

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

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

- Un microprocesseur multi-cœur 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.

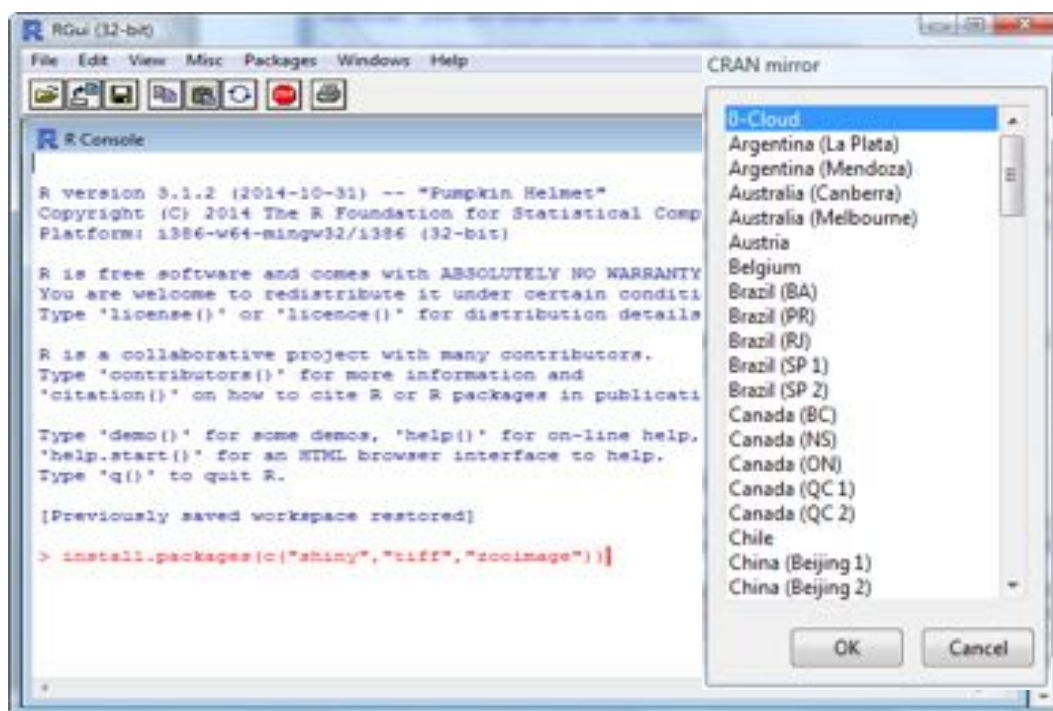
### 3.2. Installation de Zoo/PhytoImage sous Windows

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.

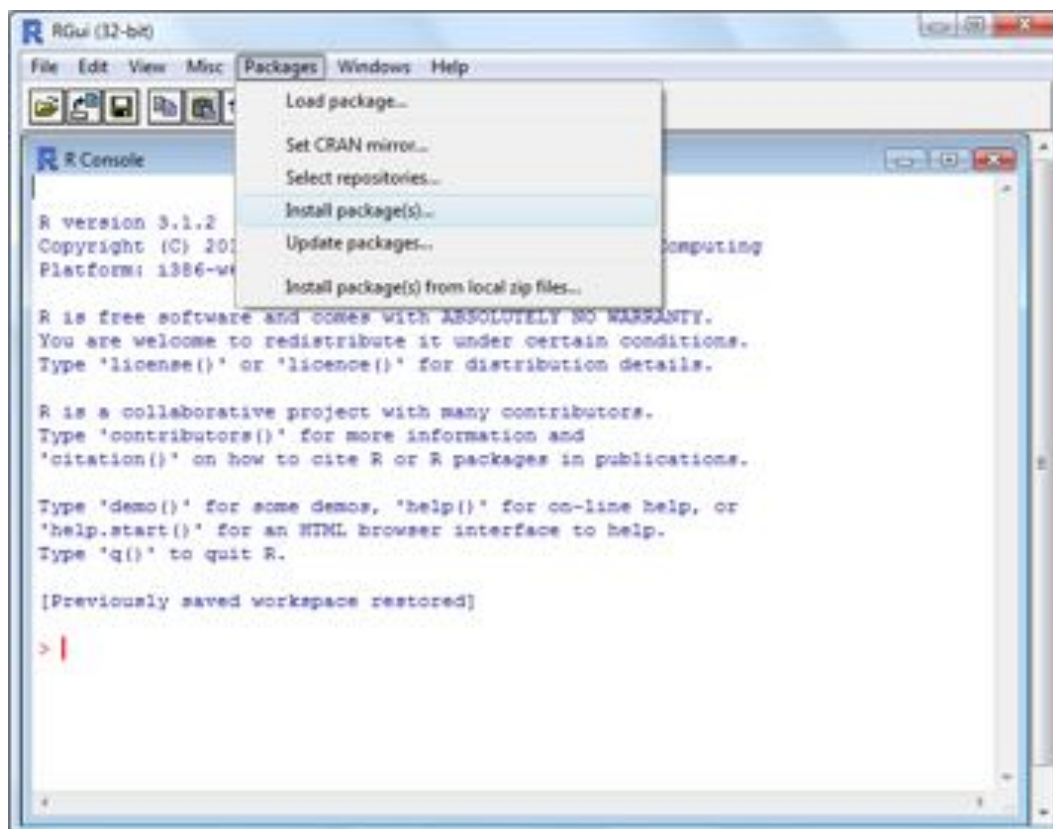
Les packages nécessaires (« shiny », « tiff » et « zooimage ») peuvent être installés directement à partir de la console R, en tapant : `install.packages(c("shiny", "tiff", "DT", "zooimage"))`. Choisissez ensuite un miroir (par défaut : 0-cloud) pour démarrer les téléchargements et les installations.



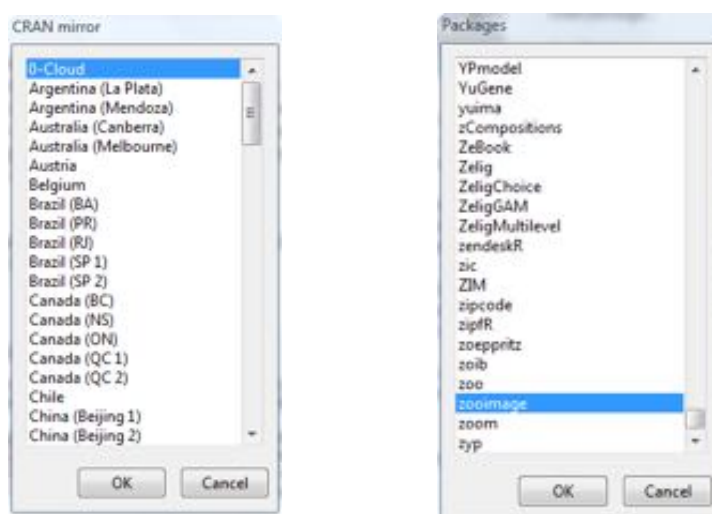
Capture d'écran de la console R pour l'installation de Zoo/PhytoImage version 5.

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.



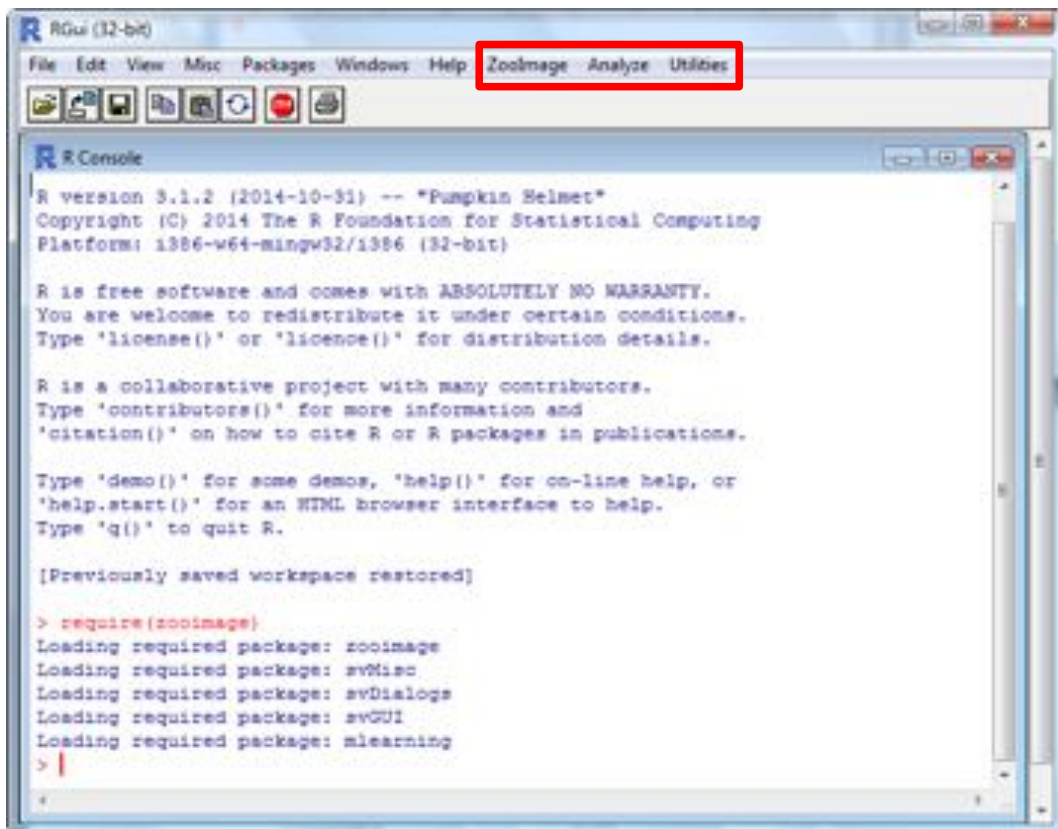
Capture d'écran pour l'installation manuelle de Zoo/PhytoImage version 5.



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.

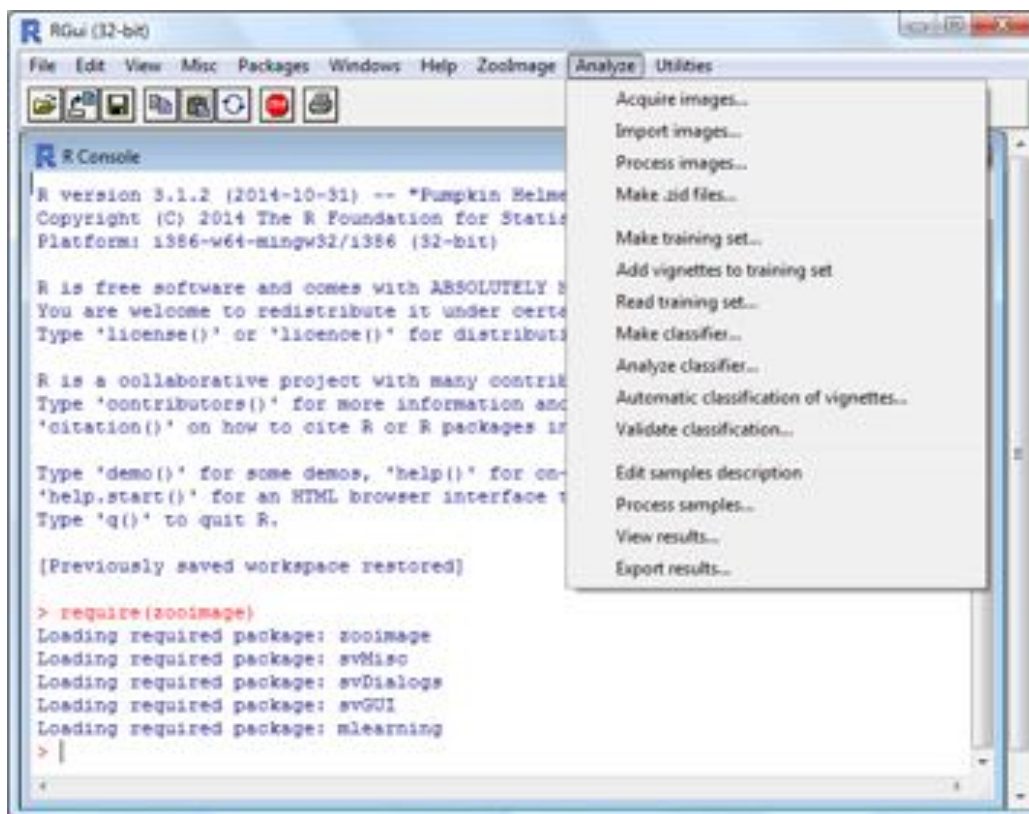


## 4. UTILISATION DES FONCTIONS A PARTIR DU MENU DE ZOO/PHYTOIMAGE

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.

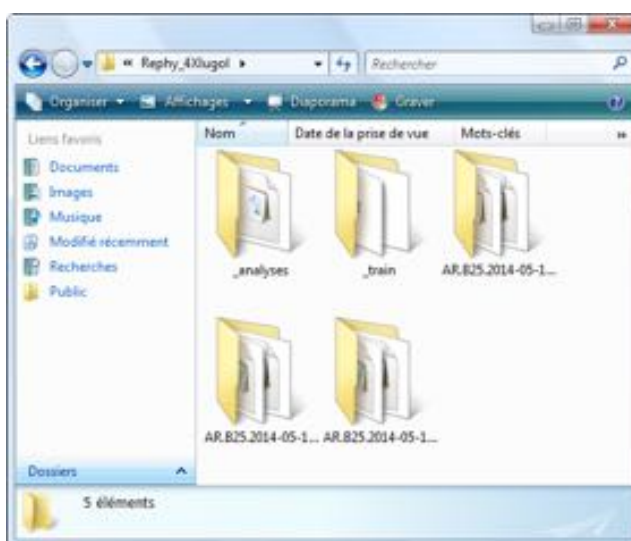
4. **Export results...** Les résultats sont écrits sur le disque dur dans un format ASCII. Ce format est lisible par d'autres logiciels (Excel, Matlab, etc.).

*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 :

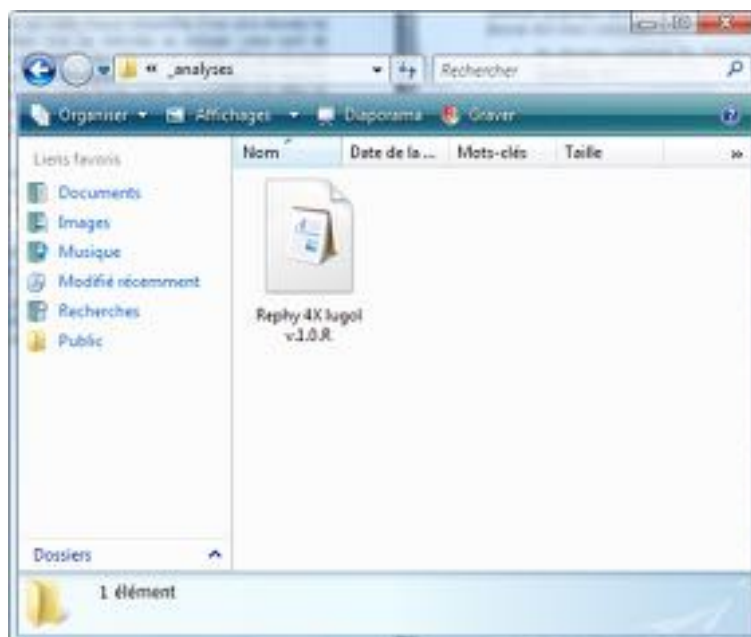
- les dossiers contenant les fichiers bruts en sortie de l'appareil d'acquisition (FlowCAM, ZooScan, etc.),



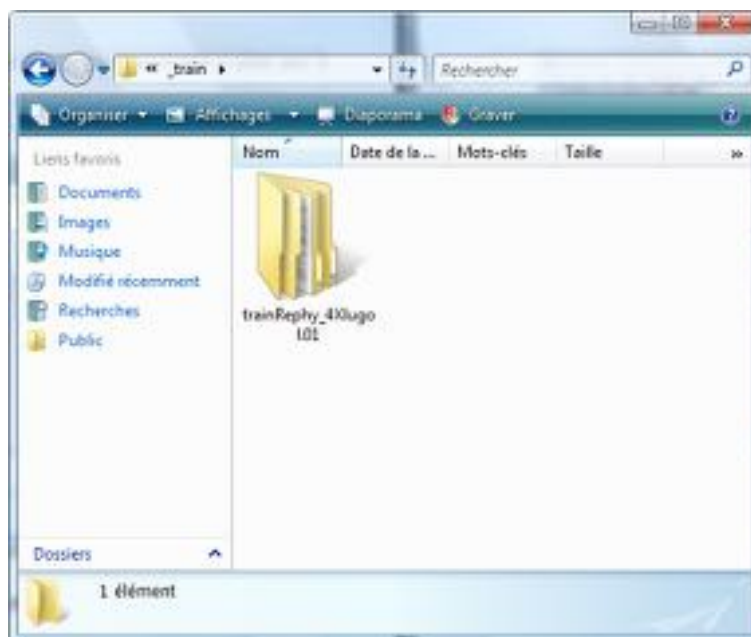
6 ECD = Equivalent Circular Diameter. (Diamètre Équivalent Circulaire)

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

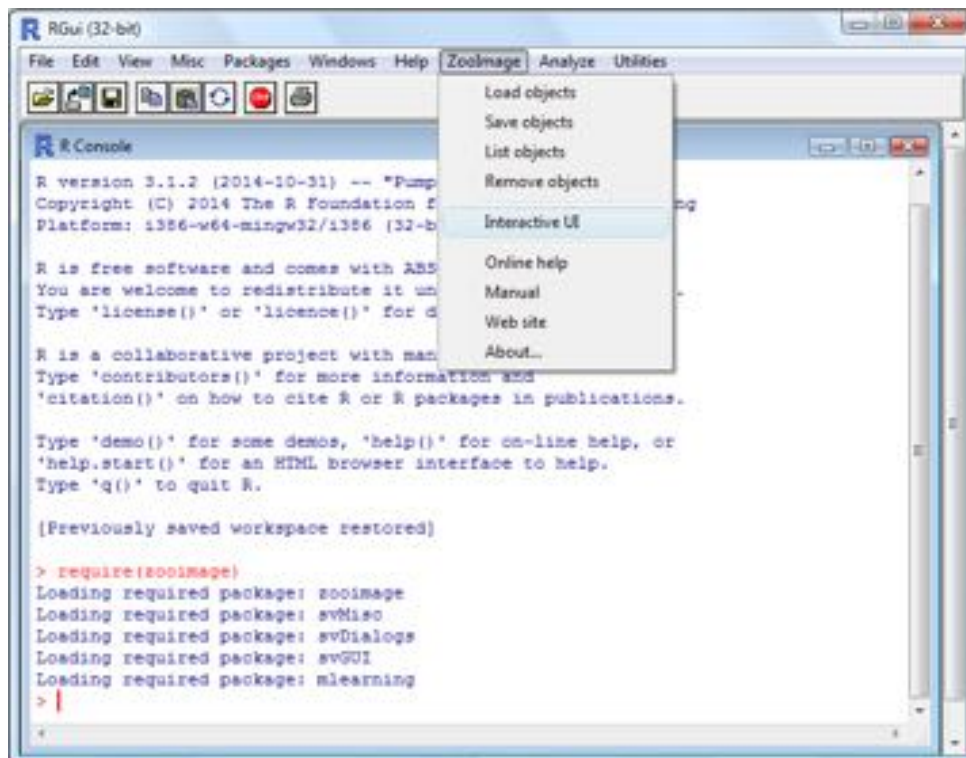
- un sous-répertoire « \_analyses » contenant les fichiers méthodes R définissant les règles d'analyse des échantillons, et qui contiendra les résultats des analyses pour chacun des échantillons,



- 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 :



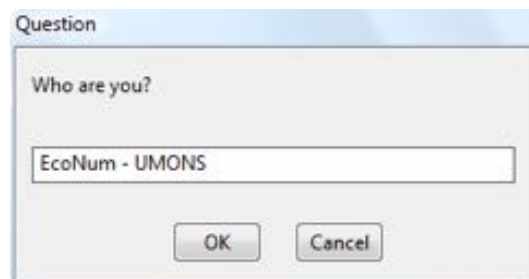
**Note :** le fichier méthode doit être édité selon les besoins et les objectifs de l'analyse. Les différents champs éditables sont présentés ci-dessous (en caractères verts et en gras dans le texte) :

- **.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)')

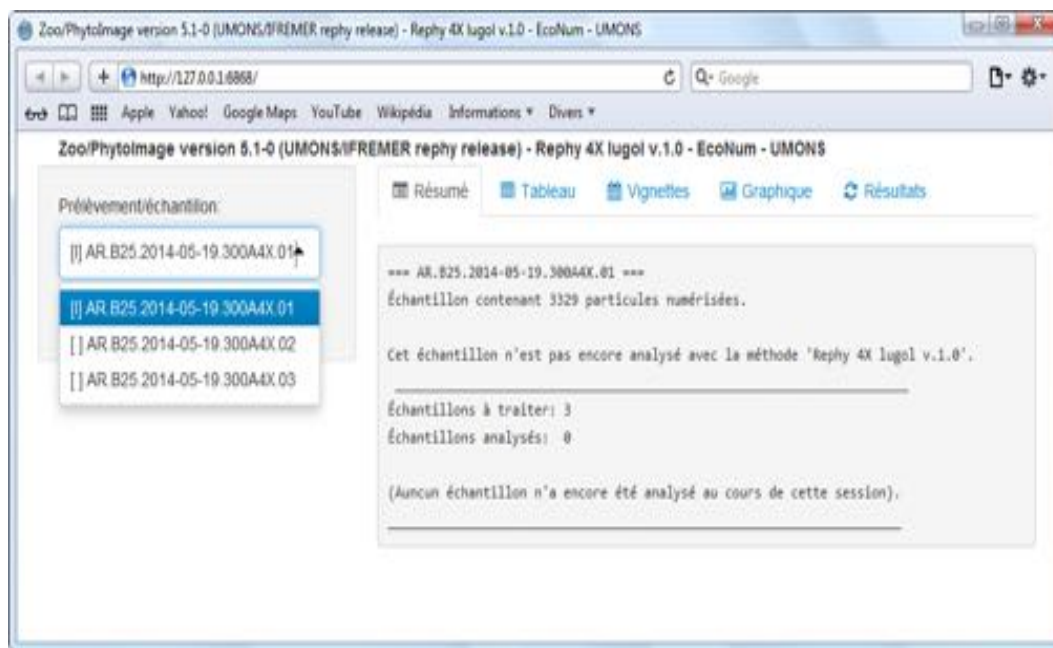
## Conversion factors for biovolume
## Biovolume calculation is P1 * ECD^P3 + P2
## TODO: fill this table, or use read.delim() on a text file
## TODO: also use number of cells per colony here...
.ZI$biovolume <- data.frame(
  Class = c("Chaetoceros_spp", "[other]"),
  P1 = c(1, 1),
  P2 = c(0, 0),
 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

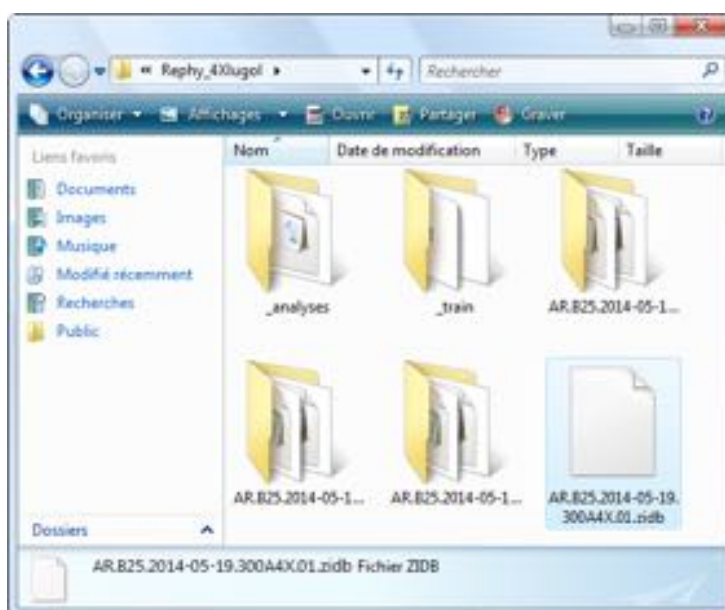
**Prélèvements/Échantillons.** Les échantillons présents dans le répertoire de travail sont listés ici. Chacun d'entre eux est précédé d'un code entre crochets.

[ ] Échantillon non importé et donc non analysé,

[ I] Échantillon importé mais non analysé,

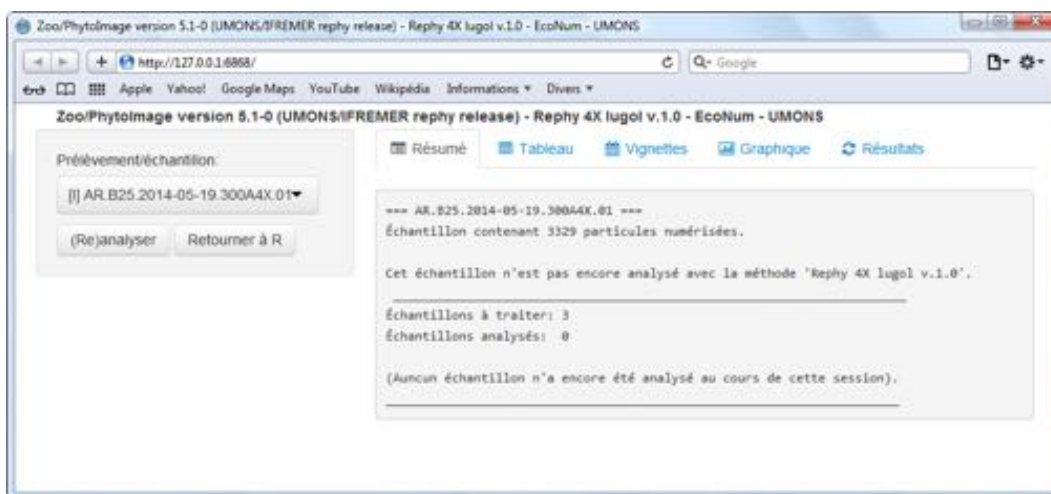
[A] Échantillon importé et analysé.

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 numérisées, état de l'analyse avec la méthode sélectionnée, nombre d'échantillons à traiter, nombre d'échantillons analysés.

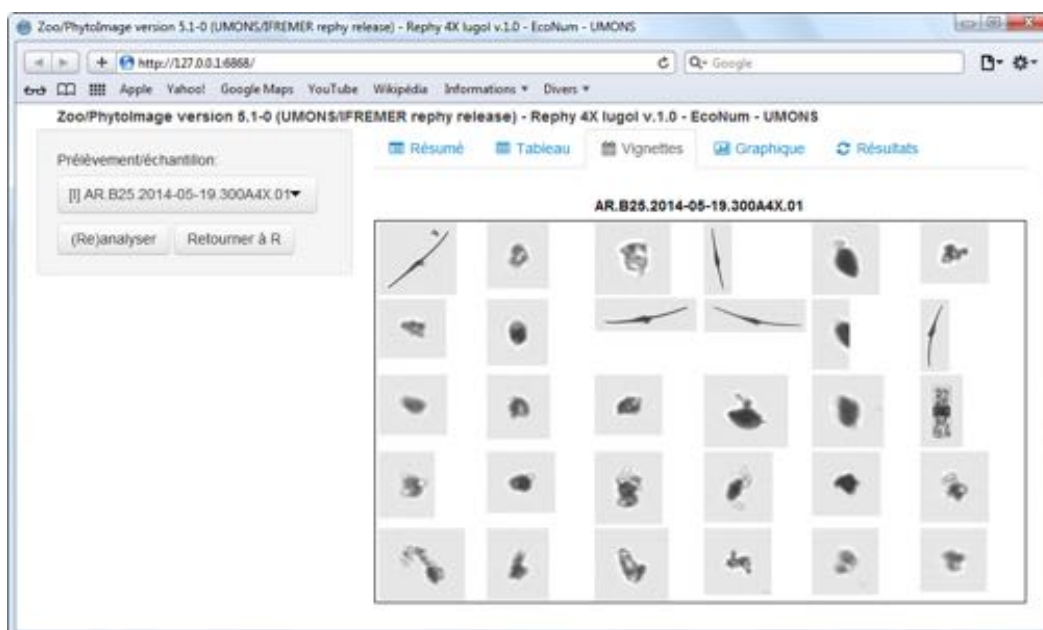


**Tableau.** Cet onglet permet une visualisation rapide et simple des mesures pour chacun des individus de l'échantillon. Il est également possible de trier par valeurs croissantes (ou décroissantes) les différentes colonnes du tableau.

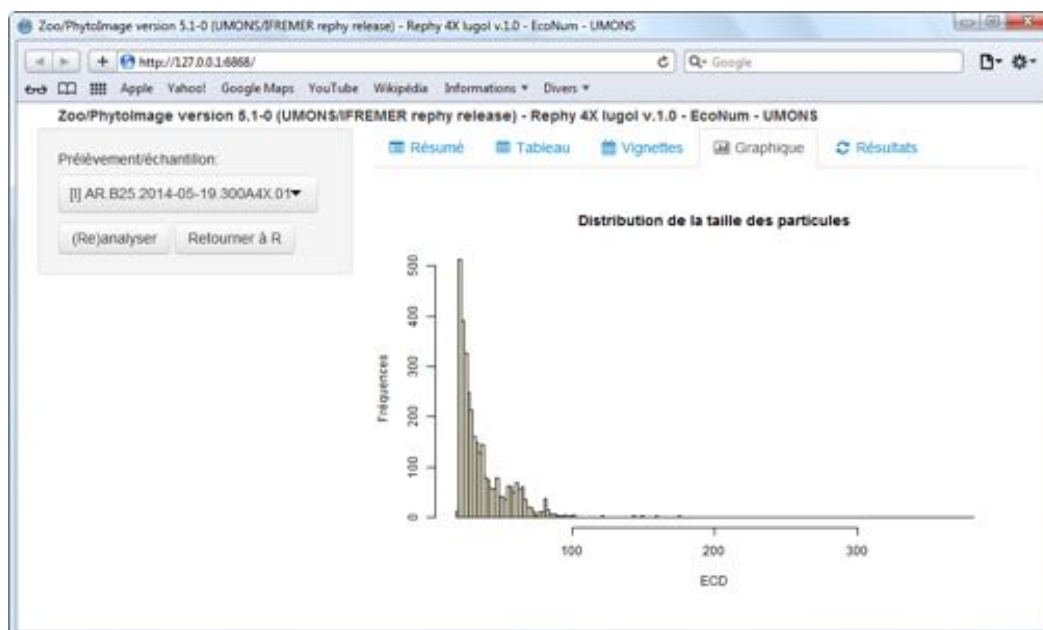
The screenshot shows the 'Tableau' tab of the Zoo/PhytoImage software. The table displays measurement data for individual samples. The columns are: Label, Item, ECD, FIT\_SaveX, FIT\_SaveY, FIT\_PixelW, FIT\_PixelH, and FIT\_C. The table contains 4 rows of data.

Label	Item	ECD	FIT_SaveX	FIT_SaveY	FIT_PixelW	FIT_PixelH	FIT_C
AR.B25.2014-05-19.300A4X.01	1	68.47136	0	0	191	217	241
AR.B25.2014-05-19.300A4X.01	2	69.16322	193	0	183	223	232
AR.B25.2014-05-19.300A4X.01	3	54.27805	378	0	211	91	160
AR.B25.2014-05-19.300A4X.01	4	33.34659	591	0	68	87	1108

**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.**

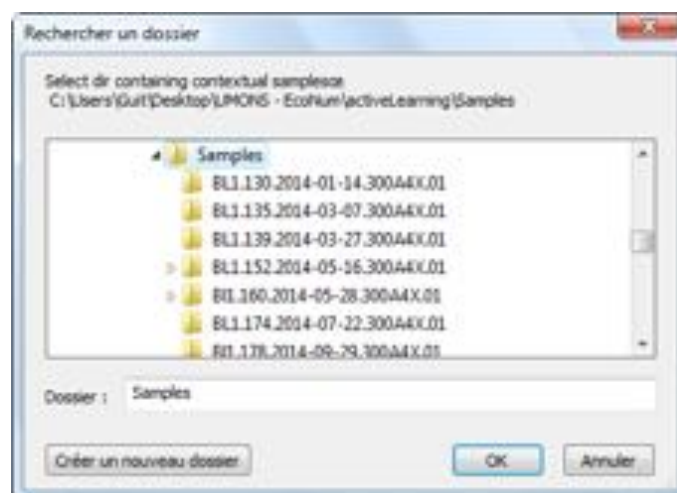
**(Re)analyser.** Lorsqu'un échantillon est importé, il est possible de l'analyser en le sélectionnant dans la liste, puis en cliquant sur le bouton « (Re)analyser ». Le processus de reconnaissance est alors exécuté sur base de la méthode R sélectionnée et de l'ensemble d'apprentissage présent dans le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

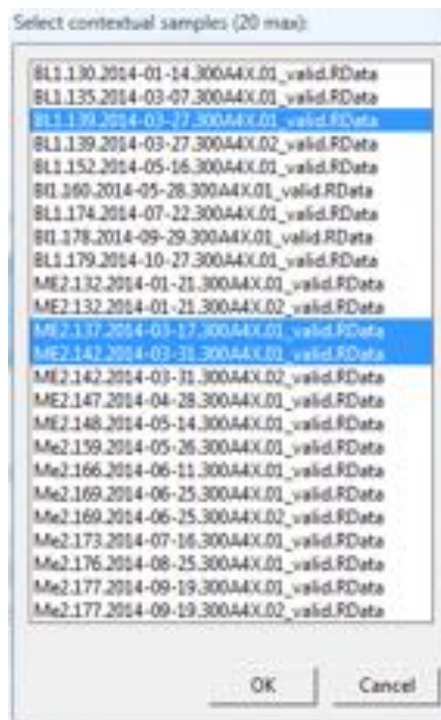
- Lors d'une première analyse avec l'interface de Zoo/PhytoImage, deux objets R sont créés dans le sous-répertoire « \_train » de votre répertoire de travail : le premier correspondant à l'ensemble d'apprentissage et le second correspondant à l'outil de reconnaissance automatique généré à partir de l'ensemble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



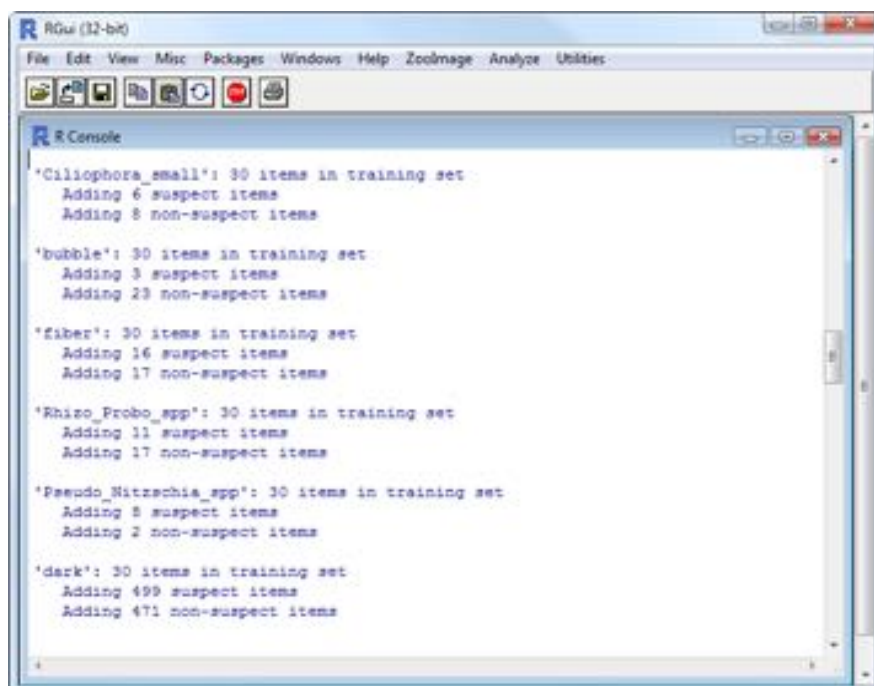
**Apprentissage actif.** Lors du traitement d'un nouvel échantillon, il est possible d'utiliser des données « contextuels » afin d'adapter l'ensemble d'apprentissage, et par conséquent, l'outil de reconnaissance associé, à l'échantillon. Selon les ensembles de données contextuels sélectionnés, l'apprentissage 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 « <echantillon>\_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 affichée. 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.

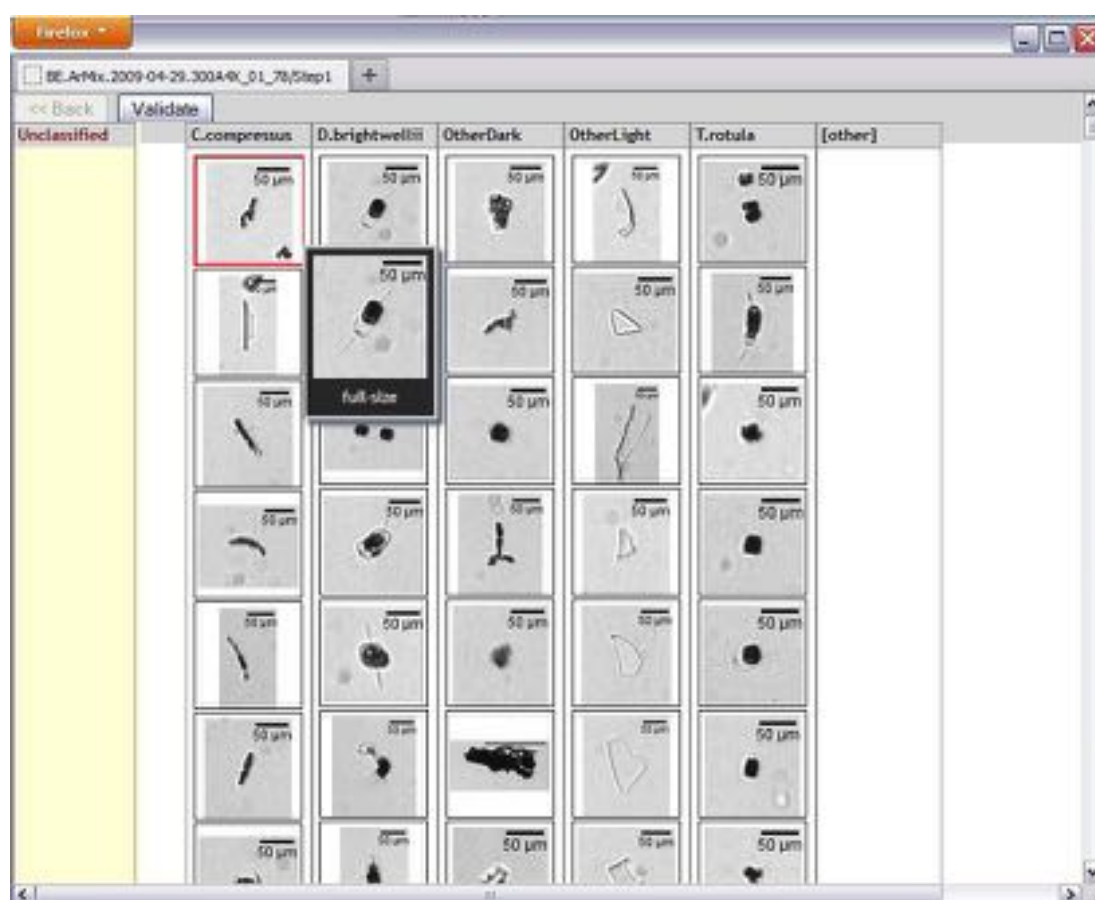




**Notes:** Les performances associées à l'apprentissage actif sont étroitement liées aux choix des échantillons contextuels.

- **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 (voir 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).

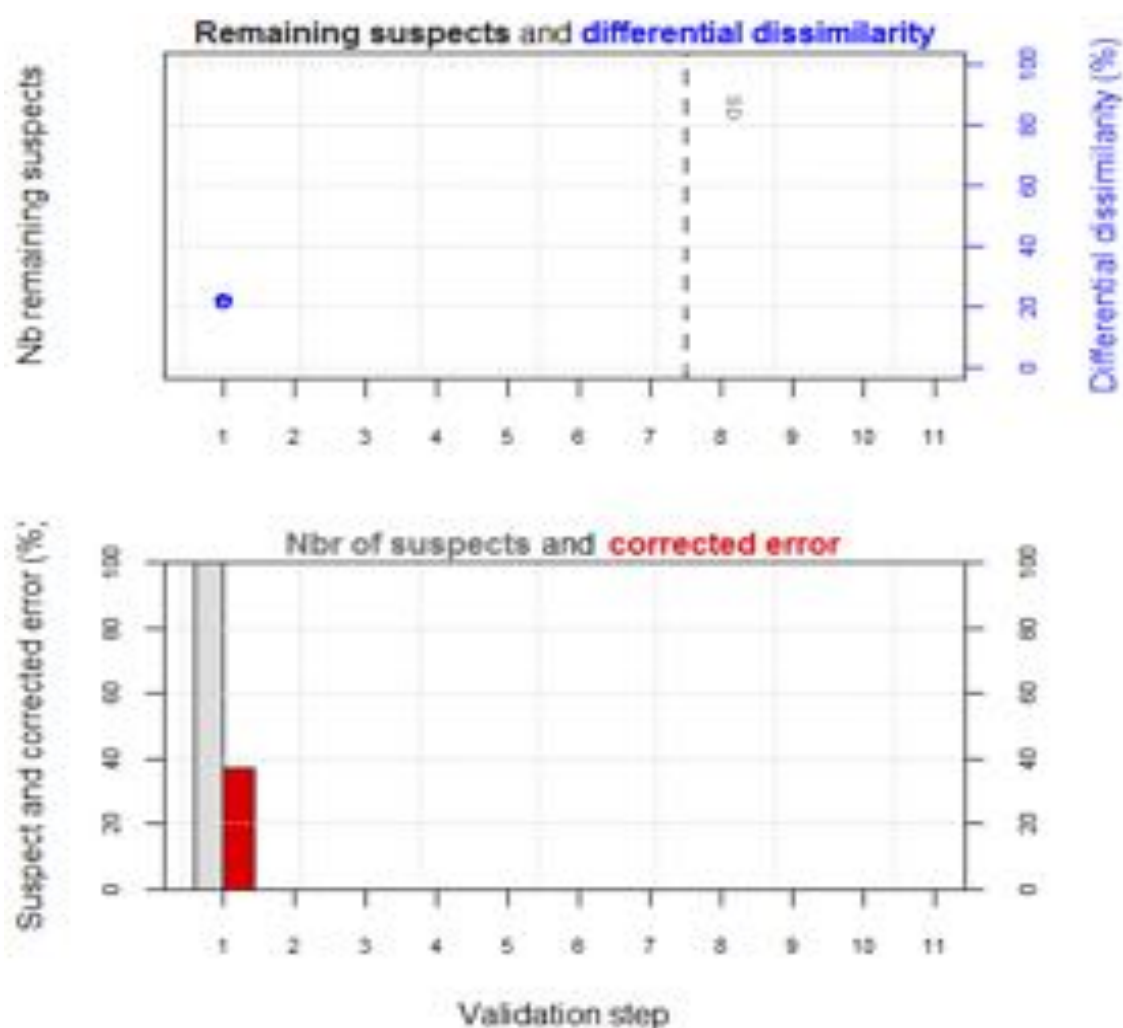


Cette page présente une première série de particules, sélectionnées aléatoirement dans l'échantillon, triées automatiquement par l'outil de reconnaissance choisi. Chaque classe est représentée par une colonne dans la page (e.g., *Ceratium\_furca*, *Ceratium\_fusus*, etc. dans l'exemple). Toutes les vignettes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

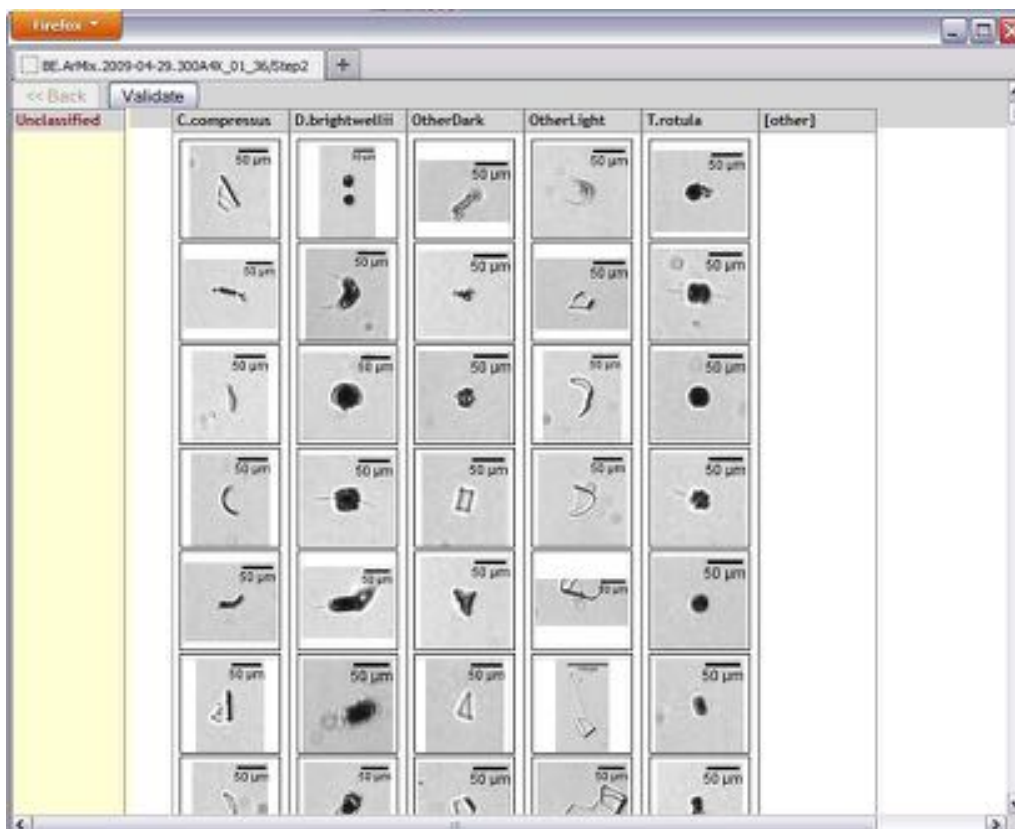
Le modèle considère également plusieurs autres aspects :

- 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

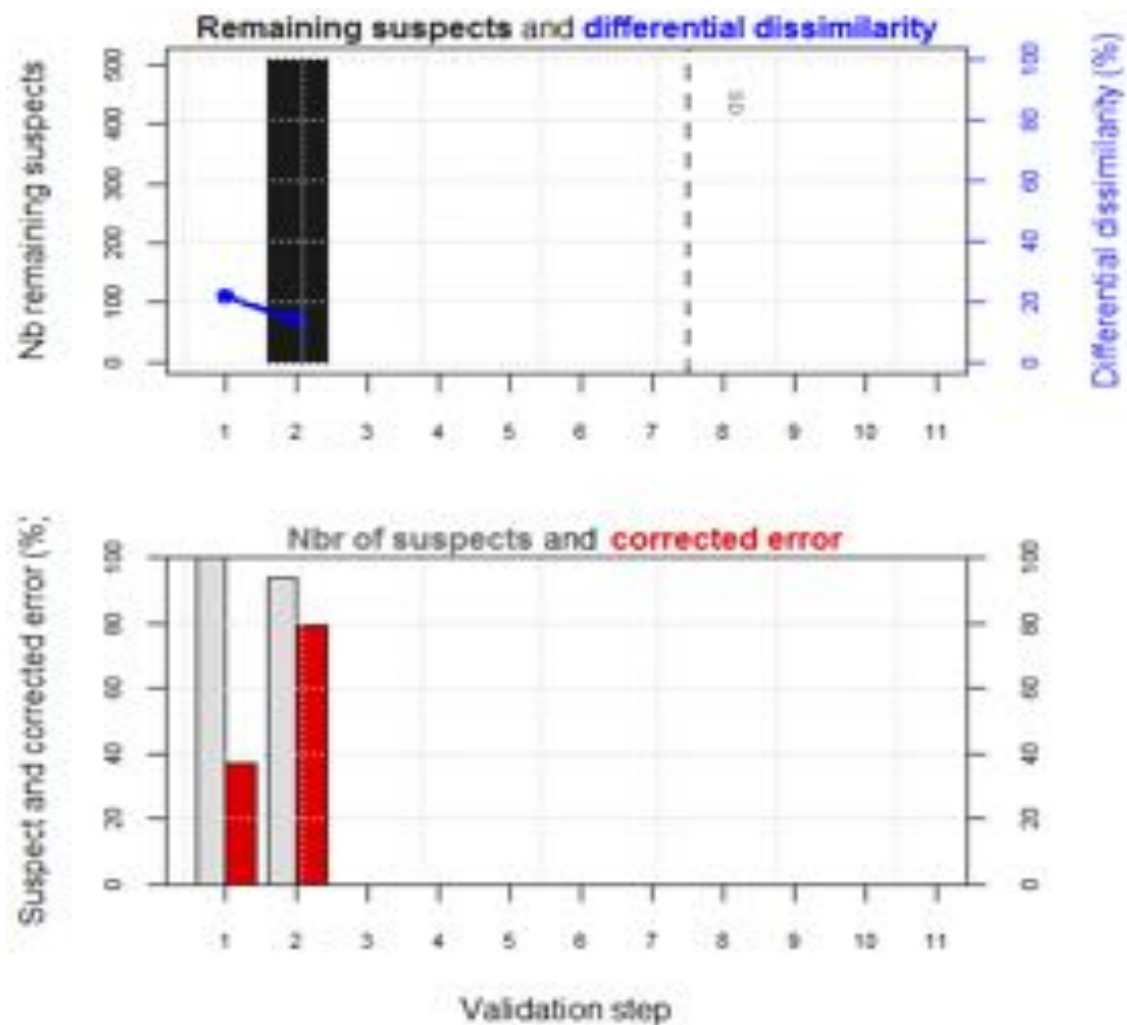
l'échantillon sous un microscope (classe A : très peu probable d'être présente, classe B : certainement présente). Indiquez juste une valeur faible (par exemple, 0.01) à la classe A et une valeur importante (par exemple, 0.99) à la classe B. Notez que les nombres que vous fournissez ne sont pas nécessairement 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 !



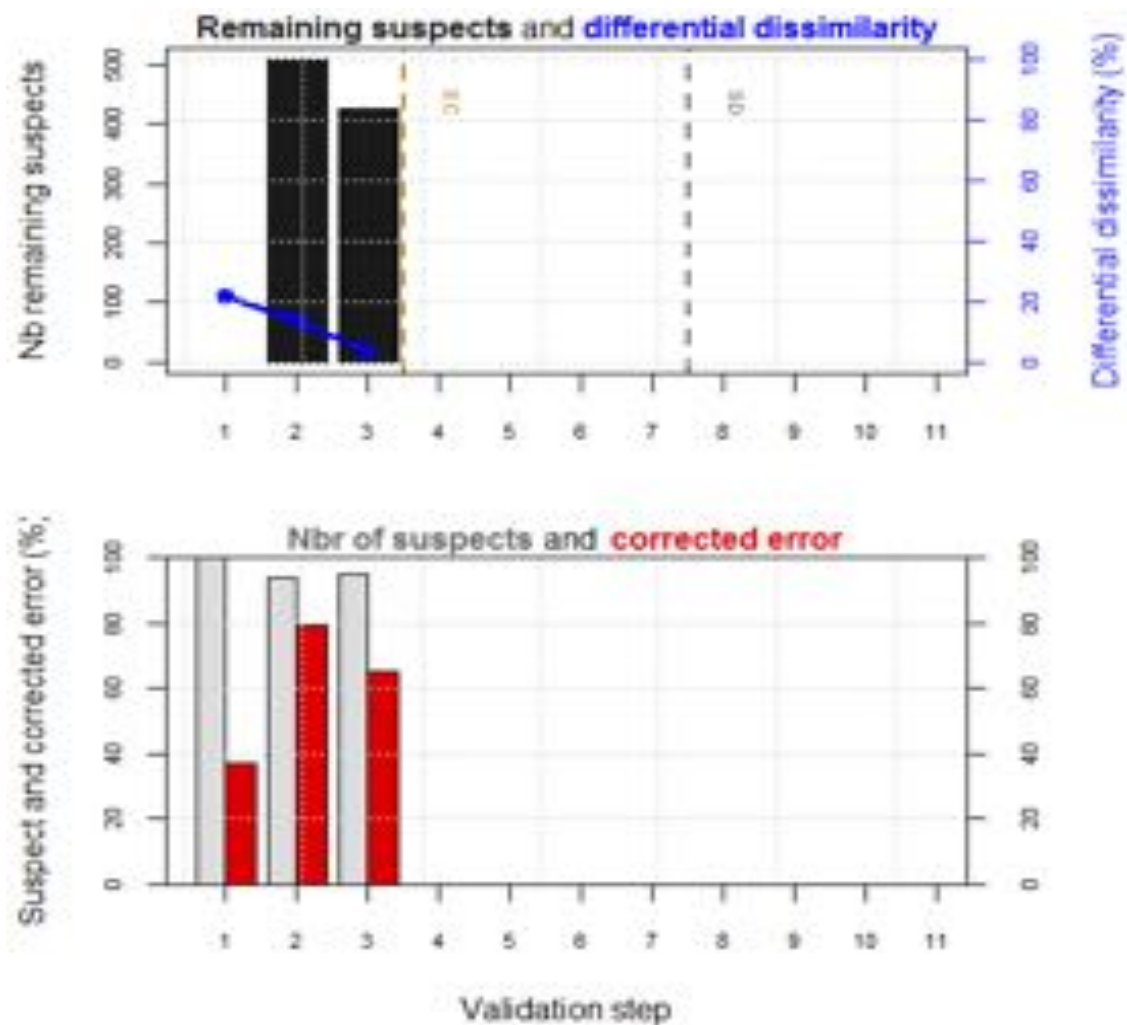
Généralement, il est assez clair que ce second ensemble présente beaucoup plus d'erreur que le précédent... et vous remarquerez également que, en effet, vous avez également beaucoup plus de particules « problématiques » (difficulté à reconnaître les particules, objets coupés, blobs avec une forme étrange, 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.



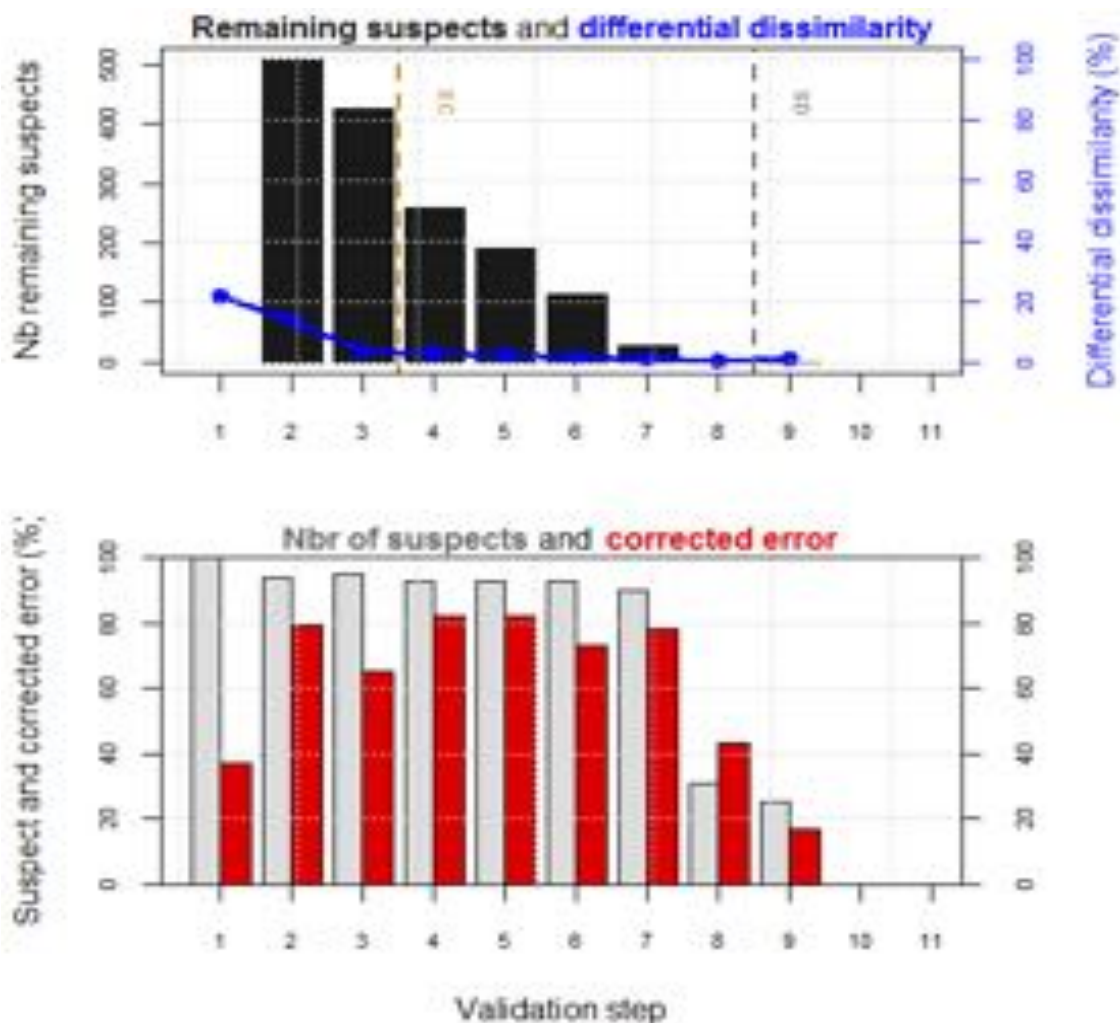
Dans le rapport, le premier graphe possède maintenant une barre grise foncée et un second point bleu. Cette barre grise permet de se faire une bonne idée du travail qu'il reste à fournir afin de valider l'ensemble des objets suspects restants. Vous remarquerez également que la dissimilarité différentielle 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 :



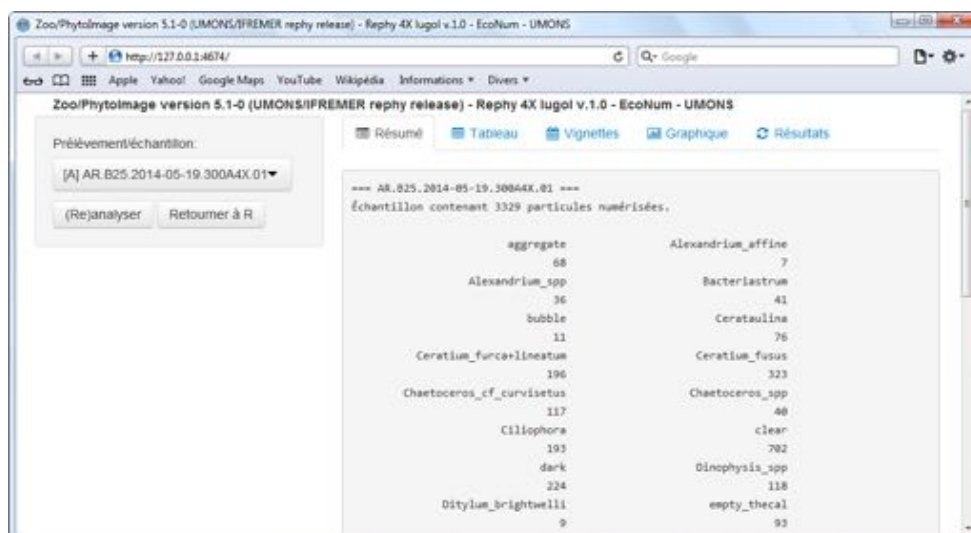
Ici, après l'étape 7, vous remarquez deux choses importantes. Premièrement, la détection des suspects correspond maintenant étroitement avec l'erreur réelle. La détection est améliorée avec la fraction de l'échantillon déjà validée qui peut être utilisée pour entraîner l'algorithme de détection. Deuxièmement,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

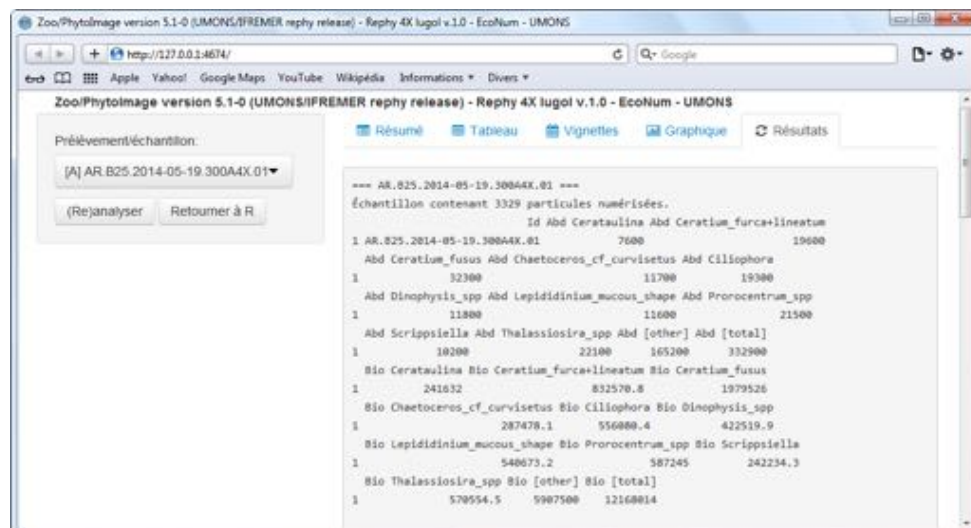
**Note :** Comme illustré sur la figure ci-dessus, il est conseillé d'effectuer une étape supplémentaire au-delà de l'indicateur SD. Celle-ci permet de confirmer la chute du nombre de particules erronées et celle de l'erreur résiduelle. De plus, dans la perspective de l'utilisation de l'échantillon en 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.



Zoo/PhytoImage version 5.1-0 (UMONS/STREMER repy release) - Rephy 4X lugol v.1.0 - EcoNum - UMONS

http://127.0.0.1:4674/

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

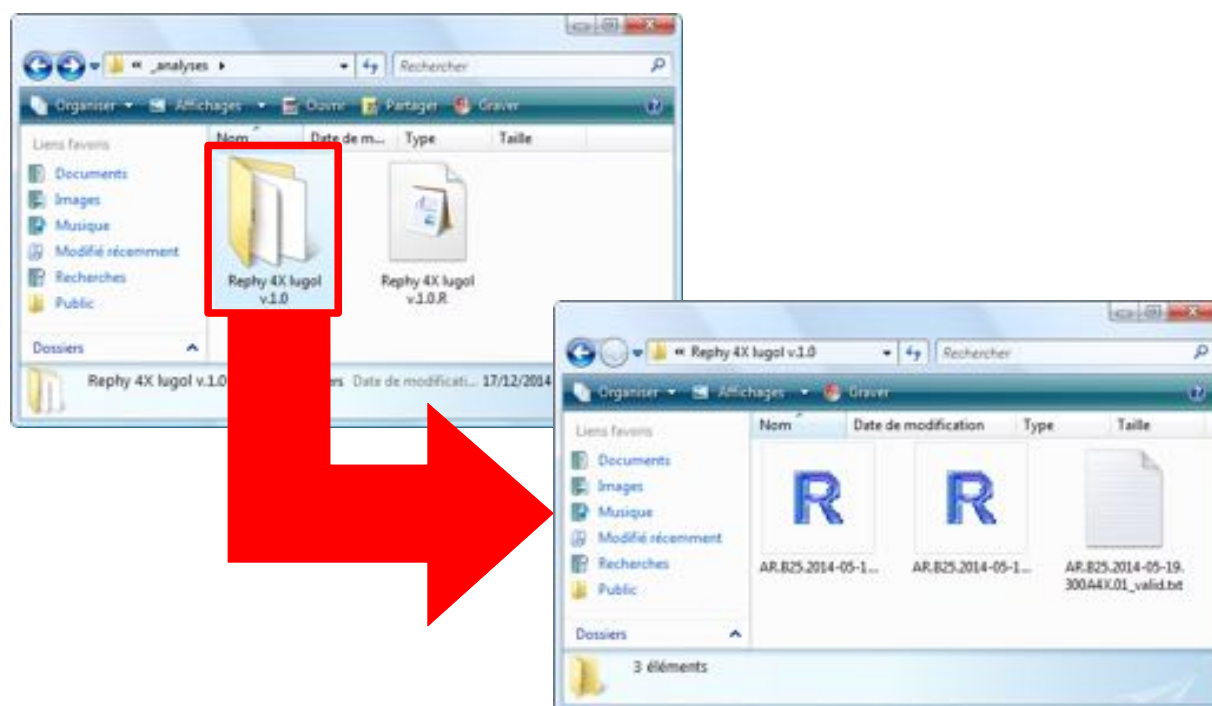
	(0,10]	(10,20]	(20,30]	(30,40]	(40,50]	(50,60]
Cerataulina	0	0	3000	4200	400	0
Ceratium_furcilineatum	0	0	0	10400	7900	900
Ceratium_fusus	0	0	0	100	900	10000
Chaetoceros_cf_curvisetus	0	100	10000	900	100	0
Ciliophora	0	0	13000	4600	300	100
Dinophysis_spp	0	0	1900	7600	1900	200
Iepididinium_mucous_shape	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_furcilineatum	300	100	0	0	0	0
Ceratium_fusus	18600	1800	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
Iepididinium_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 :

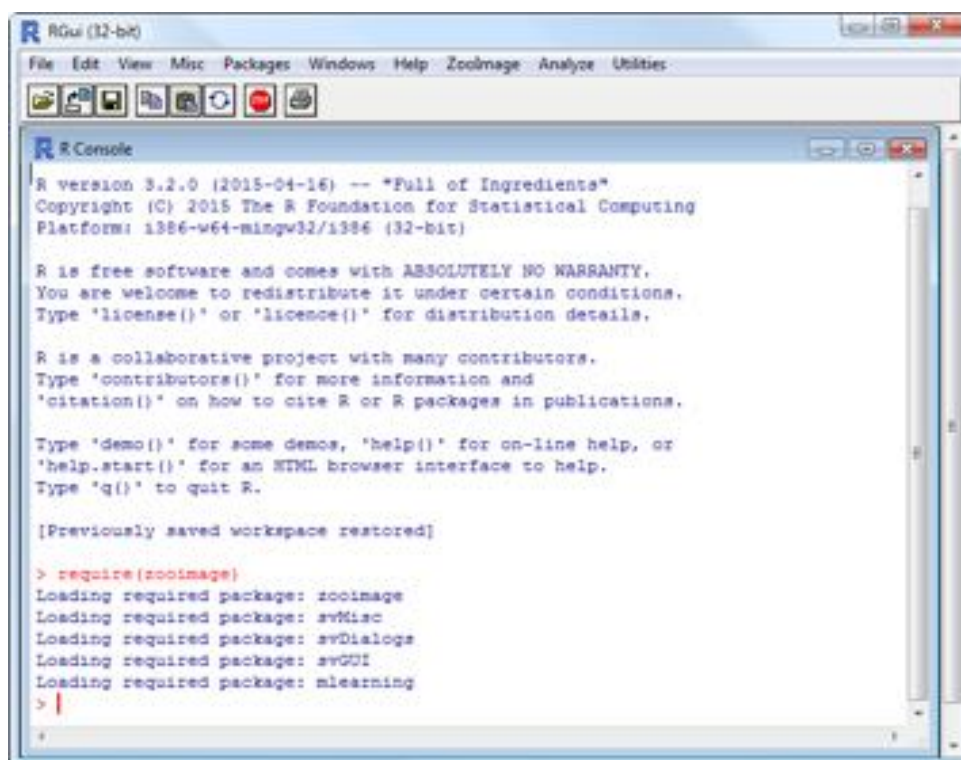
- <echantillon>\_res.RData : ce fichier peut être chargé et manipulé dans R. Il contient les abondances, les biovolumes et les spectres de taille pour chaque groupe taxonomique,
- <echantillon>\_valid.RData : ce fichier peut également être chargé et manipulé dans R. Il contient les mesures sur chacune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



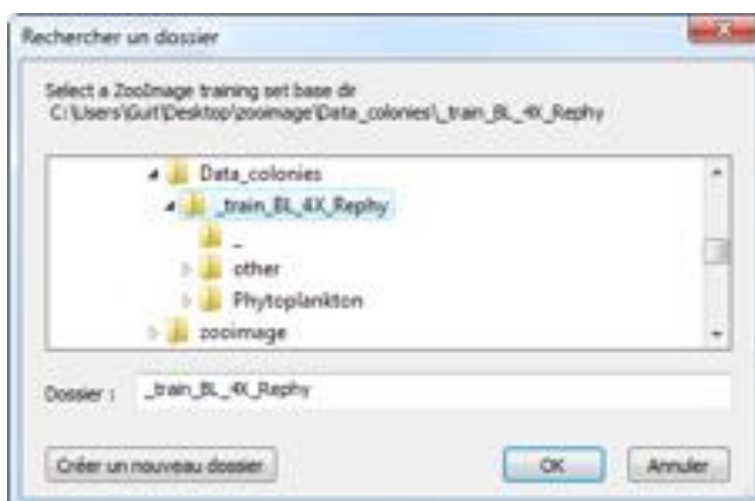
## 6. OUTIL DE DENOMBREMENT DES CELLULES EN COLONIES

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 ».



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

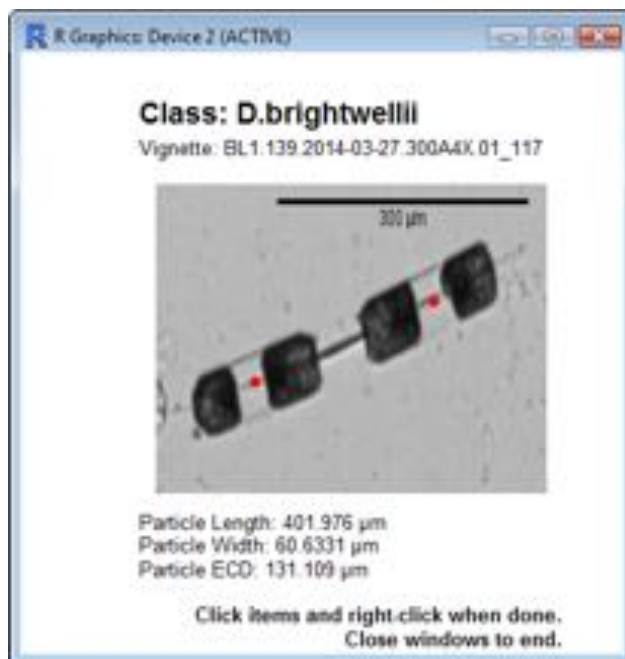
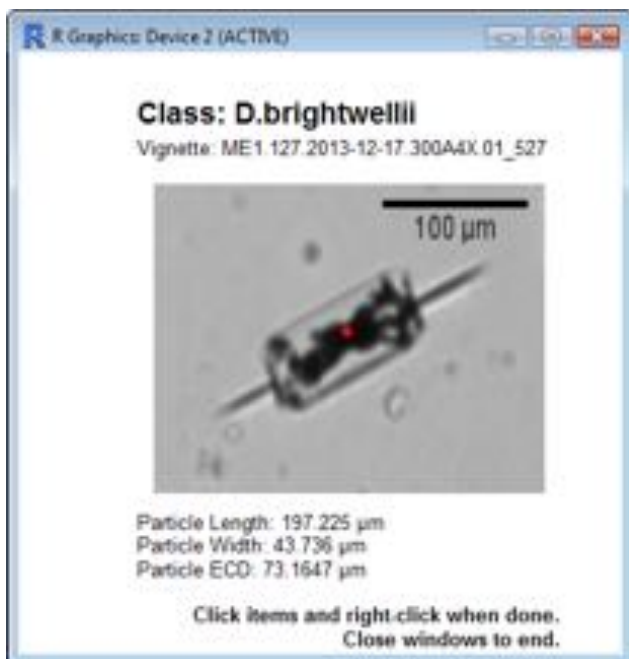




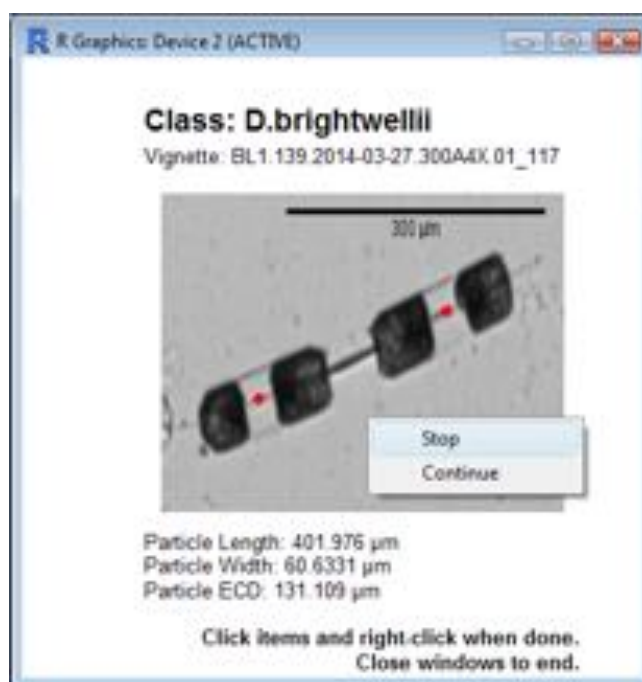
Une fois l'ensemble d'apprentissage choisi, une seconde boîte de dialogue est affichée. Celle-ci propose une liste des classes présentes dans l'ensemble d'apprentissage. Sélectionnez-en une afin de commencer les dénombrements.



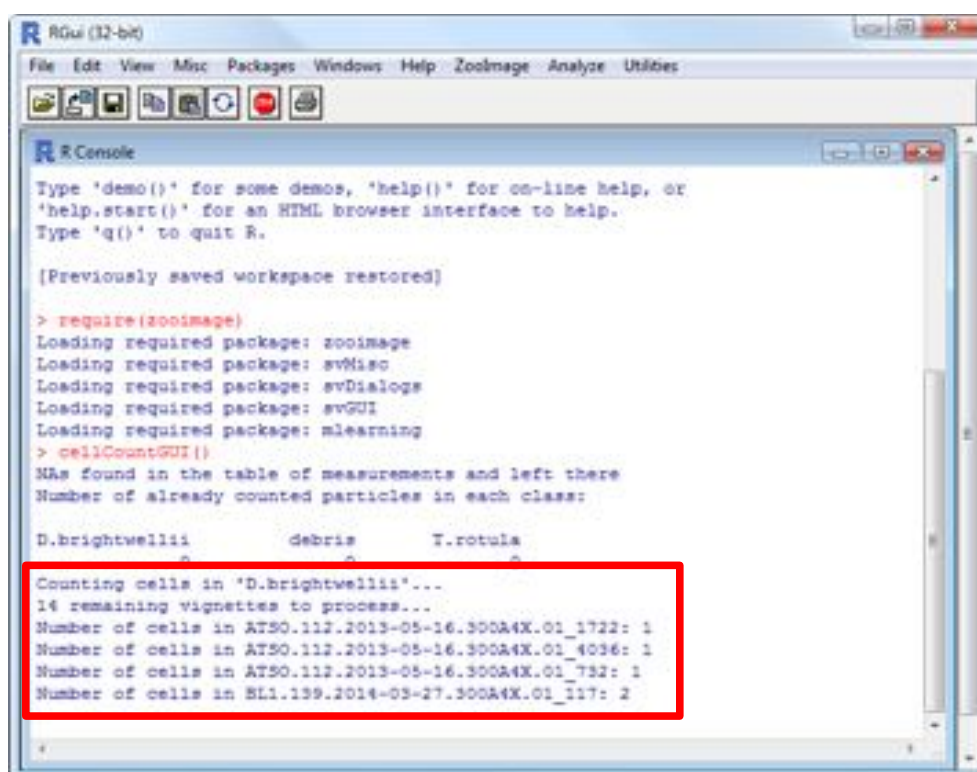
Lorsque vous cliquez sur le bouton « OK », les vignettes correspondant à la classe sélectionnée dans l'ensemble d'apprentissage vous sont présentées les unes après les autres. Sur ces figures, la classe d'appartenance, l'identifiant de la particule, ainsi que sa longueur, sa largeur et son ECD (Equivalent 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) :



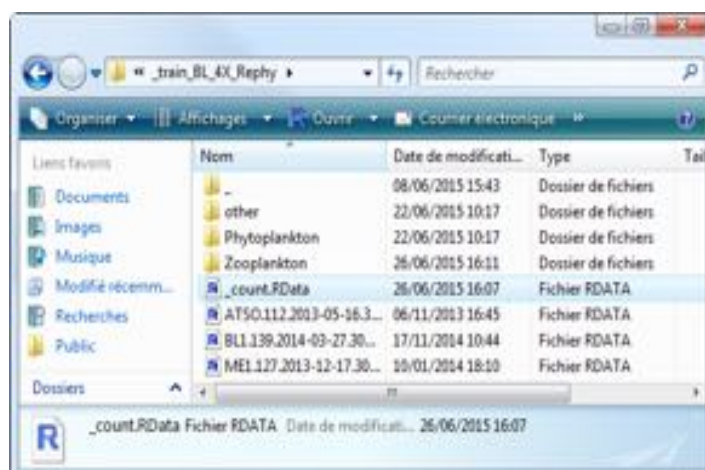
Lorsque le dénombrement des cellules est terminé, et afin de passer à la vignette suivante, il vous suffit de **cliquer droit avec la souris, et de sélectionner « STOP »**.



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 :

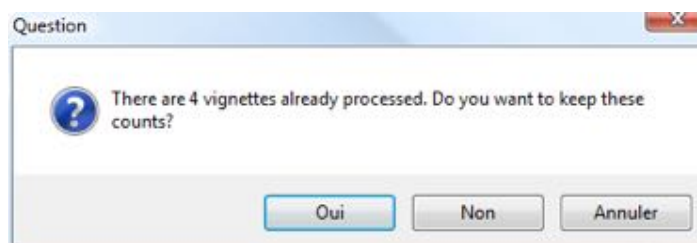
- si **le nombre de cellules est dénombrable**, l'utilisateur clique sur chacune des cellules identifiées, puis passe à la vignette suivante en cliquant droit et en sélectionnant « STOP ». **La valeur est alors enregistrée dans l'ensemble d'apprentissage.**
- si **le nombre de cellules n'est pas dénombrable** (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 enregistrée.**
- 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%20mining%20with%20R/). 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.