

Sortie marée POP

Cahier des charges des fichiers de données communs

Tout d'abord, **Que chacun note où il sera/était le jour de la sortie**; chaque année nous perdons un temps précieux à comprendre qui a travaillé où parce que plusieurs groupes nous annoncent avoir travaillé au même endroit tout simplement parce que cette info élémentaire n'avait pas été notée.

Vous avez trois informations à retenir pour caractériser votre transect sans ambiguïté : votre *espèce cible* (*Littorina littorea* ou *Thais lapillus*), le *mode* (abrité ou battu et votre *distance au chenal* que nous vous désignerons sur le terrain. Ces informations vous seront données sur le terrain, notez-les !

Le/la responsable de chaque groupe de marée sera... responsable du fait que les données demandée seront fournies à la communauté non seulement en temps et en heure mais sous le format exact demandé, ce qui permettra à tous de gagner ensuite énormément de temps dans la mise en commun et l'analyse des données de tous les groupes.

Vous devrez fournir deux fichiers, *au format texte* (.txt), format directement chargeable dans R. Dans le fichier de type « *quad* », chaque ligne décrira la position et le contenu d'un quadrat. Dans le fichier de type « *biom* », chaque ligne décrira la position et les données biométriques d'un individu, qui pourra avoir été capturé soit dans un quadrat (auquel cas les caractéristiques du quadrat seront les mêmes que dans le fichier « *quad* ») soit en chasse à vue (auquel cas les informations disponibles sur le lieu de capture seront évidemment moins précises que pour un quadrat).

Ces fichiers seront mis en ligne par mes soins sur ma page perso <http://perso.univ-rennes1.fr/denis.poinsot> dans l'onglet POP/sortie marée au fur et à mesure de leur réception (mais vous êtes évidemment libres d'organiser entre vous une mise en ligne parallèle sur f*****k ou autre dropbox).

Une fois rangés dans un répertoire, ils y seront rangés par ordre alphabétique : vous allez donc les nommer de manière standardisée pour permettre à chacun de trouver rapidement et sans erreur ce dont il a besoin.

Chaque fichier .txt aura un identifiant unique construit avec :

1. l'initiale majuscule **L** ou **T** selon que votre cible était *Littorina littorea* ou *Thais lapillus*

2. la minuscule **a** ou **b** selon que vous avez travaillé en mode abrité ou battu
3. un nombre indiquant la distance en mètres entre votre transect et le chenal.
4. la terminaison **quad** ou **biom** selon que votre fichier concerne le contenu des quadrats ou les mesures biométriques des individus.

Exemples de noms de fichiers valides :

Ta10quad.txt = groupe ayant travaillé sur **Thais** en mode **abrité**, à **10** mètres du chenal, le fichier décrit des **quadrats**.

Lb60biom.txt = groupe ayant travaillé sur **Littorina** en mode **battu**, à **60** mètres du chenal, le fichier décrit la **biométrie** des individus.

Exemples de noms de fichiers invalides :

Données Penvins.txt (strictement rien à voir avec ce qui est demandé !)

ta60mquad.txt (t en minuscules, m ajouté après le chiffre --> sera classé au mauvais endroit dans la *longue* liste alphabétique des fichiers, ce qui fera perdre du temps à tout le monde)

(...)

Mise en forme des données

Avant de voir le détail du contenu de votre fichier, voici les 28 intitulés des colonnes du fichier de type « quad » et les 32 intitulés de colonnes du fichier de type « biom » (les numéros des colonnes sont indiqués juste pour vous guider, ne les mettez pas dans votre fichier).

Fichier de type « quad »

1=group, 2=cible, 3=resp, 4=year, 5=month, 6=coef, 7=mode, 8=dist, 9=station, 10=alt, 11=surf, 12=p.roc, 13=p.moul, 14=p.huit, 15=p.bala, 16=p.alg, 17=p.eau, 18=Litlit, 19=Litsax, 20=Litobt, 21=Gibcin, 22=Gibsp, 23=Monlin, 24=Thalap, 25=Nasret, 26=Oceeri, 27=Patsp

Fichier de type « biom »

Mêmes colonnes 1 à 28 que fichier de type « quad » mais également : 28=capture, 29=haut, 30=larg, 31=peri

Ces intitulés sont à respecter absolument (distinction minuscule/majuscule comprise, pas d'accents) pour deux raisons : (1) Ils vont être vérifiés par un script R. Si un ou plusieurs intitulés diffère de l'attendu, le script échouera avec un message d'erreur. (2) vous allez avoir besoin de fusionner tous les fichiers de données en un seul avant de lancer vos analyses statistiques dans R. Si les intitulés des colonnes -- ou leur ordre -- n'est pas le même pour tous les fichiers, la fusion échouera aussi avec moult jurons et arrachages de cheveux pour trouver d'où vient l'erreur, puis d'autres jurons et cheveux perdus pour corriger la situation. Méfiez-vous particulièrement d'Excel, qui a une tendance fâcheuse à ajouter des majuscules en début de mot « à l'insu de votre plein gré ».

Voir le détail ci-dessous ; si vous suivez ce guide avec attention, personne ne sera blessé :

n°col	nom	classe R	unité	Contenu, exemple(s) corrects, #commentaires
1	group	factor	aucune	Le nom unique de votre groupe : équivalent au nom de fichier

				dont vous éliminez la partie « quad.txt » ou « biom.txt ». ex : La10 , Tb60
2	cible	factor	aucune	nom abrégé de l'espèce cible. Deux possibilités seulement : Litlit (pour <i>Littorina littorea</i>) ou Thalap (pour <i>Thais lapillus</i>) # respectez la majuscule au nom de genre
3	resp	factor	aucune	première partie de l'adresse mail étudiante du/de la responsable des données (la partie à gauche du @). ex : flore.dubois ; en cas de problème avec le fichier, vous pourrez donc demander à flore.dubois@etudiant.univ-rennes1.fr ce que c'est que ce travail de sagouin, non mais sans blagues ! # ne pas mettre la seconde partie de l'adresse mail puisque c'est la même pour tout le monde.
4	year	integer	année	L'année de prélèvement. Ici, 2016
5	month	factor	aucune	abréviation du mois de prélèvement. Ici : sept
6	coef	integer	arbitraire	coefficient de marée (108 le 19 sept, 111 le 20 sept)
7	mode	factor	aucune	mode d'exposition aux vagues (une lettre, minuscule). uniquement : a (pour abrité, côté Est à Penvins) ou b (pour battu, côté Ouest à Penvins) # ne pas écrire "abrité" ou "battu" en toutes lettres.
8	dist	integer	mètre	distance en mètre (arrondi au m) entre le départ de votre transect et le chenal. ex : 40 # ne pas écrire 40m : juste le chiffre.
9	station	integer	mètre	distance approximative (arrondie au mètre) mesurée le long de la corde matérialisant le transect, entre le lieu de prélèvement et le rivage à marée basse (point émergé le plus bas auquel vous aurez accès). ex : 75 # attention , le jour de la sortie vous mesurerez temporairement les distances par rapport au point le plus haut de votre transect (puisque vous ne pourrez pas accéder tout de suite au

				<p>point le plus bas). N'oubliez pas par la suite de mettre le vrai zéro au niveau de la marée basse.</p> <p># pour les chasses à vue (hors quadrat, donc), estimez cette distance à partir du centre de votre petite zone de chasse à vue et reportez cette valeur pour tous les animaux capturés.</p>
10	alt	numeric	mètre	<p>altitude estimée (arrondie à 0,1m) par rapport à l'altitude zéro de référence qui sera pour vous la mer à marée basse.</p> <p>ex : 1.2</p> <p># votre tableur excel va mettre des virgules. Pensez à les transformer en points décimaux dans le fichier txt (fonction chercher/remplacer tout) avant de le sauvegarder.</p> <p># pour les chasses à vue (hors quadrat), estimer l'altitude au milieu de la zone de chasse et appliquer cette valeur à tous les individus chassés dans cette zone.</p>
11	surf	numeric	mètre carré	<p>surface estimée en m2 (arrondie à 0.01m2) que vous avez réellement échantillonné dans votre quadrat, en tenant compte des bosses et fissures.</p> <p>ex : 0.35</p> <p># attention, le jour de la sortie vous estimerez cette surface en "équivalent quadrat". N'oubliez pas par la suite de convertir en m2, sachant que votre quadrat est un cercle de 50cm de diam.</p>
<p>pour les % de recouvrement qui suivent, considérer chaque « couche » comme indépendante des autres : la somme des % peut donc largement dépasser 100%</p>				
12	p.roc	integer	pourcent	<p>pourcentage estimé de substrat rocheux (arrondi au pourcent).</p> <p>ex : 95</p> <p># ne pas écrire 95% (juste le chiffre)</p> <p># ne pas considérer comme substrat rocheux du rocher recouvert de sable ou de débris.</p> <p># par contre, des moules, huîtres (sauf sous forme de débris), balanes ou algues ne modifient pas la nature rocheuse du substrat donc un rocher recouvert de moules ou d'algues reste</p>

				compté comme 100% rocheux. # pour les chasses à vue, estimer le pourcentage de recouvrement global en rocher dans la zone de chasse et l'appliquer à tous les individus chassés dans la zone.
13	p.moul	integer	pourcent	pourcentage de recouvrement du substrat par les moules ex : 25 # chasse à vue : idem que pour p.roc # ne pas écrire p.moule ou p.moules
14	p.huit	integer	pourcent	pourcentage de recouvrement du substrat par les huitres ex : 45 # chasse à vue : idem que pour p.roc # ne pas écrire p.huitres
15	p.bala	integer	pourcent	pourcentage de recouvrement du substrat et autres moules/huitres par les balanes ex : 75 # chasse à vue : idem que pour p.roc # ne pas écrire p.balanes
16	p.alg	integer	pourcent	pourcentage de recouvrement du substrat par les algues. ex : 45 # estimer ce pourcentage à partir de ce que vous voyez (pas de rocher visible = 100% de recouvrement) et non pas par rapport à la situation à marée haute (où les algues sont dressées). # chasse à vue : idem que pour p.roc # ne pas écrire p.algues
17	p.eau	integer	pourcent	pourcentage de recouvrement du substrat par de l'eau (à marée basse, évidemment : il s'agit de déceler les flaques résiduelles). ex : 100 # chasse à vue : idem que pour p.roc
(...)				

Colonnes 18--27 :

attention aux noms : majuscule au nom de genre.

demander confirmation de vos premières identifications sur le terrain.

si vous ne prélevez pas l'espèce, attention à ne pas compter les morts : coquilles vides et autres bernards l'hermite, ces derniers étant très nombreux dans les coquilles de *Thais lapillus*, *Nassaria reticulata* et *Ocenebra erinacea* en particulier.

individus de chasse à vue (donc capturés hors quadrat) : noter NA dans ces colonnes 19 à 28.

18	Litlit	integer	individu	comptage exact de <i>Littorina littorea</i> , le bigorneau commun. (comptage effectué en TP si c'est votre espèce cible, estimation la plus précise possible sur le terrain si les individus sont très nombreux mais que ça n'est pas votre espèce cible)
19	Litsax	integer	individu	nombre de <i>Littorina saxatilis</i> , la littorine des rochers
20	Litobt	integer	individu	nombre de <i>Littorina obtusata</i> (= <i>Littorina littoralis</i>), la littorine obtuse (toute ronde)
21	Gibcin	integer	individu	nombre de <i>Gibbula cinerea</i> , la gibbule cendrée
22	Gibsp	integer	individu	nombre de <i>Gibbula pennanti</i> et <i>Gibbula umbilicalis</i> (pas évidentes à distinguer sur le terrain donc une seule catégorie ici)
23	Monlin	integer	individu	nombre de <i>Monodonta lineata</i> , la grande gibbule "à dent"
24	Thalap	integer	individu	Nombre exact de <i>Thais lapillus</i> (= <i>Nucella lapillus</i>), le pourpre
25	Nasret	integer	individu	nombre de <i>Nassaria reticulata</i> (= <i>Nassarius reticulatus</i>), la nasse réticulée (allure de grenade défensive)
26	Oceeri	integer	individu	nombre de <i>Ocenebra erinacea</i> (= <i>Ocenebra erinaceus</i>), le rocher (allure de Shar-pei)
27	Patsp	integer	individu	nombre de <i>Patella sp</i> (ne distinguez pas les espèces)

Fichier de type « biom », ajouter les colonnes suivantes.

28	capture	factor	aucune	Comment l'individu a été collecté : quad s'il était dans un quadrat, vue s'il a été collecté en chasse à vue. # écrire quad et non pas quadrat # pour les individus capturés en chasse à vue, mettre NA dans les colonnes 18 à 27 (ces données n'ont de sens que si un quadrat a été utilisé).
29	haut	numeric	millimètre	taille (arrondie au dixième de mm) de la hauteur de la coquille, mesurée de l'apex au point le plus bas (pour Thais, l'extrémité du siphon) ex : 20.3 # points décimaux et non virgules
30	larg	numeric	millimètre	taille (arrondie au dixième de mm) de la plus grande largeur de la coquille mesurée *perpendiculairement* à la hauteur. ex : 12.8 # points décimaux et non virgules # attention , bien procéder selon le dessin fait en TP sinon la mesure sera faussée
31	peri	numeric	millimètre	taille arrondie au dixième de mm de la largeur du péristome ex : 5.6 # points décimaux et non virgules # nom de colonne : pas d'accent (peri et non péri) # attention , bien procéder selon le dessin fait en TP sinon la mesure sera encore plus faussée que la précédente # mettre NA si l'individu est si petit qu'une mesure précise n'est pas possible.

Autres remarques en vrac :

en cas d'estimation, donnez bien un *nombre* estimé, pas une fourchette (ou si vous estimez une fourchette sur le terrain, donnez une moyenne arrondie à un individu près pour le fichier).

les estimations sont plus fiables lorsqu'elles sont faites à plusieurs.

surtout **pas** d'estimations qualitatives du type "+++" "**beaucoup**" etc.

ne confondez pas *G. cinerea* (stries très fines) avec *G. pennanti* et *G. umbilicalis* (demandez nous !).

même remarque pour *Monodonta lineata* (=Phorcus lineatus), grande Gibbule reconnaissable à sa "dent". Si vous n'en trouvez pas, c'est que vous n'avez pas réussi à les reconnaître donc.. demandez-nous !

Par respect envers les autres groupes, c'est-à-dire pour ne pas obliger *tous* les autres groupes à traquer puis corriger *vos* erreurs de mise en forme, **vous devrez impérativement tester vos fichiers avant de me les transmettre par mail** -- c'est en particulier à ça que va servir votre script de la Penvins Cup.

Pour cela, vous les chargerez dans R (réussir cette étape garantira déjà qu'il n'y a pas de bug majeur dans la mise en forme générale du fichier), puis vous les soumettrez à un script de vérification que vous élaborerez (ce qui garantira que les intitulés de colonnes sont bien les bons, que la nature des données - chaînes de caractères, nombres entiers, nombres décimaux -- sont bien ce qui est attendu, qu'il n'y a pas de valeurs numériques aberrantes dues à des erreurs de saisie ou d'unités etc.). Si l'une ou l'autre de ces étapes échoue, il ne vous restera plus qu'à corriger les erreurs jusqu'au succès de la procédure de vérification, avant de me transmettre vos fichiers. Ne sous-estimez pas le temps que ça va prendre, que vous le fassiez "manuellement" ou via un script de vérification.

Chargement de votre fichier dans R

Tout bête, basique, mais efficace :

- 1) copiez-collez votre fichier Excel/OpenOffice dans le bloc notes.
- 2) utilisez l'éditeur du bloc notes pour remplacer vos virgules par des points décimaux
- 3) sauvegardez au format .txt
- 4) ouvrez R
- 5) saisissez les deux lignes de commande suivantes :

```
mydata=read.table(file.choose(), header=TRUE)
```

Une fenêtre s'ouvre et vous pouvez naviguer "à la souris" jusqu'à votre fichier, cliquez dessus et validez. Votre fichier est maintenant chargé dans l'objet `mydata` (que vous auriez tout aussi bien pu appeler `supercallifragilisticexpidilicious` mais c'est moins pratique).

`summary(mydata)`

R vous affiche alors un résumé du contenu de votre fichier. En particulier, les moyennes sont calculées pour toutes les colonnes considérées comme numériques. Les autres colonnes sont considérées comme des facteurs donc R vous indique simplement l'effectif de chaque modalité rencontrée. Vous pourrez déjà à ce stade détecter facilement trois types d'erreurs :

1) les colonnes d'une **classe inattendue**. Typique des colonnes de chiffres (normalement de classe integer ou numeric) qui ont été considérées par R comme des facteurs parce qu'il y traîne encore une virgule (12,5) ou une unité (95%), ou un signe (>100) voire un commentaire... Vous repèrerez tout de suite l'anomalie en constatant que votre colonne de "chiffres" n'a permis de calculer aucun des paramètres statistiques de base (moyenne, min, max...) normalement affichés dans le summary pour ce type de variable.

2) les **modalités de facteurs inattendues** : si une "litlit" -- avec une minuscule -- s'est glissée au milieu de vos centaines de "Litlit" majuscules, cela apparaîtra comme une modalité supplémentaire de ce facteur, et R vous indiquera combien de cas de ce type il a trouvé dans la colonne)

2) les **valeurs aberrantes** résultant de virgules non saisies (examinez attentivement **les valeurs min et max** en particulier : une littorine de 125mm, ça n'existe pas, une altitude de 12m à la pointe de Penvins non plus)

Bien entendu, si certaines erreurs vous échappent, elles ne pourront déjouer l'oeil d'aigle du script R que vous allez bâtir dans le cadre de la Penvins cup !

--- fin du cahier des charges ---