CHAPITRE 2

LES OUTILS DE REPRESENTATION D'UN ECHANTILLON

Lorsque l'on fait des statistiques, on travaille en général sur des **observations** sur lesquelles sont mesurées ou observées des **variables** en plus ou moins grand nombre. Une observation est l'unité élémentaire ; nous verrons qu'elle peut avoir des formes multiples : une personne ou un groupe de personnes, un animal ou une portée d'animaux, une plante, un organe ou une seule cellule. Une variable est une caractéristique ou une propriété qu'il est possible de mesurer. Mesurer quelque chose signifie que nous faisons un modèle numérique de la chose à mesurer : nous assignons un nombre à un niveau de la caractéristique à mesurer. Le poids d'une personne est une variable ; nous assignons à deux personnes de même poids la même valeur numérique.

Dans ce chapitre nous allons définir ce qu'est la structure d'un ensemble **observations / variables** que nous appellerons tableau de données (§1) ; nous verrons ensuite comment résumer un tableau de données à l'aide d'un petit nombre de caractéristiques soit unidimensionnelles (§2) soit multidimensionnelles (§3) ; nous présenterons un outil commode pour représenter le tableau de données (§4), puis nous dresserons un bilan des aspects importants de ce chapitre (§5).

1. STRUCTURE D'UN TABLEAU DE DONNEES

1.1 Questions préalables

Toute expérience ou toute enquête se fonde sur la mesure de quantités, qui sont les réalisations d'une ou de plusieurs variables. Mais avant toute analyse de ces données, il faut se poser un certain nombre de questions préalables :

- **Que sont les données ?** Sont-elles en nombre important, quelle est leur précision ?
- **Quelles sont les unités** ? Kilomètre ou mile marin, degré Celsius ou degré Farenheit, etc. ?
- **D'où proviennent-elles ?** Les valeurs sont-elles raisonnables ? Des valeurs comme 0.7 ou 44 pour le poids d'une dinde en kilogrammes indiquent-elles une simple erreur de transcription ou mettent-elles en doute l'ensemble des relevés ? Pour un relevé pluviométrique, l'emplacement du pluviomètre peut-il entraîner des variations importantes des mesures obtenues ?
- **Comment et quand ont-elles été mesurées ?** Des biais sont-ils possibles ? L'observateur sait-il arrondir les nombres ? Que signifie la température d'un jour donné, est-ce la valeur maximale, la valeur à une heure précise ? Les enregistrements sont-ils différents en fin de semaine, ont-ils été réalisés par la même personne ?

- **Comment ont-elles été acquises ?** A-t-on mesuré toutes les unités de la population, ou bien a-t-on fait un échantillonnage ? Dans ce dernier cas, le choix de l'unité expérimentale peut-il avoir une influence sur le résultat ?
- **Existe-t-il une structure logique entre les observations ?** S'agit-il du poids d'un animal mesuré à plusieurs époques ou du poids d'animaux différents ? Les unités ont-elles un rapport entre elles, comme les animaux d'une même portée ou les fruits d'une même variété ? Les observations ont-elles en commun certains facteurs d'environnement ?

C'est seulement quand on a répondu à toutes ces questions, que l'on peut pousser plus loin l'analyse.

1.2 La forme du tableau de données

Il est généralement toujours possible de présenter des données sous la forme d'un tableau rectangulaire à I lignes et J colonnes ; quelquefois nous utiliserons les notations n au lieu de I et p au lieu de J. Formellement, nous pouvons écrire qu'un tableau de données $\mathbf{X}$ est de la forme :

$$\mathbf{X} = \left[x_i^j\right], i = 1, \cdots, I \text{ (ou } n) \text{ ; } j = 1, \cdots, J \text{ (ou } p)$$

Ci-dessus, les deux indices i et j repèrent la valeur x_i^j qui se trouve à la ligne i et à la colonne j. Mais la représentation d'une ligne ou d'une colonne est différente :

- une ligne i, notée $\boldsymbol{x}_i = \left[x_i^1 \quad \cdots \quad x_i^j \quad \cdots \quad x_i^J\right]^{\mathrm{T}}$, représente une unité de base : c'est l'**observation élémentaire** sur laquelle ont été mesurées, observées ou contrôlées J caractéristiques.
- une colonne j, notée $\boldsymbol{x}^j = \left[x_1^j \quad \cdots \quad x_i^j \quad \cdots \quad x_I^j\right]^{\mathrm{T}}$, représente la valeur soit d'une **variable**, soit d'un **indicateur**[1]. La différence entre les deux provient de l'origine expérimentale de l'observation, c'est-à-dire de la façon dont on l'a obtenue : relever l'âge d'une personne fournit une variable, mais sélectionner des personnes d'un âge fixé, fournit un indicateur. C'est un problème que nous évoquerons plusieurs fois dans cet ouvrage, dans la mesure où la distinction peut ne pas être toujours évidente.

1.3 Notion de type

La valeur d'une variable, comme celle d'un indicateur, dépend de la structure mathématique qui permet de la définir. Est-elle définie sur l'ensemble des nombres réels $\mathbb{R}$ ou simplement sur l'ensemble des réels positifs $\mathbb{R}^+$? Ne peut-elle prendre qu'un nombre limité de valeurs entières ? Quelquefois, la valeur ne correspond pas à une structure numérique : on peut toutefois s'y ramener par un simple **codage**. Par exemple, pour une couleur qui peut ne présenter que les trois caractéristiques vert, jaune ou marron, on peut soit établir la correspondance suivante :

$$\{\text{vert, jaune, marron}\} \Leftrightarrow \{1, 2, 3\}$$

[1] Dans la terminologie du modèle linéaire, on emploie le mot **facteur**.

soit utiliser pour chaque réalisation une variable à deux valeurs seulement, 0 ou 1. Chaque modalité de la variable définit une variable dichotomique ; la correspondance est alors :

TABLEAU 1 - Exemple de passage d'un codage simple à un codage disjonctif complet.

	codage initial	vert	jaune	marron
vert :	1	1	0	0
jaune :	2	0	1	0
marron :	3	0	0	1

Pour un traitement statistique, on utilise ce dernier codage appelé **codage disjonctif complet**. Ainsi un objet vert sera analysé par l'ensemble des trois valeurs {1,0,0}, un jaune par {0,1,0} et un marron par {0,0,1}, comme indiqué dans le tableau 1.

Physiquement, une mesure numérique sur $\mathbb{R}$ ne peut prendre qu'un nombre fini de valeurs. En pratique, ce qui distingue les valeurs sur $\mathbb{R}$ de celles sur $\mathbb{N}$ dépend essentiellement du nombre de valeurs possibles : si ce nombre est assez grand (> 10), on pourra considérer que la mesure est définie sur $\mathbb{R}$, sinon elle l'est sur $\mathbb{N}$; le nombre d'enfants d'une famille est généralement défini sur $\mathbb{N}$ alors que la taille de la portée d'un animal prolifique peut l'être sur $\mathbb{R}$. Néanmoins, rien ne s'oppose à dire que le nombre moyen d'enfants d'une famille européenne est de 1.8, bien que cette valeur n'ait aucun sens pour une famille particulière. La limite entre la définition sur $\mathbb{R}$ ou sur $\mathbb{N}$ dépend du contexte de l'application qui devra toujours être précisé. Cette remarque ne s'applique pas à un codage ; elle n'a de sens que pour une caractéristique sur laquelle on peut faire une opération arithmétique comme la sommation, donc un calcul de moyenne. Pour une utilisation statistique, il est utile de remplacer la notion de structure mathématique par celle de type de variable, schématisée dans le tableau 2.

TABLEAU 2 - Les différents types de variables ou d'indicateurs, nature et exemples.

variable ou indicateur	Type	Nature	Exemples
quantitative	continu		*poids, taille, revenu*
	entier	discrète	*nombre d'enfants*
qualitative	polytomie ordonnée	ordinale	*(mauvais,moyen,bon), (malade,atteint,sain)*
	polytomie non ordonnée	nominale	*profession, couleur, variété végétale*
	dichotomie	binaire	*(vrai,faux],(oui,non), (présence,absence)*

1.4 Représentation d'un tableau de données

Le tableau $\mathbf{X}$ est donc constitué par deux sous-tableaux :

$$\mathbf{X} = [\mathbf{X^1} : \mathbf{X^2}]$$

$\mathbf{X^1}$ est un tableau à *I* lignes et *p* colonnes ; $\mathbf{X^2}$ est un tableau à *I* lignes et *q* colonnes. Donc $\mathbf{X}$ est un tableau à I lignes et $J = p + q$ colonnes. L'unité expérimentale n°i, la ligne i de, $\mathbf{X}$ est donc formée d'un ensemble de p variables et de q indicateurs ; les p variables définissent le **vecteur observation** $\mathbf{X_i^1}$ lui-même éventuellement repéré par un **vecteur de contrôle** $\mathbf{X_i^2}$. L'ensemble des méthodes statistiques se rattachent :

- soit à des **méthodes unidimensionnelles** pour l'analyse séparée de chaque colonne de $\mathbf{X^1}$;
- soit à des **méthodes multidimensionnelles**[2] pour l'analyse simultanée de l'ensemble des colonnes de $\mathbf{X^1}$.

Dans les deux cas, l'existence d'un vecteur de contrôle induit une structure dans le tableau de données : connue *a priori*, elle sera généralement utilisée dans l'analyse des données ; mais elle pourra aussi ne fournir que des informations non prises en compte dans l'analyse, néanmoins utiles pour l'interprétation.

Quelquefois, on peut compléter un premier tableau $\mathbf{X}$ par d'autres observations sur les mêmes J colonnes. Si ces observations ne sont pas prises en compte dans l'analyse initiale, on dira que l'on a affaire à des **observations passives** ou **observations supplémentaires**, par opposition aux premières qui sont des **observations actives**. Naturellement, on peut aussi ajouter d'autres colonnes à $\mathbf{X}$ sur les mêmes unités de base ; on parlera de **variables passives** ou **variables supplémentaires** (les secondes) et de **variables actives** (les premières). La même chose peut se produire pour des indicateurs. Schématiquement nous pouvons écrire que le tableau $\mathbf{X}$ est complété de la manière suivante :

$$\begin{bmatrix} \mathbf{X} & \mathbf{X}^{Csup} \\ \mathbf{X}^{Lsup} & / \end{bmatrix}$$

Bien que nous ne l'ayons pas indiqué, le symbole « / » peut être remplacé par des variables supplémentaires associées à des observations supplémentaires.

2. RESUMES UNIDIMENSIONNELS

Nous débuterons ce paragraphe par l'étude des variables quantitatives ; puis nous verrons ensuite de manière plus succincte au §2.5 le cas des variables qualitatives. Avant d'étudier un tableau comme $\mathbf{X^1}$, il est toujours indispensable d'analyser séparément chacune de ses colonnes ; pour des variables quantitatives, on commence par étudier séparément chaque **distribution** des variables du tableau, soit p distributions. Il existe plusieurs manières de représenter ces distributions. A titre d'exemple regardons le tableau 3 qui représente 20 valeurs de 6 variables chimiques de composition d'eau minérale. On peut étudier ce tableau de deux manières : à l'aide de graphiques ou à l'aide de résumés numériques ; nous commencerons par les premiers.

Dans ce paragraphe nous allons considérer les colonnes de $\mathbf{X^1}$ l'une après l'autre, même si ne nous interdisons pas des regards croisés ; aussi, pour éviter l'usage d'indices superflus, notons $\mathbf{x}^j$ $(j = 1, \cdots, p)$; les *n* valeurs de la colonne j (n est noté I au début du §1.4).

L'analyse des colonnes de $\mathbf{X^2}$ est aussi possible, s'il existe une structure dans le tableau de données. Elle permet de voir si cette structure induit un équilibre entre les différents groupes définis par $\mathbf{X^2}$; ainsi, lorsque $\mathbf{X^2}$ n'a qu'une colonne (alors $q = 1$), si la variable a *k* modalités, l'écriture disjonctive de $\mathbf{X^2}$ comportera *k* colonnes associées aux *k*

[2] On trouve aussi l'adjectif *multivarié,* provenant directement du *multivariate* anglais.

variables définies comme dans le tableau 1. Par exemple, si c'est une couleur : existe-t-il autant d'observations dans chaque niveau de couleur ? Nous verrons plus loin que l'existence d'un équilibre est souvent un élément important pour faire les « meilleures » statistiques. Dans la suite de ce chapitre, nous n'utiliserons pas cette possibilité de structuration, nous utiliserons la notation $\mathbf{X}$ à la place de $\mathbf{X^1}$. De plus, nous noterons n le nombre lignes (=observations, au lieu de I) et p celui des colonnes (=variables).

TABLEAU 3 - **Eaux1** : Echantillon de 20 eaux minérales non gazeuses pour 6 variables chimiques.

Eau	HCO3	SO4	Cl	Ca	Mg	Na
Aix	341	27	3	84	23	2
Bec	263	23	9	91	5	3
Cay	287	3	5	44	24	23
Cha	298	9	23	96	6	11
Cri	200	15	8	70	2	4
Cyr	250	5	20	71	6	11
Evi	357	10	2	78	24	5
Fer	311	14	18	73	18	13
Hip	256	6	23	86	3	18
Lau	186	10	16	64	4	9
Oge	183	16	44	48	11	31
Ond	398	218	15	157	35	8
Per	348	51	31	140	4	14
Rib	168	24	8	55	5	9
Spa	110	65	5	4	1	3
Tho	332	14	8	103	16	5
Ver	196	18	6	58	6	13
Vil	59	7	6	16	2	9
Vit	402	306	15	202	36	3
Vol	64	7	8	10	6	8

2.1 Graphiques

Nous verrons plus loin que les méthodes statistiques dépendent beaucoup des **suppositions** que l'on peut faire sur les variables. Il est donc indispensable de faire, avant tout calcul, des graphiques permettant de voir si ces suppositions sont plus ou moins bien respectées, c'est ce qu'on appelle une analyse exploratoire des données[3]. Ces graphiques permettent de répondre aux deux questions suivantes :

- La distribution des données est-elle sensiblement Normale, c'est-à-dire symétrique et relativement concentrée autour de la position centrale ?
- Existe-t-il des **valeurs suspectes**[4] ?

On peut faire quatre types de graphiques : un **histogramme**, un **graphique de densité**, un **graphique de Normalité**[5] et un **diagramme en boîte**[6].

[3] En anglais : *Exploratory Data Analysis* (EDA).
[4] En anglais : *outliers*.
[5] Qu'on appelle aussi *qq-plot*.
[6] Qu'on appelle aussi *boîte à moustaches* ou *Box and Whisker plot* ou *box-plot*.

Un histogramme n'est parlant que si le nombre n d'observations est suffisamment élevé (>100), aussi est-il utile de le remplacer par le graphique de densité qui en est une version lissée dont nous parlerons plus loin.

Un graphique de Normalité est un graphique des valeurs ordonnées de la variable en fonction des quantiles correspondants de la loi Normale standard (de moyenne 0 et de variance 1) : la distribution est d'autant plus proche de la Normalité que le graphique est proche d'une droite.

Un diagramme en boîte est aussi très facile à interpréter : si nous appelons $Q_{0.25}$, $Q_{0.50}$ (qui est la **médiane**) et $Q_{0.75}$ les valeurs telles que le quart, la moitié et les trois-quarts des observations leur soient inférieures (ce que nous appellerons plus loin les quartiles) on trace une boîte comme celle de la figure 1 ; les données sont celles de la variable Ca du fichier **Eaux1** (Tab.3). Les extrémités indiquées par « | » sont à des positions : pour l'inférieure à $Q_{0.25} - 1.5(Q_{0.75} - Q_{0.25})$ et pour la supérieure à $Q_{0.25} + 1.5(Q_{0.75} - Q_{0.25})$ et les valeurs qui sont extérieures sont indiquées par des « ° ». On repère ainsi très facilement les valeurs suspectes ou extrêmes, et on voit très rapidement si la distribution est symétrique. Ici $Q_{0.25} = 53.25$; $Q_{0.50} = 72.00$ et $Q_{0.75} = 92.25$. Apparaissent à droite deux observations suspectes correspondant aux valeurs 157 et 202.

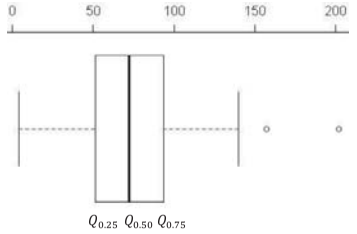

FIGURE 1. Représentation schématique d'une boîte à moustaches pour la variable Ca *du fichier* **Eaux1**.

La comparaison de plusieurs distributions est facilitée. Ainsi si nous regardons les deux premières variables du fichier de données **Eaux1** la distribution de HCO_3 apparaît assez symétrique, alors que celle de SO_4 ne l'est pas du tout. Les quatre autres boîtes à moustaches donnent une vision plus claire de la dissymétrie (Tab.4A, Fig.2).

TABLEAU 4A - **Eaux1** : Boîtes à moustaches.

```
> library(MASS); library(car); library(ade4); library(stats); library(e1071)
> Eaux1<-read.table("Eaux1.txt",h=T, row.names=7)
# Tracé des six boîtes à moustaches (Fig.2)
> boxplot(Eaux1)
```

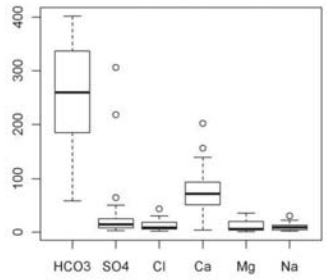

FIGURE 2. **Eaux1** : *Boîtes à moustaches des six variables.*

Un regard sur les graphiques complets de ces mêmes variables montre bien les différences entre les deux : HCO_3 a une distribution voisine de la Normale, alors que SO_4 est très dissymétrique avec un fort étalement vers les grandes valeurs (Tab.4B, Fig.3). Les graphes de Normalité vont dans le même sens : la forme de la distribution de HCO_3 est beaucoup proche de celle de la loi Normale que celle de SO_4 (Fig.3).

TABLEAU 4B - **Eaux1** : programmes permettant de faire les graphiques des figures 2 et 4.

```
#  Tracé sur un même graphe de l'histogramme et de la densité de la
  gaussienne correspondante
> Dist.forme<-function(x)
+ {
+ par(mfrow=c(2,2))
+ ## pb donnees manquantes
+ x.naomit<-na.omit(x)
+
+ hist(x.naomit, col="gray", prob=TRUE,xlim=c(min(x.naomit),
+ max(x.naomit)),main="")
+
+ curve(dnorm(x,mean=mean(x.naomit),sd=sd(x.naomit)),add=TRUE,lwd=2,
+ col="red")
+ boxplot(x.naomit)
+ iqd<-summary(x.naomit)[5] - summary(x.naomit)[2]
+ points(mean(x.naomit), col = "orange", pch = 18)
+ plot(density(x.naomit,width=2*iqd),xlab="x",ylab="",type="l",main="")
+ qqnorm(x.naomit)
+ qqline(x.naomit)
+ }
# Tracé des quatre graphiques réalisés sur les variables HCO3 et SO4
# (Fig.3)
> Dist.forme(Eaux1[,1])
> Dist.forme(Eaux1[,2])
```


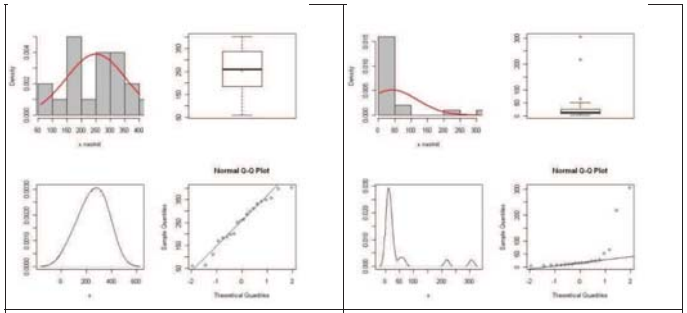

FIGURE 3. **Eaux1** : *Graphiques EDA des deux variables* HCO_3 *et* SO_4.

2.2 Paramètres numériques classiques

Pour compléter ces visions graphiques, il faut calculer des paramètres numériques traduisant le mieux possible la distribution, c'est-à-dire les caractéristiques de ces n valeurs ; on peut regrouper ces caractéristiques selon trois types d'indices : **position**, **dispersion** et **forme**. Pour chaque type, plusieurs paramètres sont possibles avec des intérêts pratiques plus ou moins importants.

2.2.1 Paramètres de position : ils définissent la tendance générale de la distribution. Deux paramètres principaux sont utilisés, avec des propriétés différentes : la moyenne et la médiane d'une variable quantitative $\boldsymbol{x}$. La **moyenne**, notée $\bar{\mathbf{x}}$, est définie par :

$$\bar{x} = \sum_{i=1}^{n} x_i / n$$

Si on a n_i valeurs x_i identiques, et si le nombre total est $n = \sum n_i$, on associe à chaque observation une **pondération** $p_i = n_i/n$, et donc $\sum p_i = 1$; alors on écrit :

$$\bar{x} = \sum_{i=1}^{n} p_i x_i$$

Naturellement, quand toutes les observations sont individualisées, $p_i = 1/n$. On appelle **variable centrée** la variable $\boldsymbol{x_c}$ de coordonnées : $x_i - \bar{x}$; la moyenne de cette variable est donc nulle.

La **médiane**, notée Me, est la valeur qui divise les n observations en deux parties égales (50 % lui sont inférieures et 50 % supérieures), elle est déterminée lorsque l'on a rangé les *n* observations par ordre croissant :

$$x_{[1]} \leq x_{[2]} \leq \cdots \leq x_{[n-1]} \leq x_{[n]}$$

alors si n est impair et égal à $2m+1$, la médiane vaut :

$$Me = x_{[m+1]}$$

et si n est pair et égal à $2m$, la médiane vaut :

$$Me = \left(x_{[m]} + x_{[m+1]}\right)/2$$

Cette formulation montre que le calcul de la médiane revient à pondérer les valeurs rangées de la façon suivante : toutes les valeurs ont une pondération nulle $p_{[j]} = 0$, sauf pour :

- n impair, alors $p_{[m+1]} = 1$ et
- n pair $p_{[m]} = p_{[m+1]} = 1/2$

En fait, la médiane est un paramètre beaucoup plus stable que la moyenne et beaucoup moins sensible qu'elle aux valeurs suspectes car les valeurs extrêmes n'ont pas de poids dans son calcul. On dit que c'est un paramètre **robuste** (Tab.5A).

TABLEAU 5A - **Eaux1** : détermination de la médiane de HCO_3.

```
> Eaux1[,1] # valeurs de HCO3 dans le fichier
  [1] 341 263 287 298 200 250 357 311 256 186 183 398 348 168 110 332 196  59 402  64
> rank(Eaux1[,1]) # rang des valeurs de HCO3 dans le fichier
  [1] 16 11 12 13  8  9 18 14 10  6  5 19 17  4  3 15  7  1 20  2
> sort(Eaux1[,1]) # valeurs rangées par ordre croissant de HCO3
  [1]  59  64 110 168 183 186 196 200 250 256 263 287 298 311 332 341 348 357 398 402
> median(Eaux1[,1]) # médiane de HCO3
  [1] 259.5
```

La dernière remarque conduit à imaginer d'autres paramètres en jouant de façon moins drastique sur les pondérations des valeurs rangées. On peut ainsi calculer une **moyenne équeutée**[7], en supprimant un certain nombre de valeurs extrêmes, en général un pourcentage de n, par exemple 5 % ; si nous prenons un pourcentage de 50 % nous obtenons la médiane. Voici ce que donnent ces calculs sur les deux variables HCO_3 et SO_4 (Tab.5B) :

TABLEAU 5B - **Eaux1** : Moyennes et moyennes équeutées des deux variables HCO_3 *et* SO_4.

HCO_3	SO_4
`> mean(Eaux1[,1])`	`> mean(Eaux1[,2])`
`[1] 250.45`	`[1] 42.4`
`> mean(Eaux1[,1],trim=0.5)`	`> mean(Eaux1[,2],trim=0.5)`
`[1] 259.5`	`[1] 14.5`
`> mean(Eaux1[,1],trim=0.05)`	`> mean(Eaux1[,2],trim=0.05)`
`[1] 252.6667`	`[1] 29.94444`

Une autre idée est celle de la **winsorization** qui ne modifie pas la pondération, mais qui donne à un certain nombre (ou à un certain pourcentage) de valeurs extrêmes la même valeur. Par exemple, on donne à $x_{[1]}$ la même valeur que celle de $x_{[2]}$ et à $x_{[n]}$ la même valeur que celle de $x_{[n-1]}$. Ces différents paramètres ont des propriétés intéressantes et, quelquefois, ils peuvent remplacer la moyenne.

Quand les observations sont discrètes ou regroupées en classes de même amplitude, on définit le **mode** comme la valeur ou la classe de fréquence maximale. Pour

[7] En anglais : *trimmed mean*.

une variable quantitative, c'est la valeur la plus probable. Mais contrairement aux autres paramètres de position, le mode peut ne pas être unique : on parlera alors de **distribution multimodale**. Cet inconvénient peut devenir un avantage, puisque la présence de deux valeurs de mode peut être l'indice d'un mélange provenant de deux populations différentes. Pour une distribution Normale : Moyenne = Médiane = Mode.

Une grande différence entre ces paramètres (en particulier entre moyenne et médiane) est un indice certain de non-Normalité de la distribution.

2.2.2 Paramètres de dispersion : ils définissent la variabilité de la distribution. Le plus utilisé est la **variance** notée s^2, définie par :

$$s^2 = var(x) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \Big/ (n-1)$$

La division par $(n-1)$ sera justifiée ultérieurement pour des raisons statistiques ; quelquefois la division est faite par n, la variance est alors la moyenne des carrés des écarts à la moyenne[8]. Nous utiliserons toujours le facteur $(n-1)$, sauf cas particulier justifié par le contexte ; de toute manière, dès que n est assez grand la différence est minime. On note s l'**écart-type** qui est la racine carrée de la variance, et s'exprime donc dans la même unité que la variable étudiée (Tab.5C).

TABLEAU 5C - **Eaux1** : variances et écart-types des deux variables HCO_3 et SO_4.

HCO_3	SO_4
`> var(Eaux1[,1],na.rm=T)`	`> var(Eaux1[,2])`
`[1] 10529.63`	`[1] 6079.516`
`> sd(Eaux1[,1])`	`> sd(Eaux1[,2])`
`[1] 102.6140`	`[1] 77.97125`

On appelle **variable centrée réduite** notée x_{CR} la variable de coordonnées : $(x_i - \bar{x})/s$; la moyenne de cette variable est donc nulle et son écart-type vaut 1.

Un autre paramètre de dispersion est l'**étendue** notée E, définie par :

$$E = x_{[n]} - x_{[1]}$$

L'étendue est donc la différence entre la valeur maximale et la valeur minimale des x_i : c'est un paramètre très sensible aux **valeurs extrêmes** ; on peut lui substituer des paramètres dérivés de la **distribution équeutée**. En pratique, on considère surtout des paramètres dérivés des **α-percentiles** définis par la valeur de la variable telle que la proportion α ($\alpha < 0.5$) des observations lui soit inférieure. Si on prend la valeur entière (arrondie à l'entier le plus proche) de $\nu = \alpha n$, alors l'α - percentile $Q_\alpha = x_{[\nu]}$. L'importance des queues de distribution peut être étudiée par l'écart inter α - percentile :

$$D_\alpha = Q_{1-\alpha} - Q_\alpha$$

[8] En anglais : *standard deviation (sd)*.

On peut démontrer, sous des conditions très larges (Kendall & Stuart, 1958), des propriétés des distributions de ces quantités ; en particulier, pour la médiane ($\alpha = 0.5$), on obtient la variance :

$$var(Me) = \frac{\pi}{2} var(\bar{x})$$

Ce résultat nous indique que la variance de la médiane est moins précise que celle de la moyenne.

Un paramètre de dispersion, souvent utilisé et beaucoup plus stable que l'étendue, est l'**écart interquartile** :

$$D_{0.25} = Q_{0.75} - Q_{0.25}$$

C'est celui-ci qui est représenté dans le graphique de la boîte à moustaches.

Nous mettons à part un autre paramètre de mesure de la variabilité, appelé **coefficient de variation**, qui est le rapport exprimé en % de l'écart-type à la moyenne :

$$CV(\%) = 100 s/\bar{x}$$

Il peut être intéressant pour comparer deux échantillons d'une même variable ; de plus, étant sans dimension, il donne une mesure absolue de la variabilité. Néanmoins, il peut être d'un usage dangereux. En effet, la valeur de la moyenne est modifiée si on ajoute une constante à la mesure d'une variable ; ce n'est pas le cas de la valeur de l'écart-type. Donc CV sera différent selon que l'on fournit une température en degré Celsius ou en degré Fahrenheit puisqu'une température de t degrés Fahrenheit correspond à $5(t-32)/9$ degrés Celsius.

2.2.3 Paramètres de forme : Leur intérêt provient principalement du fait que leur valeur théorique pour la loi Normale est connue. On les utilise donc beaucoup pour vérifier le caractère gaussien d'une distribution ce sont :

- le coefficient d'**asymétrie**[9] *ou d'obliquité* estimé par :

$$b_1 = \frac{m_3}{{m_2}^{3/2}} = \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^3\right] \Big/ \left(\frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\right)^{3/2}$$

  - < 0 si mode$> \bar{x}$; donc une queue de distribution étalée vers la gauche ;
  - $= 0$; pour une distribution symétrique ;
  - > 0 si mode$< \bar{x}$; donc une queue de distribution étalée vers la droite.

Notons qu'une forte asymétrie dans une distribution est un indicateur d'hétérogénéité au sein de la population étudiée.

- le coefficient d'**aplatissement**[10] :

$$b_2 = \frac{m_4}{{m_2}^{2}} = \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^4\right] \Big/ \left(\frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\right)^{2}$$

[9] En anglais : *skewness*.
[10] En anglais : *kurtosis*.

  - < 3 si moins aplatie qu'une loi Normale ;
  - $= 3$; si loi Normale ;
  - > 3 si plus aplatie qu'une loi Normale.

Des valeurs respectives des coefficients d'asymétrie et d'aplatissement différentes de 0 et 3 indiquent que la distribution du caractère est plus ou moins éloignée de la distribution Normale ; évidemment la taille de l'échantillon joue un rôle si on souhaite faire un test de l'écart à la Normalité (Tab.6A).

TABLEAU 6A - **Eaux1** : résumé par variable et calcul des coefficients d'asymétrie et d'aplatissement.

```
# Résumé par variable
> summary(Eaux1)
# Ecart-type, asymétrie et aplatissement
> library(e1071)
# Bibliothèque nécessaire pour les fonctions skewness & kurtosis
> for (i in 1:6  )
+   {
+   print(sd(Eaux1[, i],na.rm=T))
+   print(skewness(Eaux1[, i],na.rm=T))
+   print(kurtosis(Eaux1[,i],na.rm=T))
+   }
```

L'ensemble de ces résultats pour les six variables **Eaux1** est fourni au tableau 6B[11].

TABLEAU 6B - **Eaux1** : paramètres des 6 variables.

	HCO3	SO4	Cl	Ca	Mg	Na
$x_{[1]}$	59.0	3.00	2.00	4.00	1.00	2.00
$Q_{0.25}$	185.2	8.50	6.00	53.25	4.00	4.75
Me	259.5	14.50	8.50	72.00	6.00	9.00
$\bar{x}$	250.4	42.40	13.65	77.50	11.85	10.10
$Q_{0.75}$	334.2	24.75	18.50	92.25	19.25	13.00
$x_{[24]}$	402.0	306.00	44.00	202.00	36.00	31.00
s	102.61	77.97	10.58	48.09	11.09	7.32
b_1	-0.33	2.45	1.24	0.76	0.93	1.24
$b_2 - 3$	-1.02	4.84	1.02	0.39	-0.55	1.14

2.3 De l'intérêt des transformations des variables

Tous les paramètres présentés ci-dessus permettent de définir des indices synthétiques d'une distribution. Nous verrons ultérieurement que de nombreuses méthodes d'analyse statistique, notamment les méthodes paramétriques, se basent sur l'hypothèse de Normalité de la distribution de la variable étudiée. Même si on peut bien souvent ne pas en faire un usage exclusif en ayant recours à des méthodes statistiques qui

[11] Les valeurs des coefficients d'asymétrie et d'aplatissement n'ont ici qu'une signification limitée ; il faudrait davantage d'observations pour se rendre compte si l'éloignement à la loi Normale a ou n'a pas un sens.

ne requièrent pas la Normalité des variables (par exemple les méthodes non paramétriques) ; il est souvent bien utile de s'en rapprocher en particulier en faisant une **transformation normalisatrice** des variables. Le type de transformation à adopter dépendra bien sûr de l'allure de la distribution de fréquences des données brutes. En particulier une distribution étalée vers les grandes valeurs, donc dissymétrique vers la droite, peut suggérer que si l'on utilisait une *transformation logarithmique*, on serait plus proche du référent Normal. Ceci conduit naturellement à choisir la moyenne des logarithmes des valeurs comme paramètre de position ; celle-ci est la **moyenne géométrique** notée g_x, des valeurs de départ :

$$ln(g_x) = \frac{1}{n}\sum_{i=1}^{n} ln(x_i) \Rightarrow g_x = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

De la même façon, si c'est l'inverse qui suit une distribution Normale, un calcul du même type fournit la **moyenne harmonique**[12] notée h_x :

$$\frac{1}{h_x} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i} \Rightarrow h_x = n\left(\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1}$$

Si toutes les valeurs x_i sont positives, on peut montrer que :

$$h_x \leq g_x \leq \bar{x}$$

Il est à noter que les transformations normalisatrices courantes possèdent la propriété importante de réduire l'**hétéroscédasticité** des données, c'est-à-dire de stabiliser leur variance. Dans de nombreuses situations, la transformation normalisatrice n'opère pas une normalisation complète mais contribue à rendre la distribution de fréquences plus symétrique, ce dont on peut se satisfaire dans un certain nombre d'analyses.

Nous verrons au chapitre suivant qu'il est possible de faire un certain nombre de transformations simples avant de commencer le traitement véritable.

2.4 Estimation de la densité

Sur la figure 4, nous pouvons voir que le tracé d'un histogramme ne fournit pas de résultats très interprétables si le nombre n d'observations est trop faible. En effet pour le tracer on doit se fixer un nombre de classes, donc un intervalle de variation h dans lequel toutes les observations ont la même densité. La densité en un point est donc la valeur n_x/nh où n_x est le nombre de points de la classe. Son calcul vérifie donc la formule générale :

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

où K est la **fonction indicatrice** de l'intervalle [-1/2 , 1/2 [: c'est-à-dire qu'elle vaut 0 hors de cet intervalle et 1 à l'intérieur. Si on applique cette formule, on peut calculer

[12] Naturellement pour ces deux moyennes $x_i > 0, \forall i$.

point par point la densité ; mais la courbe qui en résulte est peu régulière. Une idée pour la régulariser consiste à remplacer la fonction indicatrice par une autre fonction que l'on appelle **noyau** ; un noyau fréquemment employé est le noyau gaussien :

$$K(u) = \frac{1}{\sqrt{2\pi}} exp(-u^2/2)$$

Ensuite, le problème important consiste à choisir la valeur de h, que l'on appelle la **fenêtre** : trop faible elle fournira une courbe irrégulière, trop grande la courbe sera lisse. Les logiciels fournissent des options standard qui fonctionnent généralement bien.

Considérons les données archéologiques **ceram18** contenant en ligne la date de la monnaie, puis le tableau de contingence croisant les groupes techniques (soit 184 lignes) et les 132 contextes (ou ensembles stratigraphiques) étudiés en colonne (Tab.7). Les groupes techniques, aussi nommés productions, se définissent par les caractéristiques techniques des récipients que sont la nature de l'argile et le choix des couvertes (glaçure, vernis, émail, engobe…) réalisés par le potier. Les contextes archéologiques correspondent aux dépôts anthropiques circonscrits dans l'espace et dans le temps, retenus pour la qualité de leur interprétation chrono-stratigraphique et fonctionnelle (niveau d'occupation d'un bâtiment, fosse dépotoir).

TABLEAU 7 - **ceram18** : données de base, tracés de 3 histogrammes et de la densité.

```
> ceram18<-read.table("ceram18.txt",h=T, row.names=1,sep="\t")
> dim(ceram18) ; monnaie=t(ceram18[1,])
> summary(na.omit(monnaie))
         monnaie
      Min.   : 353
      1st Qu.:1072
      Median :1316
      Mean   :1188
      3rd Qu.:1468
      Max.   :1643
> monnaie<-na.omit(monnaie)/100
> par(mfrow=c(2,2))
> hist(monnaie,nclass=4,xlab= "4 classes",prob=TRUE, col="lightblue",
  border="pink")
> hist(monnaie,nclass=7,xlab= "7 classes",prob=TRUE, col="lightblue",
  border="pink")
> hist(monnaie,nclass=10,xlab= "10 classes",prob=TRUE, col="lightblue",
  border="pink")
> plot(x=c(3,16),y=c(0,0.7),type="n",bty="l",xlab="Annee/100",
  ylab="Estimation de la densite ")
> rug(monnaie)
# lines(density(monnaie,width= width.SJ(monnaie,method="dpi"),n=200), lty=1)
> lines(density(monnaie,bw=0.15),col = "blue")
> lines(density(monnaie,bw=0.25),col = "green")
> lines(density(monnaie,bw=0.5),col = "red")
> lines(density(monnaie,bw=1),col = "pink")
```

L'examen des trois histogrammes (respectivement à 4, 7 et 10 classes) est peu révélateur de la forme de la distribution ; on soupçonne que la distribution de la datation des ensembles stratigraphiques (31 datés) est assez différente d'une gaussienne. L'examen des trois histogrammes est peu révélateur de la forme de la distribution ; par

contre celui de la courbe de densité (ici avec un noyau gaussien) permet de mieux déceler deux, voire trois, grandes périodes de déposition assez bien séparées (Fig.4).

Toutes les analyses précédentes, et en particulier les examens graphiques, peuvent conduire à effectuer des **transformations de variables** pour rendre plus symétriques et moins étalées certaines distributions. On pourra ainsi prendre les logarithmes, les racines carrées et ou toute autre transformation simple. Néanmoins, il faudra prendre garde à ce que l'analyse simultanée des variables peut alors s'avérer plus délicate ; non pas au niveau technique, mais au niveau de l'interprétation. L'utilisateur saura-t-il interpréter une somme, sans doute pondérée, de ces variables transformées ? Nous en reparlerons au chapitre suivant. On peut aussi envisager de travailler directement sur les rangs de toutes les variables : on remplace pour toutes les p variables, la valeur observée par son **rang** et ensuite on applique les méthodes prévues pour des variables continues. Cette pratique, qui n'est pas celle plus traditionnelle des statistiques non paramétriques sur les rangs, fournit quelquefois des résultats fort intéressants.

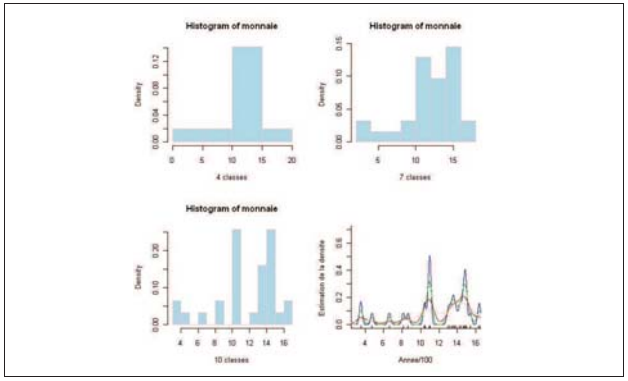

FIGURE 4. **Ceram18** *: Histogrammes et fonction de densité.*

2.5 Les variables qualitatives

Nous pouvons enfin indiquer rapidement comment faire un résumé pour des variables qualitatives. Nous n'entrerons dans les détails, mais à partir d'un exemple nous allons voir comment faire une description simple. Considérons le fichier **Poumon** (72*7) provenant d'une étude sur la byssinosis ; maladie propre aux personnes ayant travaillé sur le coton dans une atmosphère mal ventilée[13]. Le fichier comprend 72 lignes et 7 colonnes (2 variables et 5 facteurs) ; les 72 lignes du fichier représentent chacune une combinaison des niveaux des 5 facteurs (Everitt et Rabe-Hesketh, 2001). Les sept colonnes représentent :

Oui : nombre de personnes malades ;
Non : nombre de personnes non malades ;
Poussiere : Haut, Moyen, Bas ;
Race : Blanc, Autre ;
Sexe : Homme, Femme ; Tabagie : Fumeur, Non-Fumeur ;
Tempsemploi : <10 ans, 10-20 ans, >20 ans.

On peut avoir des résumés numériques (dénombrements par niveaux de facteurs, par combinaisons de niveaux de facteurs) ou des graphiques comme des histogrammes ou des représentations en « camembert »[14], cf. Tab.8.

TABLEAU 8 - Quelques analyses des données du fichier **Poumon**.

```
> Poumon<-read.table("Poumon.txt",h=T,sep="\t")
> dim(Poumon)
     [1] 72  7
> attach(Poumon)>
# Conversion de variables en facteurs
> Poumon$Poussiere<-
  factor(Poumon$Poussiere,levels=1:3,labels=c("Haut","Moyen","Bas"))
> Poumon$Race<-factor(Poumon$Race,levels=1:2,labels=c("Blanc","Autre"))
> Poumon$Sexe<-factor(Poumon$Sexe,levels=1:2,labels=c("Homme","Femme"))
> Poumon$Tabagie<-factor(Poumon$Tabagie,levels=1:2,
  labels=c("Fumeur","Non-Fumeur"))
> Poumon$Tempsemploi<-factor(Poumon$Tempsemploi,levels=1:3,
  labels=c("<10 ans","10-20 ans","> 20 ans"))
> xtabs(Oui~Race+Poussiere)
         Poussiere
     Race  1  2  3
        1 49 12 31
        2 56  6 11
> summary(xtabs(Oui~Race+Poussiere))## Avec test d'independance ...
     Call: xtabs(formula = Oui ~ Race + Poussiere)
     Number of cases in table: 165
     Number of factors: 2
     Test for independence of all factors:
            Chisq = 9.934, df = 2, p-value = 0.006963
> xtabs(Non~Race+Poussiere)
         Poussiere
     Race    1    2    3
        1  218  843 2363
        2  346  439 1045
> library(Hmisc)
```

[13] On l'appelle aussi *maladie brune des poumons* ou *fièvre du lundi*.
[14] En anglais : *pie chart*.

```
> describe(Oui~Race+Poussiere)
      Oui ~ Race + Poussiere
      3  Variables      72  Observations
--------------------------------------------------------------------------------
Oui
      n missing  unique    Mean     .05     .10     .25     .50     .75     .90
     72       0      12   2.292     0.0     0.0     0.0     1.0     3.0     4.9
    .95
    8.9
              0  1  2  3 4 5 6 8 10 12 25 31
Frequency 34 12  7 10 1 1 1 2  1  1  1  1
%         47 17 10 14 1 1 1 3  1  1  1  1
--------------------------------------------------------------------------------
Race
      n missing  unique    Mean
     72       0       2     1.5
1 (36, 50%), 2 (36, 50%)
--------------------------------------------------------------------------------
Poussiere
      n missing  unique    Mean
     72       0       3       2
1 (24, 33%), 2 (24, 33%), 3 (24, 33%)
--------------------------------------------------------------------------------
# Trouve les proportions de malade par Race et Poussiere
> prop<-tapply(Oui,list(Race,Poussiere),sum)/tapply(Non,list(Race,Poussiere),sum)
> round(prop,3)
             1     2     3
  1 0.225 0.014 0.013
  2 0.162 0.014 0.011
# Tracé de la figure 5
> barplot(prop,names= c("Haut","Moyen","Bas"),legend= c("Blanc","Autre"))
```

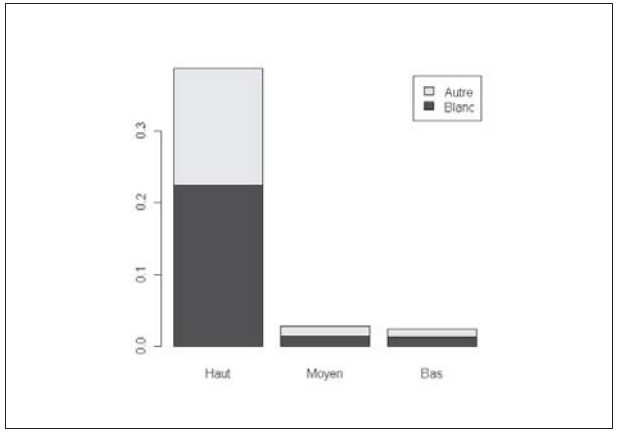

FIGURE 5. Graphique (barplot) de la proportion de malades suivant la race et le niveau de poussière.

3. RESUMES MULTIDIMENSIONNELS

3.1 Espace de représentation : point moyen $\overline{x}$

Maintenant, nous allons étudier le tableau dans son ensemble pour faire apparaître des relations entre les variables et entre les observations. Auparavant, il nous faut définir des espaces de représentation :

- celui des observations est de dimension p, noté $\mathbb{R}^p$. Chaque observation n°i est représentée par ses coordonnées sur les p axes de base de cet espace ; les valeurs de ces coordonnées sont celles des p variables de cette observation. C'est le vecteur ligne observation $\mathbf{x}_i^{\mathrm{T}} \in \mathbb{R}^p$.
- celui des variables est de dimension n, noté $\mathbb{R}^n$. Chaque variable n°j est représentée par ses coordonnées sur les n axes de base de cet espace ; les valeurs de ces coordonnées sont celles des n observations de cette variable. C'est *le* vecteur colonne variable $\mathbf{x}^j \in \mathbb{R}^n$.

Le point moyen ou **barycentre** du nuage des n points, aussi appelé **centre de gravité**[15], a comme coordonnées celles du vecteur moyenne $\overline{\boldsymbol{x}}$, calculées par :

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{1}_{n\times 1} \in \mathbb{R}^p$$

où $\mathbf{1}_{n\times 1} \in \mathbb{R}^n$ est un vecteur de n lignes formé de 1.

En retranchant à chaque colonne sa moyenne, on obtient la *matrice des données centrées* :

$$\mathbf{X_C}^{\mathrm{T}} = [\boldsymbol{x}_1 - \overline{\boldsymbol{x}} \quad \cdots \quad \boldsymbol{x}_{\mathrm{n}} - \overline{\boldsymbol{x}}] = \mathbf{X}^{\mathrm{T}} - \overline{\boldsymbol{x}}\mathbf{1}_{1\times n} = \mathbf{X}^{\mathrm{T}}\left[\mathbf{I_n} - \frac{1}{n}\mathbf{1}_{n\times 1}(\mathbf{1}_{n\times 1})^{\mathrm{T}}\right]$$
$$= \mathbf{X}^{\mathrm{T}}\left[\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n}\right]$$

où $\mathbf{1}_{n\times n}$ matrice formée de 1 et $\mathbf{I}_{\mathrm{n}} - \frac{1}{n}\mathbf{1}_{n\times n}$ sont des matrices carrées de dimension n ; la première est de rang 1 et la seconde de rang $n-1$. Ces matrices carrées sont **idempotentes** ($\mathbf{A}\times\mathbf{A} = \mathbf{A}$) et symétriques $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$; ce sont des **matrices de projection**. En transposant la dernière équation, on a : $\mathbf{X_C} = \left[\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n}\right]\mathbf{X}$.

L'instruction *R* permettant de centrer et/ou réduire les variables de $\mathbf{X}$ est :

```
> scale(x, center = TRUE, scale = TRUE)
```

Généralement, en option standard, toutes les valeurs d'une colonne sont centrées par rapport à leur moyenne et divisées par l'écart-type estimé ; les données manquantes éventuelles **NA** sont omises et la somme des carrés des écarts à la moyenne est divisée par *n-1* pour calculer l'écart-type *s*.

Si toutes les observations ont la même importance, on dit qu'elles ont le même **poids**, c'est-à-dire $1/n$. Quelquefois, il peut être utile de leur donner des poids différents,

[15] Aussi appelé *centroïde*.

p_i à chaque observation $\boldsymbol{x}_i$. Ces poids sont des nombres positifs dont la somme est égale à 1 ; ils jouent un rôle analogue à celui de fréquence. On les regroupe dans une matrice diagonale $\mathbf{D}$ de dimension n :

$$\mathbf{D} = \begin{bmatrix} p_1 & 0 & \cdots & 0 & 0 \\ 0 & p_2 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \cdots \\ 0 & 0 & \cdots & 0 & p_n \end{bmatrix} \text{ avec } \sum_{i=1}^{n} p_i = 1$$

Avec un même poids pour chaque observation, on a bien sûr $\mathbf{D} = \mathbf{I}_n/n$. Avec pondération, le vecteur moyenne est calculé par :

$$\bar{\boldsymbol{x}} = \mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{1}_{n\times 1} \in \mathbb{R}^p$$

et la matrice centrée par :

$$\mathbf{X_C} = [\mathbf{I}_n - \mathbf{1}_{n\times n}\mathbf{D}]\mathbf{X}$$

3.2 Première transformation : matrice de dispersion

Si maintenant, à partir de la matrice des données centrées, nous calculons :

$$\mathbf{S}_{p\times p} = \mathbf{X_C}^{\mathrm{T}}\mathbf{D}\mathbf{X_C} = \left[v_{kj} = \sum_{i=1}^{n} p_i\left(x_i^j - \bar{x}^j\right)\left(x_i^k - \bar{x}^k\right)\right]$$

$\mathbf{S}$ est une matrice carrée symétrique, de dimension $p \times p$; elle contient sur la diagonale les p variances déjà vues plus haut et, hors diagonale les covariances. Elle s'appelle la **matrice de variances-covariances** ou **matrice de dispersion**. Elle va jouer un rôle fondamental dans toute l'analyse linéaire des données.

Souvent on travaille sur des données non seulement centrées, mais aussi réduites. En effet, les variables ne sont pas toujours mesurées avec les mêmes unités, il peut donc être utile de se ramener à des variables plus comparables. Pour ce faire, on les transforme pour qu'elles aient la même variabilité ; il suffit de diviser les colonnes de la matrice de données par leur écart-type. Formellement, pour y arriver nous prenons les p éléments de la diagonale de $\mathbf{S}$, soit $\mathbf{diag}(\mathbf{S})$ la matrice diagonale formée par ces éléments. On a alors la matrice centrée réduite, celle que nous obtenons par la fonction `scale` de *R*, par :

$$\mathbf{X_{CR}} = \mathbf{X_C}\left(\mathbf{diag}(\mathbf{S})\right)^{-1/2}$$

Cette matrice est telle que :

$$\sum_{i=1}^{n} [x_{CR}]_i^j = 0\ ;\ \sum_{i=1}^{n}\left([x_{CR}]_i^j\right)^2 = n-1$$

Les variances sont alors toutes égales à 1. Quelquefois, il peut être utile d'avoir une matrice centrée réduite légèrement différente :

$$\mathbf{X_{CRu}} = \frac{1}{(n-1)^{1/2}}\mathbf{X_C}\left(\mathbf{diag}(\mathbf{S})\right)^{-1/2}$$

de telle sorte que les sommes des colonnes soient nulles, et les sommes des carrés des éléments de la colonne soient égales à 1 :

$$\sum_{i=1}^{n} [x_{CRu}]_i^j = 0\ ;\ \sum_{i=1}^{n}\left([x_{CRu}]_i^j\right)^2 = 1$$

Enfin, on obtient la **matrice des coefficients de corrélation linéaire.** qui résume l'ensemble des dépendances linéaires entre les p variables :

$$\mathbf{R} = \mathbf{X_{CRu}}^{\mathrm{T}}\mathbf{X_{CRu}} = \frac{1}{n}\mathbf{X_{CR}}^{\mathrm{T}}\mathbf{X_{CR}} = \frac{1}{n}\left(\mathbf{diag}(\mathbf{S})\right)^{-1/2}\mathbf{X_C}^{\mathrm{T}}\mathbf{X_C}\left(\mathbf{diag}(\mathbf{S})\right)^{-1/2}$$

Sous *R*, on peut utiliser la figure 6, appelée faute de mieux ***scatterplot***[16] pour résumer un certain nombre de caractéristiques unidimensionnelles et multidimensionnelles du corpus étudié. Les lignes de code suivantes permettent de l'obtenir sous *R* (cf. Tab. 9).

TABLEAU 9 - **Eaux1** : tracé des caractéristiques unidimensionnelles et multidimensionnelles.

```
#SCATTERPLOT
# Histogrammes sur la diagonale
> panel.hist <- function(x, ...)
+{
+    usr <- par("usr"); on.exit(par(usr))
+    par(usr = c(usr[1:2], 0, 1.5) )
+    h <- hist(x, plot = FALSE,col="lightblue")
+    breaks <- h$breaks; nB <- length(breaks)
+    y <- h$counts; y <- y/max(y)
+    rect(breaks[-nB], 0, breaks[-1], y, col="cyan", ...)
+}
# Coefficients de corrélation (en valeur absolue) sur la partie haute,
# Taille de police proportionnelle au coefficient de correlation
> panel.cor <- function(x, y, digits=2, prefix="", cex.cor, ...)
+{
+    usr <- par("usr"); on.exit(par(usr))
+    par(usr = c(0, 1, 0, 1))
+    r <- abs(cor(x, y))
+    txt <- format(c(r, 0.123456789), digits=digits)[1]
+    txt <- paste(prefix, txt, sep="")
+    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
+    text(0.5, 0.5, txt, cex = cex.cor * r)
+}
# Régression linéaire (lm) sur la partie basse ;
# d'où le signe du coefficient de correlation.
> panel.cor <- function(x, y, digits=2, prefix="", cex.cor, ...)
> panel.lm=function(x,y){
+    points(x,y)
+    abline(lm(y~x))
+    #lines(lowess(y,x),col="red")
+}
# Fig. 6
> pairs(Eaux1, diag.panel=panel.hist, cex.labels = 2,
  font.labels=2, upper.panel=panel.cor,lower.panel=panel.lm )
```

[16] On pourrait parler de *nuage de points* ou de *dispersion des points*.

La figure 6 produit une matrice de *scatterplot*s contenant sur sa diagonale l'histogramme de chaque variable ; sur sa partie inférieure, les graphiques des couples avec droite des moindres carrés et sur sa partie supérieure, les coefficients de corrélation indiqués avec une taille de police plus ou moins importante selon l'intensité de leur valeur absolue.

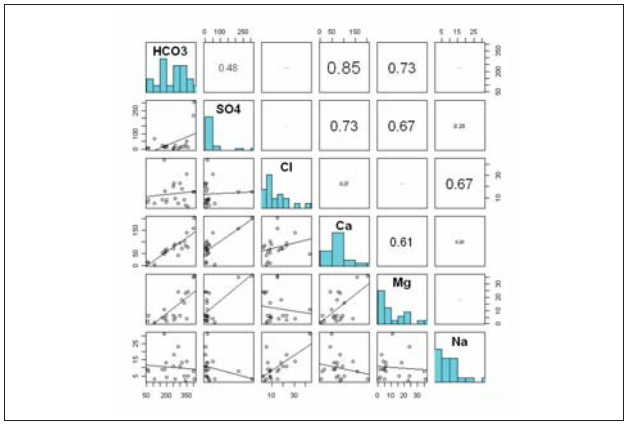

FIGURE 6. **Eaux1** *: scatterplot.*

3.3 L'espace des observations $\mathbb{R}^p$: regards sur les lignes, inertie

Un des objectifs des analyses de tableau de données comme $\mathbf{X}$, consiste bien souvent à regarder si deux lignes du tableau par exemple les deux premières, c'est-à-dire deux vecteurs ligne $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$ de $\mathbb{R}^p$sont proches ou éloignées. Il est donc important de calculer une distance entre elles. Quand les unités de mesure des variables sont les mêmes, la façon la plus simple consiste à appliquer le théorème de Pythagore qui donne le carré de la **distance** entre deux observations :

$$d^2(\mathbf{x}_1,\mathbf{x}_2) = (x_1^1 - x_2^1)^2 + \cdots + \left(x_1^p - x_2^p\right)^2 = (\boldsymbol{x}_1 - \boldsymbol{x}_2)^{\mathrm{T}}(\boldsymbol{x}_1 - \boldsymbol{x}_2)$$

En statistique, le problème est un peu plus compliqué car, bien souvent les unités de mesure sont différentes ; on peut avoir un mélange de mensurations, de poids et d'âge. Il faut donc pondérer, d'une certaine manière, les différents carrés de l'expression précédente. De plus, la distance calculée n'a de sens que pour des axes perpendiculaires ; on aurait aussi bien pu prendre des axes obliques.

On utilise donc une formulation un peu plus générale en introduisant une matrice définie positive $\mathbf{Q}$, symétrique de dimension $p \times p$. On dit que l'on a muni l'espace des

observations d'un **produit scalaire** : $\langle \boldsymbol{x}_1; \boldsymbol{x}_2 \rangle_{\mathbf{Q}} = \boldsymbol{x}_1{}^{\mathrm{T}}\mathbf{Q}\boldsymbol{x}_2$, et le carré de la distance est définie par la forme quadratique :

$$d^2(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}_1 - \boldsymbol{x}_2)^{\mathrm{T}}\mathbf{Q}(\boldsymbol{x}_1 - \boldsymbol{x}_2)$$

Prendre $\mathbf{Q} = \mathbf{I}_p$ revient à conserver les unités initiales de chaque variable et, donc, à rendre dominantes les variables de plus grande variabilité.

Généralement, on prend $\mathbf{Q} = \left(\mathbf{diag}(\mathbf{S})\right)^{-1}$ qui est la métrique inverse des variances. Les distances sont alors calculées non plus sur $\mathbf{X}$, mais sur la matrice centrée réduite $\mathbf{X_{CR}}$: on impose une variance égale à 1 à chaque variable. Utiliser une métrique diagonale quelconque :

$$\mathbf{Q_a} = \mathbf{diag}(a_1, \cdots, a_p)$$

revient à multiplier les variables par $\sqrt{a_j}$ et à utiliser ensuite la métrique $\mathbf{Q} = \mathbf{I}_p$.

On peut généraliser ce résultat à une métrique quelconque. On sait que toute matrice $\mathbf{Q}$ peut s'écrire $\mathbf{Q} = \mathbf{T}^{\mathrm{T}}\mathbf{T}$. Le produit scalaire entre deux observations s'en déduit :

$$\langle \boldsymbol{x}_1; \boldsymbol{x}_2 \rangle_{\mathbf{Q}} = \boldsymbol{x}_1{}^{\mathrm{T}}\mathbf{Q}\boldsymbol{x}_2 = \boldsymbol{x}_1{}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{T}\boldsymbol{x}_2 = (\mathbf{T}\boldsymbol{x}_1)^{\mathrm{T}}(\mathbf{T}\boldsymbol{x}_2)$$

La suite revient à utiliser la métrique $\mathbf{I}_p$ sur les données transformées $\mathbf{X}\mathbf{T}^{\mathrm{T}}$.

Regardons maintenant ce que représente, pour une matrice de données $\mathbf{X}$, munie de sa métrique $\mathbf{Q}$, la somme I_g, appelée **inertie**, du carré des distances pondérées des vecteurs observations au point moyen du nuage :

$$I_g = \sum_{i=1}^{n} p_i d^2(\boldsymbol{x}_i, \overline{\boldsymbol{x}}) = \sum_{r=1}^{n}\sum_{s=1}^{n} p_r p_s (\boldsymbol{x}_{\mathrm{r}} - \boldsymbol{x}_{\mathrm{s}})^{\mathrm{T}}\mathbf{Q}\,(\boldsymbol{x}_r - \boldsymbol{x}_{\mathrm{s}})\,;\, r < s$$

La forme de l'inertie I_g nous dit que la somme pondérée du carré des distances des observations au point moyen est égale à la somme pondérée du carré des distances entre tous les couples d'observations. Regardons ce que cela donne pour le cas d'une seule variable ($p = 1$). Avec la définition de la variance (dans la formule donnant s^2, le diviseur est $n-1$ et non pas n) I_g est égale à $((n-1)/n)\ var(\boldsymbol{x})$. La variance est donc une mesure de la cohésion interne des observations. Vérifions le sur un exemple ; avec les six observations : $(19; 23; 24; 26; 28; 30)$ de moyenne $\bar{x} = 25$ et la variance vaut 76/5 ; la somme de tous les carrés des différences entre les six valeurs vaut 456 (=76×6).

L'inertie peut aussi s'exprimer en fonction de la trace de la matrice $\mathbf{QS}$ ou de celle de la matrice $\mathbf{SQ}$:

$$I_g = tr\left(\sum_{i=1}^{n} p_i (\boldsymbol{x}_{\mathrm{i}} - \overline{\boldsymbol{x}})^{\mathrm{T}}\mathbf{Q}\,(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\right) = tr\left(\sum_{i=1}^{n} \mathbf{Q} p_i (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^{\mathrm{T}}\,(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\right) = tr(\mathbf{QS})$$

Donc :

- si $\mathbf{Q} = \mathbf{I}_p$ l'inertie est égale à la somme des p variances,
- si $\mathbf{Q} = \left(\mathbf{diag}(\mathbf{S})\right)^{-1}$, l'inertie est égale à p, le nombre de variables ; elle est indépendante du nombre de variables.

3.4 L'espace des variables $\mathbb{R}^n$: regards sur les colonnes

Regardons maintenant le **vecteur variable** $\boldsymbol{x}^j \in \mathbb{R}^n$. Que représente la proximité de deux vecteurs $\boldsymbol{x}^j$ et $\boldsymbol{x}^k$? Nous les supposons centrés et nous prenons comme métrique la matrice diagonale des poids $\mathbf{D}$ définie positive. Nous voyons que le produit scalaire entre ces deux vecteurs colonne $\langle \boldsymbol{x}^j; \boldsymbol{x}^k \rangle_{\mathbf{D}} = (\boldsymbol{x}^j)^{\mathrm{T}} \mathbf{D} \boldsymbol{x}^k = \sum_{i=1}^{n} p_i \left(x_i^j - \bar{x}^j\right)\left(x_i^k - \bar{x}^k\right)$ est proportionnel à la covariance $((n-1)/n)s_{jk}$ entre les deux variables. Le carré de la $\mathbf{D}$ norme (c'est-à-dire le carré de sa longueur) de chaque vecteur est proportionnelle à leur variance : $\left\|\boldsymbol{x}^j\right\|_{\mathbf{D}}^2 = ((n-1)/n){s_j}^2$; leur longueur est donc égale à l'écart-type. L'angle θ_{jk} entre les deux vecteurs est donné par :

$$cos\left(\theta_{jk}\right) = \frac{\langle \mathbf{x}^{\mathrm{j}}; \mathbf{x}^{\mathrm{k}} \rangle_{\mathbf{D}}}{\|\mathbf{x}^{\mathrm{j}}\|_{\mathbf{D}} \|\boldsymbol{x}^k\|_{\mathbf{D}}}$$

Le cosinus de l'angle entre deux vecteurs variable centrés est donc égal au coefficient de corrélation entre ces deux variables.

En conclusion, dans l'espace des observations, on s'intéresse aux distances entre points[17] et dans celui des variables, à l'angle entre variables.

3.5 Transformation linéaire d'un tableau de données

Les analyses que nous allons étudier ultérieurement consistent, dans la majorité des cas, à trouver des combinaisons linéaires des p variables définissant les colonnes de $\mathbf{X}$. Nous chercherons des combinaisons linéaires qui ont des propriétés particulières. Ces combinaisons linéaires seront des transformations des vecteurs $\boldsymbol{x}^j$ $(j = 1, \cdots, p)$ qui vont fournir de nouvelles variables. A chaque vecteur colonne $\boldsymbol{x}^j$ nous pouvons associer un axe de l'espace des observations $\mathbb{R}^p$; une transformation linéaire consiste à projeter les observations sur un nouvel axe de vecteur unitaire ligne $\boldsymbol{a}$. La coordonnée de l'observation i sur cet axe sera définie par $y_i = \boldsymbol{a}^{\mathrm{T}} \mathbf{Q} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) = \langle \boldsymbol{a}; \mathbf{x}_{\mathrm{i}} - \bar{\mathbf{x}} \rangle_{\mathbf{Q}}$ et l'ensemble des n coordonnées par :

$$\mathbf{y} = \mathbf{XQa} = \mathbf{Xu} = \sum_{j=1}^{p} \mathbf{x}^j \mathbf{u}_j \in \mathbb{R}^n \text{ avec } \mathbf{u} = \mathbf{Qa}$$

Nous avons donc créé une nouvelle variable avec :

- un axe de vecteur ligne unitaire $\mathbf{a} \in \mathbb{R}^p$;
- un vecteur $\mathbf{y}$ de l'espace des variables $\mathbb{R}^n$;
- une forme linéaire $\mathbf{u} \in \mathbb{R}^p$ qui est appelée **facteur**.

Si, pour simplifier, nous supposons $\mathbf{Q} = \mathbf{I}_p$ et si nous faisons q transformations de l'ensemble du tableau $\mathbf{X}_{n \times p}$ définies par la matrice $\mathbf{A} = [\boldsymbol{a}^1, \cdots, \boldsymbol{a}^q]$ telles que :

$$\mathbf{Y}_{n \times q} = \mathbf{X}_{n \times p} \mathbf{A}_{p \times q}$$

on peut montrer que la matrice de dispersion $\mathbf{S}_{\mathbf{Y}}$ de $\mathbf{Y}$ se déduit de $\mathbf{S}_{\mathbf{X}}$ celle de $\mathbf{X}$ par :

[17] Au chapitre consacré à la classification, nous donnerons d'autres types de distance. Nous donnerons aussi une extension de la notion de distance à la distance non plus entre observations mais à la distance entre tableaux en développant l'analyse procustéenne (chapitre 10).

$$\mathbf{S}_{\mathrm{Y}_{q \times q}} = \mathbf{A}^{\mathrm{T}}{}_{q \times p} \mathbf{S}_{\mathbf{X}_{p \times p}} \mathbf{A}_{p \times q}$$

Il est souvent utile d'imposer des conditions à la matrice de transformation $\mathbf{A}$:

- les q vecteurs sont unitaires : $\|\boldsymbol{a}^k\|^2 = 1; k = 1, \cdots, q$;
- les q vecteurs sont orthogonaux deux à deux : $(\boldsymbol{a}^k)^{\mathrm{T}} \boldsymbol{a}^l = 0; k \neq l = 1, \cdots, q$.

Quand $p = q$, la matrice $\mathbf{A}$ est ce que l'on appelle une **matrice orthogonale** : $\mathbf{A}\mathbf{A}^{\mathrm{T}} = \mathbf{I}_p$ et $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}_q$; la transformation est une simple **rotation** du repère de départ. Chaque vecteur observation $\boldsymbol{x}_i$ est alors dans ce cas transformé en autre vecteur observation $\boldsymbol{y}_i = \mathbf{A}^{\mathrm{T}} \boldsymbol{x}_i \in \mathbb{R}^p$. Chaque composante du nouveau vecteur est la projection de $\boldsymbol{x}_i$ sur les $p = q$ vecteurs ligne $\boldsymbol{a}^k \in \mathbb{R}^p; k = 1, \cdots, p$. Une propriété intéressante et importante est que la distance entre deux observations ligne est invariante dans une transformation linéaire :

$$(\boldsymbol{y}_r - \boldsymbol{y}_s)^{\mathrm{T}} (\boldsymbol{y}_r - \boldsymbol{y}_s) = (\boldsymbol{x}_r - \boldsymbol{x}_s)^{\mathrm{T}} \mathbf{A}\, \mathbf{A}^{\mathrm{T}} (\boldsymbol{x}_r - \mathbf{x}_s) = (\boldsymbol{x}_r - \boldsymbol{x}_s)(\boldsymbol{x}_r - \boldsymbol{x}_s)^{\mathrm{T}}$$

4. DECOMPOSITION D'UNE MATRICE DE DONNEES

Dans ce paragraphe, nous allons présenter quelques outils mathématiques d'un grand intérêt pour l'analyse d'un tableau de données. Dès maintenant, ils vont mettre en évidence la dualité entre les deux espaces de représentation. Ils permettent déjà de vérifier qu'il n'y a pas de redondance particulière entre les variables qui définissent le tableau et qu'il n'y a pas des relations trop fortes, voire fonctionnelles entre elles.

4.1 Décomposition en Valeurs Singulières (*DVS*)

A partir d'une matrice $\mathbf{X}_{n \times n} = \left[x_i^j\right]; i = 1, \cdots, n; j = 1, \cdots, p$; il est toujours possible de faire la décomposition suivante (théorème d'approximation de Schmidt (1907) et Eckart & Young (1936)) :

$$x_i^j = \theta_1 u_i^1 v_j^1 + \cdots + \theta_r u_i^r v_j^r = \sum_{k=1}^{r} \theta_k u_i^k v_j^k \; ; \; \theta_1 \geq \theta_2 \geq \cdots \geq \theta_r > 0$$

Cette décomposition s'appelle la **décomposition en valeurs singulières** de $\mathbf{X}$; les valeurs singulières de cette décomposition sont *uniques*, mais pas obligatoirement distinctes. Le nombre r de termes de $\mathbf{X}$ est le *rang* de la matrice ; il ne peut pas dépasser le plus petit des deux nombres n ou p : $r = min(n, p)$. En général, on suppose que le nombre d'observations est supérieur au nombre de variables : $n > p$, donc $r \leq p$. Si $r = p$, on dit que la matrice est de **plein rang**. Les quantités u_i^k et v_j^k sont les composantes de vecteurs unitaires $\boldsymbol{u}^k$ et $\boldsymbol{v}^k$ de dimension respective n et p :

$$\sum_{i=1}^{n} \left(u_i^k\right)^2 = \sum_{j=1}^{p} \left(v_j^k\right)^2 = 1; k = 1, \cdots, r$$

Les vecteurs $\boldsymbol{u}^k$ sont orthogonaux entre eux, de même que les vecteurs $\boldsymbol{v}^k$, soit $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_{\mathrm{r}}$.

On peut donc écrire :

$$\mathbf{X}_{n\times p} = \mathbf{U}_{n\times r}\mathbf{\Theta}_{\mathrm{r}\times\mathrm{r}}\left(\mathbf{V}_{\mathrm{p}\times\mathrm{r}}\right)^{\mathrm{T}}$$

où :

- La matrice $\mathbf{\Theta}_{\mathrm{r}\times\mathrm{r}} = \mathbf{diag}\left(\sqrt{\lambda_1},\cdots,\sqrt{\lambda_r}\right)$ est une matrice diagonale contenant les valeurs singulières de la matrice $\mathbf{X}_{n\times p}$; c'est-à-dire les racines carrées des valeurs propres positives et non nulles[18] λ_k $(k = 1,\cdots,r)$ de $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ou de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$;
- $\mathbf{U}_{n\times r} = [\boldsymbol{u}^1 \ \vdots \ \cdots \ \vdots \ \boldsymbol{u}^r]$ matrice unitaire $n \times r$, telle que $\boldsymbol{u}^k$ est vecteur propre de $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ associé à la valeur propre non nulle λ_k $(k = 1,\cdots,r)$;
- $\mathbf{V}_{p\times r} = [\boldsymbol{v}^1 \ \vdots \ \cdots \ \vdots \ \boldsymbol{v}^r]$ matrice unitaire p × r, telle que $\boldsymbol{v}^k$ est vecteur propre de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ associé à la valeur propre non nulle λ_k $(k = 1,\cdots,r)$.

La relation fournissant $\mathbf{X}$ signifie que ses lignes sont des combinaisons linéaires des lignes $\mathbf{U}$ et ses colonnes des combinaisons linéaires des colonnes de $\mathbf{V}$. Les colonnes de $\mathbf{U}$ forment donc une **base orthogonale** pour l'espace des colonnes $\mathbb{R}^n$ de $\mathbf{X}$; c'est-à-dire le sous-espace de dimension r de $\mathbb{R}^n$ engendré par les colonnes de $\mathbf{X}$. Les colonnes de $\mathbf{V}$ forment de même une base orthogonale pour l'espace $\mathbb{R}^p$ des lignes de $\mathbf{X}$.

L'avantage de cette décomposition est exploité en statistique car elle est liée au système des valeurs propres de la matrice $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ (qui est directement liée, après centrage de $\mathbf{X}$, à la matrice de dispersion $\mathbf{S}$). Les θ_k sont les racines carrées des valeurs propres non nulles de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ aussi bien que de $\mathbf{X}\mathbf{X}^{\mathrm{T}}$. Les colonnes de $\mathbf{U}$ sont les vecteurs propres de $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ et les colonnes de $\mathbf{V}^{\mathbf{T}}$ ceux de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Nous verrons que nombre de calculs statistiques font intervenir soit l'inversion soit la diagonalisation de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$; ces calculs s'expriment simplement en fonction des éléments de la décomposition singulière. Par exemple, voici deux expressions que nous utiliserons plus loin :

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \mathbf{V}\mathbf{\Theta}^{-2}\mathbf{V}^{\mathrm{T}}$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} = \mathbf{U}\mathbf{U}^{\mathrm{T}}$$

La première est une autre forme de la matrice qui intervient dans la résolution des équations du modèle linéaire. On voit apparaître le rang r de la matrice des **équations normales** $\mathbf{X}^{\mathrm{T}}\mathbf{X}$: c'est le nombre de valeurs propres non nulles, les r premiers éléments, de la matrice diagonale $\mathbf{\Theta}$. On notera que sous *R*, la fonction `svd` permet d'effectuer une *DVS*. Dans la formulation ci-dessus $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ et $\mathbf{\Theta}$ doivent être inversibles et donc $r = p$.

4.2 Inverse généralisée d'une matrice

Il est quelquefois utile de faire apparaître les relations linéaires entre les colonnes de $\mathbf{X}$; c'est ce que l'on appelle la **colinéarité**. Ceci peut arriver quand, dans un fichier de données, une variable est la somme de deux autres. Quand le rang de $\mathbf{X}_{n\times p}$ vaut r, en supposant que le nombre d'observations n est plus grand que le nombre de variables p, il existe $p - r$ relations linéaires pas toujours facilement décelables. En pratique, il peut exister une **pseudo-colinéarité** qui traduit des relations étroites entre certaines colonnes

[18] En pratique, seules les valeurs propres non nulles sont utilisées ; néanmoins, les valeurs propres nulles peuvent quelquefois avoir un intérêt en analyse des données.

de $\mathbf{X}$. Dans le premier cas, on ne peut pas inverser une matrice, dans le second le résultat est numériquement imprécis.

Pour traiter cette question, il est utile d'introduire la notion d'**inverse généralisée**, dont le calcul est simple à partir de la décomposition singulière. L'inverse généralisée, dite inverse de Moore-Penrose, d'une matrice quelconque $\mathbf{X}_{n\times p}$, notée $\mathbf{X}^{-}$, a les propriétés suivantes :

$$\begin{cases} 1 : \mathbf{X}\mathbf{X}^{-}\mathbf{X} = \mathbf{X} \\ 2 : \mathbf{X}^{-}\mathbf{X}\mathbf{X}^{-} = \mathbf{X}^{-} \\ 3 : \mathbf{X}^{-}\mathbf{X} = (\mathbf{X}^{-}\mathbf{X})^{\mathrm{T}} \\ 4 : \mathbf{X}\mathbf{X}^{-} = (\mathbf{X}\mathbf{X}^{-})^{\mathrm{T}} \end{cases}$$

Cette inverse est *unique*, et elle constitue une généralisation de l'inverse classique. Si $\mathbf{X}$ est une matrice carrée $(n = p)$ inversible, on a :

$$\mathbf{X}^{-} = \mathbf{X}^{-1}$$

Si le rang de $\mathbf{X}$ est égal au nombre de colonnes $(r = p)$:

$$\mathbf{X}^{-} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$$

Si dans la décomposition singulière de $\mathbf{X}$, on fait apparaître les valeurs singulières nulles :

$$\mathbf{X}_{n\times p} = \mathbf{U}_{n\times n}\mathbf{\Theta}_{n\times p}\left(\mathbf{V}_{p\times p}\right)^{\mathbf{T}} \text{ ; où } \theta_{r+1} = \cdots = \theta_p = 0$$

On peut démontrer que :

$$\mathbf{X}^{-} = \mathbf{V}\mathbf{\Theta}^{-}\mathbf{U}^{\mathrm{T}}$$

où $\mathbf{\Theta}^{-}$ est l'inverse généralisée de $\mathbf{\Theta}$, obtenue en prenant dans cette dernière matrice l'inverse des éléments diagonaux non nuls. Donc $\mathbf{X}^{-}$ joue donc un rôle analogue à $\mathbf{X}^{-1}$ dans la résolution de systèmes rectangulaires d'équations linéaires. Si nous avons le système $\boldsymbol{y} = \mathbf{X}\boldsymbol{z}$, où $\boldsymbol{y} \in \mathbb{R}^n$ et $\mathbf{z} \in \mathbb{R}^p$, on peut montrer que $\boldsymbol{z} = \mathbf{X}^{-}\boldsymbol{y}$ est une solution particulière de ce système.

4.3 Exemples numériques

Nous allons illustrer ces résultats par deux exemples numériques que nous réutiliserons ultérieurement.

a) $\mathbf{X}$ est de plein rang : soit la matrice $\mathbf{X}_{\mathrm{a}}$ $(n = 4; p = 2)$ et un vecteur $\boldsymbol{y}_{\mathrm{a}}$:

$\mathbf{X}_{\mathrm{a}}$		$\boldsymbol{y}_{\mathrm{a}}$
1	1	1
1	2	2
1	3	3
1	10	6

Les différents calculs des éléments de la décomposition singulière de sont présentés ci-dessous (Tab.10), ainsi que la résolution du système $\boldsymbol{y}_{\mathrm{a}} = \mathbf{X}_{\mathrm{a}}\boldsymbol{z}_{\mathrm{a}}$.

TABLEAU 10 - Exemple de calcul des matrices de la décomposition singulière de $X_{4\times2}$

```
> Xa<-c(1,1,1,1,1,2,3,10)
> dim(Xa)<-c(4,2)
> A<-svd(Xa)
# L'instruction round(,d) permet de n'imprimer que d decimals.
> round(A$d,5)
[1] 10.78332  1.31148
> round(A$u,5)
             [,1]     [,2]
[1,] -0.10489 -0.64730
[2,] -0.19670 -0.53973
[3,] -0.28851 -0.43216
[4,] -0.93117  0.32083
> round(A$v,5)
             [,1]     [,2]
[1,] -0.14108 -0.99000
[2,] -0.99000  0.14108
> round(A$d^2,5)
[1] 116.28001   1.71999
> round(eigen(t(Xa)%*%Xa)$values,5)
[1] 116.28001   1.71999
> round(eigen(t(Xa)%*%Xa)$vectors,5)
             [,1]     [,2]
[1,] 0.14108  0.99000
[2,] 0.99000 -0.14108
> t(Xa)%*%Xa
     [,1] [,2]
[1,]    4   16
[2,]   16  114
> A$v%*%diag(A$d^2)%*%t(A$v)
     [,1] [,2]
[1,]    4   16
[2,]   16  114
> A$u%*%t(A$u)# Matrice H
           [,1] [,2] [,3]  [,4]
[1,]  0.43 0.37 0.31 -0.11
[2,]  0.37 0.33 0.29  0.01
[3,]  0.31 0.29 0.27  0.13
[4,] -0.11 0.01 0.13  0.97
> Xa%*%solve(t(Xa)%*%Xa)%*%t(Xa)
# Forme directe de la matrice H
           [,1] [,2] [,3]  [,4]
[1,]  0.43 0.37 0.31 -0.11
[2,]  0.37 0.33 0.29  0.01
[3,]  0.31 0.29 0.27  0.13
[4,] -0.11 0.01 0.13  0.97
> solve(t(Xa)%*%Xa)%*%t(Xa)
# Inverse généralisée X+ de X
           [,1]  [,2]  [,3]  [,4]
[1,]  0.49  0.41  0.33 -0.23
[2,] -0.06 -0.04 -0.02  0.12
> ya<-c(1,2,3,6)
> dim(ya)<-c(4,1)
> solve(t(Xa)%*%Xa)%*%t(Xa)%*%ya
# Solution z du système ya=Xz
           [,1]
[1,] 0.92
[2,] 0.52
```


b) $\mathbf{X}$ n'est pas de plein rang : soit la matrice $\mathbf{X_b}$ ($n = 6; p = 3$), complétée par le vecteur $\boldsymbol{y_b}$:

$\mathbf{X_b}$			$\boldsymbol{y_b}$
1	1	0	7
1	1	0	36
1	1	0	2
1	0	1	13
1	0	1	44
1	0	1	18

Les différents calculs des éléments de la décomposition singulière de sont présentés ci-dessous (Tab.11), ainsi que la résolution du système $\mathbf{y_b} = \mathbf{X_b z_b}$. On notera qu'il existe aussi une fonction **ginverse** qui fournit directement l'inverse généralisée et indique que le rang de $\mathbf{X_b}$ est 2.

TABLEAU 11 - Exemple de calcul des matrices de la décomposition singulière de $\mathbf{X}_{6\times3}$.

```
> Xb<-c(1,1,1,1,1,1,1,1,1,0,0,0,0,0,1,1,1,1)
> dim(Xb)<-c(6,3)
> B<-svd(Xb)
> round(B$d,3)
[1] 3.000 1.732 0.000
> round(B$u,3)
             [,1]   [,2]   [,3]
[1,] -0.408 -0.408  0.817
[2,] -0.408 -0.408 -0.408
[3,] -0.408 -0.408 -0.408
[4,] -0.408  0.408  0.000
[5,] -0.408  0.408  0.000
[6,] -0.408  0.408  0.000
> round(B$v,3)
             [,1]   [,2]    [,3]
[1,] -0.817  0.000  0.577
[2,] -0.408 -0.707 -0.577
[3,] -0.408  0.707 -0.577
> round(B$v%*%diag(B$d^2)%*%t(B$v),0)# X'X
     [,1] [,2] [,3]
[1,]    6    3    3
[2,]    3    3    0
[3,]    3    0    3
> solve(t(Xb)%*%Xb)# Essai d'inversion de X'X
     Error in solve.qr(a): apparently singular matrix
> round(B$u[,1:2]%*%t(B$u[,1:2]),3)# matrice H
            ,1]  [,2]  [,3]  [,4]  [,5]  [,6]
[1,] 0.333 0.333 0.333 0.000 0.000 0.000
[2,] 0.333 0.333 0.333 0.000 0.000 0.000
[3,] 0.333 0.333 0.333 0.000 0.000 0.000
[4,] 0.000 0.000 0.000 0.333 0.333 0.333
[5,] 0.000 0.000 0.000 0.333 0.333 0.333
[6,] 0.000 0.000 0.000 0.333 0.333 0.333
> round(t(B$u[,1:2]%*%diag(1/B$d[1:2])%*%t(B$v[,1:2])),3)
# Inverse Moore-Penrose
             [,1]   [,2]   [,3]   [,4]   [,5]   [,6]
[1,]  0.111  0.111  0.111  0.111  0.111  0.111
[2,]  0.222  0.222  0.222 -0.111 -0.111 -0.111
[3,] -0.111 -0.111 -0.111  0.222  0.222  0.222
> yb<-c(7,36,2,13,44,18)
> dim(yb)<-c(6,1)
```

```
> round(t(yb)%*%B$u[,1:2]%*%diag(1/B$d[1:2])%*%t(B$v[,1:2]),3)
             [,1]  [,2]   [,3]
[1,] 13.333 1.667 11.667
> round(ginverse(Xb),3)
             [,1]   [,2]   [,3]   [,4]   [,5]   [,6]
[1,]  0.111  0.111  0.111  0.111  0.111  0.111
[2,]  0.222  0.222  0.222 -0.111 -0.111 -0.111
[3,] -0.111 -0.111 -0.111  0.222  0.222  0.222
attr(, "rank"):
[1] 2
```

4.4 Décomposition en valeurs singulières du triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

La plupart des méthodes d'Analyse Factorielle des Données peuvent être présentées dans un cadre commun : celui de l'extension du théorème de la *DVS* au cadre d'espaces euclidiens généraux. Elle correspond à la DVS du triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ où :

- $\mathbf{X}$ matrice $n \times p$ des données ;
- $\mathbf{Q}$ est la matrice d'un produit scalaire de $\mathbb{R}^p$ (*L'espace des individus)*, soit une matrice carrée symétrique définie positive qui définit la fonction :

$$(\boldsymbol{x}, \boldsymbol{y}) = ([x_1 \quad \cdots \quad x_p]^{\mathrm{T}}, [y_1 \quad \cdots \quad y_p]^{\mathrm{T}}) \in \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{y} = \langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbf{Q}} \in \mathbb{R}$$

- $\mathbf{D}$ est la matrice d'un produit scalaire de $\mathbb{R}^n$ (*L'espace des variables)*, soit une matrice carrée symétrique définie positive qui définit la fonction :

$$(\boldsymbol{x}, \boldsymbol{y}) = \left(\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \right) \in \mathbb{R}^n \times \mathbb{R}^n \mapsto \boldsymbol{x}^{\mathrm{T}}\mathbf{D}\boldsymbol{y} = \langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbf{D}} \in \mathbb{R}$$

DVS de $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ *:* soit les matrices réelles $\mathbf{X}_{n\times p}$ de rang r et les métriques $\mathbf{Q}_{p\times p}$ et $\mathbf{D}_{n\times n}$ de $\mathbb{R}^p$ et de $\mathbb{R}^n$. Il existe :

- Une matrice $\mathbf{U}_{n\times r} = [\boldsymbol{u}^1 \ \vdots \ \cdots \ \vdots \ \boldsymbol{u}^r]$ $\mathbf{D}$-orthonormée, dont les colonnes sont les vecteurs propres de $\mathbf{X}\mathbf{Q}\mathbf{X}^{\mathrm{T}}\mathbf{D} = \mathbf{W}\mathbf{D}$ associés aux valeurs propres $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$;
- Une matrice $\mathbf{V}_{p\times r} = [\boldsymbol{v}^1 \ \vdots \ \cdots \ \vdots \ \boldsymbol{v}^r]$ $\mathbf{Q}$-orthonormée, dont les colonnes sont les vecteurs propres de $\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}\mathbf{Q} = \mathbf{S}\mathbf{Q}$ associés aux valeurs propres $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$;
- Une matrice diagonale $\mathbf{\Theta}_{\mathrm{r}\times\mathrm{r}} = \mathbf{\Lambda}^{1/2} = \mathbf{diag}\left(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_r}\right)$ contenant les r valeurs singulières du triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

telles que $\mathbf{X}$ se décompose en :

$$\mathbf{X}_{n\times p} = \mathbf{U}_{n\times r}\mathbf{\Theta}_{\mathrm{r}\times\mathrm{r}}\left(\mathbf{V}_{\mathrm{p}\times\mathrm{r}}\right)^{\mathrm{T}} = \sum_{k=1}^{r} \sqrt{\lambda_k}\, \boldsymbol{u}^k (\boldsymbol{v}^k)^{\mathrm{T}}$$

5. BILAN

Dans ce chapitre, nous avons présenté les premiers outils de description d'un tableau de données défini par des observations et des variables. Des outils typiquement statistiques que l'on va retrouver dans tout l'ouvrage, mais aussi des outils de nature purement mathématique dont la signification en statistique apparaîtra plus loin. Pour l'instant, il est important de retenir les points les plus importants suivants :

- Avant toute analyse d'un tableau il faut bien connaître ce que représentent les observations et les variables.
- Toute étude, même celles portant sur des données multidimensionnelles, doit commencer par un examen détaillé de chaque variable, en faisant des analyses graphiques. Il ne faut pas oublier que les graphiques sont bien souvent le meilleur moyen de faire comprendre à des utilisateurs des résultats complexes ; il faut se souvenir qu'au XIX$^{\mathrm{e}}$ siècle c'est ainsi que l'illustre infirmière Florence Nightingale, plus connue sous le nom de *Dame à la lampe*, réussit à convaincre le corps médical d'améliorer ses pratiques. Première femme élue en 1859 membre de la *Royal Statistical Society*, elle est considérée comme *a true pioneer in the graphical representation of statistics* (Triola & Triola, 2012). On pourra consulter les très nombreux ouvrages sur la réalisation de graphiques (Chambers et *al.*, 1983), en particulier avec *R* (Murrell, 2005).
- En particulier, les graphiques de densité et de Normalité doivent permettre de juger l'allure des distributions pour, éventuellement, transformer les variables.
- L'analyse globale du tableau, c'est-à-dire d'une matrice $\mathbf{X}_{n\times p}$ doit être associée à la définition d'une pondération éventuelle des observations à l'aide d'une matrice diagonale $\mathbf{D}_{n\times n}$ et d'une matrice définie positive $\mathbf{Q}_{p\times p}$ qui permet de définir une métrique, c'est-à-dire comment calculer la mesure de ressemblance entre observations. Ces deux dernières matrices sont importantes. Ne pas s'en soucier revient à choisir implicitement les valeurs $\mathbf{D}_{n\times n} = \frac{1}{n}\mathbf{I}_n$ et $\mathbf{Q}_{p\times p} = \mathbf{I}_p$: ces choix ne sont pas toujours ceux qui permettent de faire l'analyse de la matrice de données la plus adaptée aux buts de l'étude.

Ce qu'il faut retenir :

- Les questions préalables à toute analyse : quelles sont les unités de mesure ? Comment les données ont été acquises ? Existe-t-il une structure *a priori* dans tableau de données. Quel est le type des variables ?
- Résumé de chaque variable : graphiques, paramètres numériques : position, dispersion, forme. Existe-t-il des valeurs suspectes et peut-on en justifier l'origine ?
- Résumé de l'ensemble du tableau de variables : centre de gravité, matrice de dispersion.
- Décomposition en valeurs singulières (*DVS*) d'une matrice de données.

PROBLEMES ET EXERCICES

1. Décrire le jeu de données **Eaux2010** dont **Eaux1** est extrait. En particulier, comparer les eaux selon leur nature (plate ou gazeuse).
2. Faire les graphiques des facteurs non analysés dans l'exemple des données **Poumon** du §2.5.
3. *Diagramme en boîte : pourquoi 1.5 ?*
   Dans un *Diagramme en boîte*, les valeurs aberrantes sont définies comme les observations en dehors de l'intervalle $I = [Q_{0.25} - 1.5(Q_{0.75} - Q_{0.25}); Q_{0.25} + 1.5(Q_{0.75} - Q_{0.25})]$. Tukey, son inventeur, justifie l'introduction de la valeur 1.5 en disant que la valeur 1 est trop petite tandis que 2 est trop grande !
   Supposer que les observations sont indépendantes et identiquement distribuées (*iid*) suivant une loi Normale $N(0,1)$
   a. Calculer les quartiles théoriques $Q_{0.25}$ et $Q_{0.75}$ associés à $N(0,1)$ en utilisant la fonction **qnorm**.
   b. En déduire l'écart interquartile $D_{0.25} = Q_{0.75} - Q_{0.25}$ et l'intervalle I.
   c. Calculer $P[X \in I]$ quand $X \sim N(0,1)$.
   d. Refaire les calculs en remplaçant 1.5 par 1 puis par 2.
   e. Commenter les résultats.
4. La robustesse de la médiane et de la moyenne peuvent être étudiées en ajoutant des observations aberrantes. A l'aide du code suivant :

```
x=rnorm(500) ;
z=5 ;
moyenne=NULL ; mediane=NULL ;
m.max=25
for( m in 1:m.max)
  { xx=c(x,rep(z,m))
  moyenne=c(moyenne,mean(xx) )
  mediane=c(mediane,median(xx))
  }
matplot(cbind(moyenne,mediane), type="l",ylim=c(-0.3,.5))
abline(h=c(mean(x),median(x)) ,col=1:2,lty=1:2)
```

   a. Simuler un échantillon de taille $n = 1000$ suivant une loi $N(0,1)$ et ajouter m fois l'observation $z = 5$.
   b. Représenter l'évolution de la médiane et de la moyenne en fonction de m.
   c. Commenter les résultats obtenus.
5. *Nombre de classes dans un histogramme*
   L'option **nclass** (ou **breaks**) de la fonction **hist** permet d'ajuster le nombre de classes n_c de l'histogramme ; par défaut fixé par la formule de Sturges à partie entière de $1 + \log_2 n$.
   a. Simuler un échantillon de de taille $n = 100$ suivant une loi $N(0,1)$
   b. Sur une même feuille graphique (**par(mfrow=c(3,3))**), tracer en faisant varier le nombre de classes l'histogramme correspond à l'échantillon simulé. On pourra prendre par exemple, **nclass**$\in \{3,5,8,10,15,20,25,30,50\}$
   c. Reprendre la question b. avec un échantillon simulé suivant une loi exponentielle de paramètre 1, puis une loi de Cauchy.

   d. Commenter les résultats.
6. Matrices partitionnées : vérifiez que l'inverse de la matrice symétrique non singulière $\mathbf{A}_{(p+1)\times(p+1)}$, avec $\mathbf{A_{11}}_{,p\times p}$, $\boldsymbol{a}_{\mathbf{12}(p\times 1)}$ et a_{22} un scalaire :

$$\mathbf{A} = \begin{pmatrix} \mathbf{A_{11}} & \mathbf{a_{12}} \\ \mathbf{a_{12}^T} & a_{22} \end{pmatrix}$$

   est donnée par :

$$\mathbf{A}^{-1} = \frac{1}{b}\begin{pmatrix} b\mathbf{A_{11}^{-1}} + \mathbf{A_{11}^{-1}a_{12}a_{12}^T A_{11}^{-1}} & -\mathbf{A_{11}^{-1}a_{12}} \\ -\mathbf{a_{12}^T A_{11}^{-1}} & 1 \end{pmatrix} \text{ où } b = a_{22} - \boldsymbol{a_{12}^T A_{11}^{-1} a_{12}}.$$

   Une matrice non singulière de la forme $\mathbf{B} + \mathbf{cc}^{\mathrm{T}}$, où $\mathbf{B}$ est non singulière a une inverse de la forme :

$$(\mathbf{B} + \boldsymbol{cc}^{\mathrm{T}})^{-1} = \mathbf{B^{-1}} - \frac{\mathbf{B^{-1}}\boldsymbol{cc}^{\mathrm{T}}\mathbf{B^{-1}}}{1 + \boldsymbol{c}^{\mathrm{T}}\mathbf{B^{-1}}\boldsymbol{c}}$$

Informations sur l'eau minérale (Chambre Syndicale des Eaux Minérales (CSEM), 2008).
Fichiers : **Eaux1** & **Eaux2010**