De plus, puisque le modèle est utilisé pour calculer un *facteur de correction* pour les objets restants, le calcul des abondances par classe deviendra assez bon.

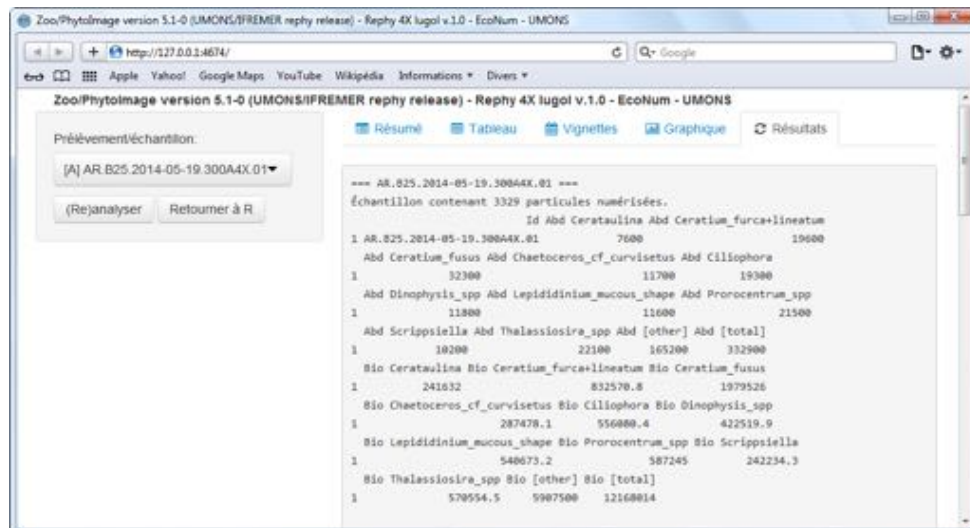
**Note :** Comme illustré sur la figure ci-dessus, il est conseillé d'effectuer une étape supplémentaire au-delà de l'indicateur SD. Celle-ci permet de confirmer la chute du nombre de particules erronées et celle de l'erreur résiduelle. De plus, dans la perspective de l'utilisation de l'échantillon en tant qu'« échantillon contextuel » dans le cadre de l'apprentissage actif, et parce que les particules suspectes sont désormais toutes validées, cette étape additionnelle permet de valider davantage de particules non suspectes, nécessaires au bon fonctionnement du processus d'apprentissage actif.

Donc, en gardant cela à l'esprit, vous pouvez raisonnablement considérer que la validation pourrait être arrêtée maintenant. 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 **Prélèvement/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.



width size spectrum:  
\$'AR.B25.2014-05-19.30044X.01'

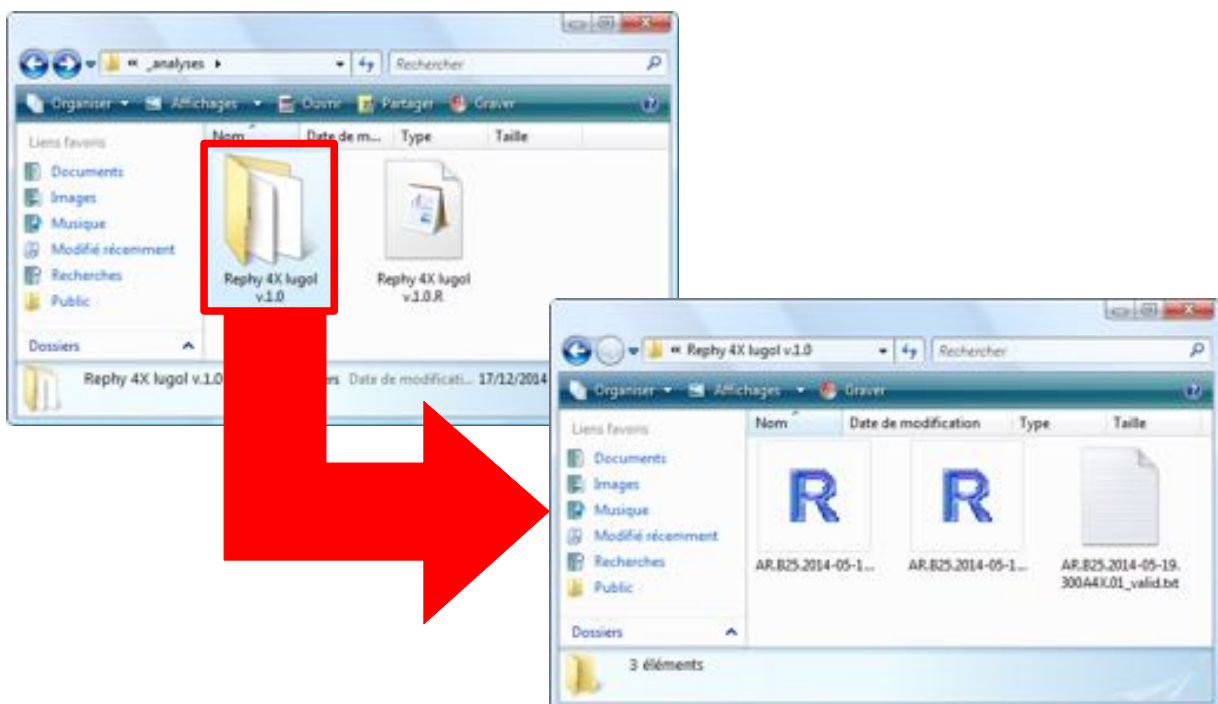
	(0,10]	(10,20]	(20,30]	(30,40]	(40,50]	(50,60]
Cerataulina	0	0	3000	4200	400	0
Ceratium_furc+lineatum	0	0	0	10400	7900	900
Ceratium_fusus	0	0	0	100	900	10000
Chaetoceros_cf_curvisetus	0	100	18000	900	100	0
Ciliophora	0	0	13000	4600	300	100
Dinophysis_spp	0	0	1900	7600	1900	200
lepididinium_mucous_shape	0	0	1400	2800	3200	2100
Prorocentrum_spp	0	0	15300	6200	0	0
Scrippsiella	0	0	9500	700	0	0
Thalassiosira_spp	0	0	17700	4200	200	0
[other]	0	1100	95000	24800	16100	16500
[total]	0	1200	169100	66500	31600	24700

	(60,70]	(70,80]	(80,90]	(90,100]	(100,110]	(110,120]
Cerataulina	0	0	0	0	0	0
Ceratium_furc+lineatum	200	100	0	0	0	0
Ceratium_fusus	18600	1000	0	0	0	0
Chaetoceros_cf_curvisetus	0	0	0	0	0	0
Ciliophora	100	100	200	100	0	0
Dinophysis_spp	100	100	0	0	0	0
lepididinium_mucous_shape	1100	600	200	100	100	0

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 :

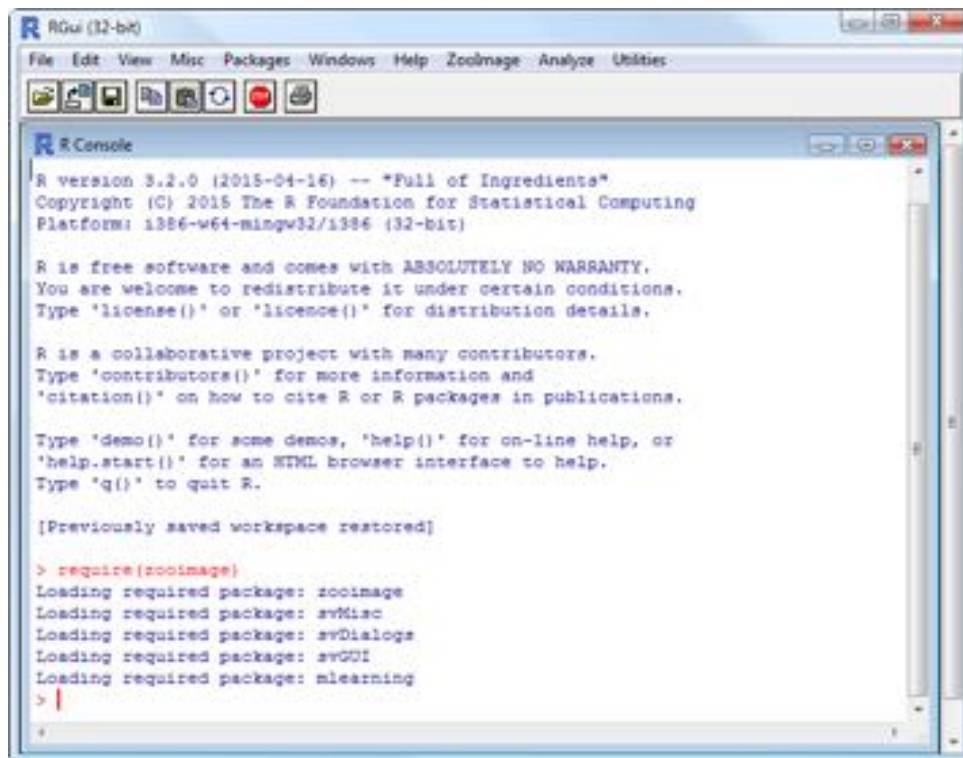
- `<echantillon>_res.RData` : ce fichier peut être chargé et manipulé dans R. Il contient les abondances, les biovolumes et les spectres de taille pour chaque groupe taxonomique,
- `<echantillon>_valid.RData` : ce fichier peut également être chargé et manipulé dans R. Il contient les mesures sur chacune 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.



## 6. OUTIL DE DENOMBREMENT DES CELLULES EN COLONIES

Un outil visuel d'aide au dénombrement des cellules dans les vignettes a été développé pour Zoo/PhytoImage. Il consiste en l'affichage de la particule et de quelques mesures élémentaires telles que la longueur, la largeur et l'ECD, puis le marquage des cellules comptées manuellement (par clics souris). Une fois la totalité des cellules dénombrée, une nouvelle entrée est créée directement et automatiquement pour chacune des particules dans l'ensemble d'apprentissage.

Pour lancer le module de dénombrement manuel des cellules sur les vignettes de l'ensemble d'apprentissage, sélectionnez l'entrée « Count cell(s) per particle » du menu « ZooImage ».



```
RGui (32-bit)
File Edit View Misc Packages Windows Help ZooImage Analyse Utilities

R Console
R version 3.2.0 (2015-04-16) -- "Pull of Ingredients"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

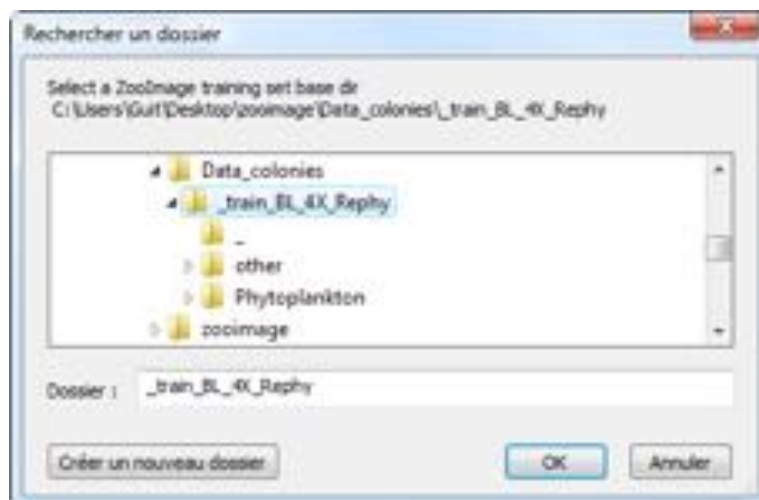
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> require(zooimage)
Loading required package: zooimage
Loading required package: svMisc
Loading required package: svDialogs
Loading required package: svGUI
Loading required package: mlearning
>
```

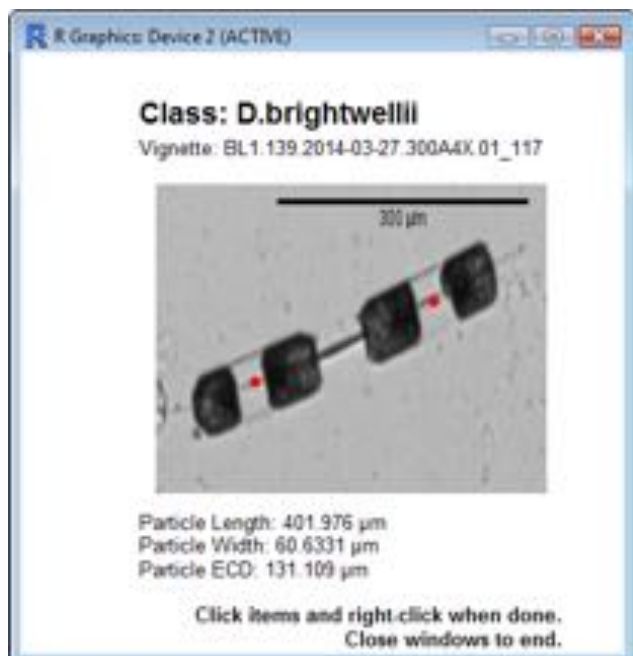
Une première boîte de dialogue vous permettant de sélectionner l'ensemble d'apprentissage à traiter, est alors affichée.



Une fois l'ensemble d'apprentissage choisi, une seconde boîte de dialogue est affichée. Celle-ci propose une liste des classes présentes dans l'ensemble d'apprentissage. Sélectionnez-en une afin de commencer les dénombrements.

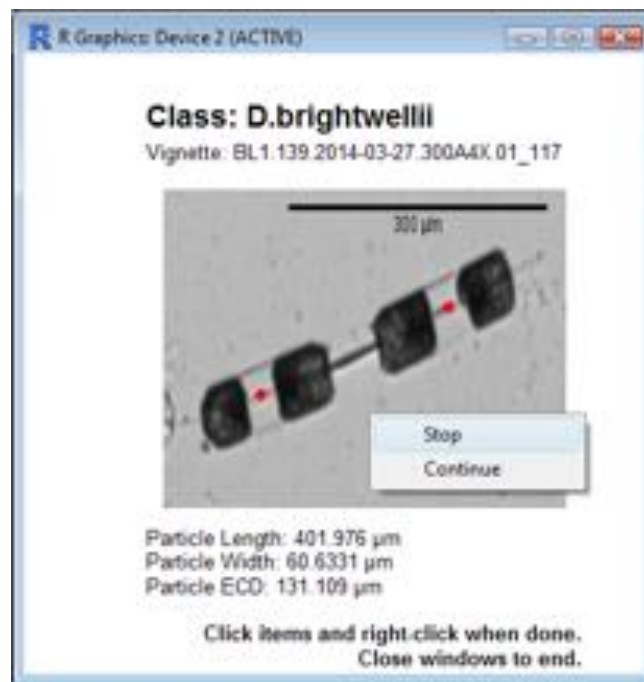


Lorsque vous cliquez sur le bouton « OK », les vignettes correspondant à la classe sélectionnée dans l'ensemble d'apprentissage vous sont présentées les unes après les autres. Sur ces figures, la classe d'appartenance, l'identifiant de la particule, ainsi que sa longueur, sa largeur et son ECD (Equivalent Circular Diameter) sont affichés. Pour dénombrer les cellules de cette vignette, il suffit alors de **cliquer à la souris sur chacune des cellules identifiées** (chaque cellule est alors marquée d'un point rouge afin d'aider l'utilisateur dans le comptage) :

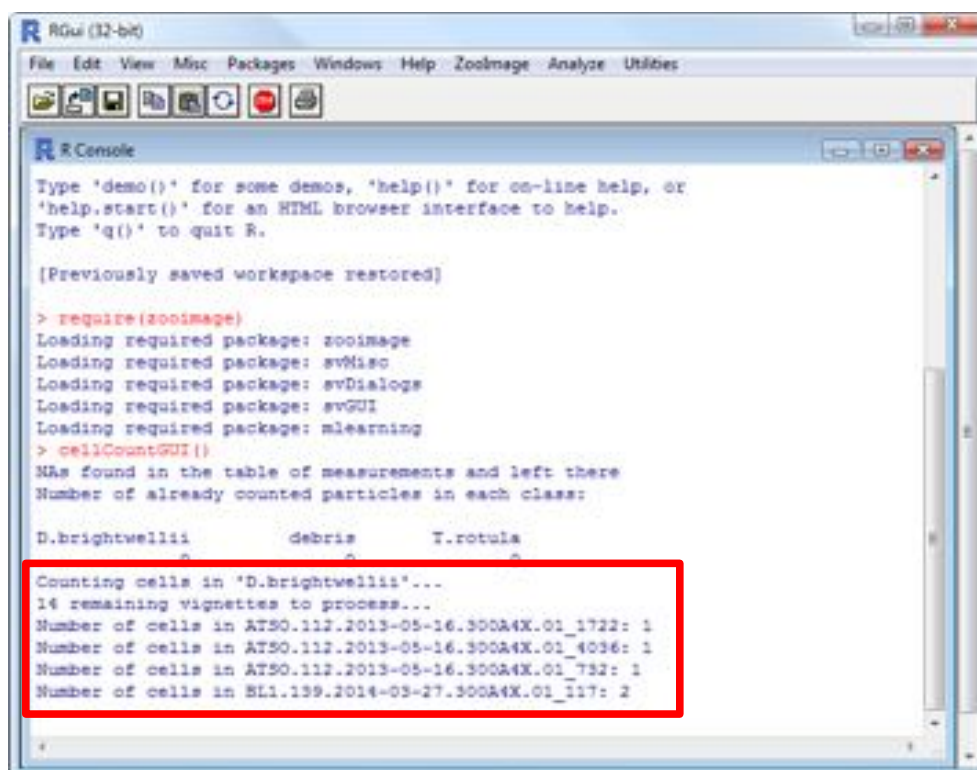




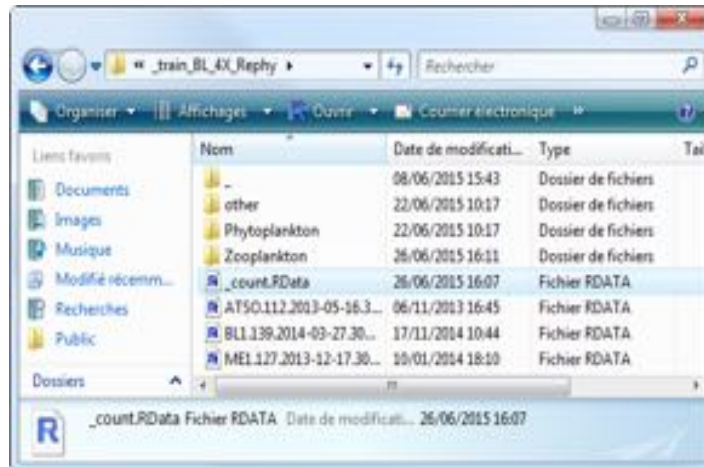
Lorsque le dénombrement des cellules est terminé, et afin de passer à la vignette suivante, il vous suffit de **cliquer droit avec la souris, et de sélectionner « STOP »**.



La vignette suivante à traiter est alors proposée à l'utilisateur et le nombre de cellules dénombrée dans la vignette précédente, ainsi que le nombre total de vignettes dénombrées dans le groupe sont affichés dans la console R.



Pour chaque vignette traitée, le nombre de cellules correspondant est directement et automatiquement sauvegardé dans l'ensemble d'apprentissage. Un nouvel objet R (au format RData) et nommé « \_count.RData » est alors créé à la racine de l'ensemble d'apprentissage.



**Remarque :** Pour le dénombrement, plusieurs possibilités s'offrent à l'utilisateur :

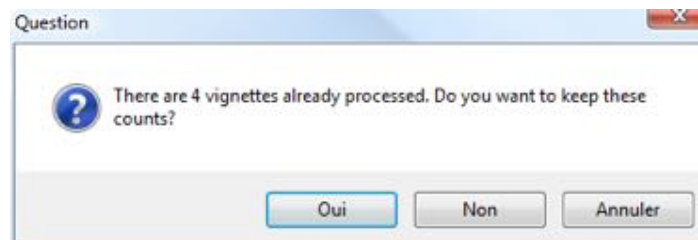
- si **le nombre de cellules est dénombrable**, l'utilisateur clique sur chacune des cellules identifiées, puis passe à la vignette suivante en cliquant droit et en sélectionnant « STOP ». **La valeur est alors enregistrée dans l'ensemble d'apprentissage.**
- si **le nombre de cellules n'est pas dénombrable** (forme complexe, mauvaise qualité d'image, etc.), l'utilisateur passe directement à la vignette suivante en cliquant droit puis en sélectionnant « STOP ». **Aucune valeur (NA) n'est alors enregistré.**
- pour stopper le processus de dénombrement, il suffit de fermer la fenêtre contenant la vignette. Le processus peut alors être repris ultérieurement.

### ***Arrêt/Reprise du dénombrement de cellules et modification de l'ensemble d'apprentissage.***

Il existe plusieurs cas pour lesquels le processus de dénombrement manuel peut être interrompu puis repris ultérieurement :

- le nombre de vignettes à traiter dans une classe est trop important,
- les cellules sur les vignettes d'une classe sont difficilement dénombrables,
- la classe de l'ensemble d'apprentissage a été modifiée (ajout de nouvelles vignettes).

Dans chacun des cas, après sélection de la classe à traiter, une boîte de dialogue permettant de conserver ou non les dénombrements effectués auparavant, est affichée.



- En sélectionnant « OUI », le processus ne propose à l'utilisateur QUE les vignettes NON traitées. Le nombre de vignettes restantes à traiter est alors affiché dans la console R.
- En sélectionnant « NON », le processus réinitialise les comptages et propose à l'utilisateur la totalité des vignettes. Le nombre de vignettes restantes à traiter est alors affiché dans la console R.

## 7. 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/](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 incluses 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 un 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 crée 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.