

Préparer des données

D. Chessel

Notes de cours cssb1

La fiche introduit un cours d'analyse des données en biologie évolutive animé sur des ordinateurs permettant l'usage de \mathbb{R} . On commence par quelques exemples d'installation de données. On utilisera ensuite des données disponibles dans les séries <http://pbil.univ-lyon1.fr/R/ppsb1.html> et suivantes.

Table des matières

1	Données - Méthodes - Procédures	2
2	galabiose : un exemple d'originalité radicale	2
3	cortes : l'espace comme donnée statistique	3
	Références	6

1 Données - Méthodes - Procédures

On réunit ici des notes pour un cours de statistique descriptive en biologie évolutive. L'intérêt de la statistique comme discipline scientifique réside dans sa position frontalière. Trois mondes s'y rencontrent.

- ★ **METH** désigne l'ensemble des méthodes, des principes mathématiques, des théorèmes, des théories, des modèles, de ce qui construit logiquement la méthodologie. On le désigne par *mathematical statistics*.
- ★ **PROC** est l'ensemble des procédures, des programmes, de la documentation et des conditions d'utilisation, de ce qui permet de faire les calculs et les graphiques, le monde du pouvoir faire, il est régi par l'informatique. On le désigne par *statistical computing*.
- ★ **DATA** est l'univers des données, des objets, des conditions d'observations, des contraintes et des objectifs. Il est dans la logique expérimentale et on y place l'analyse des données comme partie de l'expérience. Quand on est biologiste on parle de *biostatistics*.

On pourrait prendre pour symbole un triangle :

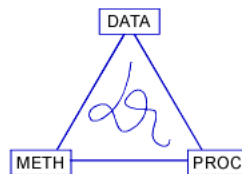



FIG. 1 – Chemin d'analyse de données !

Il est général de ne penser qu'un des éléments. Toute position est légitime : elle peut être partout stérile ou utile. Souvent encore on utilise deux des termes, le troisième est le faire valoir de leur relation. On veut faire ici la part belle au troisième terme, celui des données, en réfléchissant sur le rôle des deux autres dans la compréhension qu'on peut avoir d'objets biologiques.

La biologie évolutive a ceci de particulier qu'elle engendre de l'information numérique d'une complexité souvent en avance sur les méthodologies disponibles. C'est pourquoi elle est un domaine favorable à la statistique *ad hoc*.

2 galabiose : un exemple d'originalité radicale

On créera plusieurs dossiers de travail avec . Appeler le premier galabiose. Récupérer la fiche :

<http://pbil.univ-lyon1.fr/R/pps/pps026.pdf>

Les données [2] sont formées d'une phylogénie et d'une réponse de type (nombre de succès - nombre d'échecs). Utiliser la fiche :

<http://pbil.univ-lyon1.fr/R/querep/qri.pdf>

Étiqueter les feuilles de a à t (enlever les données manquantes), saisir dans un fichier texte l'arbre, lire le fichier texte (`readLines`) dans un vecteur de chaînes

de caractères `tre1`, charger la librairie `ade4`, transformer la chaîne en objet de la classe `phylog` (`newick2phylog`) et représenter l'objet :

```
library(ade4)
phy1 <- newick2phylog(tre1)
plot(phy1, f = 0.8)
```

Vous devez obtenir la figure 2. Implanter un *data frame* avec deux variables

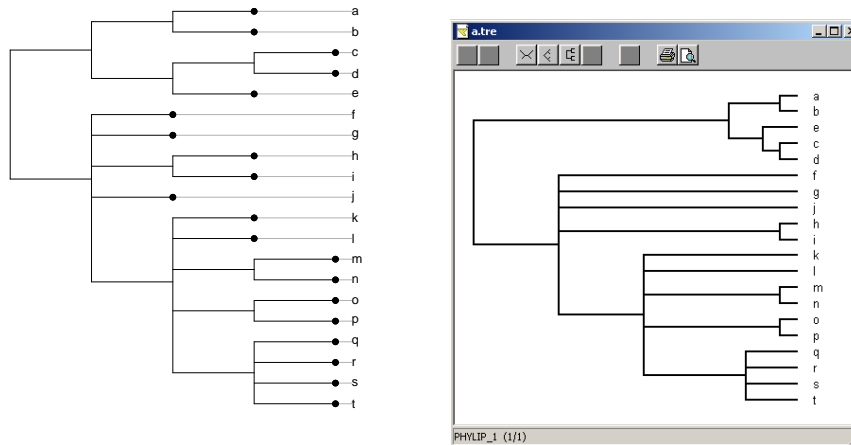


FIG. 2 – A gauche avec `ade4`, à droite avec TreeView.

oui-non, les lignes étant les feuilles de la phylogénie :

```
      a b c d e f g h i j k l m n o p q r s t
oui 0 0 0 0 0 4 0 1 0 1 7 6 3 1 2 4 8 3 24 40
non 4 3 2 14 10 1 3 0 7 0 0 0 0 0 0 1 1 3 18 10
```

La question est posée. Sauver dans un fichier `rda` la liste formée de l'arbre et du *data frame* (`save`). Le relire par un `load`. Ceci permet de communiquer un ensemble de données de complexité quelconque à un correspondant.

```
galabiose <- list(rep = tab, tre = tre1)
save(galabiose, file = "galabiose.rda")
```

Déplacer le fichier `galabiose.rda` dans un nouveau dossier de travail et vérifier que `load("galabiose.rda")` vous permet de retrouver exactement ce que vous avez voulu garder. C'est comme ça que vous communiquerez le résultat de votre rapport pour la validation de l'UE.

3 cortex : l'espace comme donnée statistique


Créer un dossier de travail `cortex`. Récupérer le fiche :

<http://pbil.univ-lyon1.fr/R/pps/pps012.pdf>

Copier dans le fichier `pdf` le petit tableau de données et le lire dans `R`. Vous devez obtenir :

```
liz
```

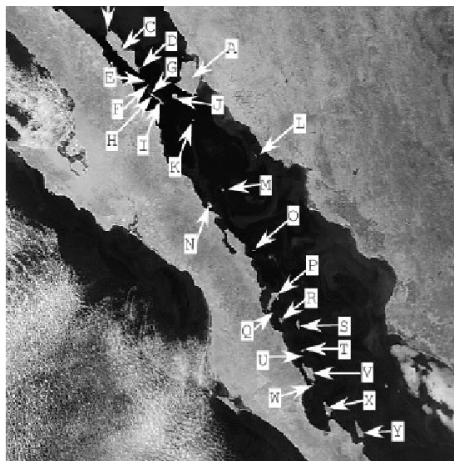
	T01	T02	T03	T04	T05	T06	T07	T08	T09	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20
A	1	1	0	1	1	1	0	1	1	1	0	0	1	1	1	1	0	0	0	1
B	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0
C	1	1	0	0	0	1	0	0	0	1	1	0	1	1	0	1	1	0	0	0
D	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0
E	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
F	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
G	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
H	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0
I	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0
J	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	1	0	0
K	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
L	0	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0
M	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
N	1	1	1	0	0	1	1	1	0	1	0	0	1	0	0	1	1	0	0	0
O	1	1	1	0	0	1	1	1	0	1	0	0	1	0	0	1	1	0	0	0
P	0	1	1	1	0	1	1	1	0	1	0	0	1	0	0	1	1	0	0	0
Q	1	1	0	0	0	1	0	1	0	1	0	1	0	0	0	1	0	0	0	0
R	0	1	0	1	0	0	1	0	0	1	0	0	0	0	0	1	1	0	0	0
S	0	1	0	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0
T	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
U	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
V	1	1	1	1	0	1	1	1	0	1	0	0	1	0	0	1	1	0	0	0
W	0	1	1	0	0	1	1	1	0	1	0	0	1	0	0	1	0	0	0	0
X	1	1	0	1	0	1	1	1	0	1	0	1	1	0	0	1	1	0	0	0
Y	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	1	1	0

C'est un archétype [1] de tableau faunistique espèces-sites en présence-absence. Récupérer le fichier <http://pbil.univ-lyon1.fr/R/donnees/cortes.pnm>. C'est une copie d'écran en noir et blanc de la figure de la fiche transformée (Image-Magick) au format pnm pour être lu dans .

```
library(pixmap)
download.file(url = "http://pbil.univ-lyon1.fr/R/donnees/cortes.pnm",
             dest = "cortes.pnm", mode = "wb")
bkg <- read.pnm("cortes.pnm")
```

Vous devez obtenir la figure :

```
par(mar = rep(0, 4))
plot(bkg)
```

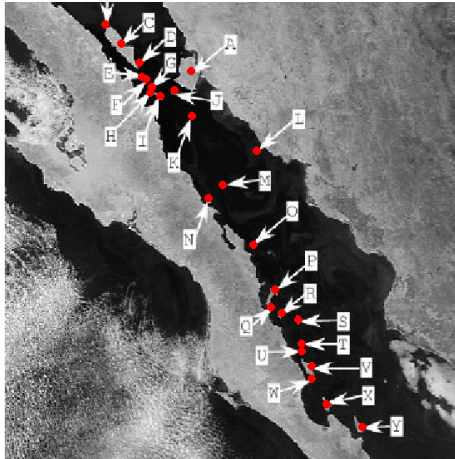


Digitaliser la position des sites :

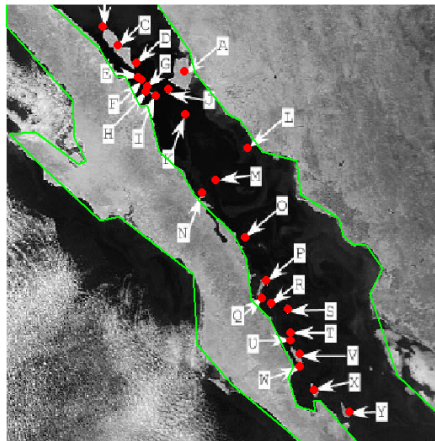
```
xy = as.data.frame(locator(26))
row.names(xy) <- LETTERS[1:25]
```

Vérifier.

```
par(mar = rep(0, 4))
plot(bkg)
points(xy, col = "red", pch = 20, cex = 2)
```



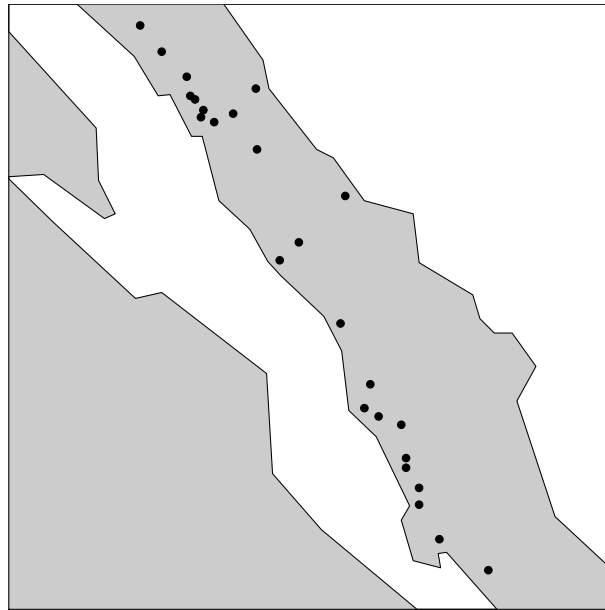
Profiter de l'affichage pour digitaliser deux polygones (voir la documentation de locator pour arrêter la saisie à la volée) :



```
pol1 <- locator(5000, type = "l", lwd = 2, col = "green")
pol2 <- locator(5000, type = "l", lwd = 2, col = "green")
```

Construire un schéma de l'espace utilisé en proposant une fonction qui affiche un fond de carte, par exemple (pour les goûts et les couleurs ...) :

```
bkgnd <- function() {
  par(mar = rep(0, 4))
  s.label(xy, xlim = c(5, 360), ylim = c(5, 360), incl = F)
  rect(0, 0, 370, 370, col = grey(0.8))
  polygon(pol2, col = "white")
  polygon(pol1, col = "white")
  points(xy, pch = 19)
  box()
}
bkgnd()
```



Ajouter enfin le code des espèces.

`codesp`

```
[1] "Coleonyx"           "Phyllodactylus"
[3] "Sceloporus orcutti"  "Sceloporus magister"
[5] "Sceloporus clarki"  "Cnemidophorus tigris"
[7] "Cnemidophorus hyperythrus" "Urosaurus"
[9] "Urosaurus ornatus"  "Uta"
[11] "Petrosaurus mearnsi" "Petrosaurus thalassinus"
[13] "Callisaurus draconoides" "Crotyphytus"
[15] "Gambelia wislizenii" "Sauromalus"
[17] "Dipsosaurus dorsalis" "Ctenosaurus hemilopha"
[19] "Sator"              "Phrynosoma solare"
```

On peut sauver alors le tableau, les coordonnées, les polygones, les étiquettes et la fonction dans un fichier `cortes.rda`. Le fichier contient ce qu'il faut pour analyser les données. Le `load` aura un effet différent. Vérifier.

```
save(liz, pol1, pol2, xy, bkgnd, codesp, file = "cortes.rda")
```

On a voulu, en isolant la préparation des données de toute autre intervention, souligner un point d'importance. Les données à traiter sont définies avant le traitement, et non pendant. Modifier les données au cours de l'analyse c'est invalider toute méthode. Contrairement à la pratique très répandue des biologistes des populations, on établit les données disponibles avant de les traiter alors que l'habitude veut qu'on extraie des variantes jusqu'à ne plus pouvoir reproduire quelque résultat que ce soit. Les données disponibles doivent être stables et on doit pouvoir indiquer clairement ce qu'on en fait. Un des premiers soucis sera de les voir ... à suivre.

Références

- [1] T.J. Case. Niche overlap and the assembly of island lizard communities. *Oikos*, 41 :427–433, 1983.

- [2] Noriko Suzuki, Jr. Laskowski, Michael, and Yuan C. Lee. Phylogenetic expression of gala1-4gal on avian glycoproteins : Glycan differentiation inscribed in the early history of modern birds. *PNAS*, 101(24) :9023–9028, 2004.