Analyses multivariées

5 décembre 2018

Ce laboratoire doit être remis le **12 décembre à 17h sur Moodle**. Dans votre réponse pour chaque question, veuillez inclure une copie du code R utilisé (s'il y a lieu) et des résultats obtenus.

1. Composition de sources minérales

Le fichier `springs.csv` contient des données tirées d'une étude de Tanaskovic et al. (2012) sur les propriétés physico-chimiques (température, pH, conductivité électrique) et la concentration d'ions de 30 sources minérales de Serbie.

```
sources <- read.csv("springs.csv")
str(sources)
```

```
## 'data.frame':    30 obs. of  14 variables:
##  $ waterSource: Factor w/ 30 levels "Banjska","Bogutovacka",..: 9 8 10 19 24 26 2 14 21 16 ...
##  $ tempCels   : num  45 25 25 48 14.4 33 25 39.1 43 35 ...
##  $ pH         : num  7.7 7.3 6.6 7 7.5 7.4 7.2 7.6 7.5 7 ...
##  $ elecCond   : int  3770 5100 677 2219 1100 722 440 265 248 324 ...
##  $ totSolid   : num  4.173 6.2 0.641 2.119 1.1 ...
##  $ Ca         : num  5.6 24 120 60 56 60 50 62 74 102 ...
##  $ Mg         : num  3.8 24 36.6 18 48.8 67 61 14 17 32 ...
##  $ Na         : num  1157 476 200 2280 226 ...
##  $ K          : num  7.9 4 22.2 30 3.5 7.1 1.5 1.2 8 2.5 ...
##  $ Cl         : num  56.8 369 89 1560 198.8 ...
##  $ SO4        : num  2 2 55 2 82 90 15 40 20 18 ...
##  $ HCO3       : num  2904 829 890 3360 573 ...
##  $ SiO2       : num  36.7 25 18 20 22 10 20 18 18 25 ...
##  $ geoStruct  : int  1 1 1 1 1 2 2 2 2 2 ...
```

Nous nous intéressons d'abord aux 8 colonnes allant de *Ca* à *SiO2*, qui représentent dans l'ordre la concentration d'ions calcium, magnésium, potassium, chlorure, sulfate, bicarbonate et silicate, toutes mesurées en mg/L.

a) Puisque ces 8 variables ont les mêmes unités, il n'est pas nécessaire de les normaliser avant d'effectuer des analyses multivariées. Toutefois, nous appliquerons une transformation logarithmique à ces données. En inspectant les données, déterminez pourquoi cette transformation est utile ici.

Note: Si un tableau de données est composé uniquement de variables numériques, vous pouvez appliquer la fonction `log` au tableau pour prendre le logarithme de toutes les variables d'un coup.

b) Réalisez une analyse en composantes principales des concentrations transformées. Que remarquez-vous au sujet des contributions des variables sur le premier axe (PC1)? Est-ce qu'il y a une propriété générale (non reliée à une variable en particulier) qui distingue les échantillons avec des valeurs négatives et positives sur cet axe?

c) Réalisez un diagramme de double projection (*biplot*) pour les composantes principales 2 et 3. Quelles variables ont le plus grand effet sur chacun de ces axes? D'après ce diagramme, donnez un exemple de paire d'ions dont les concentrations sont fortement corrélées positivement, et d'une autre paire dont les concentrations sont négativement corrélées.

d) La variable `geoStruct` du tableau de données original est une variable catégorielle représentant le type géologique du site. Comment pourriez-vous vérifier si une composante principale varie significativement entre les sites de différents types géologiques? Réalisez ce test séparément pour les composantes PC1 et PC2 et déterminez la nature des différences significatives s'il y a lieu.

e) Créez un nuage de points des composantes PC1 et PC2 qui indique aussi le type géologique pour chaque site.

2. Composition de sédiments arctiques

Le fichier `arctic.csv`, tiré d'une étude de Martinez et al. (2009), contient des données sur la composition d'échantillons provenant de différentes profondeurs (`depth`, unités non-spécifiées) d'une carotte de sédiments de l'océan Arctique. Pour chaque échantillon, les colonnes *Al* à *P* indiquent le pourcentage de la masse de l'échantillon correspondant à cet élément.

```
arctic <- read.csv("arctic.csv")
head(arctic)
```

```
##    sampleID depth   Al    Ti   Fe    Mn   Ca   Mg   Na    K     P
## 1 4C-01H-1W  0.20 8.34 0.520 5.34 0.286 4.04 1.71 1.86 2.15 0.085
## 2 4C-01H-2W  2.36 7.56 0.539 5.75 0.307 0.50 1.48 1.85 2.07 0.084
## 3 4C-02H-2W  5.62 7.84 0.468 4.67 0.149 0.46 1.13 1.70 1.90 0.075
## 4 3A-01H-4W  8.52 8.58 0.503 5.06 0.288 0.50 1.17 1.85 2.08 0.093
## 5 4C-03H-3W 12.41 7.46 0.494 5.06 0.089 0.42 1.15 1.84 2.10 0.093
## 6 2A-05X-2W 20.83 8.49 0.520 6.21 0.097 0.40 1.13 1.74 1.99 0.127
```

a) Encore une fois, puisque les variables sont sur la même échelle (% de masse), il n'est pas nécessaire de les normaliser. Réalisez une ACP avec les variables originales et une autre avec les variables normalisées, puis comparez les diagrammes de double projection. Expliquez comment le choix de normaliser ou non affecte l'importance de différents éléments sur les résultats de l'ACP.

b) À partir des variables normalisées, créez une classification hiérarchique à partir de la méthode du saut maximum et une autre à partir de la méthode de Ward. Superposez la classification en (i) 2 groupes et (ii) 4 groupes obtenue par chaque méthode aux résultats de l'ACP pour ces mêmes données. Est-ce que les groupes obtenus différent entre les deux méthodes? Est-ce que ces classifications semblent bonnes par rapport aux coordonnées de l'ACP?

Note: Pour la classification en quatre groupes, regardez jusqu'à 4 composantes principales pour juger la classification. (Ce n'est pas une règle générale, seulement une suggestion pour ce problème.)

c) Représentez graphiquement la variation des deux premières composantes principales en fonction de la profondeur (`depth`), puis l'appartenance aux groupes de la classification à 4 groupes (selon l'une ou l'autre méthode) en fonction de la profondeur. Est-ce que ces deux méthodes permettent de détecter des changements importants de composition du sédiment le long de la carotte?