Sortie marée PROJET

Cahier des charges des fichiers de données communs

Vous avez deux informations à retenir pour caractériser la position de votre transect sans ambigüité : le mode (abrité ou battu) et votre distance au chenal.

Le/la responsable de chaque groupe de marée sera... responsable du fait que les données demandées 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 trois fichiers, au format texte (.txt), format portable par excellence (et directement chargeable dans R pour ce qui concerne les données).

Dans le fichier .txt de type « quad », chaque ligne correspond à un quadrat. Elle décrira sa position, ses caractéristiques physiques et son contenu biologique (dont les effectifs par espèce).

Dans le fichier .txt de type « ind », chaque ligne correspond à un individu dont vous avez pris les mesures biométriques. Cette ligne commencera pas rappeler toutes les caractéristiques du cadrat dans lequel il a été capturé (=même information que le fichier quad !) mais elle sera complétée par son nom d'espèce abrégé, ses trois mesures biométriques et la présence éventuelle de cicatrices sur la coquille qui seraient des traces d'anciennes rencontres avec un prédateur.

Dans le fichier .txt de type « readme », vous donnerez aux futurs utilisateurs de vos données toutes les indications que vous jugerez utiles pour décrire les caractéristiques de votre transect que vous n'avez pas pu faire rentrer dans le jeu de données. Par exemple, son allure générale, la présence de failles, les difficultés rencontrées, le manque de temps (la marée monte !) ou les erreurs faites qui pourraient expliquer en particulier des données manquantes dans votre fichier. La première information à donner dans ce fichier est bien entendu l'adresse mail du/de la responsable du groupe.

La mise en ligne « officielle » de vos données sous forme .txt sera faite par mon intermédiaire, sur moodle voire directement par mail (donc, c'est à moi qu'il faudra envoyer vos trois fichiers .txt, en utilisant un titre de mail explicite, comme vu plus loin). Vous êtes cependant bien entendu libres

d'organiser entre vous une mise en ligne parallèle sur f******k ou autre dropbox, sous quelque forme que ce soit (excel, word, .txt, ça vous regarde).

Une fois rangés dans un répertoire, ces fichiers y seront rangés automatiquement 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 très simple construit avec :

1. la minuscule a ou b selon que vous avez travaillé en mode abrité ou battu
2. un nombre indiquant la distance en mètres entre votre transect et le chenal.
4. la terminaison quad, ind ou readme selon que votre fichier concerne le contenu des quadrats, les mesures biométriques des individus ou les informations libres sur votre transect et les éventuels problèmes rencontrés.

Exemples de noms de fichiers valides :
---
a10quad.txt = groupe ayant travaillé en mode abrité, à 10 mètres du chenal, le fichier décrit des quadrats.
b60ind.txt = groupe ayant travaillé en mode battu, à 60 mètres du chenal, le fichier décrit les individus.
a110readme.txt = groupe ayant travaillé en mode abrité, à 110 mètres du chenal, le fichier contient des commentaires à lire sur le transect, les difficultés éventuelles etc.

Exemples de noms de fichiers invalides :

Données Penvins.txt (strictement rien à voir avec ce qui est demandé !)
A60mquad.txt (A en majuscule, 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 N intitulés des colonnes du fichier de type « quad » et les N+5 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=transect, 2=resp, 3=date, 4=coef, 5=mode, 6=d.chenal, 7=d.mer, 8=alt, 9=surf, 10=p.roc, 11=p.moul, 12=p.huit, 13=p.bala, 14=p.alg, 15=p.encr, 16=p.eau, 17=s.fla, 18=d.fla

19 et suivantes : liste des espèces qui sera déterminée quand tout le monde aura fini le TP1 d'identification, classées dans l'ordre alphabétique, codées par 6 lettres dont une majuscule : trois premières lettres du nom de genre et trois premières lettres du nom spécifique ou bien sp. si pas de nom spécifique. Exemple : Littorina littorea = Litlit, Patella sp. = Patsp.

Fichier de type « ind »

Mêmes colonnes 1 à N que fichier de type « quad » mais également : N+1=sp, N+2=haut, N+3=larg, N+4=peri, N+5=pred , N+6=coul, N+7=text, N+8=masse

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 par la fonction rbind(), 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	transect	factor	aucune	Le nom unique de votre groupe : équivalent au nom de fichier dont vous éliminez la partie « quad.txt » ou « ind.txt ». ex : a10, b60
2	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.
3	date	factor	aucune	La gestion des dates est délicate dans R. Nous allons contourner la difficulté et vous écrirez simplement la date « en clair » mais sans espaces. ex : 14juillet1789
4	coef	integer	arbitraire	coefficient de marée (97 le 20 sept, 99 le 21 sept)
5	mode	factor	aucune	mode d'exposition aux vagues (une lettre, minuscule). ex : a # ne pas écrire "abrité" ou "battu" en toutes lettres.
6	d.chenal	integer	mètre	distance en mètre entre le départ de votre transect et le chenal. ex : 40 # ne pas écrire 40m : juste le chiffre.
7	d.mer	integer	mètre	distance approximative (arrondie au mètre) mesurée le long de la corde matérialisant le transect, entre le lieu de prélèvement et la mer à marée basse. ex : 75 # attention, le jour de la sortie vous avez mesuré temporairement les distances par rapport au point le plus haut

				de votre transect (puisque vous ne pouviez pas accéder tout de suite au point le plus bas). N'oubliez pas maintenant de mettre le vrai zéro au niveau de la marée basse, c'est le seul repère qui soit identique pour tous les groupes.
8	alt	numeric	mètre	altitude estimée (arrondie à 0,1m) par rapport à l'altitude zéro de référence qui sera pour vous la mer à marée basse. ex : 1.2 # attention, votre tableur excel va mettre des virgules. Vous connaissez maintenant la manœuvre qui permet d'en tenir compte.
9	surf	numeric	mètre carré	surface estimée en m2 (arrondie à 0.01m2) que vous avez réellement échantillonné dans votre quadrat, en tenant compte des bosses et fissures. ex : 0.35 # attention, le jour de la sortie vous estimerez cette surface en "équivalent quadrat". N'oubliez pas maintenant de convertir en m2, sachant que votre quadrat est un cercle de 50cm de diam.
pour les % de recouvrement qui suivent, considérer chaque « couche » comme indépendante des autres : la somme des % peut donc largement dépasser 100%				
10	p.roc	integer	pourcent	pourcentage estimé de substrat rocheux (arrondi au pourcent). ex : 95 # ne pas écrire 95% (juste le chiffre) # ne pas considérer comme substrat rocheux du rocher recouvert de sable ou de débris. # 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 compté comme 100% rocheux.
11	p.moul	integer	pourcent	pourcentage de recouvrement du substrat par les moules ex : 25 # ne pas écrire p.moule ou p.moules
12	p.huit	integer	pourcent	pourcentage de recouvrement du substrat par les huîtres

				ex : 45 # ne pas écrire p.huît ni p.huitres
13	p.bala	integer	pourcent	pourcentage de recouvrement du subtrat et autres moules/huitres par les balanes ex : 75 # ne pas écrire p.balanes
14	p.alg	integer	pourcent	pourcentage total 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). # ne pas écrire p.algues
15	p.encr	integer	pourcent	pourcentage de recouvrement spécifique par les algues encroûtantes ex : 5
16	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 # les flaques échantillonnées volontairement seront reconnues sans ambiguïté au fait qu'il y a des données de surface et de distance à la flaque la plus proche.
17	s.flaq	numeric	mètre carré	surface estimée de la flaque échantillonnée volontairement, en m2 (arrondie à 0.01m2) ex : 0.25 # pour les flaques échantillonnées aléatoirement, vous n'avez (en principe) pas relevé cette information, donc écrivez NA # attention, le jour de la sortie vous avez estimé cette surface en "équivalent quadrat". N'oubliez pas maintenant de convertir en m2, sachant que votre quadrat est un cercle de 50cm de diam.
18	d.flaq	numeric	mètre	distance estimée à la flaque la plus proche (arrondie à 0.1m près)

				ex : 2.5 # pour les flaques échantillonnées aléatoirement, vous n'avez (en principe) pas relevé cette information, donc écrivez NA

(…)

Colonnes 19—N (N sera connu quand la série des TP d'identification sera terminée) :

attention aux noms : majuscule au nom de genre.
attention à ne pas compter les morts : coquilles vides et autres bernards l'hermite, ces derniers étant très nombreux dans les coquilles de *Thais*, *Nasses* et *Ocenebra* en particulier.

Les espèces seront rangées par ordre alphabétique et codées sur 6 lettres : les trois premières lettres du nom de genre et les trois premières lettres du nom spécifique (ou bien « sp. » s'il est inconnu)
ex : Thais lapillus --> Thalap, Patella sp. --> Patsp.
Du fait que les espèces ont parfois de nombreux alias, nous vous fournirons évidemment une liste « officielle » des codes à utiliser lorsque la liste des espèces sera connue.

Pour les fichier de type « ind », il suffira de reproduire pour chaque individu les N colonnes caractérisant son quadrat (ce sont celles du fichier quad) et les cinq colonnes suivantes :

N+1	sp	factor	aucune	Le code à six lettres de l'espèce de l'individu. Ex : Thalap, Patsp.
N+2	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 les espèces à siphon comme Thais, nasses et Ocenebra, l'extrémité du siphon) ex : 20.3 # points décimaux et non virgules
N+3	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 schéma indiqué en TP sinon la mesure sera différente d'un groupe à l'autre
N+4	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.
N+5	pred	factor	aucune	Présence de « cicatrices » sur la coquille qui indiqueraient une tentative de prédation (cassures à nouveau scellées, trous ronds): "oui" ou "non"
N+6	coul	factor	aucune	"clair", "sombre" ou "rayures", cette dernière catégorie à utiliser pour les Thais comportant des bandes alternées noires et blanches
N+7	text	factor	texture	"lisse" (ex. Littorines), "rugueux" (ex. Nasses) ou "bosses" (Rochers)
N+8	masse	numeric	masse	Masse à 0,01 ou 0,02g près selon la balance utilisée

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.

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 toutes les vérifications nécessaires (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) sauvegardez au format .txt
4) ouvrez R
5) saisissez la ligne de commande suivante :

```
Mydata <- read.table(file.choose(), header=TRUE, dec=",")
```

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 supercallifragilistiquexpidélilicieux 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 un double point décimal (..) 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 20m à la pointe de Penvins non plus.

Attention, si certaines erreurs vous échappent, elles ne pourront déjouer l'oeil d'aigle du script R que les masters vont bâtir dans le cadre de la Penvins cup, et vous aurez à craindre la colère du peuple.

Quand envoyer vos fichiers ?
Avant le lundi 23 octobre 2017 à minuit.

Comment les envoyer ?
Par mail à denis.poinsot@univ-rennes1.fr, avec un titre de mail simple mais *explicite* :
Exemple de bon titre : UE projet : voici les données du groupe a60

Et si notre script de la Penvins cup n'est pas entièrement fini à ce moment-là ?
Aucun problème. En fait il est CERTAIN qu'il ne sera pas encore fini, et ça n'est pas grave du tout car sur vos petits fichiers personnels, vous pouvez facilement faire les vérifications nécessaires « manuellement ». Votre script de la Penvins cup donnera toute sa mesure lorsque vous devrez re-vérifier massivement tous les fichiers des autres groupes (ce qui ne sera pas du luxe, je vous le garantis), et vous pourrez l'utiliser pour cela bien avant qu'il soit achevé puisqu'il sera très probablement composé de fonctions qui sont autant de modules indépendants. La date de remise de la Penvins cup sera bien plus tardive que la remise des fichiers, pour que vous ayez amplement le temps de tester les performances de votre script face à de vrais fichiers, et donc corriger ses lacunes éventuelles.

Bon courage à tous !

--- fin du cahier des charges ---