

Exercices d'initiation à la manipulation de données avec R

Gilles San Martin - Centre Wallon de Recherche Agronomique (CRA-W)

14 November 2015

Contents

Exercices de prise en main de R	2
Exercices de manipulation de formats de données	3
Exercice 1	3
Exercice 2	4
Exercices d'importation de données	5
Exercices subsidiaires (pas de R)	5
Exercices sur l'extraction de données (Subscripting)	7
Exercices de manipulation de caractères	8
Exercice 1	8
Exercice 2	10
Structures de contrôle : fonctions, boucles, exécutions conditionnelles,...	11
Exercice 0	11
Exercice 1	11
Exercice 2	12
Exercice 3	12
Exercice 4	12
Exercice 5	13
Agrégation et reshaping	14
Exercice 1	14
Exercice 2	15

Exercices de prise en main de R

Le but de cette série d'exercices est de se familiariser avec l'utilisation basique de R, d'acquérir le réflexe de chercher dans l'aide et de comprendre et utiliser la vectorisation. On vous demandera donc par moment d'utiliser des fonctions qui n'ont pas été vues dans la partie théorique. Pas de panique, cherchez simplement dans l'aide...

- 1) Calculer les racines carrées arrondies à 2 décimales de tous les nombres de 1 à 100 (fonctions `round` et `sqrt`)
- 2) Calculer la surface des cercles de rayon variant de 0 à 250 cm par pas de 10 cm (26 cercles donc) (fonction `seq`)
- 3) Ecrivez les phrases suivantes : “Un cercle de 0 cm de rayon a une surface de 0 m²”, “Un cercle de 10 cm de rayon a une surface de 0.03 m²”, etc... pour tous les cercles de rayon variant de 0 à 250 cm par pas de 10 cm (fonction `paste`)
- 4) Tracer un graphique montrant la relation entre le périmètre et la surface d'un cercle. Ajoutez un titre explicite au graphique
- 5) Utilisez les 3 vecteurs ci-dessous : “jour”, “mois”, “année” pour créer un vecteur nommé “date” sous la forme 28/9/2012. (fonction `paste`) Essayez de comprendre comment sont construits ces 3 vecteurs (regardez l'aide des fonctions utilisées: `floor`, `runif`, `set.seed`).

```
set.seed(123)
jour <- floor(runif(30, 1, 31))
mois <- floor(runif(30, 1, 12))
année <- floor(runif(30, 1900, 2012))
```

- 6) Créez 2 variables aléatoires de distribution normale (fonction `rnorm`) ayant chacune une moyenne différente. Réalisez ensuite un test t de student pour comparer les moyennes de ces deux variables (cherchez dans l'aide la fonction adéquate...). Faites un boxplot horizontal de ces deux variables sur le même graphique. Vous devez pour ce faire d'abord les coller côte à côte dans un même objet avec la fonction `cbind` (par exemple `cbind(a,b)`)
- 7) Créez une variable aléatoire avec une distribution de Poisson (une fonction similaire à `rnorm` et `runif`). Tracez un histogramme de cette variable de couleur rouge.
- 8) Voici deux vecteurs représentant le nombre de mâles et de femelles observées dans une population et dans 5 catégories de couleurs : noir, jaune, rouge, bleu, blanc. Tracez un barplot représentant côte à côte les valeurs des mâles et des femelles pour chaque couleur (et avec le nom des couleurs dans les étiquettes de l'axe x). Pour ce faire commencez par coller les 2 vecteurs de manière à former une matrice à 2 lignes et 5 colonnes (fonction `rbind`). Ajoutez les 5 couleurs dans les noms de colonnes (fonction `colnames`) et ensuite, tracez le graphique.

```
males <- c(3, 5, 7, 9, 10)
femelles <- c(0, 2, 5, 14, 16)
```

Exercices de manipulation de formats de données

Exercice 1

L'exercice suivant n'est pas très réaliste. Le but est de vous faire manipuler dans des cas simples les formats de données. Voici 2 vecteurs de 10 nombres a (10, 20, 30, etc) et b (5, 5, 5, etc).

```
a <- seq(10,100,10)
b <- rep(c(5,10), c(7,3))
```

- 1) Faites la somme de chaque vecteur (somme a = 550, somme b = 65)
- 2) Concaténez ces deux vecteurs l'un à la suite de l'autre de façon à obtenir un vecteur de longueur 20 et dont la somme vaut 615
- 3) Collez les nombres de ces vecteurs a et b 2 à 2 (pour obtenir 105, 205, 305 etc) et sauvez le résultat dans un nouveau vecteur "ab". Ce vecteur a une longueur de 10 éléments. Quel est son mode ?
- 4) Faites la somme de ce nouveau vecteur (réponse = 29865).
- 5) Faites un graphique de a en fonction de b (plot(a~b))
- 6) Transformez le vecteur b en facteur et sauvez ce résultat dans un nouveau vecteur appelé "f"
- 7) Faites un graphique de a en fonction de f
- 8) Faites la somme de f (ça ne fonctionne pas...)
- 9) Retransformez f en numérique pour pouvoir en faire la somme (vous devrez passer par un format intermédiaire avant de le transformer en numérique). Vous devriez obtenir le même résultat que sum(b) soit 65 (et pas 13).
- 10) A partir du vecteur a, créez une matrice de 5 colonnes et 2 lignes avec sur la première ligne les valeurs 10, 20, 30, etc... (et pas 10, 30, 50 etc...). Sauvez cette matrice dans un objet appelé "mat".
- 11) Ajoutez comme nom de colonnes col1, col2, col3 etc... à cette matrice.
- 12) Au moyen de l'opérateur \$ essayez d'extraire la colonne 2 (mat\$col2). Que faut-il faire pour que cela fonctionne ? (dans quel format faut-il transformer mat?)
- 13) Créez un objet à 2 colonnes et 3 lignes contenant dans la première colonne les valeurs A, B et C et dans la deuxième colonne 3, 6 et 9.
- 14) Faites la somme de la deuxième colonne de l'objet ainsi créé

Exercice 2

Dans cet exercice plus réaliste, vous devrez à de nombreuses reprises passer d'un format de données à un autre (dates, character, numeric,...) pour pouvoir effectuer les opérations demandées.

Chargez le fichier "biche.txt" avec la commande suivante : `d <- read.table("biche.txt")` après avoir défini votre répertoire de travail où se trouve le fichier "biche.txt" avec la fonction `setwd`, par exemple : `setwd("/home/gilles/exercices")` (NB : vous devez utiliser des slashes "/" et pas de backslashes "\" dans le chemin du répertoire).

Ce fichier contient la date, l'heure et les points gps (coordonnées x y) de position d'une biche. Les colonnes `date0`, `time0`, `xt0`, `yt0`, contiennent pour chaque ligne les positions et temps de l'observation (point GPS) précédente. L'objectif général est de déterminer la période de l'année (saison) pour chaque point et de calculer la vitesse de déplacement entre deux points successifs.

- 1) Visualisez le contenu de votre jeu de données avec `summary(d)`, `head(d)` et `str(d)`. Identifiez le type de chaque variable.
- 2) Passez les colonnes `time` et `time0` en format date-temps (fonction `strptime`). Vous devrez au préalable coller la date et le temps dans une même variable (fonction `paste`). Utilisez la fonction `summary` pour visualiser la transformation du type de données
- 3) Créez une colonne pour l'année et le mois (par exemple avec la fonction `strftime`). Pour rappel, vous pouvez ajouter une colonne "mois" au dataframe "d" par exemple comme ceci : `d$mois <- c(1, 3, 5, 7)` Utilisez la fonction `head` pour visualiser les premières lignes du jeu de données
- 4) Sur base de votre colonne "mois", créez une colonne "saison" divisée en 3 périodes : nourrissage de janvier à avril (inclus), "été" de mai à septembre (inclus) et "chasse" de octobre à décembre. Vous pouvez utiliser la fonction `cut` pour ce faire. Faites un `summary()` de cette nouvelle colonne. Vous devriez avoir 2192 données en période de nourrissage, 969 en été et 1265 en période de chasse.
- 5) Calculez le temps écoulé entre 2 points gps successifs (2 lignes successives). Lorsque vous faites la différence entre deux heures (classe `POSIXct`), Vous obtenez une variable au format "difftime". Vous pouvez voir les unités utilisées au moyen de la fonction `units`. Voir l'aide pour `difftime` pour plus d'infos. Les 3 premières valeurs sont : NA, 3.800833 et 7.999167 heures.
- 6) Calculez pareillement la distance parcourue. Les coordonnées x et y sont des lambert belges (coordonnées projetées) en mètres. Pour Rappel, le théorème de Pythagore pour un triangle rectangle : $\text{Hypoténuse}^2 = \text{base}^2 + \text{hauteur}^2$ Les 3 premières valeurs sont NA, 0.149308406 et 0.286062930 km
- 7) Calculez la vitesse en km/h entre deux points successifs (attentions aux unités). Quelle est la vitesse maximale ? (réponse : 3.312303 km/h et pas NA ! Regardez dans l'aide en cas de besoin...)

Exercices d'importation de données

Le but de ces exercices est de se familiariser avec les cas les plus fréquents d'importation de données dans R (données dans un tableur ou dans un fichier texte). Les premiers exercices sont assez simples mais on a progressivement et volontairement introduit de nombreux pièges. L'objectif n'est point de jouer du plaisir sadique de vous voir patager mais bien de vous confronter aux problèmes fréquents que l'on rencontre pour importer ses données. Quand vous devez importer un fichier texte, ouvrez le d'abord dans un bon éditeur de texte (par exemple avec Rstudio) et regardez comment il est structuré (est-ce qu'il y a des entêtes de colonne, des valeurs manquantes, comment sont séparés les champs, quels sont les séparateur de décimales, etc. ...).

Pour les fichiers xls, exportez les en format texte en respectant les recommandations vues dans la partie théorique. Après l'importation visualisez systématiquement votre jeu de données avec `summary()` et en imprimant le jeu de données.

Les deux derniers exercices sont subsidiaires et ne sont pas à proprement parler des exercices de R, l'essentiel du travail devant se faire dans le tableur. Ils montrent des fichiers Excel bien structurés et clairs mais assez mal adaptés à l'analyse et à l'importation dans R (ou tout autre logiciel). Le but est de montrer qu'il vaut mieux structurer ses fichiers de données brutes en réfléchissant au traitement des données.

- 1) Importez le fichier `data1.txt`. Calculez la moyenne de la variable "haraxy" (réponse : 5.85)
- 2) Importez le fichier `data2.txt`. Faites la moyenne de la variable "adabip" (=31.77) Utilisez la fonction `plot` pour faire un graphique de adabip en fonction de site : `plot(adabip ~ site, data=mydata)`. Vous obtenez un boxplot. Faites de même pour adabip en fonction de date : `plot(adabip ~ date, data=mydata)`. Vous n'obtenez pas un boxplot parce que date n'est pas dans le bon format. Transformez la date en facteur de façon à ce que : `plot(adabip ~ date, data=mydata)` vous donne un boxplot.
- 3) Importez le fichier `data2bis.csv`. Avec la ligne de code suivante (où "mydata" est le nom de votre jeu de données), calculez la moyenne d'adabip pour le site "l'île de Niverlée". Vous devez obtenir 33.35.
`mean(mydata[mydata$site == "L'île de Niverlée", "adabip"])`.
Si ça ne fonctionne pas, n'oubliez pas de regarder le jeu de données que vous avez importé pour essayer de comprendre la source du problème.
- 4) Importez le fichier `data3.xls`. Calculez la moyenne de adabip (=31.77) et adadec (=10.42)
- 5) Importez le fichier `data4.xls`. Avec la fonction `mean`, calculez la moyenne de calqua (=7.19) et oencon (=3.95) (NB : Inspiré de faits réels. ...)
- 6) Importez le fichier `data5.xls` et vérifiez que l'importation s'est faite correctement. Si l'importation ne s'est pas faite correctement, essayez d'importer le fichier sans la dernière colonne et essayez de comprendre l'origine du problème

Exercices subsidiaires (pas de R)

- 7) Importez les données contenues dans le fichier "insectes_feuilles.xls". Calculez la proportion de feuilles F5 où les insectes étaient présents tous traitements et toutes variétés confondus NB : l'essentiel du travail se fait ici dans le tableur
L'essentiel du travail consiste ici à réorganiser le fichier xls dans le tableur pour qu'il soit exploitable. Il faut éliminer les sous-totaux, fusionner les deux feuilles de calcul, rajouter une colonne "traitement" et une colonne "variété. Cette manière d'encoder les données n'est pas idéale pour le traitement. Il vaut mieux en général encoder les données brutes et faire les statistiques descriptives à part (par exemple avec des tableaux croisés dynamiques/pilotes de données) Lors de l'importation dans R on prendra garde au fait que les cases vides correspondent à des données manquantes (argument `na.string` de `read.table`).
- 8) Importez les données contenues dans le fichier "météo.xlsx" issu d'un site internet. Calculez la température minimale moyenne et la quantité totale de précipitations sur la période. NB : l'essentiel du travail se fait ici dans le tableur
Ici aussi il y a pas mal de boulot à faire dans le tableur pour pouvoir sauver ce fichier dans un format texte exploitable. Il faut défusionner les cellules fusionnées, ensuite faire un tri sur la date pour séparer les lignes contenant les heures de prise de mesure des mesures proprement-dites. Il faut éliminer les unités "°C" et

“mm” avec des chercher-remplacer. Attention dans certains cas (Excel 2007 ?) après un rechercher remplacer sur “°C”, il reste un caractère invisible qu’il faut également éliminer.

Exercices sur l'extraction de données (Subscripting)

Chargez le fichier “ladybirds.txt”. Il s'agit de données de comptage de coccinelles sur 3 espèces d'arbres (colonne “tree”) : pins (P), tilleuls (L cfr “Lime”), érables (M cfr “Mapple”) dans 3 types de paysages (landsc) : Urbain (U), Suburbain (S) et Rural (R), sur une série de sites et à 4 dates. Les colonnes après la quatrième correspondent aux espèces de coccinelles (nombre d'individus).

- 1) Sélectionnez toutes les colonnes sauf la 6, 9 et 12
- 2) Sélectionnez les 10 premières lignes et uniquement les colonnes correspondant aux espèces
- 3) Sélectionnez les données de l'espèce “haraxy” sur pin.
- 4) Ne gardez que les observations (lignes) où l'espèce “anaoce” était présente
- 5) Ne gardez que les observations (lignes) “rurales” où “anaoce” était présente
- 6) Sélectionnez les données des sites PS1, MU2, LS3, PU2, LR3, LR2, MR1
- 7) Sélectionnez toutes les données sauf celles de ces sites (PS1, MU2, LS3, PU2, LR3, LR2, MR1)
- 8) Sélectionnez les données sur Pins à la date 1 et sur feuillus (tilleuls et érables) pour toutes les dates sauf la date1
- 9) Sélectionnez les données collectées sur érable ou tilleul en milieu suburbain et pour les espèces haraxy, calqua, exoqua, harqua et aphobl
- 10) Sélectionnez 10 lignes aléatoirement (à l'aide de la fonction sample)
- 11) Sélectionnez une ligne sur 5 (à l'aide de la fonction seq)
- 12) Transformez les données d'abondance en données binaires (présence/absence). (n'écrasez pas le jeu de donnée)
- 13) Remplacez les données de plus de 150 individus par des valeurs manquantes
- 14) Remplacez les données d'abondance d'espèce du site PS1 à la date 3 par des valeurs manquantes
- 15) Multipliez les valeurs des sites sur pins par 5/4
- 16) Réordonnez le jeu de données de manière à ce qu'on ait d'abord la colonne date puis site puis landsc puis tree suivies des colonnes espèces
- 17) Ne gardez dans le jeu de données que les espèces pour lesquelles on a plus de 100 individus observés au total. (utilisez la fonction colSums)
- 18) Triez les lignes par tree, landsc et date (dans cet ordre). Vous devrez utiliser la fonction order.
- 19) Comment faire pour que la colonne landsc se classe dans cet ordre : U, S, R ? (utilisez la fonction factor et le paramètre labels= pour changer l'ordre des niveaux du facteur landsc. Ensuite réutilisez la même ligne de code que ci-dessus.)
- 20) Réordonnez colonnes du jeu de données de manière à ce que les espèces soient présentées de la plus abondante à la moins abondante (ie harqua, adabip, myroct, adadec, etc. . .)

Exercices de manipulation de caractères

Quelques rappels :

- les fonctions `grep` et `grepl` ont un argument `ignore.case` qui permet d'ignorer la casse des caractères quand on le désire
- dans les expressions régulières, si vous voulez rechercher par exemple un point “.” il faudra chercher “\.” car le point a une signification particulière dans la syntaxe des expressions régulières.

Exercice 1

Voici un jeu de données sur lequel on va travailler (espèce, nombre d'individus, site)

```
d <- data.frame(
  sp = c("COCCINELLA SEPTEMPUNCTATA", "Coccinella septempunctata",
        "Coccinella Septempunctata Brucki", "Adalia spp", "COCCINELLIDAE",
        "coccinella hieroglyphica", "ADALIA BIPUNCTATA REVIELERI",
        "Adalia bipunctata bipunctata", "Adalia decempunctata", "Formicidae",
        "Harmonia spp.", "Adalia bipunctata", "Coccinella quinquepunctata",
        "Coccinella sp.", "Myrmica speciosus", "NOMADA BISPINOSA",
        "NOMADA sp.", "Formicinae", "Apoidea", "Dinocampus coccinellae", "COLEOPTERA",
        "Aphaenogaster spinosa", "MYRMICA SP.", "Lasius niger", "Formicinae"),
  n = c(1, 5, 10, 33, 2, 1, 4, 6, 1, 3, 1, 1, 2, 1, 9, 3, 1, 1, 3, 2, 6, 1, 2, 1, 3),
  site = c("Mazée", "Namur", "Res. Nat de Matagne", "Treignes",
           "res nat du coupu tienne", "Vaucelles", "Rés. Nat. Roche Madou", "Vierves",
           "natoye", "Olloy", "Le Mesnil", "resnat du Chamousias", "Natoye", "Mazée",
           "Réserve Naturelle de la Haie Gabaux", "Dourbes", "Doische",
           "Oignies", "Mazée", "Vaucelles", "Rés. nat. de la Montagne de la Carrière",
           "Regniessart", "Vireux", "Haybes", "res nat Al Florée")
)
summary(d)
d
```

Rappels sur la nomenclature des noms d'espèces et la taxonomie.

Le nom scientifique d'une espèce est toujours composé du nom de genre suivi du nom d'espèce. Par exemple : *Adalia bipunctata*, *Adalia* étant le nom de genre, *bipunctata*, le nom d'espèce. Lorsqu'on ne sait pas identifier une espèce jusqu'à l'espèce, on indique en general le nom de genre suivi de “sp” ou “spp” (“species”). Par exemple *Adalia sp.* Lorsqu'il y a trois noms à la suite (pex *Adalia bipunctata bipunctata*), le 3ème désigne une sous-espèce. Lorsqu'on a un seul nom terminant par “inae”, “idae”, “oidea”, “era”,... il s'agit de niveaux taxonomiques supérieurs au nom de genre (sous-famille, famille, ordre,...)

- 1) Dans la colonne `sp`, la casse des noms scientifiques n'est pas uniforme. Mettez tous les caractères en majuscule.
- 2) Sélectionnez les données du genre *Adalia* (n=5)
- 3) Sélectionnez les données correspondant au genre *Coccinella* (attention aux parasites !). (n = 6) Comment faire pour éviter de sélectionner *Dinocampus coccinellae* (un parasite de coccinelles...) sans changer la casse des caractères ?
- 4) Sélectionnez les données des genres *Adalia* et *Coccinella* (n = 11)
- 5) Transformez le vecteur de noms scientifiques “sp” de manière à avoir la première lettre en majuscule et les suivantes en minuscule. Vous pouvez simplement utiliser ici la fonction `substring` (avec `paste`, `toupper` et `tolower`). Les expressions régulières ne sont nécessaires (il est aussi possible d'utiliser les expressions régulières mais c'est plus compliqué dans ce cas).

- 6) Sélectionnez uniquement les 6 données qui n'ont PAS été identifiées au moins jusqu'au genre (familles, sous-familles, etc...). Piste : il s'agit des éléments de la colonne "sp" contenant au moins une espace (utilisez `regexpr` ou `grepl`).
- 7) Sélectionnez les données qui ont été identifiées au niveau générique ou infra générique (genre espèces, sous-espèces). NB : ne cherchez pas midi à 14 heures. Il suffit normalement de rajouter ou de changer un seul caractère à la ligne de code précédente.
- 8) Créez un code composé des 3 premières lettres du genre suivies des 3 premières lettres de l'espèce. Pour les données identifiées au niveau supragénérique remplacez le code obtenu par une chaîne de caractère vide.
2 pistes :
 - a) utilisez une combinaison de `substring` et de `regexpr` pour trouver la position du premier caractère espace
 - b) utiliser `gsub` et les possibilités de capture des expressions régulières
- 9) Sélectionner les données qui ont été identifiées au niveau de la sous-espèce. (NB : utilisez `grep` ou `grepl` mais pas `gregexpr`)
- 10) Pour ces espèces identifiées au niveau de la sous-espèce, transformer le nom de manière à ne garder que le Genre et l'espèce. Utilisez `gsub` et les possibilités de capture des expressions régulières
- 11) Sélectionnez uniquement les données qui ont été identifiées au niveau générique mais pas spécifique (ie *Adalia* sp., *Adalia* spp, etc...). NB : il y en a 5, pas 7 !
- 12) Sélectionnez toutes les données observées dans des réserves naturelles. (NB il y en a 7, pas 9) Dans la colonne "site" les données observées dans des réserves naturelles sont systématiquement indiquées mais de manière peu uniformisée (res. nat, résnat, etc...)
- 13) Sélectionnez les données du genre *Adalia* observées dans des réserves naturelles (2 données). Il faut impérativement utiliser ici `grepl`, la version de `grep` qui retourne un vecteur logique

Exercice 2

Voici un vecteur de coordonnées géographiques en Degrés Minutes Secondes (DMS).

```
DMS <- c("50°45'58.32\"N", "50°51'12.73\"S", "4°32'5.66\"E", "3°34'53.66\"W",  
         "2°28'7.26\"O", "50°36'23.88\"", "50°36'23\"", "50°36'23.65\"' 'N (N2)",  
         "50°36'23.65\"' 'N (N2)", "50°36'07,26\"N", " 50 ° 45 ' 58,32 \" N")
```

On veut le transformer en degrés décimaux ($DD = D + M/60 + S/(60*60)$). Les 4 première coordonnées sont correctes et cohérentes. Les suivantes n'ont pas toutes été encodée de la même manière (NB : inspiré de faits réels...)

Utilisez les expressions régulières pour extraire les valeurs correspondant aux degrés, aux minutes, aux secondes et à l'hémisphère (par défaut on considère qu'on est dans l'hémisphère nord et à l'est du méridien de Greenwich).

Conseil : Commencez par transformer les 4 premières valeurs puis coplexifiez progressivement vos expressions pour prendre en compte tous les cas particuliers...

Vous devriez obtenir les valeurs suivantes :

```
c(50.7662, -50.853536, 4.534906, -3.581572, -2.468683, 50.606633, 50.606389, 50.606569, 50.606569, 50.602017, 50.7662)
```

Structures de contrôle : fonctions, boucles, exécutions conditionnelles,...

Exercice 0

Petits exercices simples de mise en jambes...

- 1) Créez une fonction “salut” qui écrit simplement “Bonjour”. Cette fonction n’a pas d’arguments : salut() écrira “Bonjour”
- 2) Ajoutez un argument “qui” pour pouvoir choisir un nom quelconque: salut(qui = “Gilles”) écrira “Bonjour Gilles”, salut(“Louis”) écrira “Bonjour Louis”
- 3) Ajoutez un argument “english” de type vrai/faux permettant de choisir la langue : salut(qui = “Gilles”, english = FALSE) écrira “Bonjour Gilles” et salut(“Philippe”, english = TRUE) écrira “Hello Philippe”
- 4) Au moyen d’une boucle for, calculez pour chaque ligne de la matrice suivante, la moyenne et l’écart type (sd). Stockez ces valeurs dans une nouvelle matrice à 2 colonnes et 10 lignes. NB : en pratique on ne devrait pas utiliser de boucle pour ce genre de calculs (on utiliserait rowMeans et apply).

```
mat <- matrix(c(1:40), 10, 4, byrow=TRUE)
```

- 5) Utilisation de la fonction merge. On a un dataset “obs” qui contient un nombre d’observations pour une série d’espèces (sp) identifiées par un code à 6 lettres.
On veut ajouter une colonne à ce dataset avec le nom scientifique complet correspondant. Les correspondances code - nom scientifiques se trouvent dans le dataset “tax”.
Attention certains codes de la table obs n’ont pas de correspondance dans “tax”. Veuillez malgré tout à ne pas perdre de données dans la table “obs”.

```
obs <- data.frame (
  sp = c("cocund", "cocsep", "cocsep", "scysut", "psyvig", "adabip", "rhychr", "adabip",
        "psyvig", "cocsep"),
  nbr = c(2,3,46,3,4,8,9,20,4,3)
)

tax <- data.frame(
  code = c("adabip", "cocsep", "cocund", "psyvig", "subvig", "epiarg"),
  taxon = c("Adalia bipunctata", "Coccinella septempunctata",
            "Coccinella undecimpunctata", "Psyllobora vigintiduopunctata",
            "Subcoccinella vigintiquatuorpunctata", "Epilachna argus")
)
```

Exercice 1

- Construisez une fonction qui crée un code composé des 3 premières lettres du nom de genre et des 3 premières lettres du nom d’espèce. L’exercice a déjà été fait précédemment il suffit ici de l’encapsuler dans une fonction. Pour rappel les lignes de code suivant permettaient de faire le travail (en dehors d’une fonction) :
`esp_position <- regexpr(" ", x)`
`code <- paste(substring(x, 1, 3), substring(x, esp_position + 1, esp_position + 3), sep="")`
N’oubliez pas de créer un vecteur exemple de noms scientifiques pour tester votre fonction.
- Modifiez ensuite cette fonction de façon à ce qu’on puisse choisir le nombre de lettres prises en compte
- Modifiez encore votre fonction de façon à ce que l’utilisateur puisse choisir de mettre toutes les lettres du code en majuscule
- Modifiez votre fonction de façon à ce que l’utilisateur puisse choisir de mettre toutes les lettres du code en majuscule ou en minuscule ou de pas y toucher

Exercice 2

- Chargez le fichier “ladybirds_aggr.txt”. Changez les valeurs des deux premières colonnes pour des valeurs plus explicites : tree : M = Mapple, L = Lime, P = Pine ; landsc : U = Urban, S = Suburban, R = Rural (le plus simple est d'utiliser la fonction `levels()` <-)
- Réordonnez le tableau selon l'espèce d'arbre (tree) et ensuite du paysage (landsc) et dans l'ordre inverse de l'ordre alphabétique.
- Tracez un barplot de la première ligne du tableau représentant l'abondance des différentes espèces de coccinelles sur Pins en milieu urbain. Ajoutez dans le titre l'espèce d'arbre et le paysage correspondant (paramètre “main” de la fonction plot).
- A l'aide d'une boucle “for” répétez ensuite le même graphique pour chaque ligne du tableau en adaptant les titres. Juste avant la boucle for placez la commande suivante : `par(mfrow = c(3,3), mar = c(4,3,3,1), las=2, cex = 0.7)` mfrow divise la fenêtre graphique en 9 parties, mar ajuste la taille des marges, las force les étiquettes à être perpendiculaires à l'axe et cex diminue la taille des caractères (on détaillera ces options plus loin dans la partie sur les graphiques).
Ne vous tracassez pas trop de la mise en forme des graphiques qui sera abordée plus loin.
- Faites ensuite un barplot pour une espèce de coccinelle en fonction des différentes combinaisons d'arbre et de paysage (on va travailler sur les colonnes, avec une bare par ligne du tableau de données). Vous devrez ajouter vous-même les étiquettes au moyen de l'argument “names.arg” de la fonction barplot. Pour créer ces étiquettes, collez les valeurs des deux premières colonnes et séparez les par un caractère “\n” qui représente un retour à la ligne.
- Faites ensuite un graphique similaire pour chacune des 12 espèces au moyen d'une boucle. Avant la boucle for, ajoutez la ligne de code suivante : `par(mfrow = c(4,3), mar = c(4.5,2,3,1), las=2, cex = 0.65)`
N'oubliez pas d'ajouter les titres appropriés.

Exercice 3

Chargez les fichiers biche1.csv (séparateur : tabulations) et ephemerides.csv (séparateur : point virgule).

Le fichier “biche” contient les dates, heures et localisations GPS d'une biche. Le fichier éphémérides contient les heures de coucher (sunrise) et de coucher de soleil (sunset) à Uccle.

On veut pour chaque position de la biche, déterminer si il faisait jour ou nuit au moment de l'enregistrement.

Il faut commencer par formater correctement les colonnes date, sunset et sunrise du fichier d'éphémérides et les colonnes date et time du fichier biche. Attention, les éphémérides sont en heure civile pour l'Europe centrale (tz=“CET”) alors que le fichier biche est en temps universel (tz=“UTC”). Afin que les comparaisons et calculs soient corrects, il faut donc bien indiquer le fuseau horaire au moyen de l'argument tz au moment où on formate les temps (avec strptime). Pour certaines opérations, R retourne un message d'avis quand les temps impliqués ne sont pas dans le même tz mais les comparaisons sont néanmoins correctes.

Fusionnez ensuite les deux tables (merge) et créez une colonne jour/nuit. Il y a 155 données de jour et 288 données de nuit dans ce fichier. Vous pouvez le vérifier par un summary() sur votre colonne jour/nuit après l'avoir transformée en facteur.

Exercice 4

Un des but de l'exercice précédent était d'utiliser la fonction merge. Cependant, il serait plus efficace dans ce cas précis de pouvoir déterminer la période de la journée sans passer par un fichier extérieur. C'est ce que permet de faire la fonction sunriset du package maptools qui calcule l'heure de coucher ou de lever de soleil sur base des coordonnées géographiques et de la date. La fonction crepuscule permet également de déterminer les heures de pénombre.

Ecrivez une fonction qui détermine pour n'importe quelle date-heure en un point précis si il s'agit du jour ou de la nuit. Vous devez utiliser les colonnes longitude et latitude et la méthode appelée “## S4 method for signature ‘matrix,POSIXct’” dans l'aide de sunriset. Attention, par défaut le résultat de sunriset est donné dans le fuseau horaire local. Si le fuseau horaire du vecteur date-temps reçu en argument par sunriset est bien spécifié, la valeur retournée par sunriset sera dans ce fuseau horaire. Il peut y avoir de légères différences entre les résultats de sunriset et le fichier ephemerides.csv, ce dernier étant calculé pour Uccle (il y a 6 différences). Vous pouvez utiliser les latitude et longitude de l'observatoire d'Uccle comme valeurs par défaut à votre fonction (long = 4.357886, lat = 50.79873)

Exercice 5

Au moyen d'une boucle, modifiez le code des deux exercices précédents de façon à répéter ces opérations pour les 10 fichiers `biche1.csv`, `biche2.csv`, ... Votre programme devrait fonctionner pour n'importe quel nombre de fichiers.

Utilisez la fonction `list.files` pour faire la liste des fichiers visés et la stocker dans un vecteur.

Utilisez ensuite ce vecteur pour lire successivement les différents fichiers dans une boucle et ajoutez leur la colonne `jour/nuit`.

Ajoutez en plus une colonne identifiant chaque biche (chaque fichier).

Vous pouvez stocker chaque jeu de données dans les éléments d'une liste.

Agrégation et reshaping

NB1 : dans ces exercices, on utilisera jamais de boucle explicite.

NB2 : par “créer une fonction à la volée”, on entend créer une fonction à usage unique qui n’est pas sauvée dans un objet. On utilise typiquement ce genre de fonctions dans les fonctions de la famille apply.

Exercice 1

Chargez le jeu de données “ladybirds.txt”. On a déjà utilisé à plusieurs reprises un jeu de données similaire. Les 4 premières colonnes contiennent : l’espèce d’arbre (tree : Pine, Mapple ou Lime), le paysage (landsc : Urban, Suburban, Rural), un code identifiant le site (site) et la période d’échantillonnage (date). Les colonnes suivantes contiennent des abondances de coccinelles identifiées par un code à 6 lettres.

- 1) faites le total du nombre d’individus pour chaque espèce et par site. Utilisez la fonction aggregate et sauvez le résultat dans un dataset appelé “site”. Gardez dans ce jeu de données les colonnes tree, landsc et site dont vous aurez besoin dans les exercices suivants.
- 2) Au moyen de la fonction apply, calculez le nombre moyen et l’écart-type de chaque espèce de coccinelles entre les sites (autrement dit : la moyenne et l’écart-type des colonnes espèces du dataset “site”).
NB : on pourrait utiliser ici la fonction colMeans pour la moyenne mais il n’y a pas d’équivalent pour l’écart-type.
haraxy : moyenne = 6.62, sd = 12.22
- 3) Utilisez les fonctions lapply et puis sapply pour obtenir le même résultat que la fonction précédente (la présentation peut être différente mais pas les chiffres). Le jeu de donnée étant un data.frame, on peut le traiter comme une liste dont les éléments sont les colonnes du data.frame, ce qui permet d’utiliser les fonctions sapply et lapply.
- 4) Au moyen de la fonction apply, calculez le nombre total de coccinelles (toutes espèces confondues) pour chaque site. (on veut donc faire la somme des lignes du dataset “site”)
NB1 : on devrait de préférence utiliser ici la fonction rowSums
NB2 : impossible ici d’utiliser sapply ou lapply car on travaille sur les lignes
- 5) Ecrivez une fonction qui calcule l’indice de diversité de Simpson d’un site sur base du nombre d’individus de chaque espèce. Cet indice se calcule comme 1 moins la somme du carré des abondances relatives de chaque espèce. Calculez ensuite cet indice de simpson pour chaque site. (indice pour le site LR1 = 0.76)
- 6) En partant du dataset “site”, calculez l’abondance moyenne et l’écart type pour chaque combinaison d’arbre et de paysage.
Vous devrez utiliser la fonction aggregate. Le résultat est un tableau à 9 lignes et 14 colonnes.
- 7) Faites la même opération que dans la question précédente mais avec un résultat arrondi à 2 décimales (vous pouvez le faire uniquement pour la moyenne). Essayez de répondre à cette question en utilisant les 2 approches différentes suivantes :
 - Appliquez simplement la fonction round au résultat de la fonction aggregate de la question précédente (moins les 2 premières colonnes).
 - Au lieu d’utiliser la fonction mean seule dans l’aggregate, créez à la volée une fonction qui calcule la moyenne et l’arrondi à 2 décimales.
- 8) Si vous le désirez, explorez les possibilités de la fonction cast du package reshape. Cette fonction permet de faire tout ce que aggregate et tapply font mais avec une syntaxe beaucoup plus simple.
Vous devrez commencer par passer le jeu de données en format “long” avec la fonction melt()

Exercice 2

Chargez le jeu de données “dff.csv” et visualisez sa structure. Il s’agit d’un jeu de données faunistiques (de 15000 lignes) typique des jeu de données récoltés par des naturalistes volontaires : espèces observée (sp), position géographique (utm1, utm5 : mailles de 1 et 5 km², x, y : coordonnées géographiques), date (dat2), la plante hôte (plante), le nombre d’individus (n) et diverses infos comme la méthode de capture (capt), le stade de développement (devl), le type de support (micr).

- 1) Combien a-t-on d’observations par espèce ?
Essayez d’arriver au même résultat avec la fonction table (la plus simple dans ce cas) et aggregate. Pour aggregate utilisez la fonction “length” pour compter le nombre d’observations.
(réponse : sp_01 : 1133 obs, etc)
- 2) Combien a-t-on de données par espèce ? On définira la donnée comme une combinaison unique de sp x utm1 x dat2.
Utilisez la fonction “unique” pour agréger les valeurs identiques suivie de table.
(réponse : sp_01 : 946 données etc)
- 3) Quels sont les 30 carrés utm de 5 km² (utm5) avec le plus d’espèces ? (réponse : 31UER775775 et 16 autres carrés avec 9 espèces)
- 4) Extrayez l’année, le mois et le jour de dat2 (utilisez simplement substring) et stockez les dans 3 nouvelles colonnes du jeu de données
- 5) Faites un graphique (par exemple barplot) de l’évolution du nombre total d’observations au cours des années (240 observations en 2011).
Faites de même un graphique montrant la phénologie de toutes les espèces confondues par mois (17 observations en Janvier).
- 6) Faites un graphique phénologique par décade (toutes espèces confondues). Vous pouvez par exemple utiliser la fonction cut pour subdiviser le vecteur “jours” en 3 groupes et coller ce nouveau vecteur avec le vecteur mois. (3ème décade de Mars : 55 données)
- 7) Faites un tableau espèces x décade. Appliquez la fonction barplot à chaque ligne de ce tableau (fonction apply).
Avant la ligne de code apply, insérez la commande suivante qui divise la fenêtre graphique en 9 zones et modifie les marges :
par(mfrow=c(3,3), mar = c(3,3,1,1))
NB : il n’y a pas de manière très directe d’ajouter un titre explicite aux graphiques en utilisant apply (en récupérant les noms de colonne). Dans ce cas une boucle peut être plus adaptée.
- 8) Calculez le nombres d’observations par année pour chaque espèce (sp_1 en 2011 : 94 données). Utilisez deux approches différentes pour obtenir les mêmes chiffres (mais présentés différemment) :
 - fonction table (le plus simple et le plus direct dans ce cas) : vous obtiendrez une table espèces x années à 9 lignes x 14 colonnes
 - fonction aggregate avec comme fonction d’agrégation “length”. Les mêmes valeurs seront dans un tableau à 3 colonnes.
- 9) De façon similaire à la question précédente, calculez le nombre maximum d’individus(max(n)) observés pour chaque année et chaque espèce avec les deux approches possibles. NB : Vous devrez utiliser tapply à la place de table ici.
- 10) Trouvez la date d’observation la plus précoce pour chaque année et chaque espèce. (= 215 en 1998 pour sp_1) Commencez par créer une colonne au format Date. Utilisez ensuite la fonction aggregate et min.
Transformez ensuite la colonne date du résultat de votre aggregate en nombre de jours depuis le premier janvier de chaque année. Vous pouvez utiliser pour ce faire la fonction “format” avec pour argument format “%j” à transformer ensuite en numérique (voir aide de strptime). Il n’est pas nécessaire de passer en format POSIX.
Réorganisez ensuite le jeu de données obtenu en un tableau espèces x années. Unstack ne fonctionne pas

bien ici à cause des valeurs manquantes pour certaines années. Vous pouvez utiliser soit la fonction `cast` du package `reshape`, soit la fonction `tapply` avec comme fonction d'agrégation la somme ou la moyenne.

Comme il n'y a qu'une seule valeur par combinaison d'espèce x année cette opération ne changera pas les valeurs et ne fera que réorganiser le jeu de données.

NB : Vous pouvez aussi essayer `tapply` à la place de `aggregate` dès le début mais le format `date` est alors directement transformé en numérique et les manipulations sont un peu plus compliquées. Pour pouvoir transformer cette table en format `date`, il faut d'abord transformer la table en `data.frame` avec `as.data.frame.matrix`.

Ensuite on peut transformer ces valeurs en appliquant `as.Date` à chaque colonne avec `lapply` et en spécifiant comme origine le 1er janvier 1970.

- 11) Ajoutez au jeu de données un vecteur de dates en nombre de jours depuis le premier janvier de cette année. (`format(x, format = "%j")`) Sur base de ce nouveau vecteur, trouvez les 5 dates les plus précoces par année et par espèce et calculez en la moyenne (pour chaque année et chaque espèce). Vous pouvez utiliser `aggregate` ou `tapply` au choix. Vous devrez construire une fonction personnalisée qui trie les dates, extrait les 5 premières et en calcule la moyenne (sans oublier de gérer les NA).
 - 12) Pour chaque espèce faites une liste des carrés de 5 km² (`utm5`) où cette espèce a été observée au moins 3 fois. Ensuite comptez les nombres de ces carrés pour chaque espèce. Commencez par faire une table `utm5 x sp`. Sauvez les noms de lignes (noms des `utm5`) dans un vecteur. Ensuite avec l'aide de la fonction `apply`, récupérez pour chaque espèce les noms des `utm5` pour lesquelles le nombre d'observations est > 3 (vous devrez construire une fonction à la volée). Le résultat sera une liste. Avec les fonctions `lapply` puis `sapply` comptez le nombre d'`utm5` pour chaque espèce. Visualisez les différences de format entre ces deux fonctions.
 - 13) Pour chaque espèce calculez le nombre de carrés `utm5` où l'espèce a été vue au moins 0, 1, 2, 3, ... fois. Appliquez simplement (avec `apply`) la fonction `table` à chaque colonne de votre tableau `utm5 x sp`
 - 14) Créez un tableau espèces de plantes x espèces d'insectes donnant le nombre d'observations pour chaque combinaison. Éliminez les combinaisons plante x espèce pour lesquelles on a seulement une observation (transformez les 1 en 0) et ensuite éliminez les plantes pour lesquelles on a aucune observations (pex en créant un vecteur logique sur base de `rowSums`). Combien y a-t-il de plantes hôtes par espèce sur base de ce tableau ? (131 plantes pour `sp_1`)
 - 15) Pour chaque espèce, établissez la liste des dix plantes hôtes les plus utilisées avec le nombre d'observations pour chaque plante. Vous devez d'abord transformer votre tableau plante x espèce en `data.frame`. Si votre tableau est encore au format "table" vous devez utiliser `as.data.frame.matrix`. Si il est déjà au format "matrix" (ça dépend des opérations faites sur votre table d'origine) vous pouvez utiliser juste `as.data.frame` ou `as.data.frame.matrix`. Ensuite vous pouvez utiliser `lapply` pour appliquer à chaque colonne du tableau plante x espèce une fonction personnalisée qui trie les abondances par ordre décroissant et extrait les 10 premières. Avec une simple fonction comme celle-là vous aurez les nombres d'observations mais vous aurez perdu les noms des lignes (plantes). Il faut donc ajouter un argument à la fonction qui reçoit les noms des plantes et les ajouter au sein de votre fonction personnalisée. Vous pouvez ensuite modifier votre fonction de manière à ajouter par exemple une colonne de % d'observations (`nb obs / sum (nb obs)`) et à la fin le nombre total d'observations et le nombre total d'espèces de plantes.
 - 16) On veut calculer les indices d'Ellenberg pour chaque espèce. Les indices d'Ellenberg sont des valeurs indicatrices (lumière, température, humidité, richesse du sol, ...) pour une série d'espèces de plantes. On calcule en général un indice synthétique pour un site en multipliant l'abondance des plantes du site par leurs indices d'Ellenberg puis en prenant la somme de ces valeurs divisée par l'abondance totale (une valeur par indice). Il s'agit donc de la moyenne des indices des plantes présentes sur le site pondérée par leur abondance. On peut calculer également ces indices pour les biotopes occupés par des insectes.
- Commencez par charger le fichier `Ellenberg.txt`, ensuite, fusionnez les noms de lignes de votre tableau plantes x espèces avec le dataset `Ellenberg` (`merge`) afin que les lignes de ce dataset `Ellenberg` correspondent parfaitement aux lignes de votre tableau plantes x espèces.

- Calculez ensuite les indices d'Ellenberg pour sp_1 uniquement. Vous pouvez multiplier la matrice d'Ellenberg par le vecteur sp_1, diviser par le nombre total d'observations pour sp_1 et faire la somme de chaque colonne. Attention cependant le nombre total d'observations pour sp_1 ne devrait prendre en compte que les observations sur les plantes pour lesquelles l'indice d'Ellenberg est disponible. Le dénominateur est donc différent pour chaque indice d'Ellenberg.
- Généralisez ensuite ces calculs pour toutes les espèces (ie les colonnes du tableau plantes x espèces) par exemple au moyen de la fonction sapply.