

Guide d'utilisation de la plateforme PAMPA WP2

Aide au calcul, à la représentation et l'analyse des métriques relatives à la biodiversité et aux ressources

Yves REECHT (Yves.Reecht@ifremer.fr), Romain DAVID, Jérémie HABASQUE et Bastien PREUSS

Résumé : Ce document vise à aider les gestionnaires et scientifiques à utiliser la plateforme WP2 afin de calculer des métriques biodiversité et ressources à partir de leur jeu de données. Il les guidera également dans l'utilisation de l'interface pour générer des graphiques et conduire des analyses statistiques.

Ce guide présente la version de la plate-forme à la fin du projet PAMPA. Des évolutions sont d'ores et déjà envisagées, mais ne sont pas évoquées dans ce document. Le retour des utilisateurs sera crucial pour l'évolution de la plate-forme.

Informations importantes : ce document contient des avertissements importants (ayant la même apparence que ce cadre) – notamment en ce qui concerne l'installation de la plateforme – que vous êtes invités à lire attentivement avant installation et/ou utilisation.

Ce document doit être cité comme suit :

Reecht Y., David R., Habasque J. et Preuss B. Guide d'utilisation de la plateforme PAMPA WP2 – Aide au calcul, à la représentation et l'analyse des métriques relatives à la biodiversité et aux ressources. PAMPA/WP2/Meth/4. Version du 24/08/2011. 53 p.

Table des matières

1. Présentation.....	1
A. Obtenir de l'aide.....	2
B. Mise à jour.....	2
C. Rapport de bug.....	3
2. Données analysées.....	4
A. Types de données d'entrée.....	4
B. Format des fichiers de données.....	4
Quelques rappels importants :	4
3. Installation de l'environnement de travail et configuration.....	5
A. Prérequis : installation de R.....	5
B. Installation de la plateforme sous Windows.....	5
C. Création du dossier de travail.....	6
D. Configuration.....	7
i. Dossier de travail et fichiers de données.....	7
ii. Langue des noms de variables.....	8
iii. Catégories benthiques supplémentaires.....	8
4. Chargement de l'interface.....	9
A. Démarrage de l'application.....	9
B. Packages.....	9
5. Importation des données.....	11
A. Dossiers et fichiers de données.....	11
B. Importation des fichiers par défaut.....	11
C. Informations de chargement.....	12
D. Opérations « spéciales » lors du chargement et calculs divers.....	13
i. Estimation des tailles d'après les classes de tailles.....	13
ii. Année de campagne.....	13
E. Sélection et re-calcul.....	14
F. Informations sur les données.....	15
i. Test du référentiel espèces.....	15
ii. Plan d'échantillonnage.....	16
iii. Informations par « espèce ».....	16
iv. Informations par unité d'observation.....	17
6. Sous-interfaces standard de sélection des métriques/facteurs.....	18
Note :	19
7. Graphiques.....	20
A. Métriques agrégées par unité d'observation par espèce (et éventuellement par classe de taille).....	20
i. Boîtes à moustaches ou Boxplots.....	21

ii. Diagrammes en barres ou Barplots.....	22
B. Métriques agrégées par unité d'observation (et éventuellement par classe de taille)	23
i. Boîtes à moustaches ou Boxplots.....	23
ii. Diagrammes en barres ou barplots.....	24
iii. Remarques.....	25
Classes de tailles :	25
C. Remarques générales sur les graphiques.....	27
i. Rang d'utilisation du facteur « statut de protection » :	27
ii. Options graphiques.....	27
iii. Options graphiques supplémentaires (cachées).....	29
D. Cartes (démonstration sur données Nouvelle-Calédonie).....	31
8. Analyses statistiques.....	33
A. Modèles linéaires.....	33
i. Métriques agrégées par unité d'observation par espèce (et éventuellement par classe de taille).....	33
ii. Métriques agrégées par unité d'observation (et éventuellement par classe de taille)	36
iii. Résultats.....	37
Informations sur le modèle.....	38
détails sur les facteurs significatifs et leurs coefficients	38
Valeurs prédites.....	39
Comparaisons multiples (2 facteurs).....	39
Comparaisons multiples (1 facteur).....	41
iv. Graphiques diagnostiques et valeurs aberrantes.....	41
GLMs et graphiques diagnostiques.....	43
(Log-)LMs et graphiques diagnostiques.....	44
B. Arbres de régression multivariée.....	45
i. Métriques agrégées par unité d'observation par espèce (et éventuellement par classe de taille).....	46
ii. Métriques agrégées par unité d'observation (et éventuellement par classe de taille)	46
iii. Résultats Graphiques.....	47
iv. Résultat « texte ».....	48
Rappel du modèle :	48
Résumé de l'arbre sous forme de texte :	48
Détails :	48

1. Présentation

Un programme de calcul d'indicateurs de ressources et biodiversité a été développé sous le logiciel R dans le cadre du projet PAMPA. Ce document a pour but de guider l'utilisateur (gestionnaire et scientifique) dans l'utilisation de ce programme. Le programme fonctionne sous Windows et Linux.

La définition et le détail du calcul des métriques figure dans le document : « Métriques biodiversité et ressources PAMPA/WP2/Meth/1 ».

L'équipe coordinatrice du projet fournit aux utilisateurs :

- le programme R de calcul des métriques.
- l'aide aux utilisateurs pour le formatage des données.
- le référentiel espèces.

L'interface est constituée d'un menu « déchirable » (comprendre « qui s'ouvre »), contenant plusieurs rubriques. Les rubriques correspondent à différentes étapes d'analyse de vos jeux de données. Chacune d'entre elle contient des sous menus correspondant à des types d'analyses et de graphiques.

Calcul d'indicateurs PAMPA WP2

Données Sélection et recalcul Graphiques Statistiques Outils Aide Quitter...

Ci dessous l'aide contextuelle

ETAPE 2 : Vous pouvez restreindre votre sélection de données (menu "Sélection et recalcul") ou commencer les traitements standards.

Suivi des opérations

C:/PAMPA existe

Patientez, chargement des données en cours ...

INFO : Les métriques par unités d'observations ont été calculées sur l'ensemble du jeu de données importé

INFO : Les fichiers .csv:

- Contingence
- PlanEchantillonnage
- Métriques par unité d'observation (UnitobsMetriques.csv)
- Métriques par unité d'observation pour les espèces présentes (ListeEspecesUnitobs.csv)
- Métriques par unité d'observation / espèce (UnitobsEspeceMetriques.csv)
- Métriques par unité d'observation / espèce / classe de taille (UnitobsEspeceClasseTailleMetriques.csv)

ont été créés

Espace de travail : C:/PAMPA

Aire Marine Protégée : CB ; type d'observation : TRUVC

{Type de fichier}	Source	{Nb enregistrements}	{Nb champs}
{Fichier d'unités d'observations}	CB_UnitobsUVC_Couronne.txt	353	35
{Fichier d'observations}	CB_ObsUVC_Couronne_cor.txt	6592	10
{Référentiel espèce}	{PAMPA-WP1-Meth-5-RefSpMed 191010.txt}	857	127

Critères de sélection

Tout

-> Nombre d'espèces dans le fichier d'observation : 59

-> Nombre d'unités d'observations dans le fichier d'observations : 353

Restaurer les données Fermer les graphiques

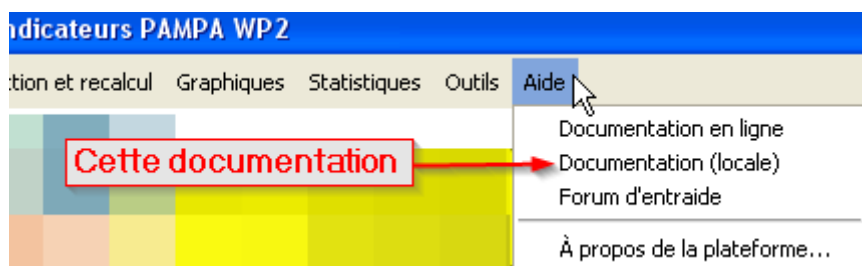
Le reste de l'interface (en dessous des menus) est constituée de :

1. deux champs de texte, dont une aide contextuelle en jaune et un champ d'information en blanc intitulé « suivi des opérations » qui vous permet de suivre le déroulement de vos principales analyses.

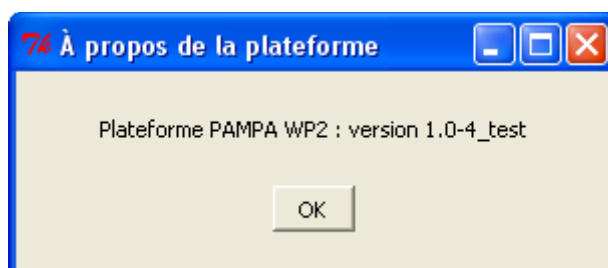
2. trois zones d'information supplémentaires vous permettent en un coup d'œil de visualiser l'état du jeu de données que vous traitez :
 - une zone texte indiquant l'emplacement de votre zone de travail, l'AMP considérée et le type d'observation en cours de traitement.
 - un tableau présentant les fichiers observation, unités d'observation et référentiel que vous avez saisis, avec leur nom, leurs nombres de champs et d'enregistrements. Plus, le cas échéant, une quatrième colonne indiquant le nombre d'éléments (observations) de la sélection.
 - une zone « critère de sélection » qui affiche le critère que vous avez sélectionné le cas échéant et le nombre d'espèces et unités d'observation restantes dans votre fichier d'observations.
3. boutons pour des opérations courantes sur la dernière ligne : restaurer les données originales (après avoir procédé à des sélections) et fermer l'ensemble des graphiques ouverts.

A. Obtenir de l'aide

En plus de la présente documentation, il est possible d'accéder à un forum d'entraide ainsi qu'à une documentation en ligne (sur le wiki du projet ; pas à jour actuellement, mais chacun peu participer à son élaboration) par le menu « Aide » :



« À propos de la plateforme... » permet de consulter le numéro de la version utilisée :

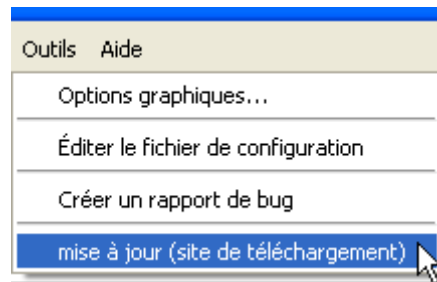


Ce numéro est à préciser pour obtenir de l'aide sur le forum ou lors de l'émission d'un rapport de bug (voir page suivante).

Il permet également de savoir si une mise à jour peut-être faite.

B. Mise à jour

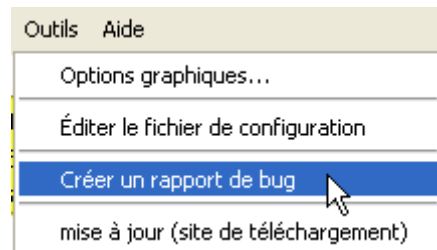
Les nouvelles versions de l'interface sont publiées sur la page de téléchargement du wiki, accessible par le menu « Outils » :



La liste des modifications entre deux versions peut également être consultée sur le wiki : <http://projet-pampa.fr/wiki/doku.php/wp2:changelog>.

C. Rapport de bug

Afin d'assurer une maintenance efficace du programme, il est fortement conseillé de faire remonter à l'équipe de développement tout bug constaté. Un modèle de rapport est accessible par le menu « Outils » :



Il doit être envoyé par e-mail à developpeur-wp2@projet-pampa.fr.

2. Données analysées

A. Types de données d'entrée

Plusieurs types de données – qui nécessitent des traitements particuliers – sont supportés par la plateforme :

- **UVC** (*underwater visual censuses* – comptages visuels sous-marins) : utilisé pour la faune et les habitats sous-marins.
- **LIT** (*Line Intercept Transect*) : un type d'UVC utilisé pour le benthos et l'habitat.
- **SVR** : Stations Vidéos Rotatives, avec séparation des rotations en secteurs. Nécessite un traitement particulier, et notamment des interpolations des secteurs non-valides (pas analysables).
- Données de pêche (capture et effort), avec plusieurs sous-catégories :
 - **DEB** : débarquements.
 - **EMB** : échantillonnage par observateur(s) embarqué(s).
 - **PSCI** : pêches scientifiques.
 - **PecRec** : enquêtes de pêche récréative.

Un seul type de donnée peut-être analysé à la fois, en raison, notamment, des différences dans les traitements et – dans certains cas – dans les unités des métriques calculées.

B. Format des fichiers de données

Quelque soit le type de données traitées, l'information doit être répartie en 3 fichiers texte (.txt) distincts :

1. table contenant les unités d'observation.
2. table contenant les observations (comptages, captures, etc.)
3. le référentiel espèces (Méditerranée ou Outre-Mer) fourni par l'équipe coordinatrice.

Le format d'entrée des données pour exécuter les programmes de calcul est standardisé (Cf. Formats de données – PAMPA/WP2/Meth/3). Des problèmes peuvent se poser lorsque le format des données n'est pas respecté.

Quelques rappels importants :

- Le séparateur décimal des fichiers texte est le point.
- Le séparateur de champ des fichiers texte est la tabulation.
- Les fichiers doivent être préalablement inspectés pour ne contenir ni espaces (à remplacer par des *underscores*, « _ »), ni « ; ».

3. Installation de l'environnement de travail et configuration

A. Prérequis : installation de R

Le logiciel R peut être récupéré en vous rendant sur la page du projet <http://cran.cict.fr/> (vous pouvez choisir un [autre miroir¹](#) plus proche de chez vous). Vous y trouverez le programme et/ou les instructions d'installation pour votre système.

Il est pour l'instant conseillé d'installer la [version 2.11.1](#) qui se trouve être la plus récente, permettant le lancement de l'interface à l'aide des raccourcis (voir section suivante).

B. Installation de la plateforme sous Windows

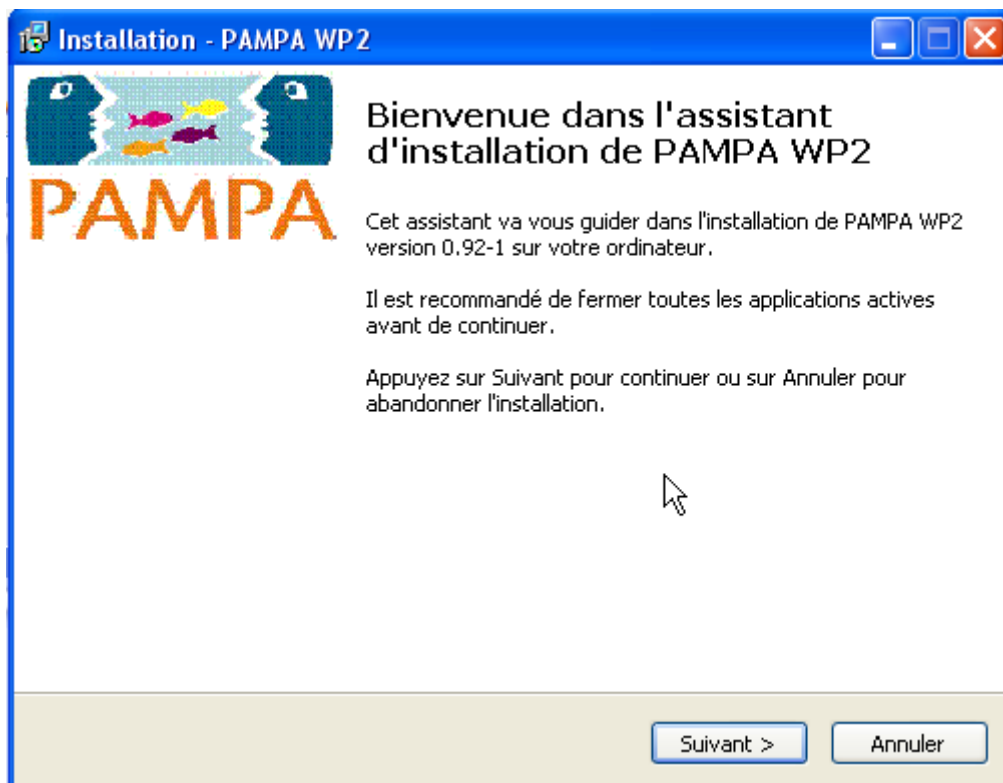
La plateforme est dotée d'un installeur pour Windows, qui place les scripts dans les dossiers adéquats et crée des raccourcis pour lancer R et l'interface graphique.

Attention : cet installeur étant pour l'instant très basique, il écrase les fichiers existants sans demander confirmation. Vous devez donc soit copier à un autre endroit, soit renommer les fichiers d'une ancienne installation que vous avez modifiés par vous même, si vous souhaitez ne pas perdre les informations qu'ils contiennent.

Si le nom de l'installeur contient « **update** » le fichier `config.r` (qui doit contenir les noms de vos fichiers de données) sera cependant conservé. Si ce n'est pas le cas, il est recommandé de le sauvegarder. Vous pourrez ensuite copier votre configuration dans le fichier nouvellement installé (*ne pas le remplacer, il vous manquerait des lignes de code nécessaires au bon fonctionnement de la plateforme*).

En exécutant le fichier « <Setup|Update>_PAMPA_WP2_<version>.exe » vous serez dirigé vers une procédure d'installation classique (sous Windows) :

¹ Copie à l'identique du site R (documentation, logiciel, paquets optionnels,...).



Les interfaces de chaque étape sont explicites et sont dotées de choix par défaut qui doivent suffire pour la plupart des cas.

Attention : ne modifiez pas le chemin d'installation par défaut (« C : / PAMPA / Exec »), la plateforme ne serait pas fonctionnelle... à moins que vous ne soyez prêt à faire vous même les modifications qui s'imposeraient dans le code.

Il est prévu d'assouplir, dans un avenir proche, la gestion des dossiers, mais cette fonctionnalité n'est, à l'heure actuelle, pas implémentée.

Une fois l'installation terminée, vous disposez de raccourcis pour lancer R et l'interface graphique, ainsi que d'accès directs aux documentations et un modèle de rapport de *bug* :



C. Création du dossier de travail

La plateforme utilise un dossier de travail dans lequel elle va chercher les répertoires de données et de résultats. Par défaut, il s'agit de `C : / PAMPA /`, mais vous pouvez en choisir un autre (et même en avoir plusieurs pour l'analyse de différents jeux de données). Vous devez créer, dans ce répertoire, un dossier « Data » qui contiendra vos jeux de données.

Il est ensuite nécessaire de renseigner ce dossier de travail dans le fichier de configuration « `C : / PAMPA / Exec / config.r` » (voir la section suivante, « 3.D. Configuration »).

D. Configuration

i. Dossier de travail et fichiers de données

Le dossier d'installation contient un fichier dédié à la configuration des fichiers de données et du répertoire de travail.

[C:/PAMPA/Exec/config.r](#)

En paramétrant correctement ce fichier, vous pourrez recharger vos données en un clic à chaque lancement de l'interface !!

Il est conseillé afin de recharger les données en un seul clic, d'indiquer au programme le nom de vos fichiers d'unité d'observation, d'observation et de référentiel espèces.

Dans ce fichier de configuration, changer les valeurs correspondante comme indiqué ci-dessous, en y mettant vos noms de fichier (en respectant les minuscules et majuscules !).

Le dossier de travail peut également être renseigné à ce niveau-là. Les lignes éditables ressemblent donc à ce qui suit :

```
#### CB
SiteEtudie <- "CB" # Identification du site.
fileName1 <- "CB_UnitobsUVC_Couronne.txt" # Unités d'observation.
fileName2 <- "CB_ObsUVC_Couronne.txt" # Fichier d'observations.
fileName3 <- "refEspècesMED.txt" # Référentiel espèce.
nameWorkspace <- "C:/PAMPA/WD" # Répertoire de travail.
```

En utilisant le système de commentaire de R (voir, par exemple, les descriptions de type fichier dans l'exemple ci-dessus), vous pouvez conserver la trace de plusieurs configurations et y revenir lorsque vous le souhaitez. Toute ligne commençant par « # » ne sera pas traitée (considérée comme du commentaire) :

```
## SiteEtudie <- "CB"
## fileName1 <- "unitobs_CB_peche_pro.txt"
## fileName2 <- "obs_biodiv_CB_peche_pro.txt"
## fileName3 <- "refEspècesMED.txt"
## nameWorkspace <- "C:/PAMPA/Pêche_pro"
```

Après chaque modification, n'oubliez pas de commenter (faire précéder de « # ») les lignes de configuration devenues inutiles.

Ces fichiers peuvent alors être (i) directement chargés depuis le menu « Données » en cliquant sur l'entrée « Dossiers et fichiers par défaut », (ii) modifiés avec l'entrée « Choix des dossiers et fichiers de données ».

Ce fichier faisant parfois l'objet de modifications profondes, vous ne devriez pas l'écraser avec une ancienne version après une nouvelle installation. Depuis la version 1.1-0, l'installation fait toutefois une sauvegarde de votre ancien « config.r » pour pouvoir par la suite copier votre configuration (cf. encadrés ci-dessus) dans le nouveau fichier installé.

ii. Langue des noms de variables

Par défaut les noms usuels de variables sont en français. Afin de permettre la production de graphiques destinés aux publications scientifiques, il a été ajouté la possibilité d'avoir les noms de variables en anglais sur ceux-ci. Pour cela, il faut se rendre à la fin du fichier de configuration, à

```
## Noms d'usage des variables des principales tables de données
## (référentiels compris) :
assign("varNames",
      read.csv(paste(basePath, "/Exec/NomsVariables_fr.csv", sep=""),
              header=TRUE, row.names=1, stringsAsFactors=FALSE),
      envir=.GlobalEnv)
```

et remplacer `"/Exec/NomsVariables_fr.csv"`

par `"/Exec/NomsVariables_en.csv"` puis recharger les données.

Ceci n'affecte que les noms des variables, le reste des titres restant en français. Dans l'attente de l'implémentation d'un véritable système d'internationalisation, les titres doivent être supprimés en utilisant `options(P.graphPaper=TRUE)`, voir section 7.C.iii. Options graphiques supplémentaires (cachées), page 29.

iii. Catégories benthiques supplémentaires

Il est maintenant possible d'ajouter facilement des catégories benthiques supplémentaires, correspondant à des regroupements des catégories existantes. Celles-ci sont définies dans un fichier « `corresp-cat-benth.csv` » (dans `C:/PAMPA/Exec`) qui peut être édité avec un tableur de type Excel. Ce tableau donne les correspondances entre le champ « `Cat_benthique` » et les nouvelles catégories, et se présente comme suit :

	A	B	C	
1	Cat_benthique	CategB_general	CategB_groupe	
2	AA	PLANT	AA	
3	ACB	CV	ACR	
4	ACD	CV	ACR	
5	ACE	CV	ACR	
6	ACS	CV	ACR	
7	ACT	CV	ACR	
8	CA	PLANT	NACR	
9	CB	CV	NACR	
10	CBL	CV	CBL	
11	CE	CV	NACR	
12	CF	CV	NACR	
13	CHL	CV	NACR	

La première colonne doit contenir obligatoirement les valeurs du champ « `Cat_benthique` » du référentiel espèce. Les suivantes – non limitées en nombre – représentent les catégories agrégées. Des nouveaux facteurs seront alors disponibles dans l'interface, sous les noms affichés en première ligne (éviter les espaces et caractères spéciaux), pour produire des graphiques, faire des analyses statistiques ou effectuer des sélections sur le jeu de données.

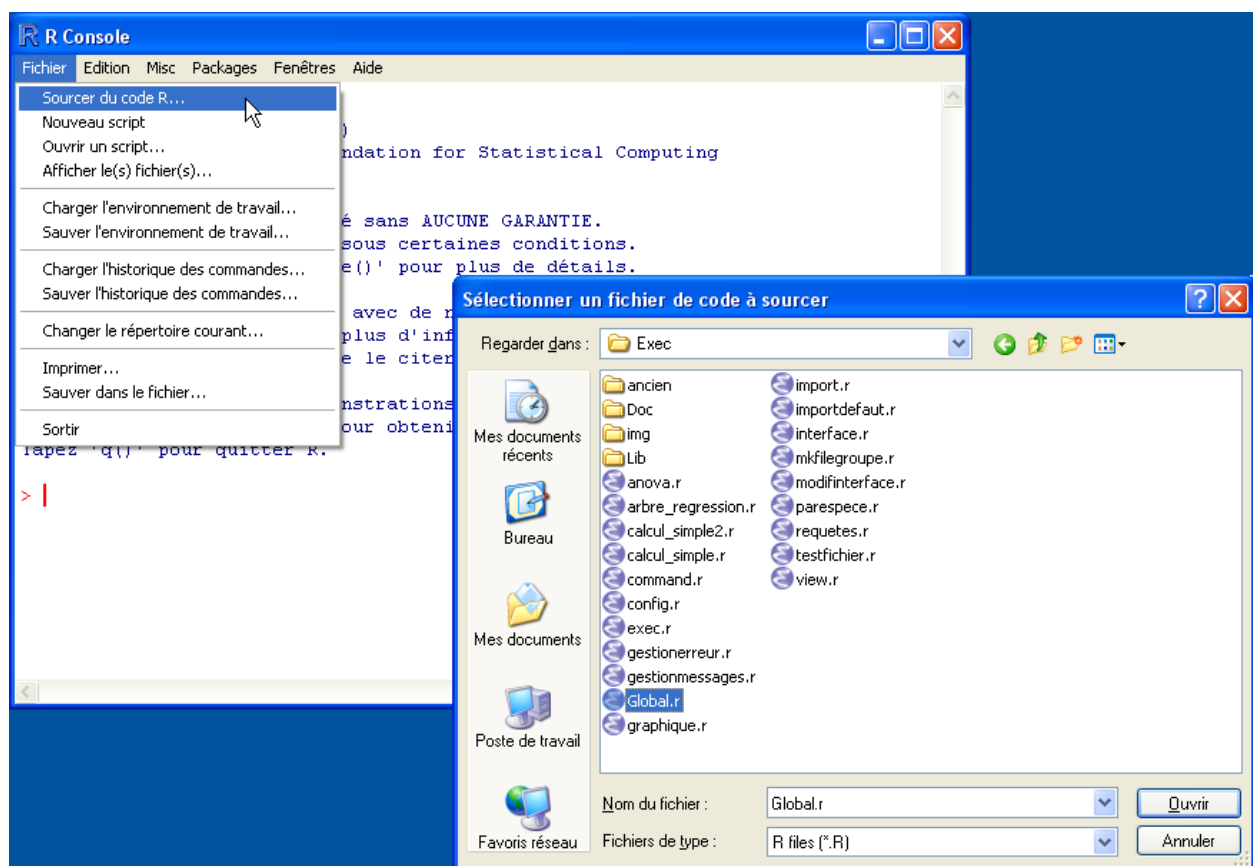
4. Chargement de l'interface

A. Démarrage de l'application

Comme expliqué plus haut, vous disposez sous Windows de raccourcis pour le chargement de l'interface.

Si toutefois ils n'étaient pas fonctionnels sur votre machine, vous pouvez charger cette interface manuellement après avoir lancé R :

- Soit à l'aide du menu de la console R :
 1. Menu « Fichier » cliquez sur « Sourcer du code R... ».
 2. ...en sélectionnant le fichier « `C:/PAMPA/Exec/Global.r` ».



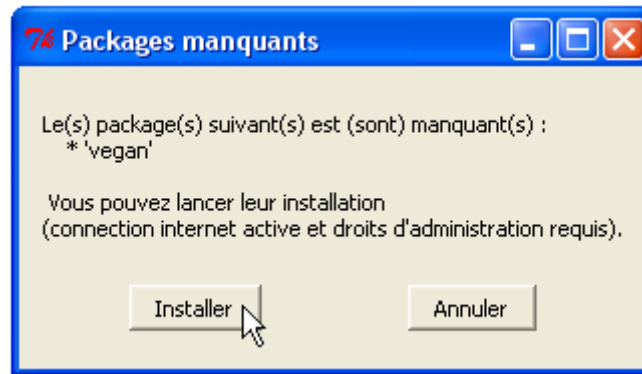
- soit en collant la ligne suivante dans la console R :

```
source("C:/PAMPA/Exec/Global.r")
```

B. Packages

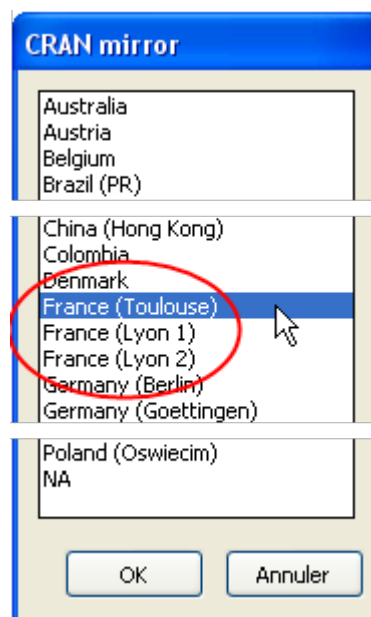
R présente un système basé sur l'utilisation de paquet ou *packages* optionnels dédiés à des tâches précises. Cette plateforme nécessite l'usage d'un certain nombre d'entre eux.

Si des *packages* manquent à l'appel, la plateforme propose leur installation :



Pour l'installation de *packages*, vous devez avoir les droits d'administration et être connecté à internet.

Si vous choisissez l'installation, R va probablement vous demander de choisir un dépôt² pour le téléchargement des *packages* :



Choisissez un dépôt le plus près de chez vous possible.

Si vous annulez ou si une erreur se produit durant l'installation, la liste des *packages* à installer manuellement est affichée sur la console R :

```
> source("C:\\PAMPA\\Exec\\Global.r")
Erreur dans switch(res, ok = invisible(sapply(requiredPack, library, character.only = TRUE))), :
  Vous devez installer manuellement le(s) package(s) :

  * 'vegan'
> |
```

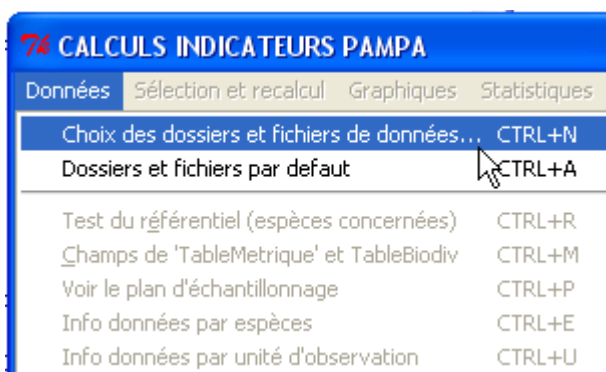
Notez que dans ce cas, plutôt que de procéder à une installation manuelle, vous pouvez également relancer la plateforme dans des conditions qui permettent l'installation (*i.e.* avec les droits d'administration et une connexion internet).

² espace sur internet qui centralise le stockage des *packages*.

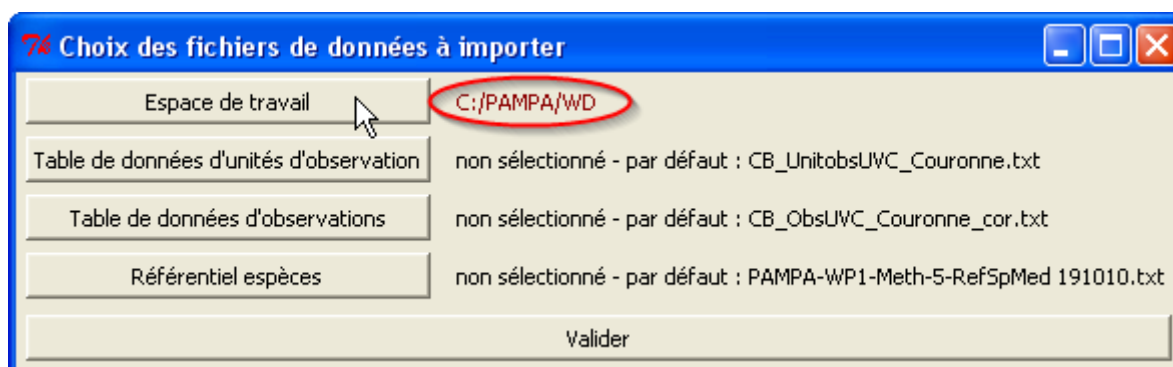
5. Importation des données

A. Dossiers et fichiers de données

L'entrée « Choix des dossiers et fichiers de données » permet de choisir le dossier de travail et les fichiers de données :



Elle tient compte du répertoire de travail qui peut être configuré dans le fichier « [config.r](#) » (variable « nameWorkspace »). Cliquer sur l'un des boutons vous permettra de choisir un nouvel espace de travail (premier bouton) ou fichier de données/référentiel (les trois suivants) :



Le dossier de travail est celui qui contient le répertoire « Data » (contenant lui-même les fichiers de données). Vous ne devez donc pas sélectionner le répertoire « Data » mais son répertoire parent.

Les modifications faites à l'aide de cette boîte de dialogue ne seront cependant conservées que dans la session courante. De même, un chargement des « Dossiers et fichiers par défaut » (voir section suivante) les écrasera.

Il est donc conseillé de renseigner les noms de fichiers de données dans le fichier de configuration, comme cela est expliqué dans la section 3.D. Configuration.

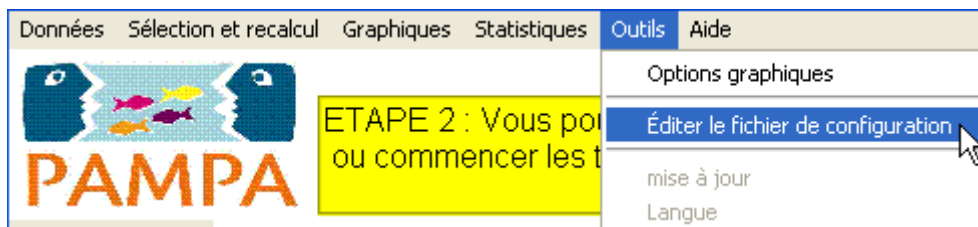
B. Importation des fichiers par défaut

Ainsi, lorsque les fichiers par défaut sont correctement configurés, ils peuvent être directement chargés sans passer par l'étape de choix des fichiers un à un. Il suffit d'utiliser la seconde entrée du menu « Données » ou bien le raccourci « Ctrl+A » :



Lors d'une modification du fichier de configuration, il n'est pas nécessaire de relancer l'interface pour une prise en compte de la nouvelle configuration. À chaque fois que ce mode de chargement est utilisé, le fichier de configuration est à nouveau lu sur le disque dur.

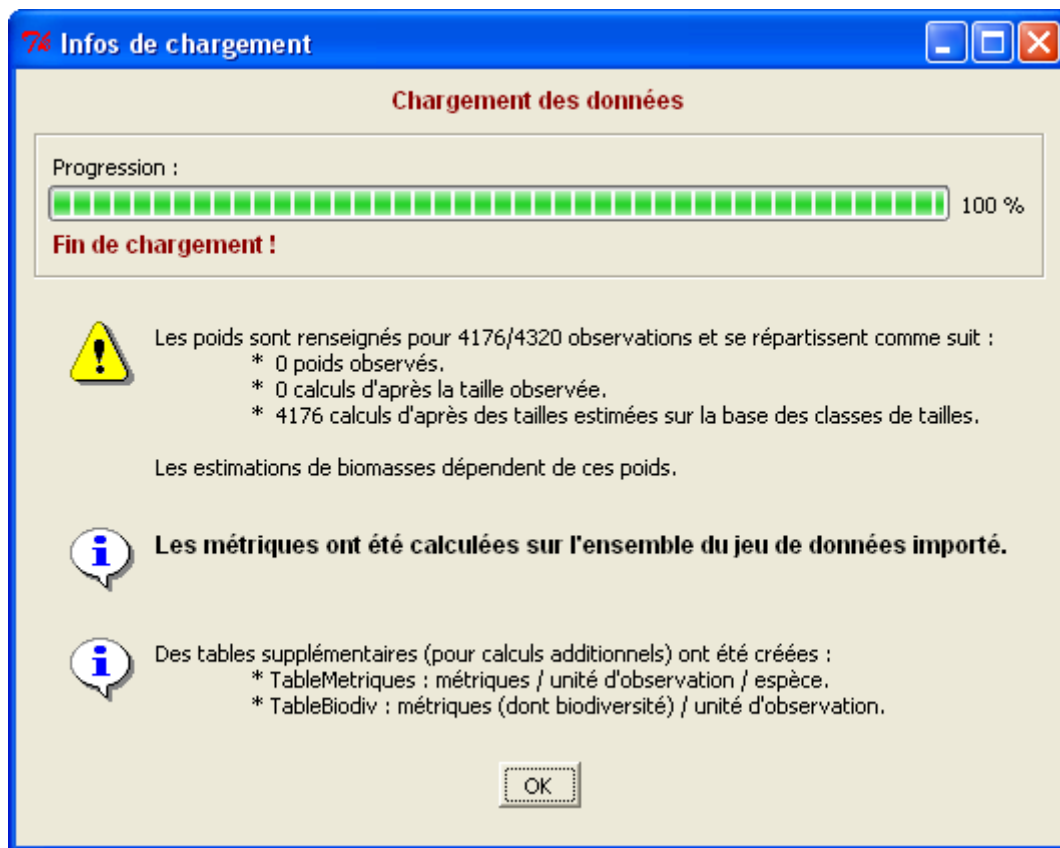
Pour modifier les dossier et fichiers par défaut, vous pouvez directement ouvrir le fichier « `C:/PAMPA/Exec/config.r` » depuis le menu « outils » de l'interface principale :



Une fois la nouvelle configuration enregistrée, il vous suffira de recharger les « Dossier et fichiers par défaut ».

C. Informations de chargement

Au cours du chargement des données, une fenêtre d'information apparaît, qui résume les instructions exécutées et en cours, et attire votre attention, au fur et à mesure, sur les éléments importants :



Lorsque le bouton « OK » apparaît, c'est que le chargement est terminé.

La barre de progression ne constitue qu'une information grossière et ne donne pas une estimation précise par rapport au temps total de chargement.

D. Opérations « spéciales » lors du chargement et calculs divers

i. Estimation des tailles d'après les classes de tailles

Les tailles précises ne peuvent pas être estimées avec toutes les méthodes d'observations *in situ*. En l'absence de relevé des tailles, celles-ci sont maintenant estimées d'après les classes de taille (si celles-ci sont renseignées et s'y prêtent, bien entendu). Les formats de classe pris en compte sont pour l'instant de la forme "5-10", "5_10", "40-", "40_", "_5", etc. Les classes ouvertes vers le bas sont considérées comme "0-<valeur>" et pour celles ouvertes vers le haut, la taille est supposée être la borne inférieure (e.g. "40-" -> 40, faute de pouvoir faire mieux). Pour tout le reste, la taille est estimée comme la moyenne des deux bornes.

ii. Année de campagne

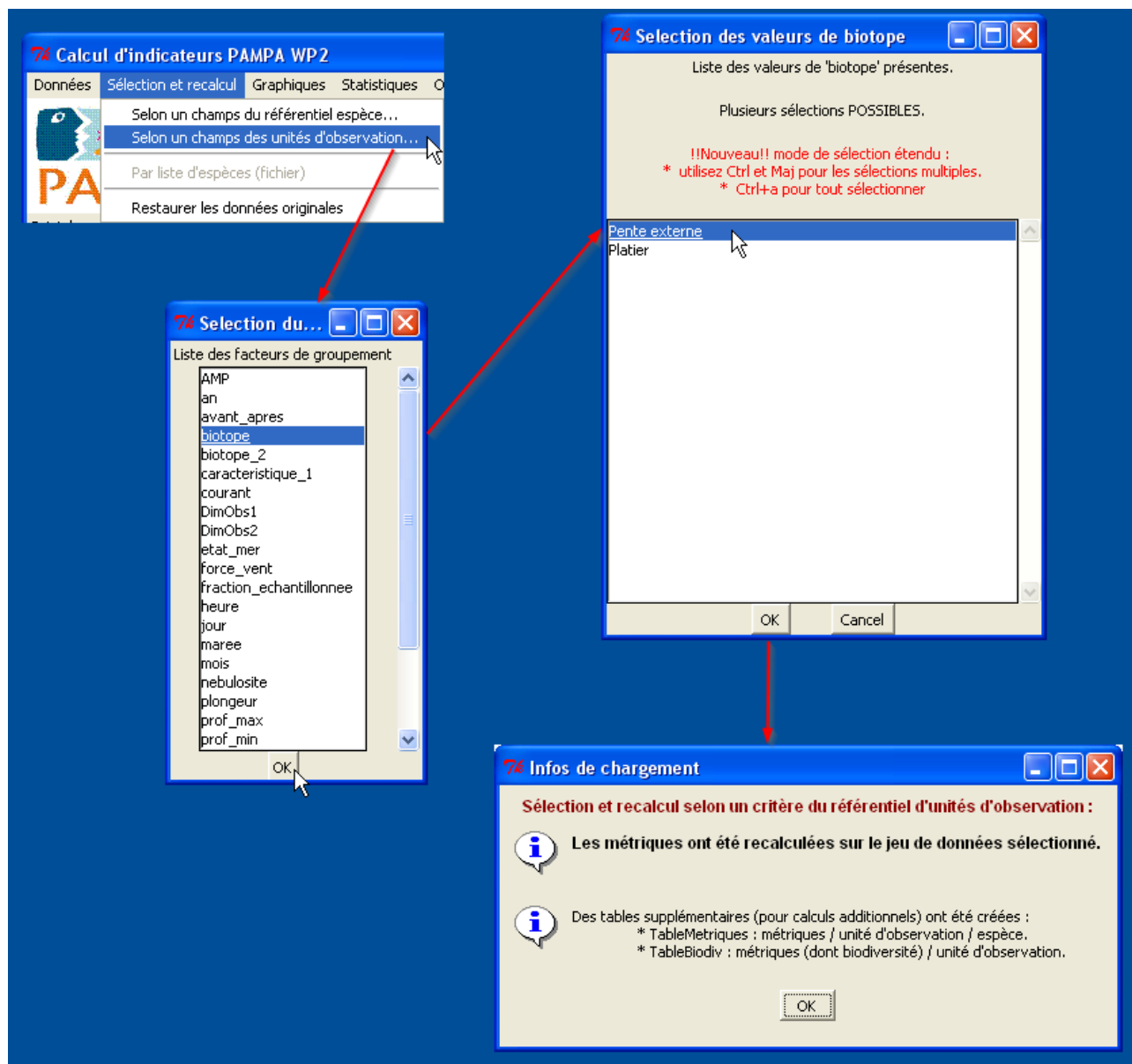
Certaines campagnes d'acquisition de données peuvent être à cheval sur deux années civiles, notamment dans l'hémisphère austral. Il peut donc dans ce cas être préférable de substituer une « année de campagne » à l'année civile comme facteur explicatif de la variabilité des métriques. Il a donc été convenu de placer dans ce cas l'année de campagne dans la colonne « *caractéristique_2* » du référentiel des unités d'observations, sous la forme « C<année> » ou bien « c<année> » (e.g. "C2004", "c1999",...). Le fait de placer ce type d'informations précisément dans cette colonne assure une bonne prise en compte de leur temporalité lors d'analyses statistiques (cf.

Comparaisons multiples (2 facteurs)).

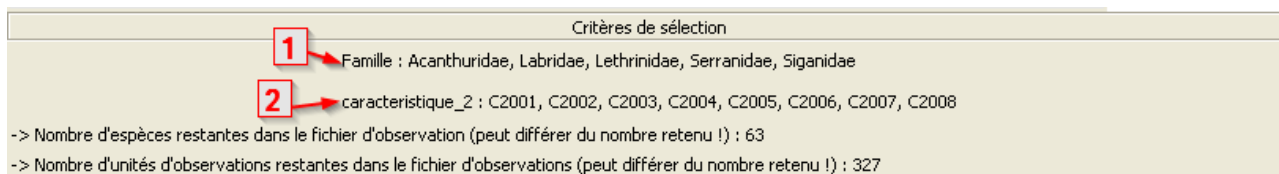
Le champ « caractéristique_2 » est renommé en « annee.campagne » s'il suit le format indiqué ci-dessus ("C<année sur quatre chiffre>").

E. Sélection et re-calcul

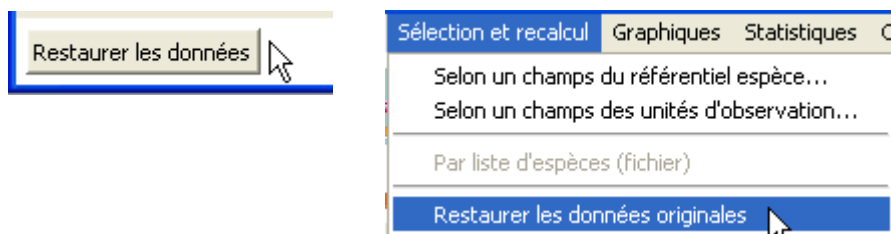
Les « Sélections et recalcul », accessibles par le menu du même nom, vous permettent de réduire le jeu de données à certaines observations, suivant un critère du référentiel espèces et/ou du référentiel d'unités d'observation :



Ces sélections peuvent être imbriquées, c'est à dire que plusieurs critères peuvent être appliqués à la suite. Dans ce cas, la zone d'information sur les critères de sélection (dans l'interface principale) garde la trace de toutes les sélections actives :



Pour restaurer vos données originales sans recharger vos jeux de données, et ainsi gagner du temps, vous pouvez utiliser soit le bouton sur l'interface principale, soit l'entrée dans le menu « Sélection et recalcul » :

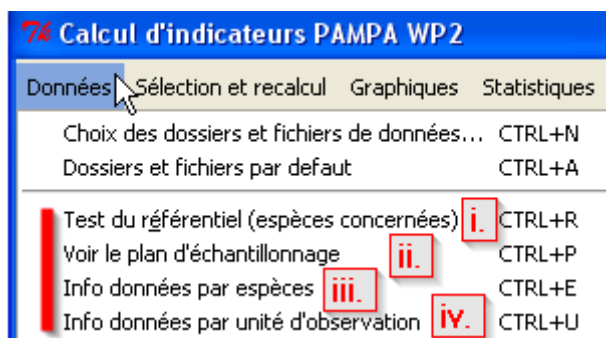


Il n'est pas possible ni prévu de pouvoir annuler juste un niveau de sélection.

Il est cependant envisagé de remplacer, à l'avenir, ce système par un système intégré aux sous-interfaces de production de graphiques et d'analyses. Les sélections seront donc faites juste avant ces opérations et donc modifiables « à la volée » en permanence.

F. Informations sur les données


Une fois les données chargées, des informations sur celles-ci sont accessibles depuis le menu « Données » :



i. Test du référentiel espèces

Donne des informations sur le taux de remplissage des champs du référentiel espèces pour les espèces de la sélection en cours :

74 Informations sur PAMPA-WP1-Meth-5-RefSpMed 191010.txt



Taux de renseignement des champs de PAMPA-WP1-Meth-5-RefSpMed 191010.txt pour le jeu de données (tient compte des sélections)
CB_ObsUVC_Couronne_cor.txt

Nombre de champs de PAMPA-WP1-Meth-5-RefSpMed 191010.txt : 50
 Nombre d'espèces référencées pour CB : 258
 Nombre d'espèces du jeu de données CB_ObsUVC_Couronne_cor.txt : 59

Enregistrer en CSV
 Fermer

Vous pouvez copier-coller ce tableau dans Excel

{Champ #}	Nom	{Nb de valeurs}	{% renseigné}
1	code_espece	59	{100 %}
2	GrSIH	59	{100 %}
3	CodeSIH	30	{50.85 %}
4	IssCaap	57	{96.61 %}
5	TaxoCode	45	{76.27 %}

ii. Plan d'échantillonnage

Affiche un plan d'échantillonnage basique (nombre de stations par an et par statut de protection) :

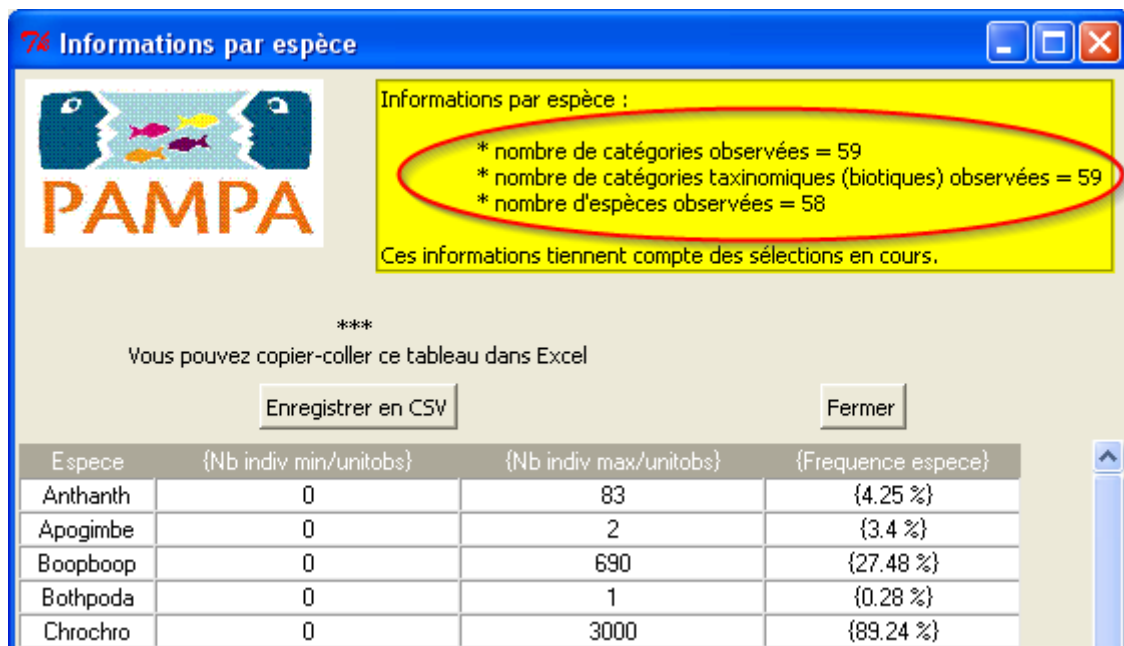
74 Plan d'échantillonnage

{	Statut de protection	HR	RE
Année			
	1995	24	46
	1998	24	48
	2001	24	47
	2004	24	48
	2007	24	44

Des routines de mise en forme d'un plan d'échantillonnage à la carte (choix des facteurs par l'utilisateur) sont en cours de développement.

iii. Informations par « espèce »

Cette entrée du menu permet d'obtenir des informations sur les « espèces » (les guillemets sont de mise, voir l'encadré qui suit) présentes dans la sélection en cours :



À savoir :

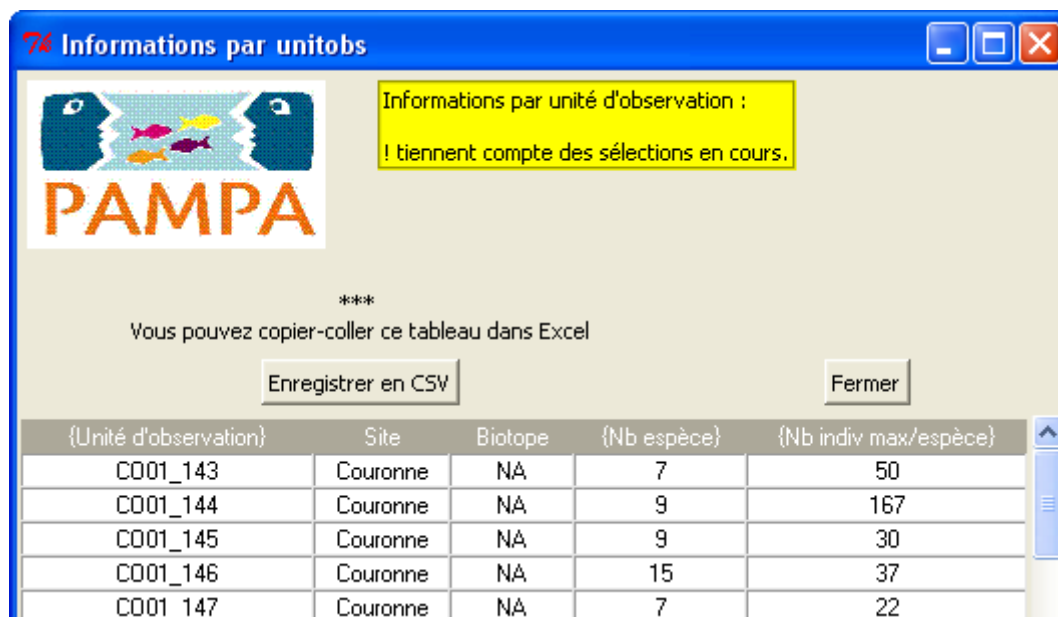
- les nombres minimaux et maximaux d'individus observés par unité d'observation.
- la fréquence d'occurrence (pourcentage d'unité d'observation où l'espèce est présente).

Les catégories présentes dans le champ « Espèce » correspondent à celles du champ « code_espece » du référentiel espèces, présentes dans les observations (en tenant compte de la sélection en cours).

Il peut s'agir (i) de catégories abiotiques, (ii) de groupes taxonomiques supérieurs à l'espèce (e.g. identification au niveau du genre) ou (iii) d'espèces. Le cadre d'information jaune permet d'estimer le nombre de catégories pour chaque niveau.

iv. Informations par unité d'observation

Fournit des informations pour chaque unité d'observation de la sélection en cours :

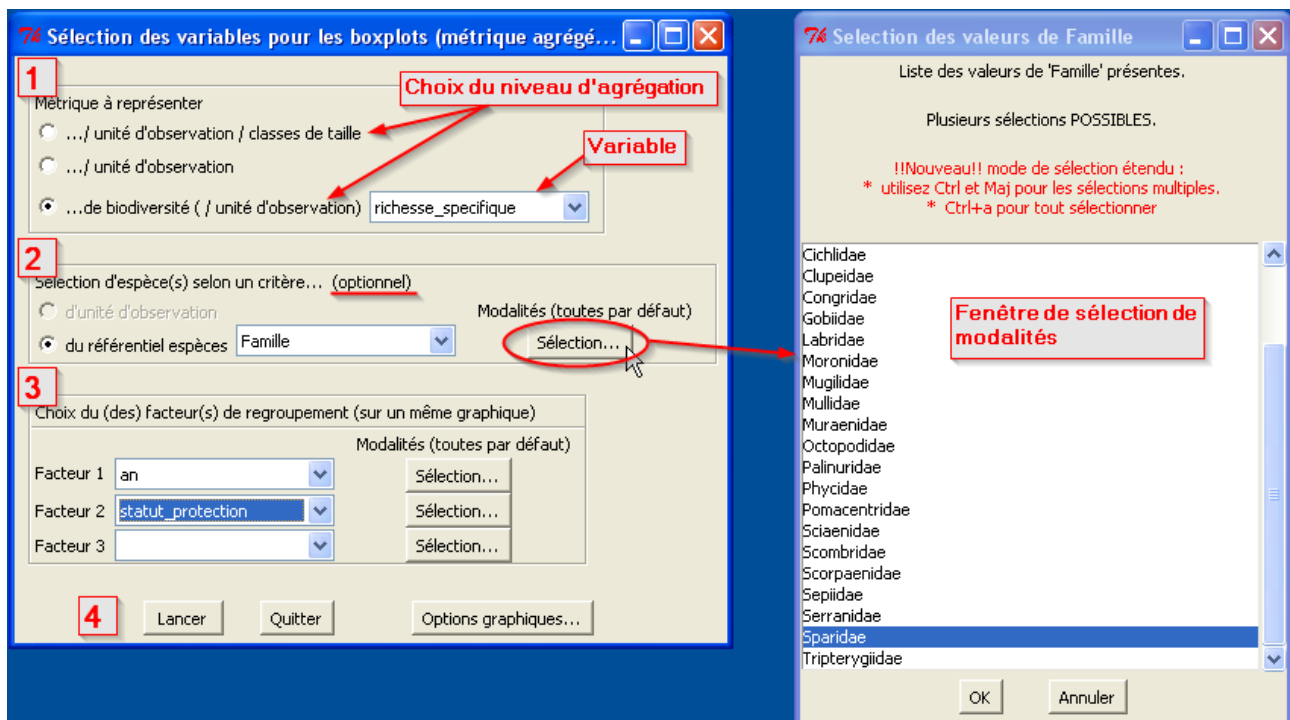


6. Sous-interfaces standard de sélection des métriques/facteurs

L'essentiel des analyses et productions de graphiques est basé sur la construction d'une métrique et le choix des facteurs explicatifs. Il a donc été créé un type d'interface destiné à cette tâche.

Toutes sont basées sur un même squelette, mais présentent de légères variations en fonction du type de traitement à appliquer à la sélection et du niveau d'agrégation de la métrique.

L'organisation générale de ces interfaces est succinctement présentée ici car seront présentes dans toutes les sections qui suivent :



Cette interface se compose de quatre parties principales (dont trois cadres) :

1. un cadre de choix de la variable expliquée et de son niveau d'agrégation. Dans certains cas, ce cadre ne permet pas de choix (variable fixée) mais il reste présent.
2. un cadre de restriction des observations, qui constitue la seconde caractéristique d'une métrique (qui sera dans cet exemple la « richesse spécifique – par unité d'observation – de la famille des sparidés »).
3. un cadre de choix des facteurs explicatifs. L'interface ne limite pas le nombre de facteurs qui peuvent être choisis, mais certains traitements (e.g. les modèles linéaires) n'en autorisent qu'un certain nombre. Le fait de laisser un facteur non renseigné entre deux facteurs renseignés n'a aucune conséquence.

Choix du (des) facteur(s) de regroupement (sur un même graphique)

Modalités (toutes par défaut)

Facteur 1	an	Sélection...
Facteur 2		Sélection...
Facteur 3	statut_protection	Sélection...
Facteur 4		Sélection...



4. une zone de boutons de contrôle.

Note :

La cohérence entre type de métrique, les choix de facteurs (deuxième et troisième cadre) ainsi que le type de traitement, est vérifiée lorsque vous cliquez sur « Lancer » (valable pour toutes les interfaces de ce type dans ce qui suit) :

The screenshot shows the 'Sélection des variables pour les boxplots (métrique/espèce...)' dialog box. It has several sections: 'Métrique (/espèce/unité d'observation) à représenter' with two radio buttons, 'Créer un graphique par facteur...' with two radio buttons and a dropdown, and 'Choix du (des) facteur(s) de regroupement' with four dropdowns. A red circle highlights the 'du référentiel espèces' radio button and the 'Famille' dropdown. A red oval highlights the 'statut_protection' dropdown in the factors section. Overlaid on this is a smaller 'Vérification des sélections' dialog box with a red 'X' icon and the text 'Vous devez sélectionner une métrique'. Below this, a yellow warning icon is shown with the text 'Attention : représentation d'une métrique par espèce mais 'espece' ou 'code_espece' n'est pas utilisé comme facteur'. At the bottom of the main dialog are 'Lancer', 'Quitter', and 'Options graphiques...' buttons.

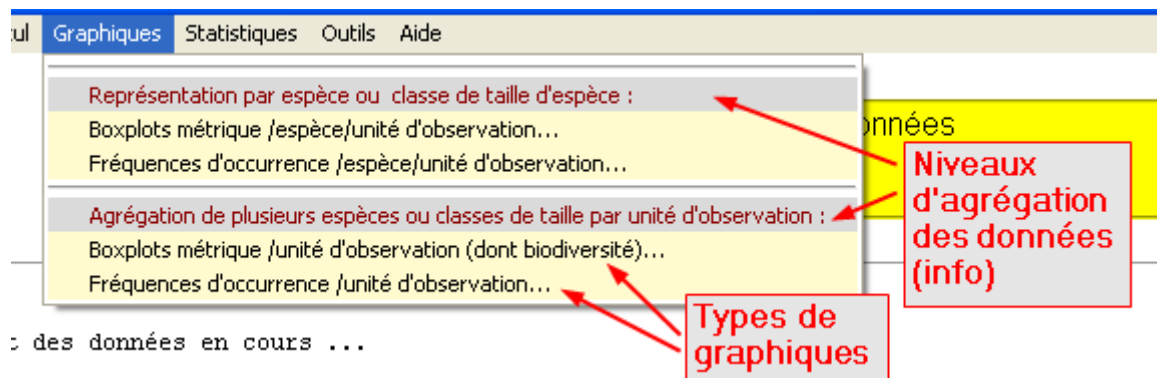
Les messages peuvent correspondre à :

-  des erreurs bloquantes. Rien ne sera fait.
-  des avertissements signalant que le traitement va être appliqué (après avoir cliqué sur « OK ») mais qu'il est probable que les résultats ne correspondent pas à ce que vous souhaitez représenter/analyser.

7. Graphiques

Des interfaces de création de graphiques ont été créées pour faciliter la production en série de graphiques suivant un modèle prédéfini.

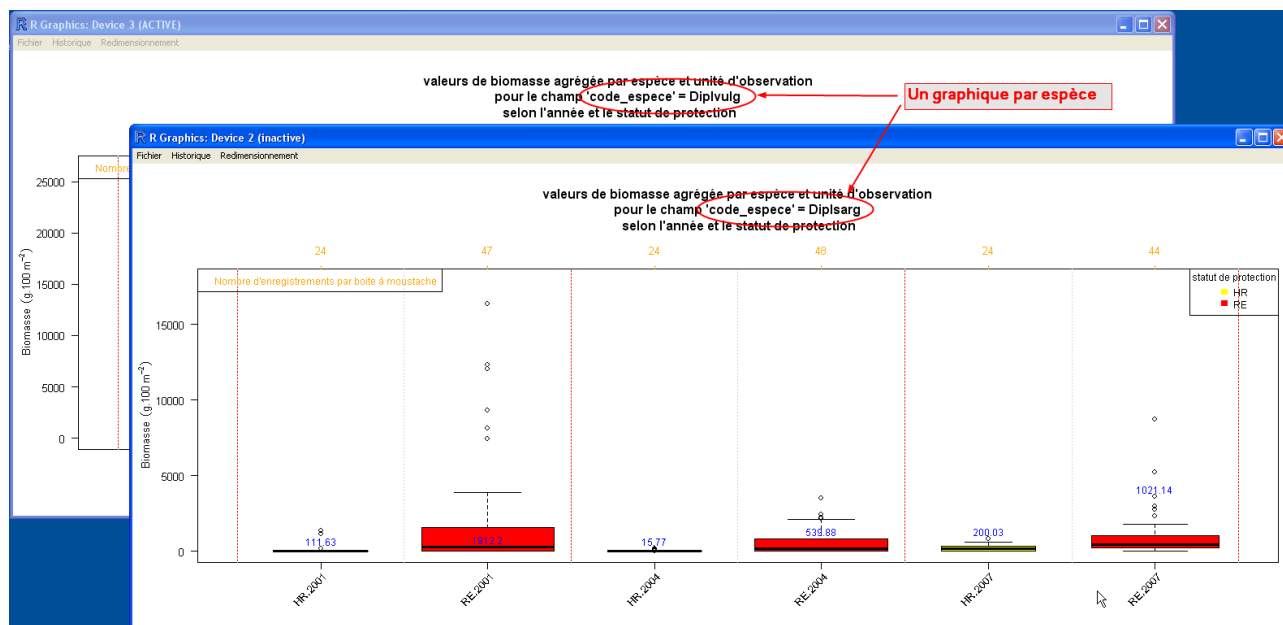
Plusieurs sous menus du menu « Graphiques » donnent accès à différents types de graphiques et niveaux d'agrégation des données :



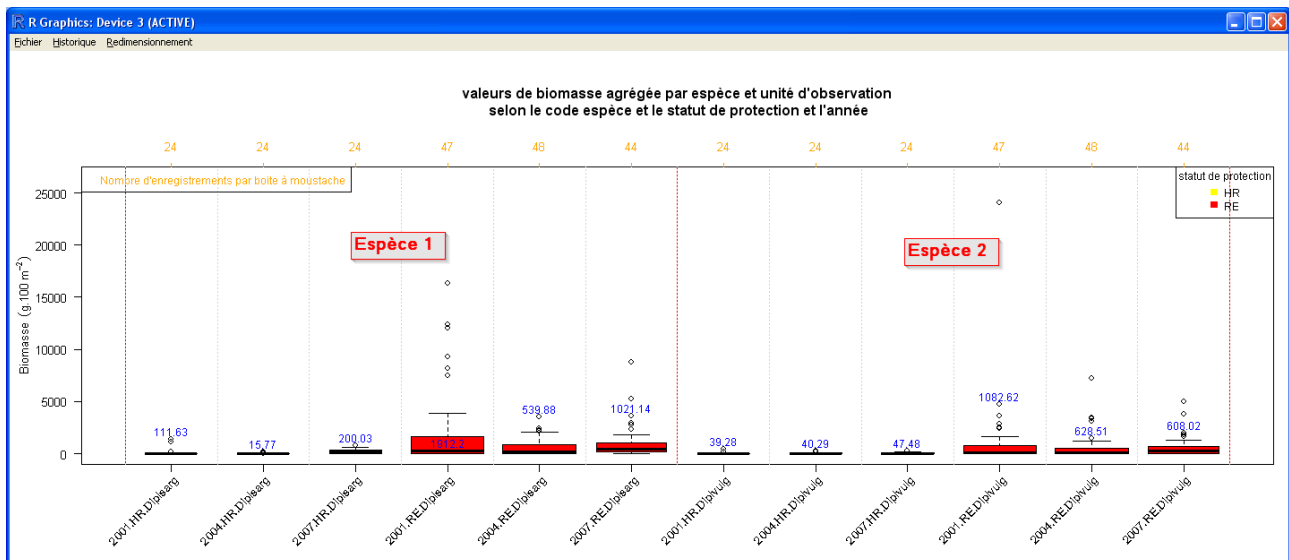
A. Métriques agrégées par unité d'observation par espèce (et éventuellement par classe de taille)

Celles-ci sont à utiliser lorsque l'on souhaite :

- représenter une espèce (ou classe de taille d'espèce) par graphique :



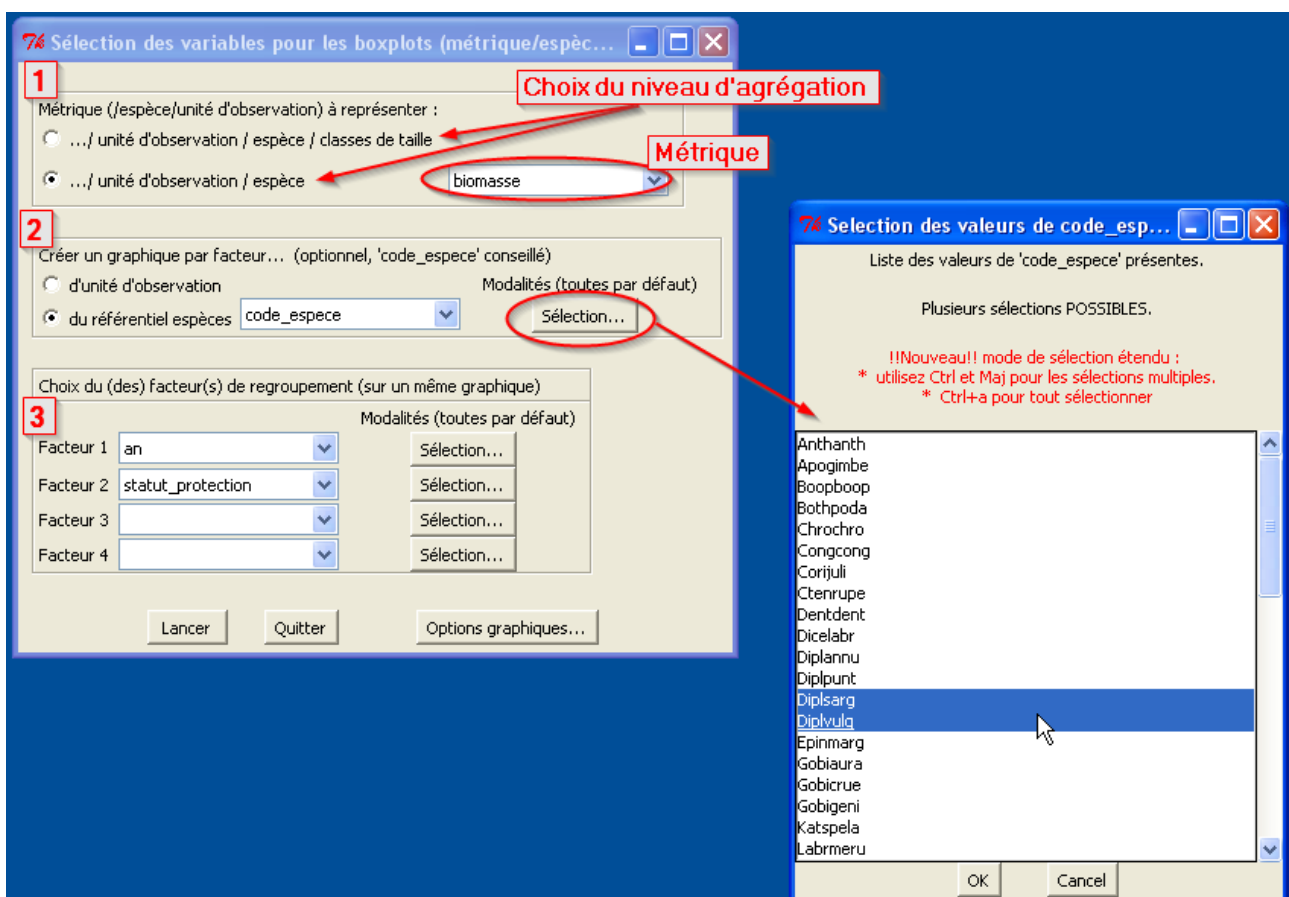
- utiliser l'espèce (plus éventuellement la classe de taille) comme « facteur de regroupement » des *boxplots* (ou *barplots*) :



i. Boîtes à moustaches ou *Boxplots*

Adaptés pour la représentation de la distribution de la plupart des métriques (comme la biomasse dans les exemples ci-dessus).

L'interface se présente comme suit :



1. cadre de sélection de la métrique et du niveau d'agrégation (espèce ou classe de taille d'espèce). Deux niveaux d'agrégation possibles.
2. cadre de sélection du **facteur (optionnel) de séparation des graphiques**. Un nouveau graphique est créé pour chaque modalité (suivant éventuellement une sélection) du facteur choisi ici.

Si vous sélectionnez le champ « espece » (ou « code_espece ») et ne procédez pas à une sélection de celles-ci, vous obtiendrez un graphique par espèce, c-à-d probablement de très nombreux graphiques !

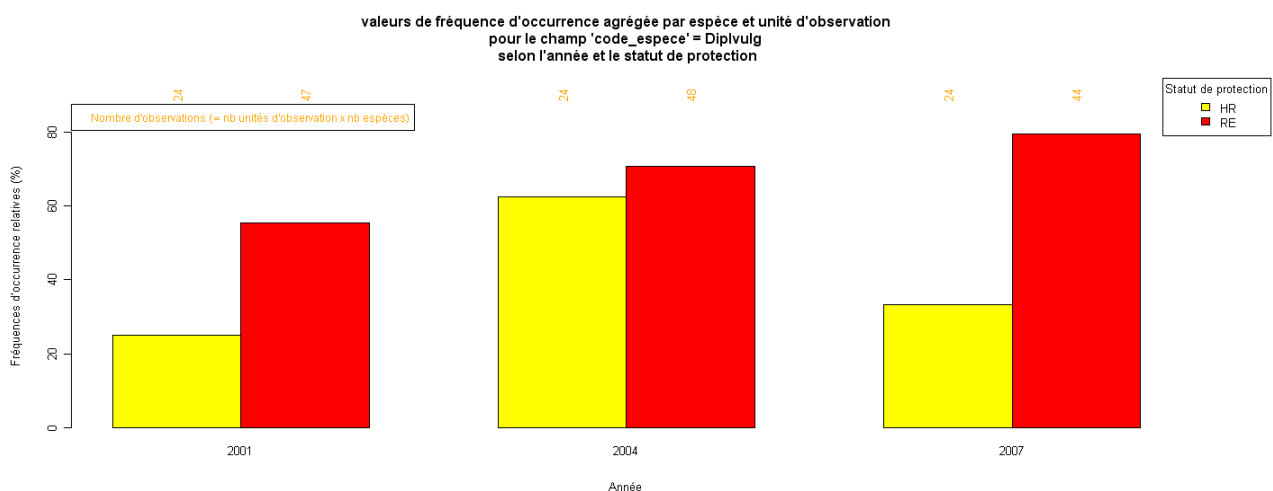
- cadre de sélection du(des) facteur(s) de regroupement et éventuellement sélection des modalités retenues (création d'une boîte à moustache pour chaque combinaison des modalités des facteurs). Un facteur – au minimum – doit être sélectionné dans ce cadre.

ii. Diagrammes en barres ou *Barplots*

Uniquement utilisés pour représenter les fréquences d'occurrence car les *boxplots* ne sont pas adaptés pour celles-ci.

L'interface est donc similaire sauf que la métrique ne peut être choisie :

Ce qui donne (pour une des espèces sélectionnées) :



En l'état actuel des choses, il n'est pas possible de représenter la fréquence

d'occurrence agrégée /**classe de taille**/espèce/unité d'observation.

Cette fonctionnalité sera implémentée à l'avenir.

B. Métriques agrégées par unité d'observation (et éventuellement par classe de taille)

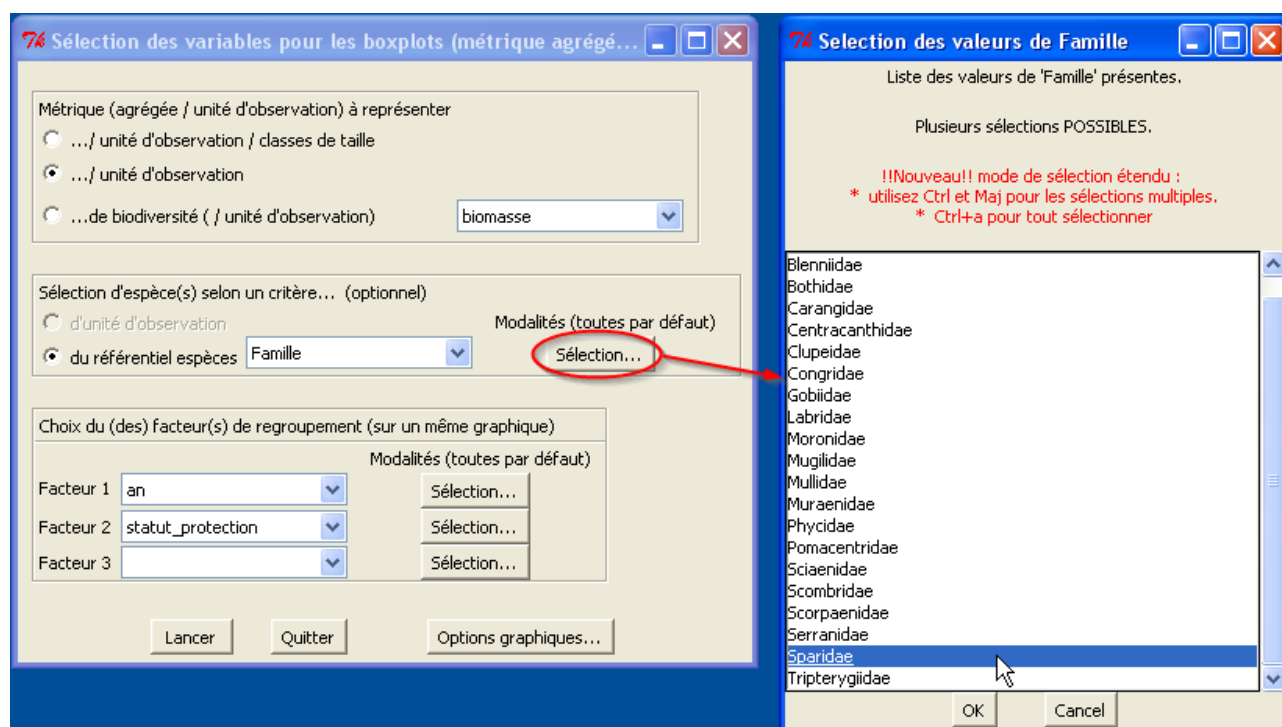
Dans ce cas-ci, la métrique est recalculée pour chaque unité d'observation, après une éventuelle sélection d'individus selon un critère du référentiel espèces ou bien une sélection de taille (classes de tailles P, M et G pour « petits », « moyens » et « grands »).

Dès lors que plusieurs espèces sont amenées à être représentées dans une même boîte à moustache d'un *boxplot*, ou la même barre d'un *barplot*, vous devriez avoir recours à ce niveau d'agrégation des métriques.

i. Boîtes à moustaches ou *Boxplots*

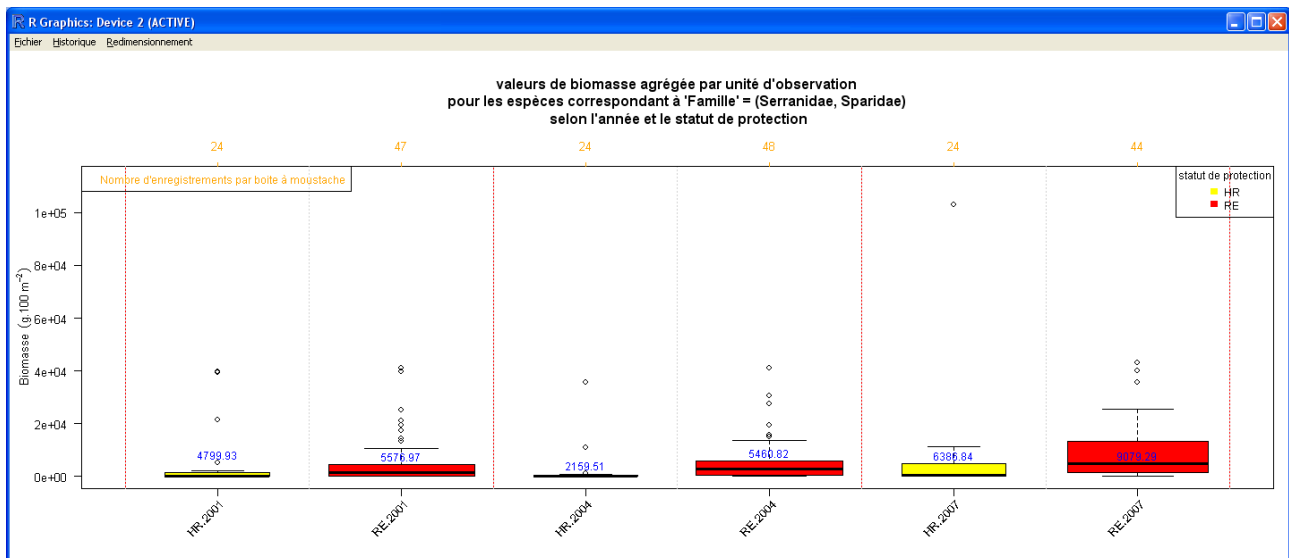
L'interface est proche de celle pour les métriques agrégées par unité d'observation par espèce, mais le second cadre n'a pas la même fonction. Il sert ici à faire une sélection des individus à conserver, généralement selon un critère du référentiel espèce (sinon sur un critère de classe de taille). Les données seront ensuite agrégées par unité d'observation pour tous les individus correspondant aux critères. Si ce cadre est laissé vide, les données de l'ensemble des espèces et classes de tailles seront agrégées par unité d'observation et représentées en fonction des critères de regroupement (troisième cadre).

Par exemple pour avoir la biomasse d'une famille en fonction de l'année et du statut de protection :



Ceci fonctionne également pour avoir la biomasse de plusieurs familles à la fois (e.g. biomasse agrégée de *Sparidae* et *Serranidae*).

Avec cette sous-interface, un seul graphique est produit à la fois :

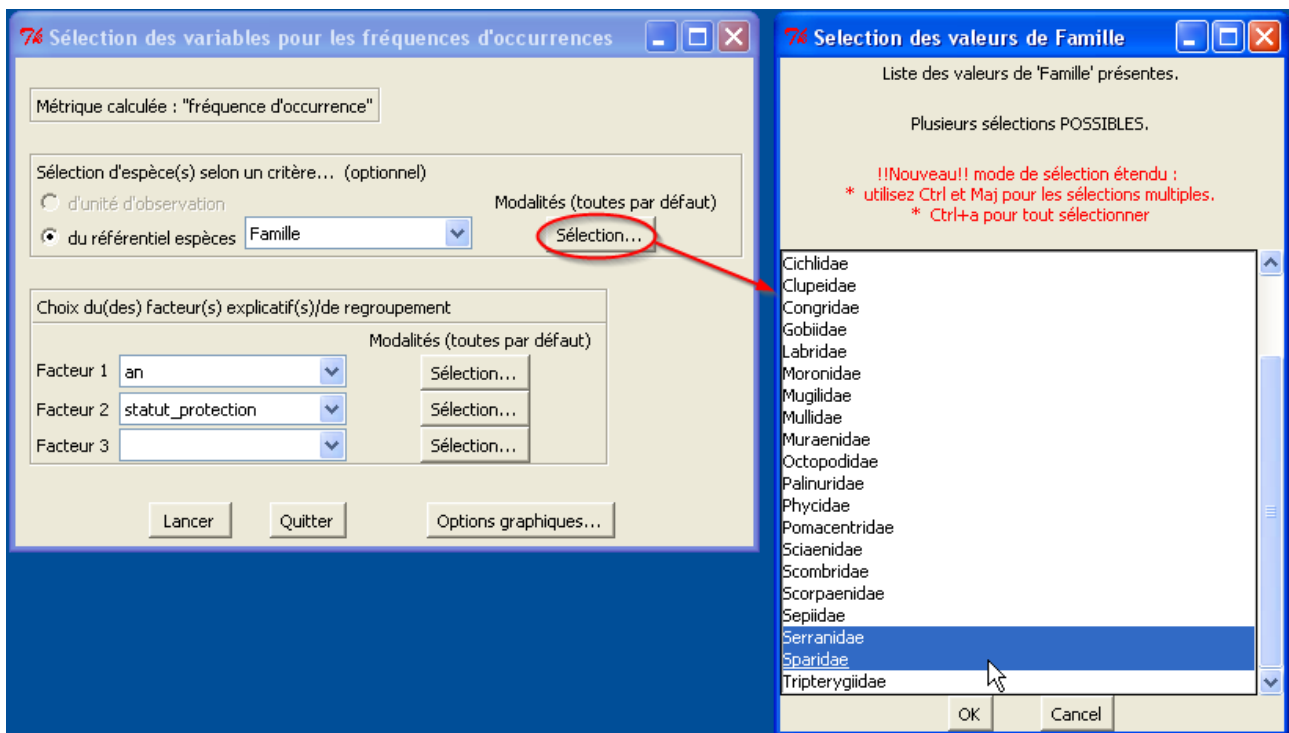


Dans le premier cadre, des indices de biodiversité (calculés au niveau de l'unité d'observation) peuvent également être sélectionnés (troisième « bouton radio » puis choix de la métrique).

Le facteur du second cadre est optionnel : si aucun facteur n'y figure ou aucune modalité n'est sélectionnée, toutes les espèces (ou classes de taille) sont conservées.

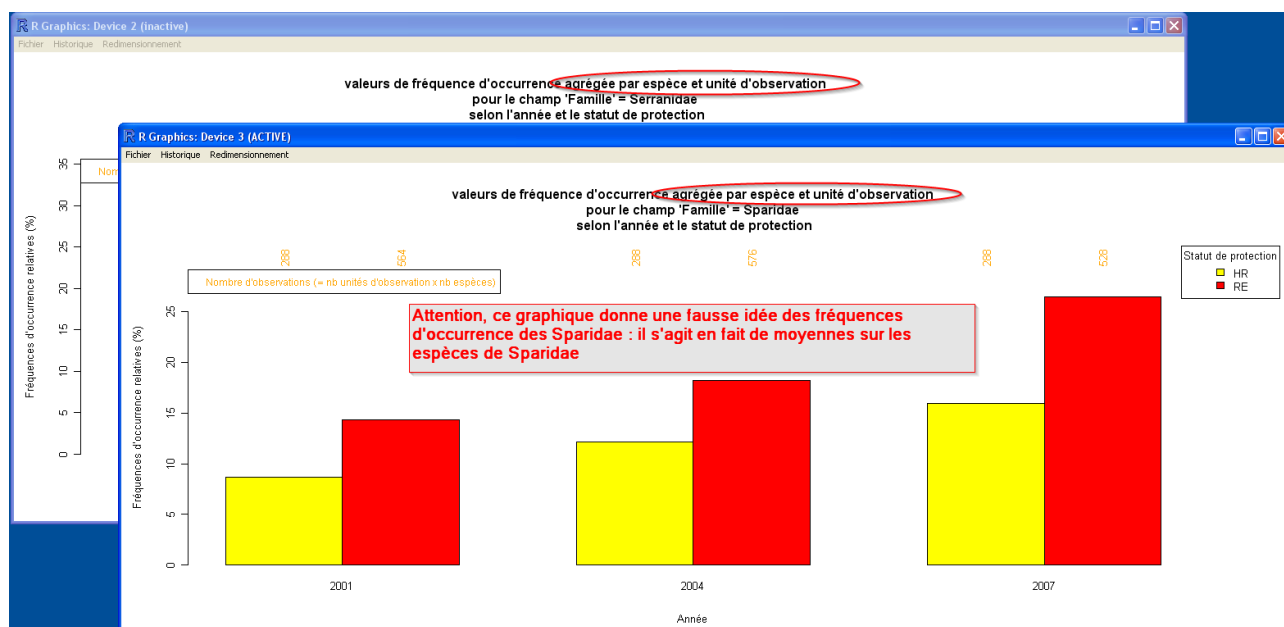
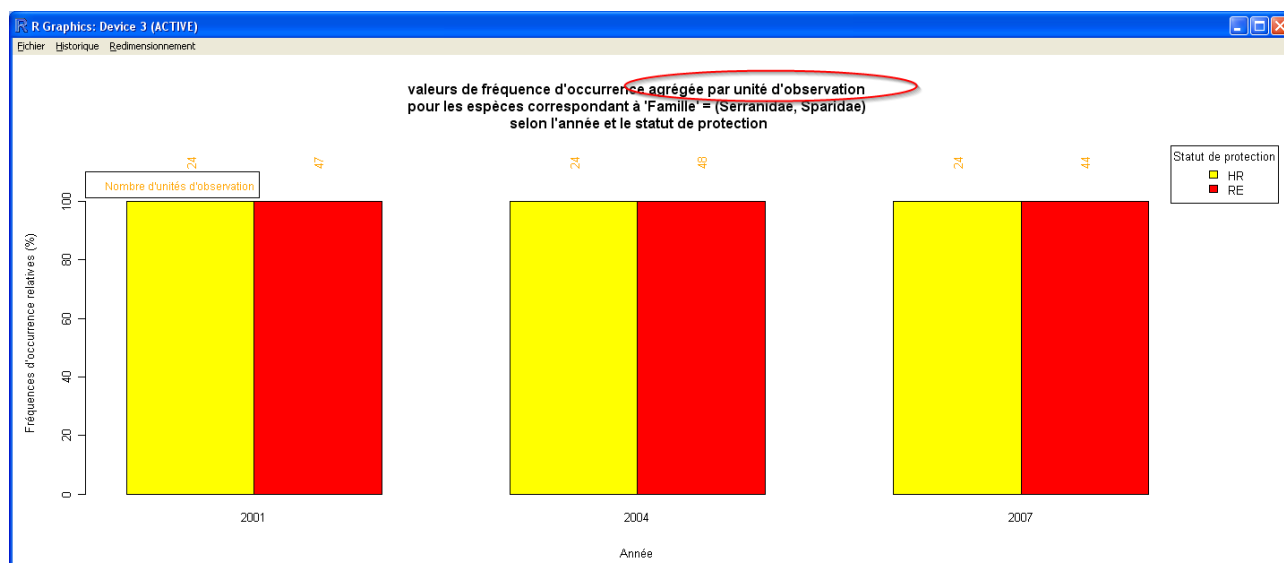
ii. Diagrammes en barres ou *barplots*

Le principe de l'interface est le même que dans le cas des *boxplots* pour les métriques agrégées par unité d'observation (et éventuellement par classe de taille), à l'exception du premier cadre, où la métrique ne peut être choisie. Le second cadre est ici aussi destiné à procéder à une sélection d'individus selon un critère du référentiel espèce ou de classe de taille :



Cet exemple donne un (premier) graphique sans grand intérêt puisque l'une ou l'autre des deux familles au moins est représentée dans chaque unité d'observation (fréquence

d'occurrence de 100 %). Il montre cependant que c'est bien la fréquence d'occurrence de l'ensemble des familles, et non la « moyenne » des espèces qui les composent (seconds graphiques obtenus avec l'interface pour les métriques agrégées par espèce et unité d'observation) qui est représentée :



En l'état actuel des choses, il n'est pas possible de représenter la fréquence d'occurrence agrégée /**classe de taille**/unité d'observation.

Cette fonctionnalité sera implémentée à l'avenir.

iii. Remarques

Classes de tailles :

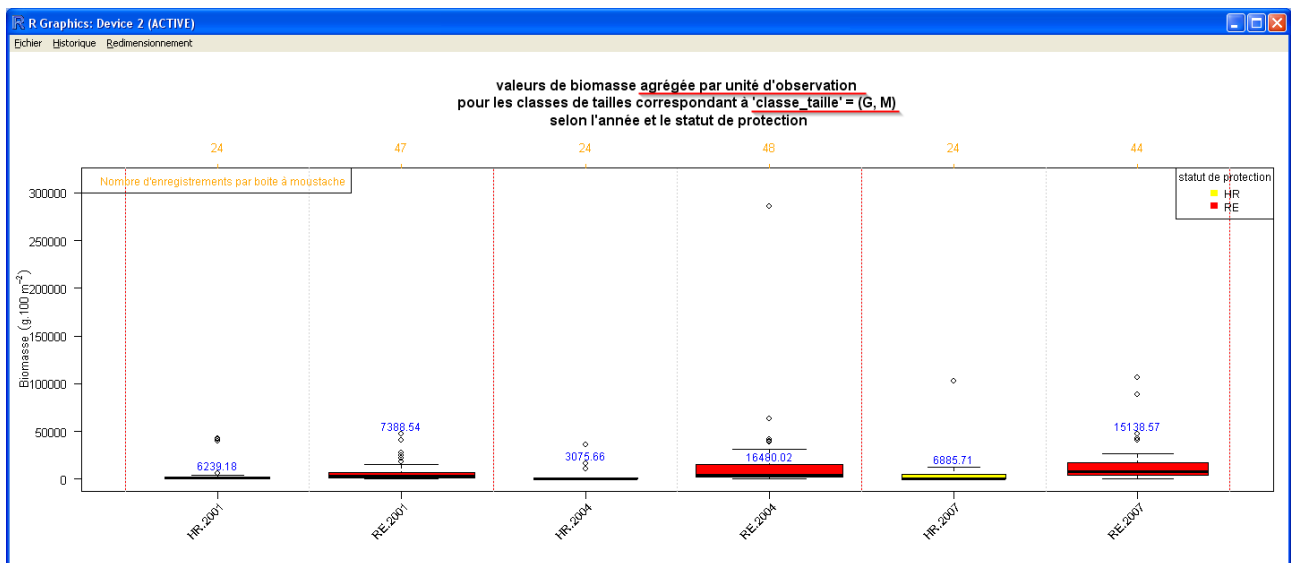
Si une métrique « .../unité d'observation/classe de taille » est sélectionnée dans le premier cadre (*boxplots* uniquement pour l'instant) et que le facteur « classe_taille » n'est utilisé nulle part, l'avertissement suivant sera affiché :



Et le résultat (moyenne sur les classes de tailles) ne correspondra certainement pas à ce qui est souhaité.

Dans le cas contraire, deux cas peuvent être distingués :

1. le facteur « classe_taille » est utilisé dans le second cadre (sélection des individus) et la métrique sera alors agrégée par unité d'observation, après sélection des classes de tailles souhaitées :

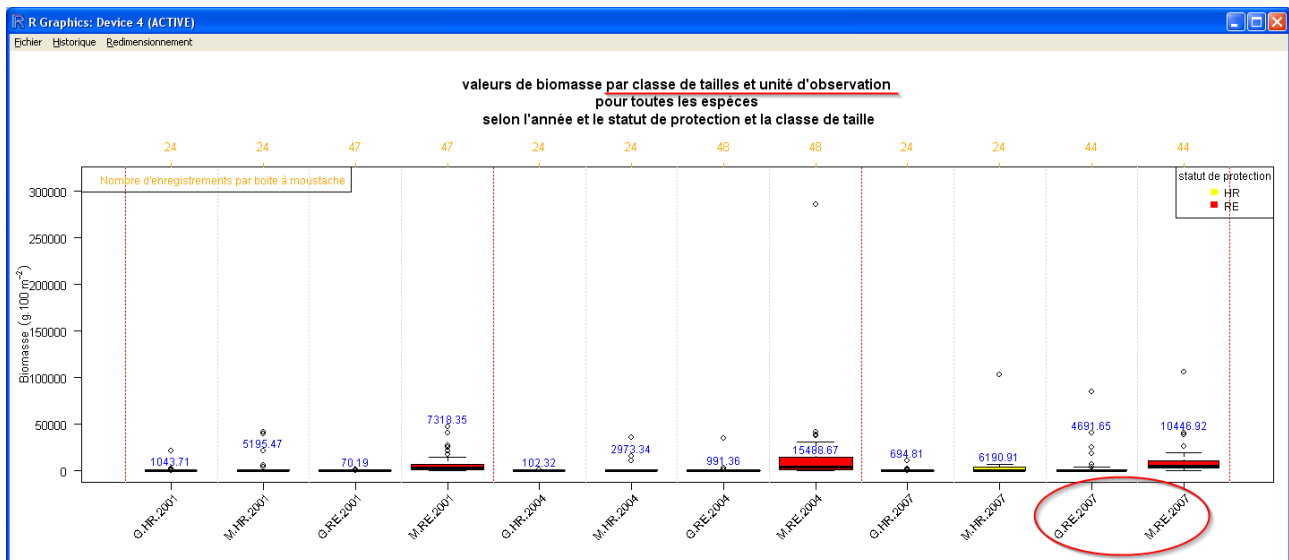


2. le facteur « classe_taille » est utilisé dans le troisième cadre (facteurs de regroupement), la métrique est donc agrégée par unité d'observation par classe de taille et les boîtes à moustaches sont séparées par classe de taille :

Choix du (des) facteur(s) de regroupement (sur un même graphique)

Modalités (toutes par défaut)

Facteur 1	an	Sélection...
Facteur 2	statut_protection	Sélection...
Facteur 3	classe_taille	Sélection...
Facteur 4		Sélection...



(Notez qu'en faisant la somme des moyennes – en bleu – pour les groupes cerclés de rouge, on retrouve bien les 15138,57 g.100m⁻² du précédent graphique pour les zones en réserve et l'année 2007.)

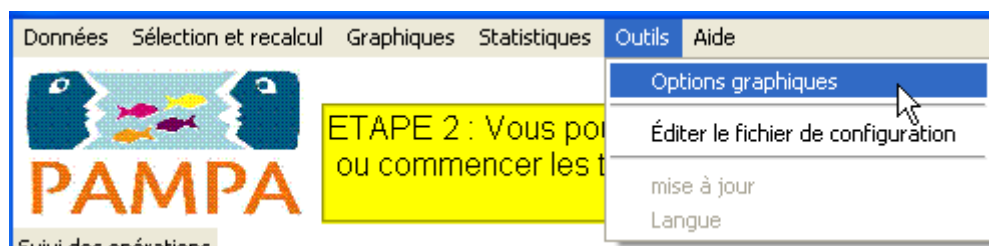
C. Remarques générales sur les graphiques

i. Rang d'utilisation du facteur « statut de protection » :

Que ce soit pour les *boxplots* ou les *barplots*, il est recommandé de toujours mettre le statut de protection comme deuxième facteur de regroupement (troisième cadre) afin que les codes couleurs correspondent à des niveaux de protection. Ceci n'est bien évidemment pas obligatoire, mais permet de mettre en lumière l'effet de la protection.

ii. Options graphiques

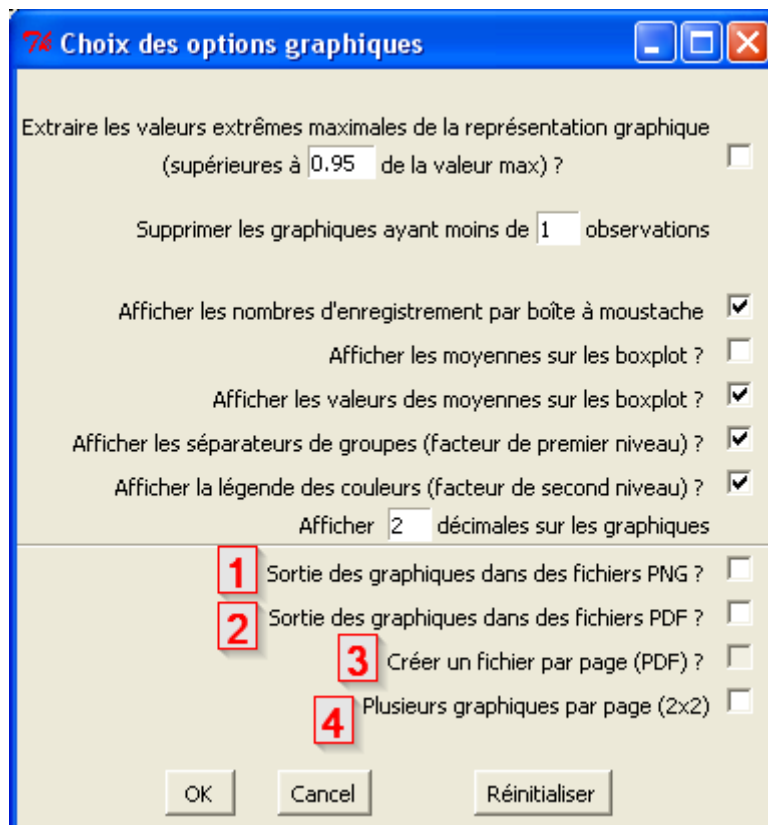
Les options graphiques peuvent être modifiées à partir, soit de l'interface principale, soit des sous-interfaces de création de graphiques ; respectivement :



et dans les interfaces standard de sélection des métriques :



qui ouvrent une sous-interface de gestion des options graphiques (ici avec les options par défaut) :



La plupart des choix sont parlants. Les options numérotées en rouge – sous le séparateur – concernent les périphériques graphiques :

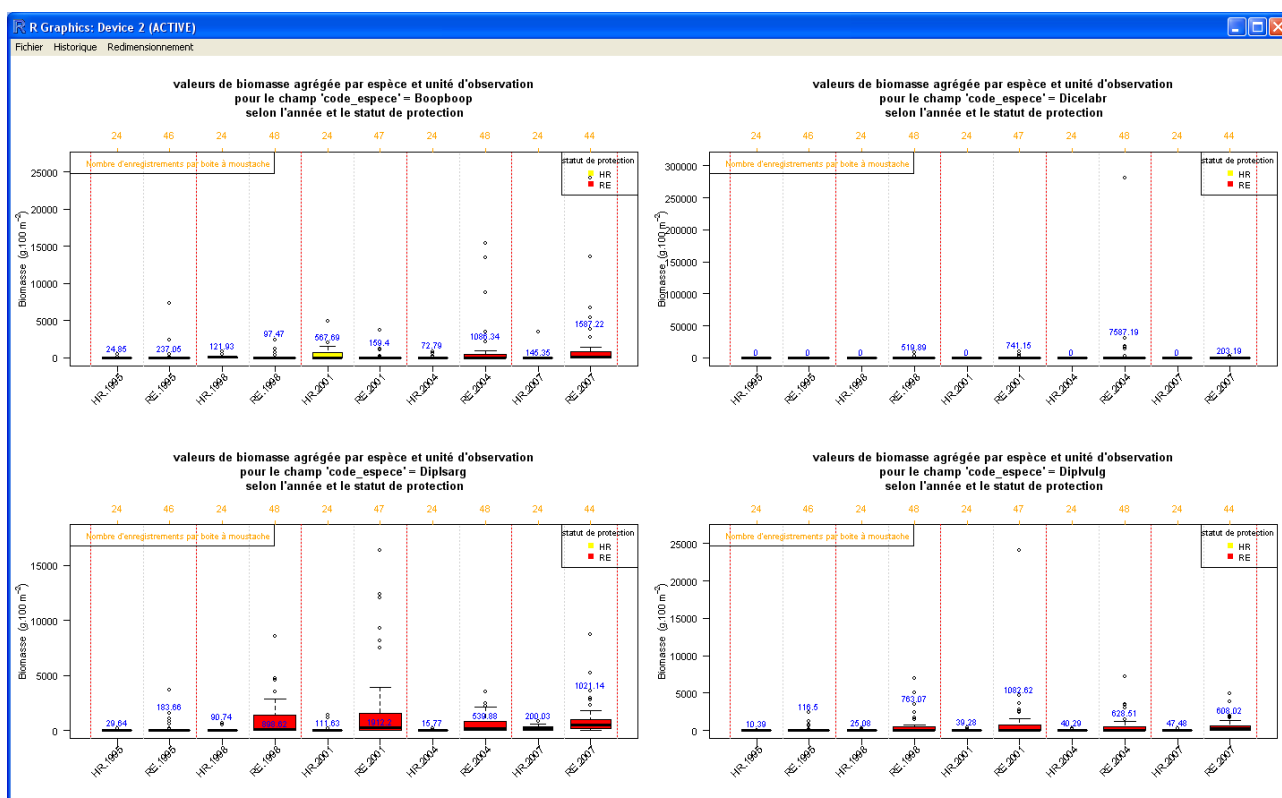
1. Le graphique n'est pas affiché mais un fichier au format **PNG** (insérable dans tout type de document, ou presque) est créé dans le dossier « <dossier de travail>/FichiersSortie » avec un nom aussi explicite que possible, contenant :
 - le type de graphique (boxplot dans l'exemple ci-dessous).
 - la métrique (boxplot).
 - le niveau d'agrégation (Agr-CL+unitobs).
 - d'éventuels facteurs de séparation des graphiques/sélections d'individus + les modalités sélectionnées (Famille(Lethrinidae)).
 - le(s) facteur(s) de regroupement (statut_protection-classe_taille).

Ce qui donne par exemple :

```
boxplot_biomasse_Agr-CL+unitobs_Famille(Lethrinidae)_statut_protection-classe_taille.png
```

2. Le graphique n'est pas affiché mais un fichier au format **PDF** est créé dans le dossier « <dossier de travail>/FichiersSortie ». Option mutuellement exclusive avec l'option 1.
3. Option active uniquement si 2. est active, et utile en cas de création de graphiques par lots (métriques agrégées par espèce...). Force les graphiques PDF d'un même lot à être créés dans des fichiers séparés. Si l'option est inactive, ils sont créés dans des pages séparées d'un même fichier.
4. Pour tout type de périphérique graphique. Permet de placer jusqu'à quatre

graphiques sur la même page/périphérique graphique. De même que l'option 3., celle-ci n'est utilisée que pour le traitement de graphiques par lots (i.e. pour les métriques agrégées par espèce, séparés par « `code_espece` »).



Les choix d'option ne seront effectifs qu'après avoir cliqué sur le bouton « OK ». Le bouton « Réinitialiser » permet de retrouver les options par défaut.

Les options graphiques personnalisées sont persistantes pour la session en cours, c'est-à-dire qu'elles resteront les mêmes après fermeture puis ouverture d'une quelconque sous-interface, « sélection et recalcul », rechargement de données, etc.

Elles sont en revanche réinitialisées aux valeurs par défaut à chaque chargement de l'interface principale.

iii. Options graphiques supplémentaires (cachées)

Un certain nombre d'options supplémentaires – modifiables dans la console R uniquement – ont été ajoutées afin de contrôler plus finement le rendu des graphiques (en particulier pour permettre la production de graphiques destinés à la publication dans des revues scientifiques). Celles-ci sont présentées ci-dessous sous la forme `<nom d'option>` (`<type de donnée>`, `<valeur par défaut>`):

- **P.graphPaper** (booléen, FALSE) : si la valeur est TRUE, les titres sont supprimés et des graphiques plus compacts (moins de place dans un document) sont produits.
- **P.warnings** (booléen, TRUE) : si la valeur est TRUE, les avertissements pour petits effectifs et le troncature du graphique (valeurs extrêmes) sont affichés sous forme de texte en haut du graphique.
- **P.colPalette** (chaîne de caractères, "heat") : nom d'une palette de couleurs prédéfinie. La palette par défaut correspond aux couleurs de la plupart des

graphiques de cette documentation. L'autre valeur possible ("gray") produit des graphiques en niveaux de gris. Lorsque cette option est changée, elle n'est effective qu'après avoir lancé la commande :

```
> makeColorPalette.f()
```

- **P.graphWMF** (booléen, FALSE) : lorsque la valeur est TRUE, si les graphiques sont affichés à l'écran et sous Windows uniquement, ils sont également sauvegardés dans des fichiers .wmf placés dans le dossier de résultats.
- **P.pointMoyenneCol** (chaîne de caractères, "blue"), **P.valMoyenneCol** (chaîne de caractères, "blue") et **P.sepGroupesCol** (chaîne de caractères, "red") permettent de choisir respectivement les couleurs des points de moyennes, des valeurs de moyennes et des séparateurs de groupes.
- **P.pointMoyenneCex** (numérique, 1) : multiplicateur de taille des points de moyenne.
- **P.pointMoyennePch** (entier, 18) : type de point de moyenne.
- **P.cex** (numérique, 1) : multiplicateur de taille de police des graphiques.
- **P.ncolGraph** (entier, 2) et **P.nrowGraph** (entier, 2) : respectivement les nombres de colonnes et lignes lorsque plusieurs graphiques sont affichés dans une même fenêtre/un même fichier (traitement par lot).

Ces options peuvent être changées en exécutant dans la console R une commande du type :

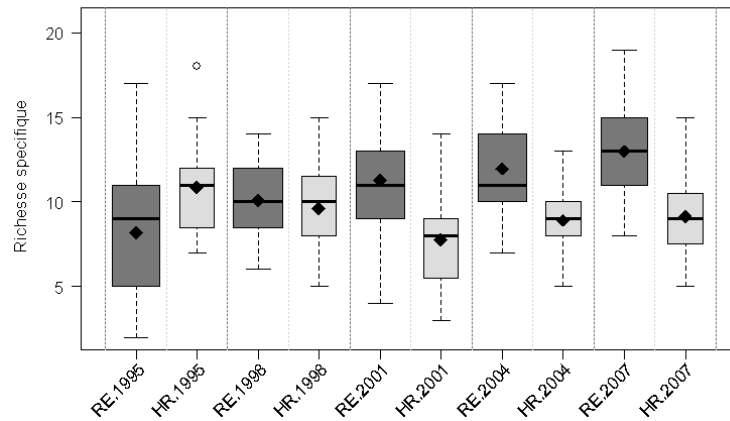
```
> options(<nom d'option> = <nouvelle valeur>)
```

Par exemple, les commandes suivantes, exécutées après le chargement de la plateforme :

```
options(P.colPalette="gray", P.pointMoyenneCol = "black",
        P.sepGroupesCol = "#6f6f6f", P.valMoyenneCol = "black",
        P.NbObsCol = "black",
        P.legendeCouleurs = FALSE, ## Accessible par l'interface !
        P.valMoyenne = FALSE, P.NbObs = FALSE, P.pointMoyenne = TRUE,
        P.pointMoyennePch = 18, P.pointMoyenneCex = 2, P.cex=1.1,
        P.graphWMF=TRUE, P.warnings=FALSE, P.graphPaper=TRUE)

makeColorPalette.f()      ## nécessaire pour la prise en compte
                          ## de la nouvelle palette de couleurs.
```

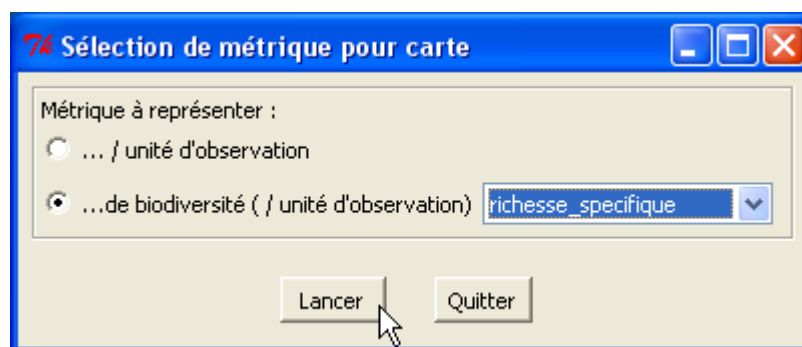
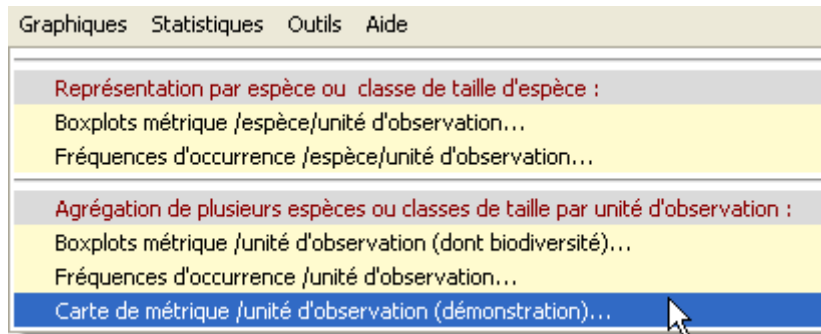
permettent d'obtenir des graphiques qui ressemblent à ce qui suit :



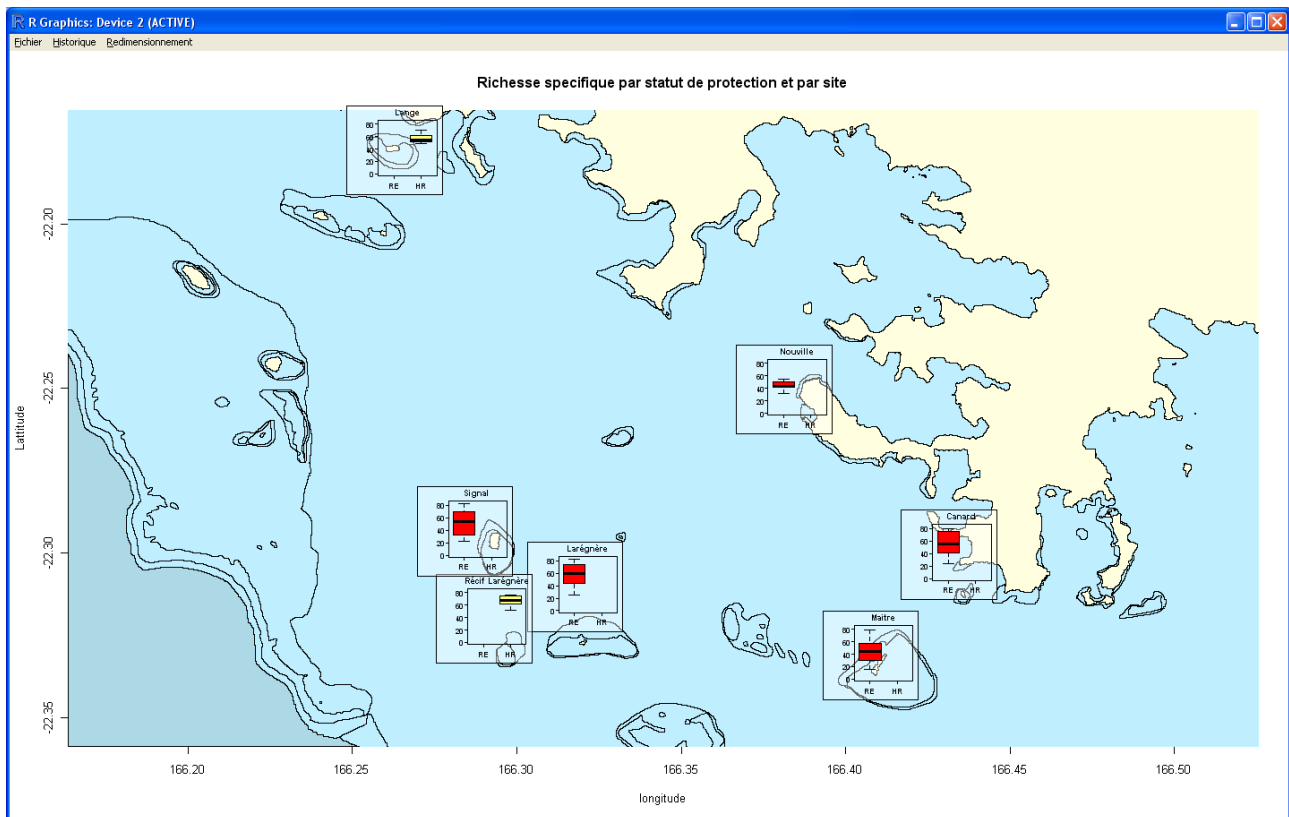
Ces commandes d'options peuvent également être placées dans le fichier de configuration : <C:/PAMPA/Exec/config.r>.

D. Cartes (démonstration sur données Nouvelle-Calédonie)

Afin de donner un avant goût des possibilités de représentation spatiale des données – qui seront développées ultérieurement – une démonstration a été créée pour les sites de Nouvelle-Calédonie (pour d'autres sites, l'option est absente) :



Elle ne permet de représenter que les métriques agrégées par unités d'observations, sous forme d'un *boxplot* par site avec l'effet statut de protection. Le centre de chaque *boxplot* se situe au barycentre des positions de stations du site correspondant.

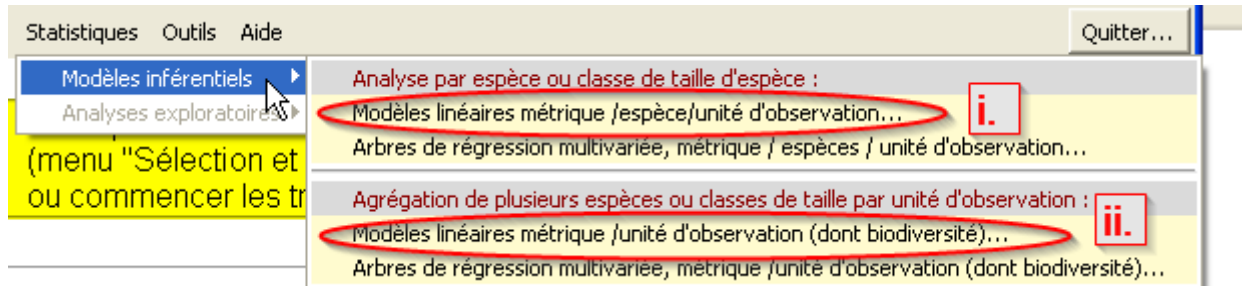


Il est toutefois possible de limiter la métrique à une espèce ou un groupe d'espèces à l'aide des « sélections et recalcul ».

8. Analyses statistiques

A. Modèles linéaires

Sous le menu « Statistiques > Modèles inférentiels » se trouvent les analyses pour les comparaisons temporelles et spatiales :



Les boîtes de dialogue de choix des métriques/indicateurs et facteurs sont très similaires à celles développées pour les graphiques. Ce sont d'ailleurs les mêmes critères qui doivent mener au choix de métriques agrégées par espèce par unité d'observation ou par unité d'observation uniquement (plus éventuellement par classe de taille). Se reporter aux explications concernant les choix des niveaux d'agrégation pour les graphiques.

Les différents modèles statistiques implémentés sont les modèles linéaires simples (ANOVAs & régressions ; sur données log-transformées ou non) et les modèles linéaires généralisés (GLMs ; famille Gamma, binomiale négative et binomiale, selon la nature de la métrique).

En dépit de leur plus grande complexité théorique, les GLMs produisent ici des résultats similaires et tout aussi aisés à interpréter que ceux des modèles linéaires simples.

Dans de nombreux cas, ils permettent l'obtention de résultats plus robustes lorsque les modèles linéaires simples ne peuvent être validés (voir la section 8.A.iv., page 41).

Pour les modèles linéaires, les facteurs explicatifs (troisième cadre de la sélection des variables) doivent présenter au moins deux modalités chacun. Sélectionner une seule modalité ou un facteur n'en contenant qu'une produira donc une erreur.

i. Métriques agrégées par unité d'observation par espèce (et éventuellement par classe de taille)

Le principe de fonctionnement de l'interface est le même que pour les graphiques avec le même niveau d'agrégation des métriques à cela près que les options graphiques n'y sont pas disponibles :

74 Sélection des variables pour les modèles linéaires (métriq...

Métrique expliquée :

☐ .../ unité d'observation / espèce / classes de taille

☒ .../ unité d'observation / espèce biomasse

Séparer les analyses par facteur... (optionnel)

☐ d'unité d'observation Modalités (toutes par défaut)

☒ du référentiel espèces code_espece Sélection...

Choix du(des) facteur(s) explicatif(s)

Facteur 1 an Sélection...

Facteur 2 statut_protection Sélection...

Facteur 3 Sélection...

Lancer Quitter Options graphiques...

En cliquant sur « Lancer », vous lancez un module de sélection de la distribution théorique pour les données de la métrique choisie. La distribution théorique va orienter vers le type d'analyse adéquat. Pour des données continues, les choix sont entre Anova (modèle linéaire), Anova sur données log-transformées (modèle linéaire sur données log-transformées) et Modèle Linéaire Généralisé (GLM) de la famille Gamma. Les sorties de ces trois types de modèles se présentent sous des formes similaires.

Selon la nature de la métrique sélectionnée, le nombre de choix peut être variable. Pour les données de comptage (discrètes, *eg.* « nombre »), la loi binomiale négative est également disponible.

Remarque : dans le cas des données de présence/absence

Métrique expliquée :

☐ .../ unité d'observation / espèce / classes de taille

☒ .../ unité d'observation / espèce

pres_abs

biomasse

densite

nombre

poids

poids_moyen

pres_abs

taille_moy

– qui permettent de conduire des analyses correspondant aux fréquences d'occurrence pour la partie graphique – vous n'aurez pas de choix de modèle, la distribution binomiale étant automatiquement retenue.

La distribution qui s'ajuste le mieux à vos données selon le critère d'information d'Akaike (AIC) – lequel doit être le plus petit possible – est déjà présélectionnée dans l'interface ci-après. Pour en changer, il vous suffit de cliquer sur le graphique ou le bouton correspondant à votre choix.

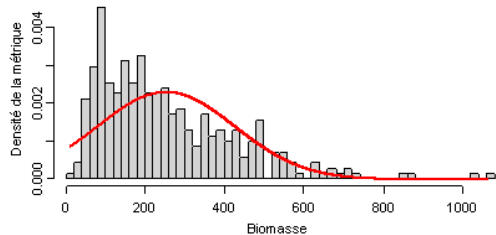
Choix de distribution théorique de la métrique 'biomasse'

INFO :

Cette fenêtre vous permet de choisir la distribution la plus adaptée pour faire vos analyses.

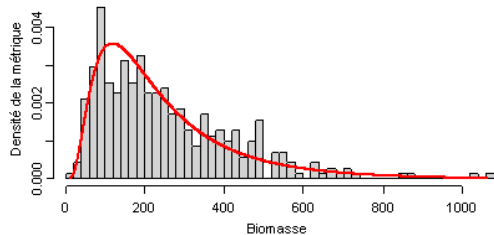
La distribution (courbe rouge) s'ajustant le mieux à vos données (histogramme) d'après le critère d'information de Akaike (AIC ; doit être le plus petit possible) est pré-sélectionnée.

Comparaison avec la loi Normale



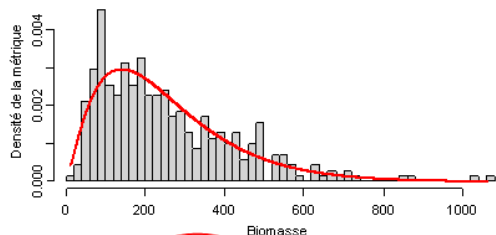
☐ loi Normale (AIC=4642). **Modèle : ANOVA**

Comparaison avec la loi log-Normale



☐ loi log-Normale (AIC=4515). **Modèle : ANOVA, données log-transformées**

Comparaison avec la loi Gamma



☒ loi Gamma (AIC=4508). **Modèle : GLM, famille 'Gamma'**

OK

Annuler

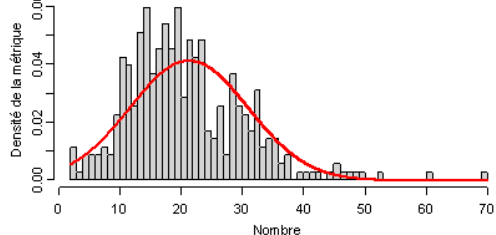
Choix de distribution théorique de la métrique 'nombre'

INFO :

Cette fenêtre vous permet de choisir la distribution la plus adaptée pour faire vos analyses.

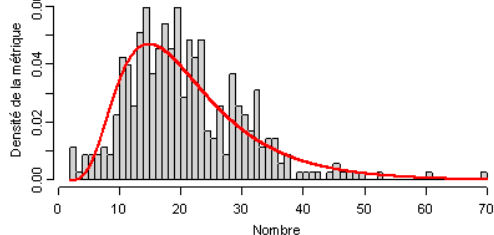
La distribution (courbe rouge) s'ajustant le mieux à vos données (histogramme) d'après le critère d'information de Akaike (AIC ; doit être le plus petit possible) est pré-sélectionnée.

Comparaison avec la loi Normale



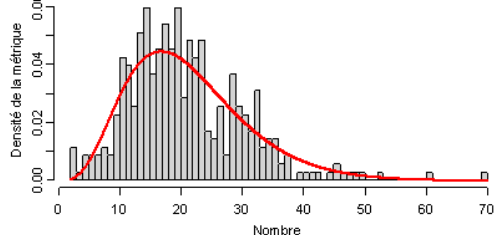
☐ loi Normale (AIC=2605). **Modèle : ANOVA**

Comparaison avec la loi log-Normale



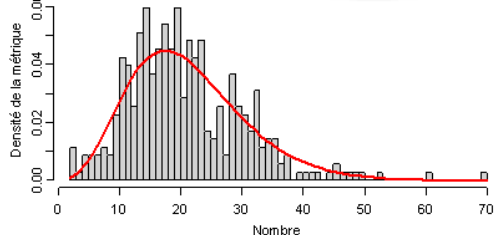
☐ loi log-Normale (AIC=2602). **Modèle : ANOVA, données log-transformées**

Comparaison avec la loi Gamma



☐ loi Gamma (AIC=2566). **Modèle : GLM, famille 'Gamma'**

Comparaison avec la loi Binomiale négative



☒ loi Binomiale négative (AIC=2561). **Modèle : GLM, famille 'Binomiale négative'**

OK

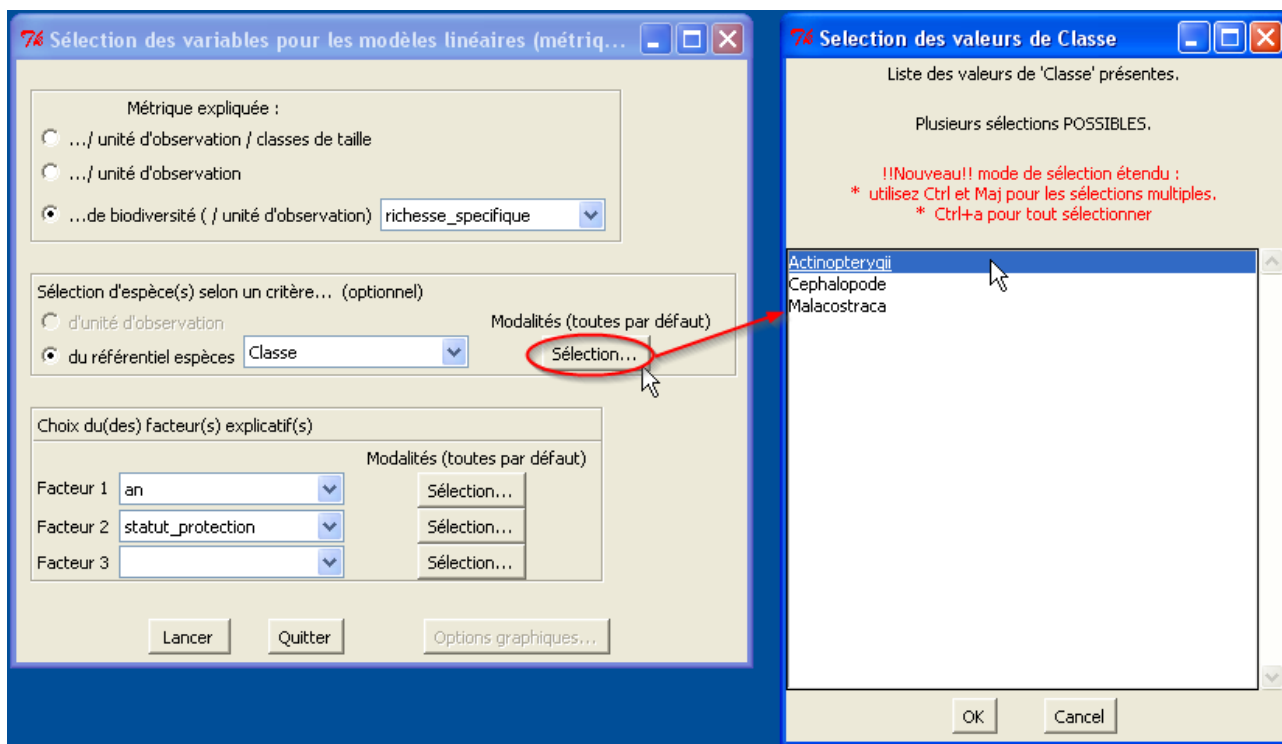
Annuler

En cliquant sur « OK » vous lancez l'analyse correspondant à la distribution choisie.

Dans les cas où les lois log-normale et binomiale négative sont toutes deux disponibles et donnent des ajustements de qualités relativement similaires, il est préférable de sélectionner la seconde.

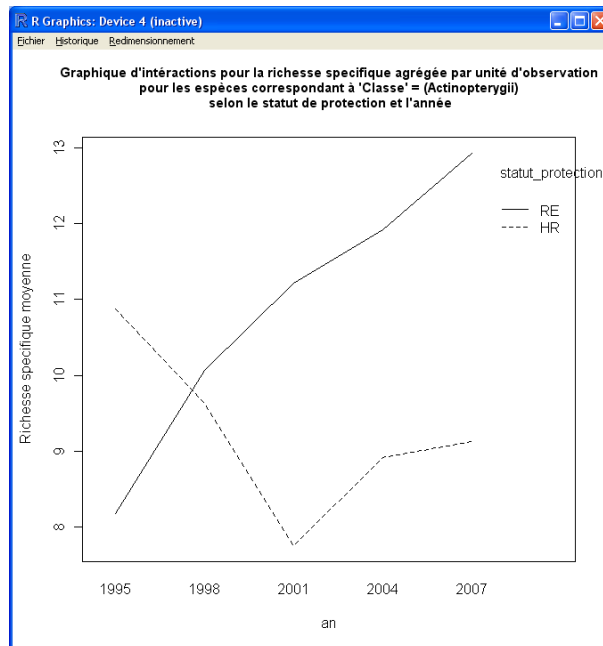
ii. Métriques agrégées par unité d'observation (et éventuellement par classe de taille)

Ici aussi, l'interface ressemble à celle des *boxplots* pour les mêmes niveaux d'agrégation et le principe reste le même. Par exemple pour analyser les effets « année » et « statut de protection » sur la richesse spécifique des **poissons** :



iii. Résultats

Si deux facteurs sont sélectionnés, un *interaction plot* est produit, qui peut servir de support visuel à l'interprétation des comparaisons multiples :



Attention : Un interaction plot n'est pas lié à un modèle, il s'agit d'une représentation graphique des données, ici des moyennes, pour chaque combinaison de niveaux des facteurs sélectionnés.

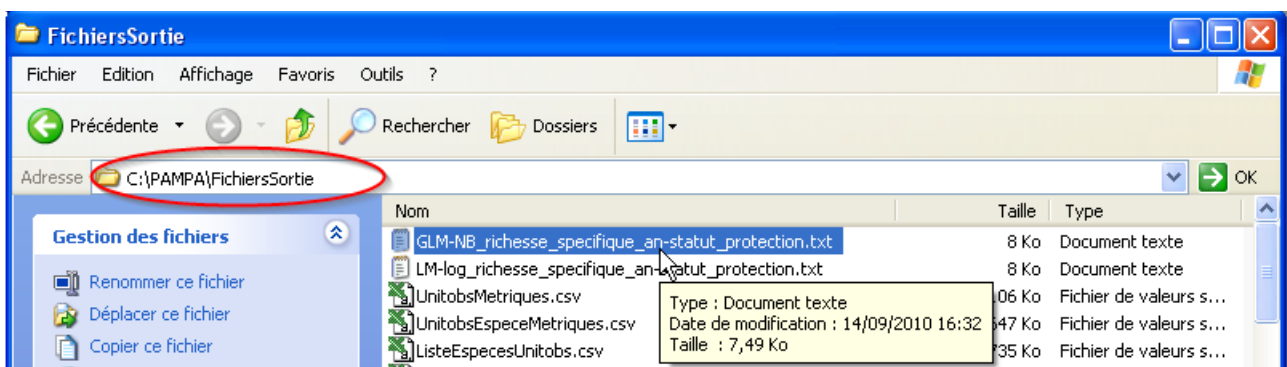
Les résultats de l'analyse sont stockés dans le dossier C:/PAMPA/FichiersSortie/ dans un fichier texte de la forme

<préfixe>_<métrique>_[<facteur de séparation/de sélection>(<modalité>)]<facteur 1>[-<facteur 2>...].txt

(les parties entre [] sont optionnelles).

Avec comme préfixe :

- LM pour le modèle linéaire simple (ou ANOVA),
- LM-log pour le modèle linéaire sur données log-transformées,
- GLM-NB pour le glm avec la distribution binomiale négative.
- GLM-Ga pour le glm avec la distribution Gamma.



Ce fichier contient plusieurs parties détaillées dans les sous-sections suivantes.

Informations sur le modèle

Modèle ajusté :

```
glm.nb(formula = richesse_specifique ~ an * statut_protection,  
data = tmpDataMod, init.theta = 280.5474544, link = log)
```

(ici un glm avec la distribution binomiale négative, sur la richesse spécifique, avec l'année et le statut de protection comme facteurs).

Dans le cas d'un modèle linéaire, les statistiques globales du modèle sont également affichées dans cette partie :

Modèle ajusté :

```
lm(formula = log(biomasse) ~ an * statut_protection, data = Data)
```

Statistique de Fisher Globale et R^2 :

R^2 multiple : 0.0826 ; R^2 ajusté : 0.0585

F-statistique : 3.430 sur 9 et 343 DL, P-valeur : 0.0004646

détails sur les facteurs significatifs et leurs coefficients

Table d'analyse de la déviance :

Modèle : Binomiale négative(280.5475), lien : log

Réponse : richesse_specifique

Termes ajoutés séquentiellement (premier au dernier)

	Df	Deviance	Resid. Df	Resid. Dev	P(> Chi)
NULL			374	1499.16	
an	9	1044.86	365	454.30	< 2.2e-16 ***
statut_protection	2	17.98	363	436.33	0.0001249 ***
an:statut_protection	18	88.36	345	347.97	2.839e-11 ***

Significativités des paramètres

(seuls ceux correspondant à des facteurs/intéractions significatifs sont représentés) :

	z value	Pr(> z)
(Intercept)	10.0449	< 2.2e-16 ***
an2000	5.1418	2.721e-07 ***
an2001	0.9085	0.3635972
an2002	2.7556	0.0058577 **

```

an2003                3.7057 0.0002108 ***
an2004                3.4696 0.0005213 ***
an2005                3.4386 0.0005848 ***
an2006                3.0279 0.0024624 **
an2007                3.1885 0.0014303 **
an2008                7.3832 1.545e-13 ***
statut_protectionPP   6.2847 3.286e-10 ***
statut_protectionRE   5.6507 1.598e-08 ***
an2000:statut_protectionPP -4.8547 1.205e-06 ***
an2001:statut_protectionPP -4.8689 1.122e-06 ***
...

```

Valeurs prédites

Valeurs prédites par le modèle :

```

1999:HR  1999:PP  1999:RE  2000:HR  2000:PP  2000:RE  2001:HR  ...
8.00000  32.00000  26.73333  24.36364  30.50000  36.35294  9.80000  ...
2003:PP  2003:RE  2004:HR  2004:PP  2004:RE  2005:HR  2005:PP  ...
14.41667  17.76190  17.33333  19.50000  17.16667  17.00000  15.88889  ...
2007:RE  2008:HR  2008:PP  2008:RE
15.27778  38.20000  33.55556  39.44444

```

Valeur prédite par le modèle pour chaque combinaison des modalités des facteurs explicatifs de l'analyse. À part pour les ANOVAs sur données log-transformées (Log-LM) – pour lesquelles elles sont données dans l'échelle logarithmique – ces valeurs prédites le sont dans l'échelle d'observation des données (*i.e.* non-transformées).

Comparaisons multiples (2 facteurs)

Lorsque deux facteurs sont sélectionnés :

Comparaisons multiples :

Attention : les estimations de différences sont sur les logarithmes :
 $(\log(A) - \log(B))$

Comparaisons pour les différences spatiales (statut de protection) par année :

Hypothèses linéaires :

	Estimate	Std. Error	z	value	Pr(> z)
1999 : PP - HR == 0	1.386294	0.220584	6.285	<1e-04	***
2000 : PP - HR == 0	0.224635	0.092732	2.422	0.336	
2001 : PP - HR == 0	0.135514	0.131672	1.029	1.000	
2002 : PP - HR == 0	0.145634	0.108530	1.342	0.990	

```

2003 : PP - HR == 0 -0.234257 0.112120 -2.089 0.613
...

Comparaisons pour les différences temporelles par statut de protection :

Hypothèses linéaires :
      Estimate Std. Error z value Pr(>|z|)
HR : 2008 - 1999 == 0 1.56339 0.21175 7.383 < 0.001 ***
HR : 2008 - 2007 == 0 0.87025 0.07992 10.889 < 0.001 ***
HR : 2007 - 2006 == 0 0.03390 0.09462 0.358 1.00000
HR : 2006 - 2005 == 0 -0.09453 0.09873 -0.957 0.99994
HR : 2005 - 2004 == 0 -0.01942 0.10957 -0.177 1.00000
...
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(P-valeurs ajustées -- méthode 'single-step')

```

Ces résultats indiquent par exemple une richesse spécifique significativement supérieure ($p < 0.0001$) dans la zone de protection partielle (PP) par rapport à l'extérieur de la réserve (HR) en 1999. Et une augmentation significative hors réserve sur la période 1999-2008.

Les facteurs temporels sont traités d'une façon différente de tous les autres. Les comparaisons temporelles ne sont faites qu'entre le début et la fin de la série (e.g. 2008 - 1999) et entre deux « pas de temps » successifs (e.g. 2007 - 2006, 2006 - 2005 mais pas 2007 - 2005).

Sont considérés comme facteurs temporels :

- le champ « an »
- le champ « annee.campagne » qui correspond à « caractéristique_2 », renommé, s'il suit le format suivant : "C<année sur quatre chiffre>" (e.g. "C1998").

Lorsque le modèle est un GLM ou un LM sur données log-transformées, un avertissement est affiché au début des comparaisons multiples, qui précise dans quel espace de transformation sont estimées les différences. Dans le cas des GLMs de la famille Gamma, la fonction de lien est l'inverse :

```

Attention : les différences sont estimées dans la fonction de lien
(inverse) :
      (1/A) - (1/B)  =>  *inversion du signe des différences*

```

Dans ce cas (et uniquement celui-là), le signe des différences est changé par rapport aux différences dans l'espace des observations.

Par exemple :

```

Hypothèses linéaires :
              Estimate Std. Error z value Pr(>|z|)
RE - HR == 0  -0.2864      0.1186  -2.415   0.0157 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(P-valeurs ajustées -- méthode 'single-step')

```

signifie que la métrique est significativement plus élevée en réserve (RE) que hors réserve (HR).

Comparaisons multiples (1 facteur)

Lorsqu'un seul facteur est sélectionné, des comparaisons sont faites

- entre modalités successives et entre état final et initial si le facteur est temporel (voire encadré ci-dessus).
- entre toutes les paires de modalités sinon.

```

-----
Comparaisons des modalités :
    Attention : les estimations de différences sont sur les
logarithmes :
    (log(A) - log(B))

Facteur 'caracteristique 2' (temporel) :

Hypothèses linéaires :
              Estimate Std. Error z value Pr(>|z|)
C2008 - C1999 == 0  0.38385      0.05323   7.212  <0.001 ***
C2008 - C2007 == 0  0.90064      0.05105  17.641  <0.001 ***
C2007 - C2006 == 0 -0.01081      0.05926  -0.182    1.000
C2006 - C2005 == 0 -0.07685      0.05815  -1.322    0.810
C2005 - C2004 == 0 -0.08583      0.06832  -1.256    0.847
C2004 - C2003 == 0  0.14008      0.06733   2.081    0.280
C2003 - C2002 == 0  0.04923      0.05376   0.916    0.971
C2002 - C2001 == 0  0.07912      0.06025   1.313    0.815
C2001 - C2000 == 0 -0.99037      0.06098 -16.241  <0.001 ***
C2000 - C1999 == 0  0.37864      0.05891   6.427  <0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(P-valeurs ajustées -- méthode 'single-step')

```

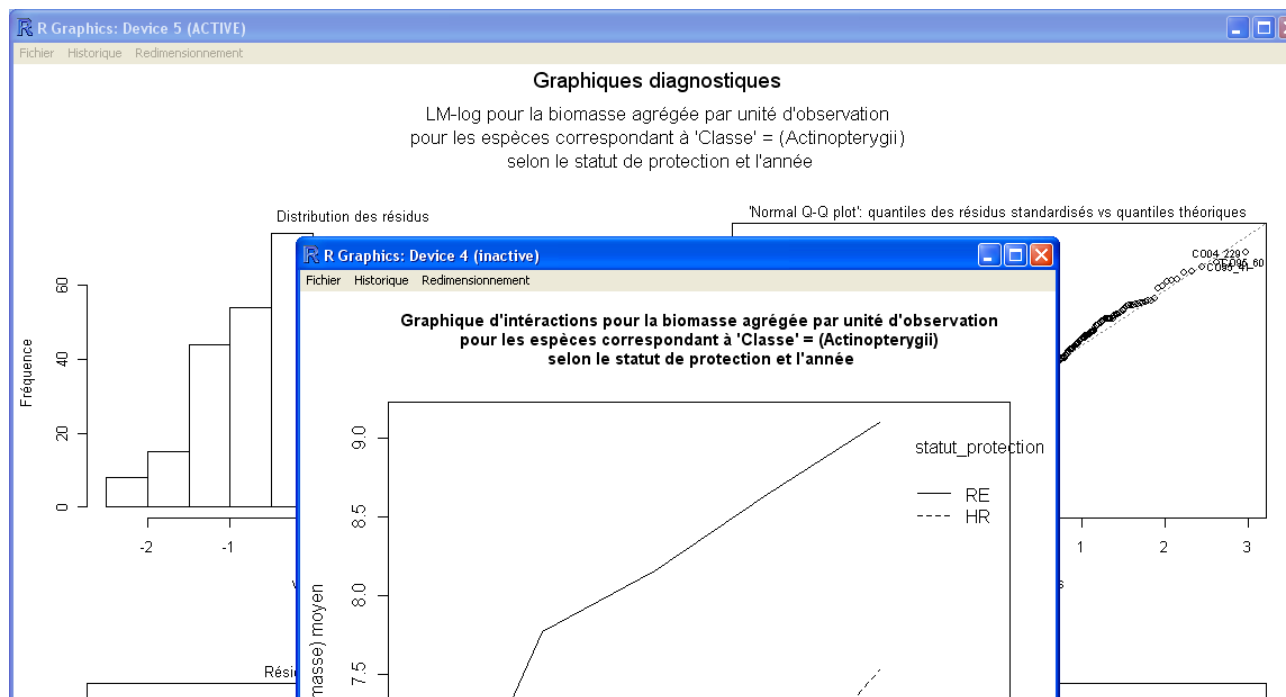
Dans cette exemple, la métrique apparaît, par exemple, significativement plus élevée pour l'année de campagne 2000 que pour l'année de campagne 1999.

iv. Graphiques diagnostiques et valeurs aberrantes

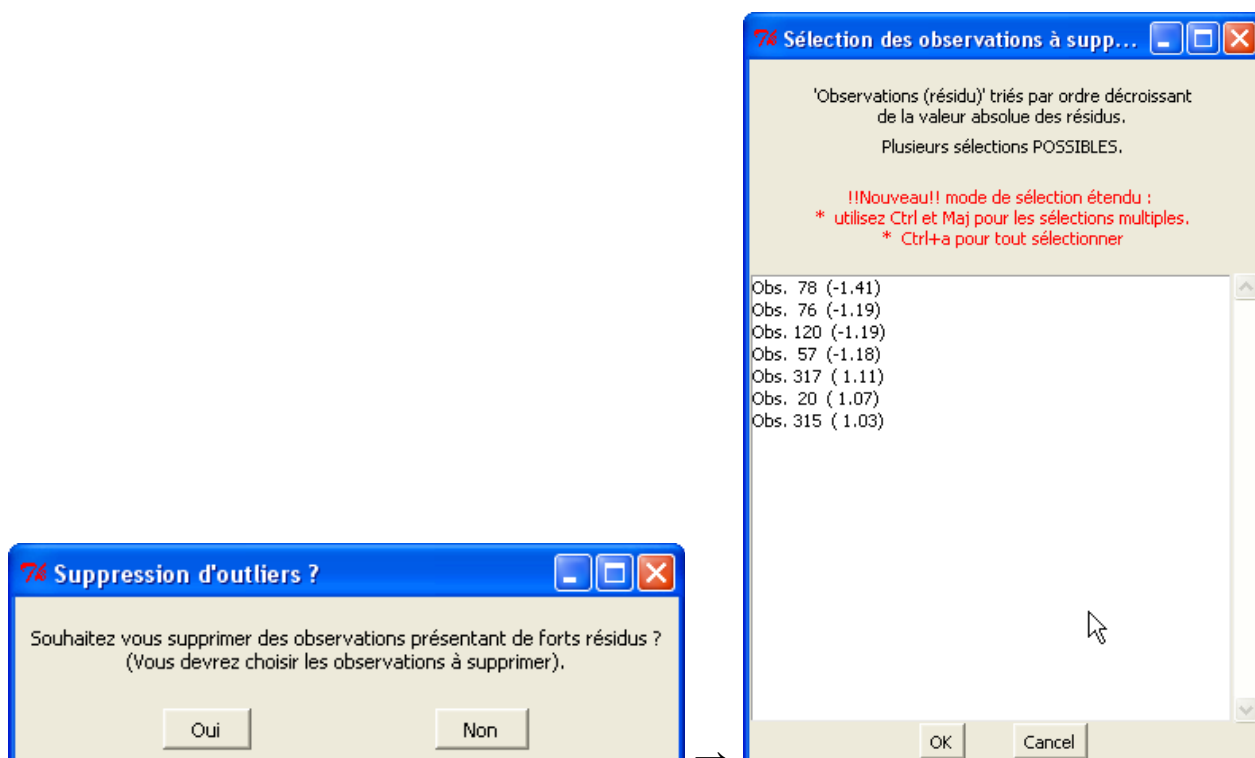
Des graphiques diagnostiques sont créés à la fin de chaque analyses. Ils permettent de tester la qualité de l'ajustement et la pertinence du modèle utilisé.

Ces graphiques sont particulièrement utiles avec les modèles linéaires simples (avec données log-transformées ou non).

Ces graphiques présentent maintenant des titres explicites qui permettent de ne pas les confondre lorsque plusieurs analyses sont lancées simultanément :



Si le jeu de données contient des observations à écarter des analyses, il est possible relancer les mêmes analyses en supprimant les observations provoquant les plus forts résidus :



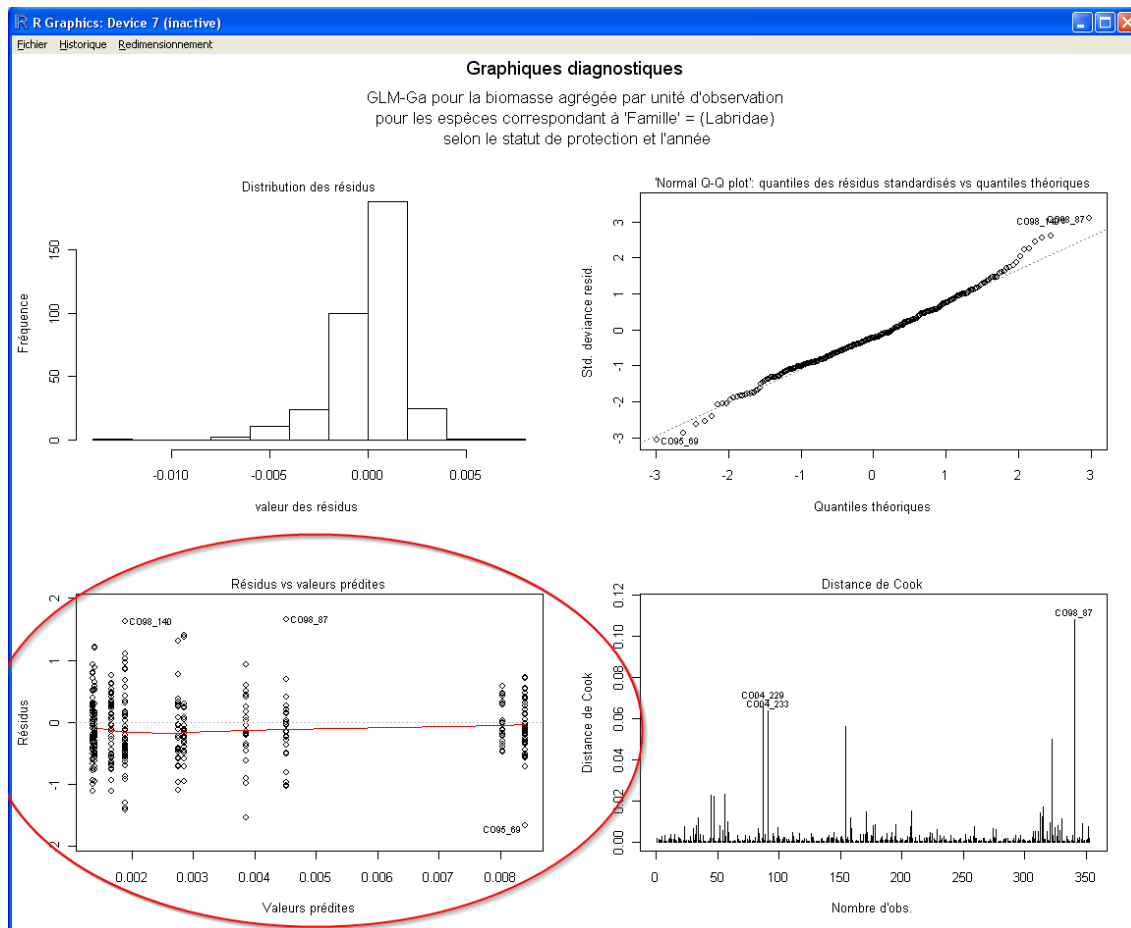
Le fichier de résultats obtenu après suppression des observations aura le même nom, plus un suffixe « _(red) » (pour « réduit »).

GLMs et graphiques diagnostiques

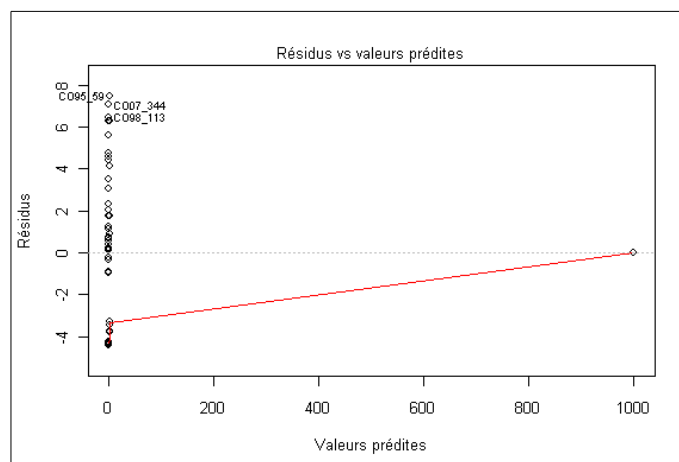
Pour les GLMs, la normalité des résidus – évaluée à l'aide du « Normal Q-Q plot » – en particulier, n'est pas nécessairement attendue. Ne pas en tenir compte.

Dans ce cas, le graphique qui présente le plus d'intérêt pour évaluer la qualité d'ajustement du modèle est celui des résidus en fonction des valeurs prédites. Si l'ajustement est bon, il ne doit pas y avoir de tendance.

Voici par exemple un ajustement correct :



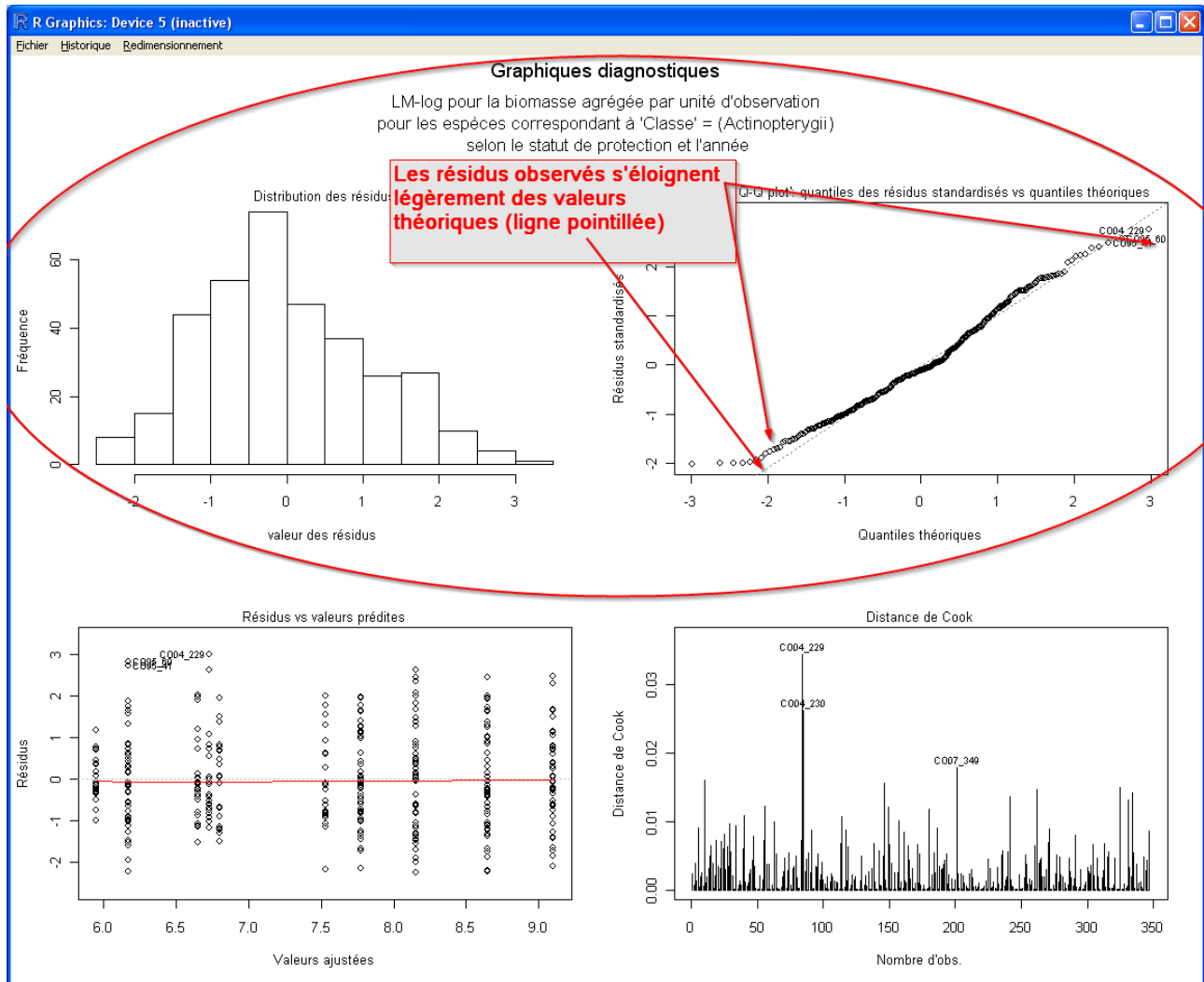
Tandis que le suivant est mauvais :



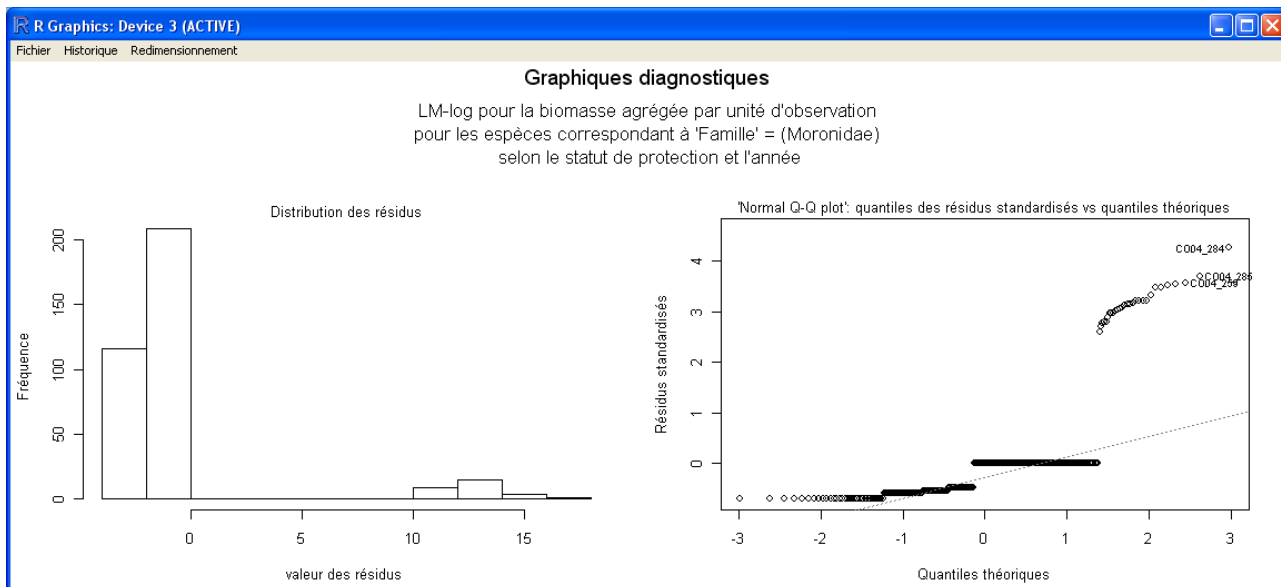
(Log-)LMs et graphiques diagnostiques

Ici aussi, l'absence de tendance dans la représentation des résidus en fonction des valeurs prédites constitue un critère de bon ajustement. Mais à cela vient s'ajouter un critère de normalité des résidus du modèle, sans laquelle les p-valeurs risquent d'être mal estimées.

Dans l'exemple ci-dessous – un ajustement « somme-toute correct » – la distribution des résidus est légèrement dissymétrique, comme le montrent les deux graphiques de la première ligne :



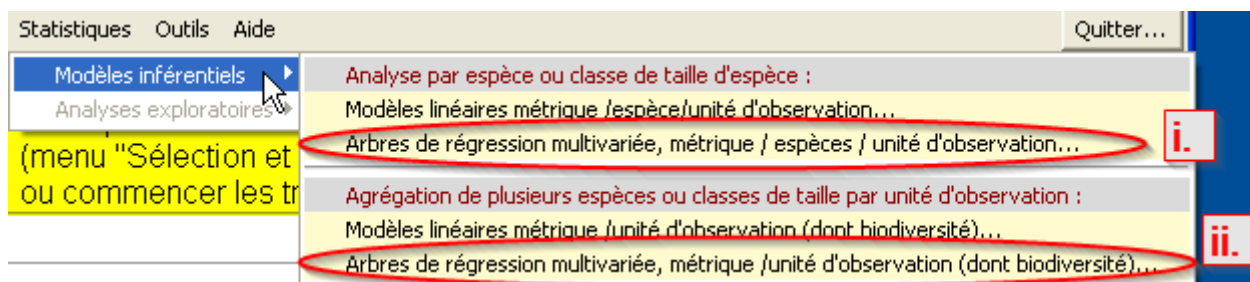
Le modèle est acceptable dans ce cas là ; rien d'aussi mauvais que dans l'exemple suivant (distribution bimodale) :



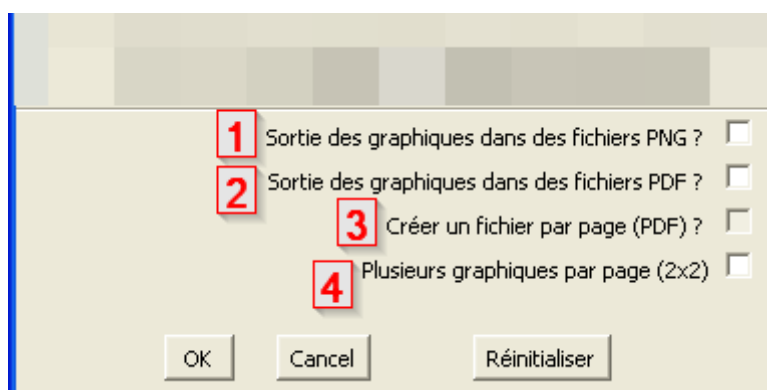
Ce dernier modèle ne pourrait être validé !

B. Arbres de régression multivariée

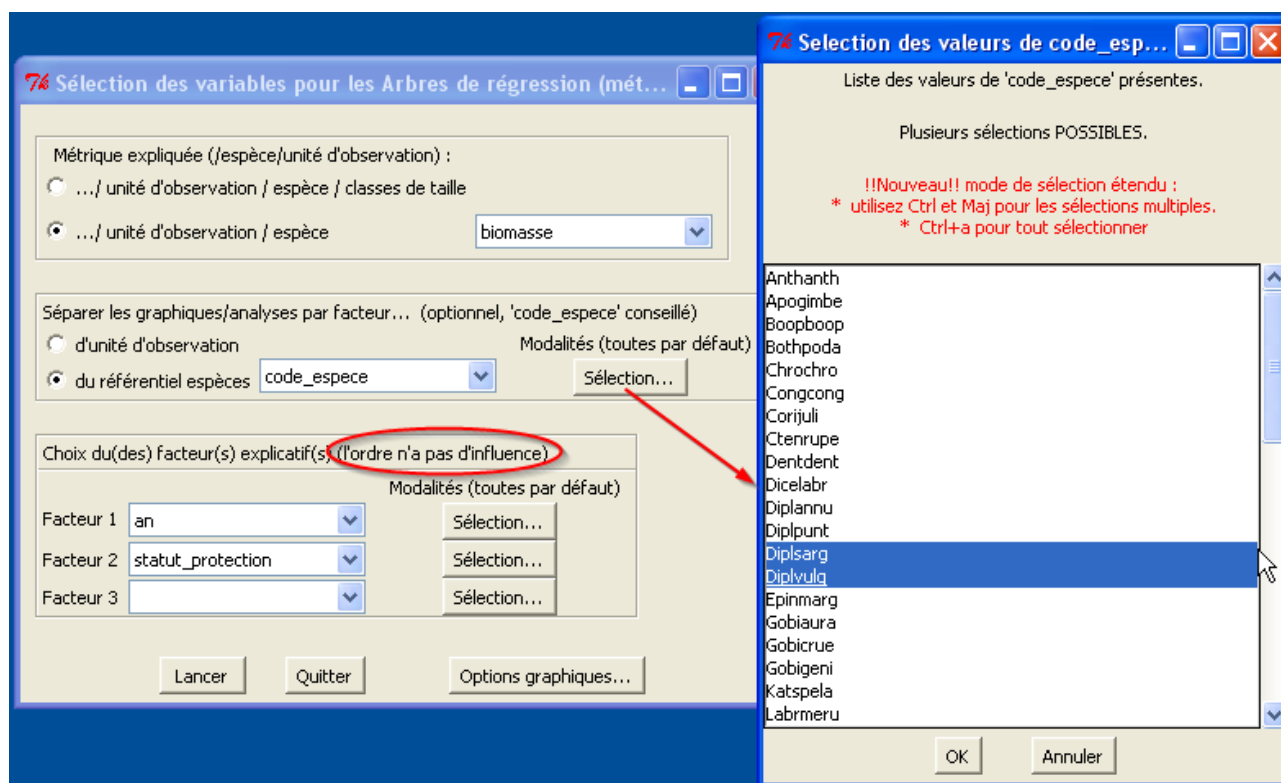
Les arbres de régression multivariés entrent dans la catégorie des modèles inférentiels



mais produisent à la fois des sorties graphiques et des fichiers textes. Les options graphiques seront donc accessibles dans l'interface standard de sélection des métriques et facteurs. Cependant seules les quatre dernières (se rapportant aux périphériques graphiques ; type de fichier, etc.) sont utiles pour ces graphiques.



i. Métriques agrégées par unité d'observation par espèce (et éventuellement par classe de taille)



À quelques détails près (textes d'information), l'interface et son fonctionnement sont similaires à ceux des autres traitements sur les métriques agrégées par espèce. Et en particulier, leur principal intérêt est – ici aussi – de permettre de lancer en une fois plusieurs analyses/graphiques, séparés selon le facteur du second cadre.

Notez toutefois qu'ici, l'ordre des facteurs n'a pas d'influence, contrairement aux précédents types de graphiques/analyses.

ii. Métriques agrégées par unité d'observation (et éventuellement par classe de taille)

De la même manière, l'interface et son fonctionnement sont similaires à ceux des autres traitements sur les métriques agrégées par unité d'observation (et éventuellement par classe de taille). Ces analyses peuvent également être menées sur des indices de diversité.

74 Sélection des variables pour les Arbres de régression (mét...

Métrique expliquée (agrégée / unité d'observation) :

☐ .../ unité d'observation / classes de taille
☐ .../ unité d'observation
☒ ...de biodiversité (/ unité d'observation)

richesse_specifique

Sélection d'espèce(s) selon un critère... (optionnel)

☐ d'unité d'observation
☒ du référentiel espèces

Modalités (toutes par défaut)

Sélection...

Choix du(des) facteur(s) explicatif(s) (l'ordre n'a pas d'influence)

Modalités (toutes par défaut)

Facteur 1 an

Facteur 2 statut_protection

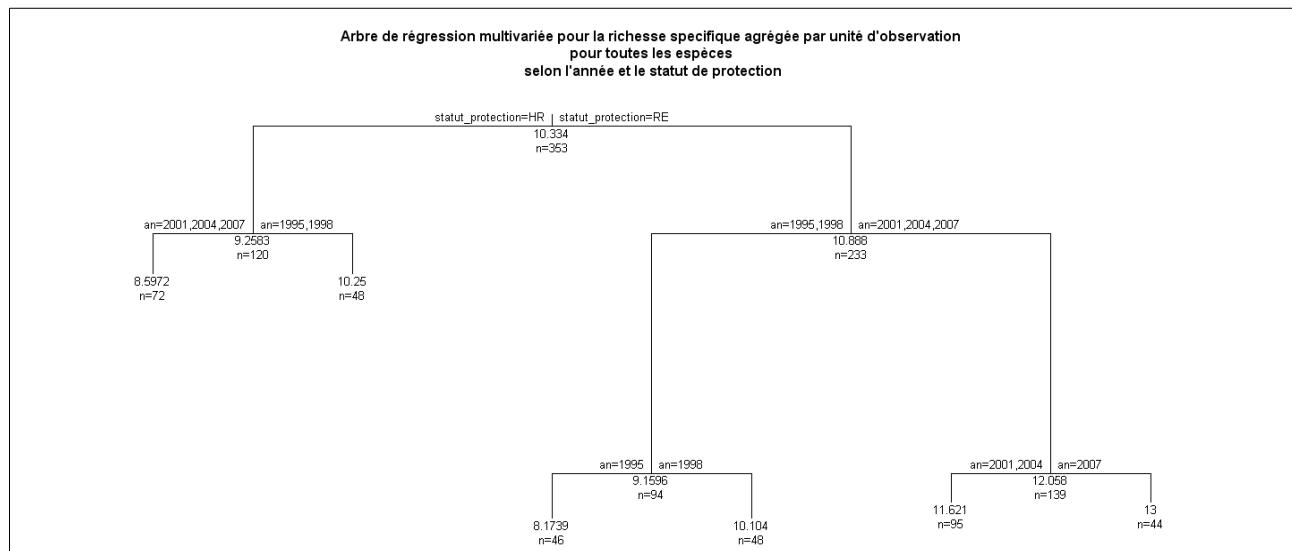
Facteur 3

Sélection... Sélection... Sélection...

Lancer Quitter Options graphiques...

iii. Résultats Graphiques

Le dernier exemple ci-dessus donne le graphique en arbre qui suit :



À chaque nœud est indiqué :

- en partie supérieure, les règles de dichotomie en sous groupes (le cas échéant).
- en partie inférieure :
 - la moyenne de la métrique dans le groupe correspondant.
 - n=le nombre d'« individus » (nombre de valeurs agrégées par unité d'observation et éventuellement espèce et/ou classe de taille).

Si une sortie sous forme de fichier a été sélectionnée dans les options graphiques, le nom

de fichier est construit selon les mêmes règles que pour les *boxplots* ou *barplots*, à la différence du préfixe, ici « MRT_ » (*Multivariate regression tree* : acronyme anglais pour arbre de régression multivariée).

iv. Résultat « texte »

Les résultats de l'analyse sont stockés dans le dossier C:/PAMPA/FichiersSortie/ dans un fichier texte de la forme

```
MRT_<métrique>_[<facteur de séparation/de sélection>(<modalité(s)>)]<facteur 1>[-<facteur 2>...].txt
```

(les parties entre [] sont optionnelles).

Il se compose de trois parties :

Rappel du modèle :

```
Appel :  
rpart(formula = richesse_specifique ~ statut_protection + an,  
      data = tmpData)
```

Résumé de l'arbre sous forme de texte :

Résultat général :

n= 353

noeud), division, n, déviance, yval
* indique un noeud terminal

```
1) root 353 3714.5550 10.334280  
  2) statut_protection=HR 120 900.9917 9.258333  
    4) an=2001,2004,2007 72 475.3194 8.597222 *  
    5) an=1995,1998 48 347.0000 10.250000 *  
  3) statut_protection=RE 233 2603.0990 10.888410  
    6) an=1995,1998 94 958.6064 9.159574  
      12) an=1995 46 692.6087 8.173913 *  
      13) an=1998 48 178.4792 10.104170 *  
    7) an=2001,2004,2007 139 1173.5400 12.057550  
      14) an=2001,2004 95 720.3579 11.621050 *  
      15) an=2007 44 396.0000 13.000000 *
```

Celui-ci donne, pour chaque nœud, la déviance en plus des informations figurant sur le graphique.

Détails :

Détails :

Appel :

```
rpart(formula = richesse_specifique ~ statut_protection + an,  
      data = tmpData)  
n= 353
```

	CP	nsplit	rel error	xerror	xstd
1	0.09172264	0	1.0000000	1.0066223	0.07796905
2	0.02356097	2	0.8165547	0.8463447	0.06296443
3	0.02117945	3	0.7929938	0.8353084	0.06232848
4	0.01539395	4	0.7718143	0.8334635	0.06058923
5	0.01000000	5	0.7564204	0.8405566	0.06143505

Noeud #1: 353 observations, param de complexité=0.09172264
mean=10.33428, MSE=10.52282

fils Gauche=2 (120 obs) fils Droit=3 (233 obs)

Division initial:

statut_protection divisé en GD, improve=0.05665951, (0
manquant)

an divisé en GGGDD, improve=0.05569685, (0
manquant)

Noeud #2: 120 observations, param de complexité=0.02117945
mean=9.258333, MSE=7.508264

fils Gauche=4 (72 obs) fils Droit=5 (48 obs)

Division initial:

an divisé en DDGGG, improve=0.08731737, (0 manquant)

...