Exercice libre

Analyser des données trouvées sur Internet

Résumé

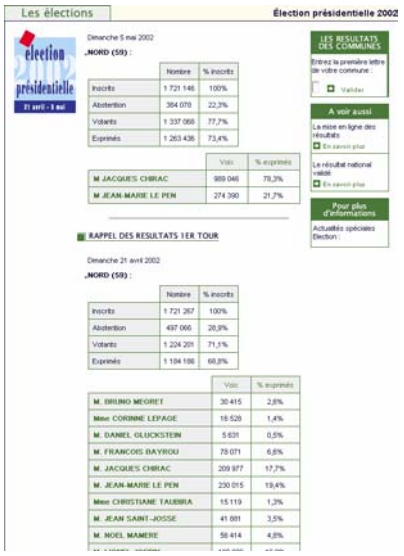
Sur un exemple initié par Sophie GROSZ (INSA 4° année Filière BIM) dans son rapport d'analyse des données (2004), la fiche indique quelques unes des opérations fondamentales dans les exercices ouverts du type : "Rédiger un rapport sur un jeu de données trouvé sur Internet en utilisant quelques unes des méthodes statistiques étudiées". L'exercice est difficile et demande de choisir un point de vue. On illustre plusieurs manières différentes d'aborder cette question.

Plan

1. Acquérir des données originales

La recherche d'un jeu de données est une entreprise qui peut prendre beaucoup de temps. L'énergie consacrée à la mise en forme de sa découverte peut suffire : l'argument est "j'ai trouvé un ensemble de données qui peut servir à d'autres ; voilà ce que j'ai fait pour y arriver". Un exemple illustre cette stratégie.

Les données sont disponibles sur le site du Ministère de l'Intérieur, de la sécurité intérieure et des libertés locales :



http://www.interieur.gouv.fr/avotreservice/elections/presid2002/

Il y a une page par département. On ne considère que les 94 départements de la France continentale. Une fonction permet de lire une page et d'extraire l'informations numérique :

```
"lecture" <- function(k) {
       url = "http://www.interieur.gouv.fr/avotreservice/elections/presid2002/"
       cha <- as.numeric(k)
       if (nchar(cha)<3) cha <- paste("0",cha,sep="")
       if (nchar(cha)<3) cha <- paste("0",cha,sep="")
       url=paste(url,cha,sep="")
       dw=readLines(url)
       n3 <- grep("%",dw)-2
       dw <- dw[n3]
       n3 <- grep("EDEDED",dw)
       dw <- dw[-n3]
       dw <- sub("\<td class=\\\"marine11\\\" align=\\\"center\\\"\>","",dw)
       dw <- sub("\</td\>","",dw)
       dw <- gsub(" ","",dw)
       dw <- dw[-c(1,2)]
       dw <- as.numeric(dw)
       dw
}
```

Le reste n'est que manipulations élémentaires. On obtient presi021 et presi022. Le premier a 20 colonnes et 94 lignes :

```
 dim(presi021)
[1] 94 20
 presi021 [c(1,13,58,74),]
    inscrits abstentions votants exprimes Megret Lepage Gluksten Bayrou Chirac
D1    338220       89003  249217   240652   8425   5258      1077  18614  41348
D13  1144969      306821  838148   815890  40628  12823      2979  47530 138189
```

```
D59  1721267      497066 1224201  1184186  30415  16528     5631  78071 209977
D75  1081420      322390  759030   744943   7616  18285     2498  58924 178841
    Le_Pen Taubira Saint.Josse Mamere Jospin Boutin   Hue Chevenement Madelin
D1   52617    4527        9718  12339  30418   2966  5134       14665   12580
D13 182778   14078       32097  41563 114323   6472 39032       44947   31371
D59 230015   15119       41881  56414 199036  12232 57639       54733   40684
D75  69658   28189        3905  55050 148624  11119 16299       49342   48403
    Laguiller Besancenot
D1      11562       9404
D13     40220      26860
D59     85680      50131
D75     22714      25476
```

Les noms des variables sont explicites. Le second a 6 colonnes et 94 lignes :

```
dim(presi022)
[1] 94  6
```

```
 presi022 [c(1,13,58,74),]
    inscrits abstentions votants exprimes Chirac Le_Pen
D1    338181       64364  273817   258986 204029  54957
D13  1145020      218918  926102   880502 638688 241814
D59  1721146      384078 1337068  1263436 989046 274390
D75  1081313      184502  896811   872242 784741  87501
```

Les deux data.frame sont sauvegardés dans des fichiers texte. On les trouvera à :

http://pbil.univ-lyon1.fr/R/donnees/presi2002tour1.txt
http://pbil.univ-lyon1.fr/R/donnees/presi2002tour2.txt

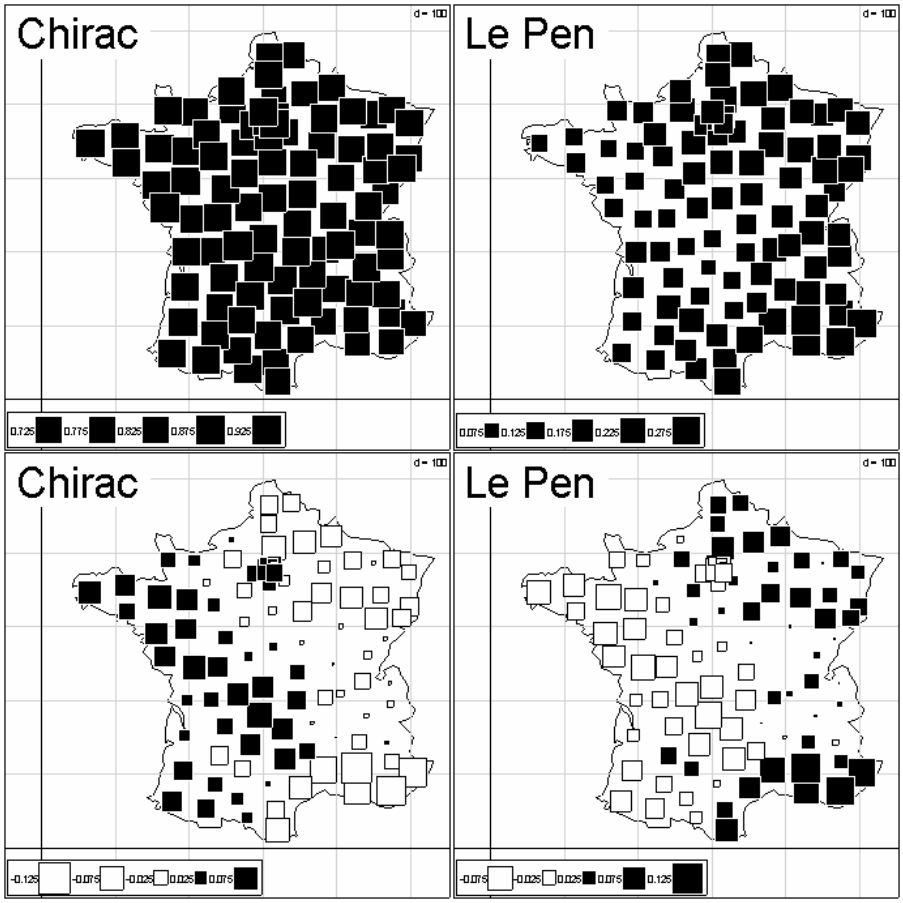
```
download.file("http://pbil.univ-lyon1.fr/R/donnees/presi2002tour1.txt",
    "tour1.txt")
    trying URL `http://pbil.univ-lyon1.fr/R/donnees/presi2002tour1.txt'
    Content type `text/plain' length 11452 bytes
    opened URL
    downloaded 11Kb

download.file("http://pbil.univ-lyon1.fr/R/donnees/presi2002tour2.txt",
    "tour2.txt")
    trying URL `http://pbil.univ-lyon1.fr/R/donnees/presi2002tour2.txt'
    Content type `text/plain' length 4286 bytes
    opened URL
    downloaded 4286 bytes
```

```
 tour1=read.table("tour1.txt")
 tour2=read.table("tour2.txt")
 dim(tour1)
[1] 94 20
 names(tour1)
 [1] "inscrits"    "abstentions" "votants"     "exprimes"    "Megret"
 [6] "Lepage"      "Gluksten"    "Bayrou"      "Chirac"      "Le.Pen"
[11] "Taubira"     "Saint.Josse" "Mamere"      "Jospin"      "Boutin"
[16] "Hue"         "Chevenement" "Madelin"     "Laguiller"   "Besancenot"
 dim(tour2)
[1] 94  6
 names(tour2)
[1] "inscrits"    "abstentions" "votants"     "exprimes"    "Chirac"
[6] "Le.Pen"
```

Une illustration permet de vérifier le bon état de l'information transmise.

```
data(elec88)
par(mfrow=c(2,2))
w1 = tour2$Chirac ; w2 = tour2$Le_Pen
w3 = w1 + w2 ; w4 = sum(w1)/sum(w3)
w1 = w1/w3 ; w2=w2/w3
s.value(elec88$xy,w1 ,contour=elec88$contour,sub="Chirac", csub=3, possub="topleft")
```

```
s.value(elec88$xy,w2 ,contour=elec88$contour,sub="Le Pen", csub=3, possub="topleft")
s.value(elec88$xy,w1-w4 ,contour=elec88$contour,sub="Chirac", csub=3, possub="topleft")
s.value(elec88$xy,w2-1+w4, contour=elec88$contour,sub="Le Pen", csub=3, possub="topleft")
```



Scores des deux candidats du second tour des élections présidentielles de 2004.

En haut valeurs brutes, en bas valeurs centrées. Une structure spatiale très forte.

Dans cet exemple, l'effort a porté sur le rassemblement de l'information. On a remarqué la transparence remarquable de l'information publique et le soin porté à son affichage. La transformer en un tableau de données a été l'objectif et cet objectif est atteint. Ce qu'on pourra faire avec ces données est une autre affaire.

Ce qui est essentiel dans cet exemple est le caractère parfaitement fiable des résultats numériques publiés. On peut faire l'hypothèse que ces résultats sont obtenus à l'aide de procédures transparentes, compilées avec les plus grandes vérifications, diffusées et contrôlées : elles ont force de loi. Elles ne sont pas discutables. Au contraire, on pourra évidemment trouver sur le réseau toutes sortes de résultats farfelus dans tous les compartiments : conception, acquisition, destination, reproduction, diffusion … A chacun de se méfier.

La prudence est une vertu cardinale : elle doit être constante. Une introduction prudente, c'est bien :

Introduction

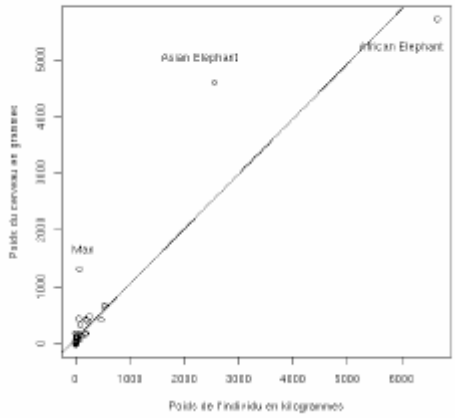
Le jeu de données utilisé est issu d'une étude publiée en 1976 dans *Science* par ALLISON, TRUETT, CICCHETTI et DOMENIC intitulée **Sleep in Mammals : Ecological and Constitutional Correlates**, Science 1976 Nov 12;194(4266) :732-4 (je n'ai pas eu accès à l'article). Ces données, disponibles à `http://lib.stat.cmu.edu/datasets/sleep`, portent entre autre sur la morphologie et la composition du sommeil de 62 espèces de mammifères. Pour chaque espèce, on dispose des informations suivantes :

- poids d'un individu en kilogrammes
- poids du cerveau en grammes
- nombre d'heures de sommeil paradoxal (aussi connu sous l'appélation Rapid-Eye-Movement-sleep) par jour
- nombre d'heures de sommeil sans rêve (ou *Low-Wave*) par jour
- quantité totale de sommeil journalïer (c'est la somme des deux précédents)
- durée de vie maximale en années
- nombre de jours de gestation
- index de prédation ; echelle de 1 (qui a peu de prédateurs) à 5 (qui a beaucoup de prédateurs).
- index d'exposition pendant le sommeil ; echelle de 1 à 5. Un animal qui dort par exemple dans un antre aura un indice d'exposition de 1.
- index de danger de la part d'autres animaux ; echelle de 1 à 5. (dépend entre autre des indices précédents)

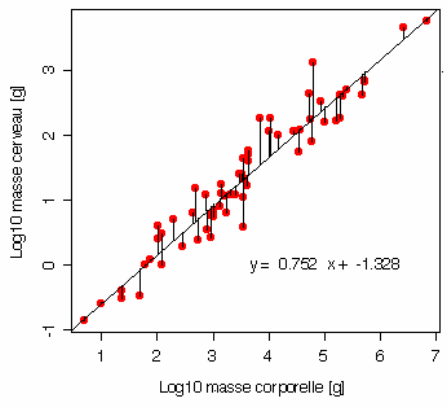
Je n'ai pas de détails par rapport à l'établissement de ces données. Pour les données quantitatives il s'agit probablement pour chaque espèce de moyennes établies à partir de plusieurs individus. J'espère que ces données, tout comme la classification des espèces en différents groupes ont été réalisées par des personnes compétentes. Étant donné qu'elles ont donné lieu à une publication dans *Science*, je vais leur faire confiance.

Mais plus loin :

FIG. 4 – Poids du cerveau en fonction du poids de l'individu



Tomber sur un des jeux de données les plus célèbres de la statistique, présent dans deux librairies de R, fondement pédagogique de la morphométrie et ne s'apercevoir de rien, voilà qui manque sérieusement de prudence (*http://pbil.univ-lyon1.fr/R/fichestd/tdr333.pdf*) :

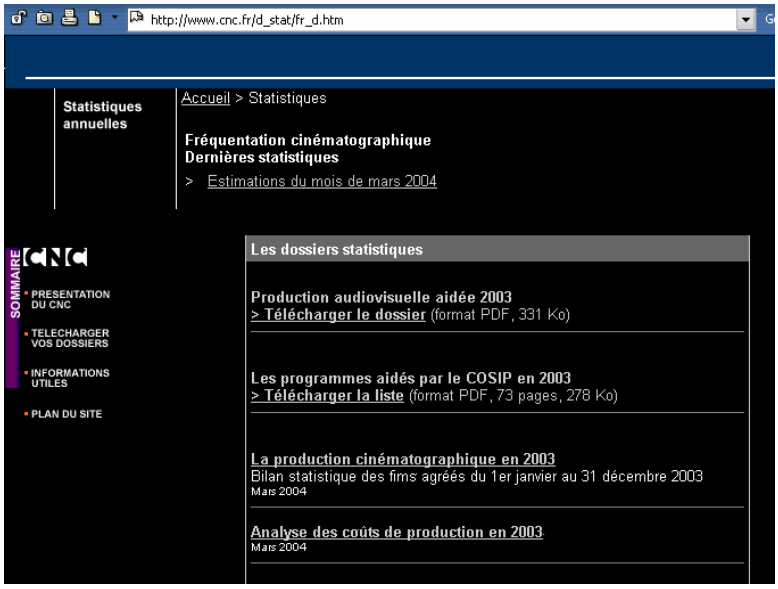


2. Citer ses sources

En général, les données seront issues d'une publication. On devra porter un minimum d'intérêt biologique, sociologique, politique, culturel, sportif, ludique, … à l'information conservée. Merci d'éviter l'éternel problème de la criminalité aux Etats-Unis, ou d'exprimer de la passion pour les salaires des joueurs de base-ball en 1982.

L'essentiel est de citer ses sources avec précision et de vérifier les droits d'accès. Sophie GROSZ propose d'accéder au serveur **`http://www.cnc.fr/`** du Centre National de la Cinématographie (**CNC**) :

Le lien STATISTIQUES renvoie à un ensemble important de documents :



Le document http://www.cnc.fr/cncinfo/288/index.htm permet de télécharger le tableau :

La géographie du cinéma

Equipement et résultats d'exploitation en 2002 par département(1)

Département	Population (millions)(2)	Entrées (millions)	Evolution des entrées 2002/2001	Recettes guichets (M€)	Recette moyenne par entrée (€)	Indice de fréquentation	Séances (milliers)	Communes équipées	Etab. actifs	Salles actives	Fauteuils	Etab. Art et Essai	Multipl. actifs
01 AIN	0,515	0,769	-2,2%	4,065	5,29	1,49	26	16	20	35	6 288	12	
02 AISNE	0,536	0,731	-5,2%	3,942	5,39	1,36	28	14	15	38	7 403	8	
03 ALLIER	0,345	0,499	-6,5%	2,789	5,59	1,45	17	7	11	28	3 956	4	
04 ALPES DE HTE PROVENCE	0,140	0,453	-1,7%	2,262	5,00	3,24	17	13	15	23	3 480	7	
...													
92 HAUTS DE SEINE	1,429	3,978	+5,5%	21,707	5,46	2,78	107	33	39	89	20 762	24	2
93 SEINE SAINT DENIS	1,383	4,803	-0,5%	25,543	5,32	3,47	127	26	32	97	21 168	16	4
94 VAL DE MARNE	1,227	4,052	+8,6%	23,967	5,91	3,30	122	30	36	89	23 448	17	2
95 VAL D'OISE	1,105	2,160	-1,3%	11,581	5,36	1,95	66	21	22	56	12 939	7	1
FRANCE	58,52	184,461	-1,5%	1 027,865	5,57	3,15	5 625	1 632	2 146	5 257	1 070 522	932	106

Ce document porte la mention CNC Info n°288 –septembre 2003.

Pour être vraiment utile, un exercice libre doit porter sur des données publiques, utilisables et communicables sans autre contrainte que le devoir d'indiquer les sources.

CNC info La lettre du CNC est la lettre mensuelle d'information du CNC. Elle a pour vocation de diffuser l'actualité du Centre à l'ensemble des professionnels, des institutionnels et de la presse. Son organigramme sur le numéro d'avril 2004 porte la mention "Reproduction autorisée avec mention d'origine". Nous pouvons donc étudier ce tableau :

LA LETTRE DU CNC • n°13 - avril 2004

Une publication du Centre national de la cinématographie — 12, rue de Lübeck
75784 Paris Cedex 16 — Tél : 01 44 34 36 95
Fax : 01 44 34 34 73 — www.cnc.fr
Directeur de la publication : David Kessler
Coordination générale : Milvia Pandiani-Lacombe
Secrétaire de rédaction : Marc-Antoine Chaumien
Comité de rédaction : Eric Busidan, Marc-Antoine Chaumien, Benoît Danard, Steeve Desgagné, Julien Ezanno, Caroline Jeanneau, Eric Le Roy, Catherine Merlhiot, Milvia Pandiani-Lacombe, Olivier Wotling.
Stagiaire : Lisa Bentes
Conception graphique : Redline. **Impression :** GMK
Dépôt légal à parution - ISSN : 1762-4789

3. Préparer les données

3.1. Préparer un fichier texte

Préparer à partir de ce document un fichier texte avec séparateur tabulation. Pour étudier clairement les composantes spatiales de la fréquentation cinématographique sur le territoire de la métropole enlever les deux lignes des département de Corse. Les variables sont pour chacun des départements :

popu	La population (en million d'habitants)
entr	Le nombre d'entrées (en millions)
evol	L'évolution du nombre d'entrées par rapport à l'année 2002
rece	Les recettes aux guichets (en millions €)
prix	Les recettes moyennes par entrée (rapport des recettes sur le nombre d'entrées)

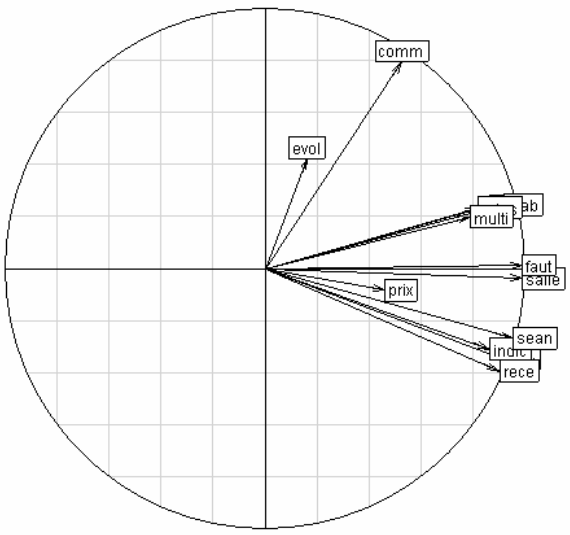
indic	L'indice de fréquentation (rapport du nombre d'entrées sur le nombre d'habitants)
sean	Le nombre de séances (en milliers)
comm	Le nombre de communes équipées de cinémas
etab	Le nombre d'établissements actifs
salle	Le nombre de salles actives
faut	Le nombre de fauteuils
artes	Le nombre d'établissements Art et Essai
multi	Le nombre de multiplexes actifs

Lire le fichier dans R en conservant les noms de lignes et charger les données **elec88** de la librairie **ade4** pour disposer des outils de représentation cartographique par département. On doit obtenir un objet **cnc**. Vérifier ses propriétés :

```
names(cnc)
 [1] "popu"  "entr"  "evol"  "rece"  "prix"  "indic" "sean"  "comm"  "etab"
[10] "salle" "faut"  "artes" "multi"

 dim(cnc)
[1] 94 13

 row.names(cnc)
 [1] "D1"  "D2"  "D3"  "D4"  "D5"  "D6"  "D7"  "D8"  "D9"  "D10" "D11" "D12"
[13] "D13" "D14" "D15" "D16" "D17" "D18" "D19" "D21" "D22" "D23" "D24" "D25"
[25] "D26" "D27" "D28" "D29" "D30" "D31" "D32" "D33" "D34" "D35" "D36" "D37"
[37] "D38" "D39" "D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48" "D49"
[49] "D50" "D51" "D52" "D53" "D54" "D55" "D56" "D57" "D58" "D59" "D60" "D61"
[61] "D62" "D63" "D64" "D65" "D66" "D67" "D68" "D69" "D70" "D71" "D72" "D73"
[73] "D74" "D75" "D76" "D77" "D78" "D79" "D80" "D81" "D82" "D83" "D84" "D85"
[85] "D86" "D87" "D88" "D89" "D90" "D91" "D92" "D93" "D94" "D95"
```

Vérifier que les données contiennent un effet taille trivial.

```
s.corcircle(dudi.pca(cnc,scann=F)$co)
```



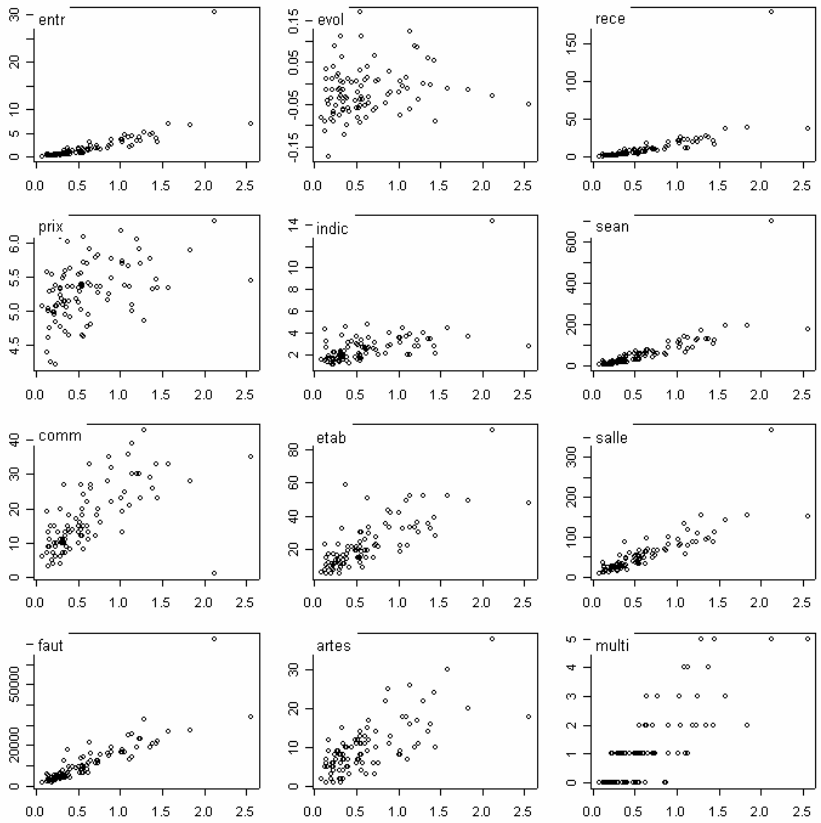
3.2. Poser une question

L'analyse statistique n'est pas destinée à nous apprendre que si la population augmente d'un département à l'autre, le nombre de salles de cinéma aussi, donc le nombre de fauteuils aussi et donc le nombre d'entrées aussi. Il est peu vraisemblable qu'au milieu d'une zone rurale déserte de grande surface on trouve des multiplexes qui restent vides.

La question est : y a-t-il une géographie de la fréquentation cinématographique ?

3.3. Enlever une évidence

Le lien entre la population et les autres variables est évident, ce qui ne signifie pas qu'il suffit de diviser par le nombre habitants *a priori* pour faire le tour de la question.

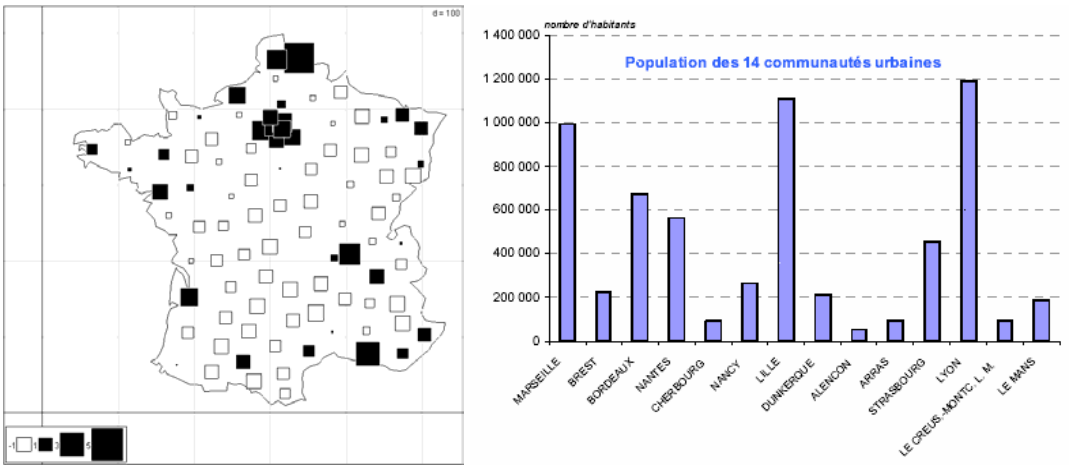


Utiliser la fonction **`sco.quant`** pour représenter le lien avec la première variable de chacune des autres. **Observer** la présence d'un département très peuplé qui présente une fréquentation cinématographique complètement disproportionnée par rapport à la prévision simple. **Identifier** ce département et celui qui est encore plus peuplé mais qui ne présente pas cet effet.

```
elec88$lab["D75"] #"Paris"
elec88$lab["D59"] #"Nord"
```

La fréquentation cinématographique de Paris est un phénomène totalement étranger à la logique du reste du territoire. **On continue l'étude avec 93 lignes**. **Vérifier** que vous pouvez cartographier une variable dans la nouvelle configuration (par exemple la population).

```
s.value(xy,scalewt(cnc1$pop),cont=elec88$contour)
```

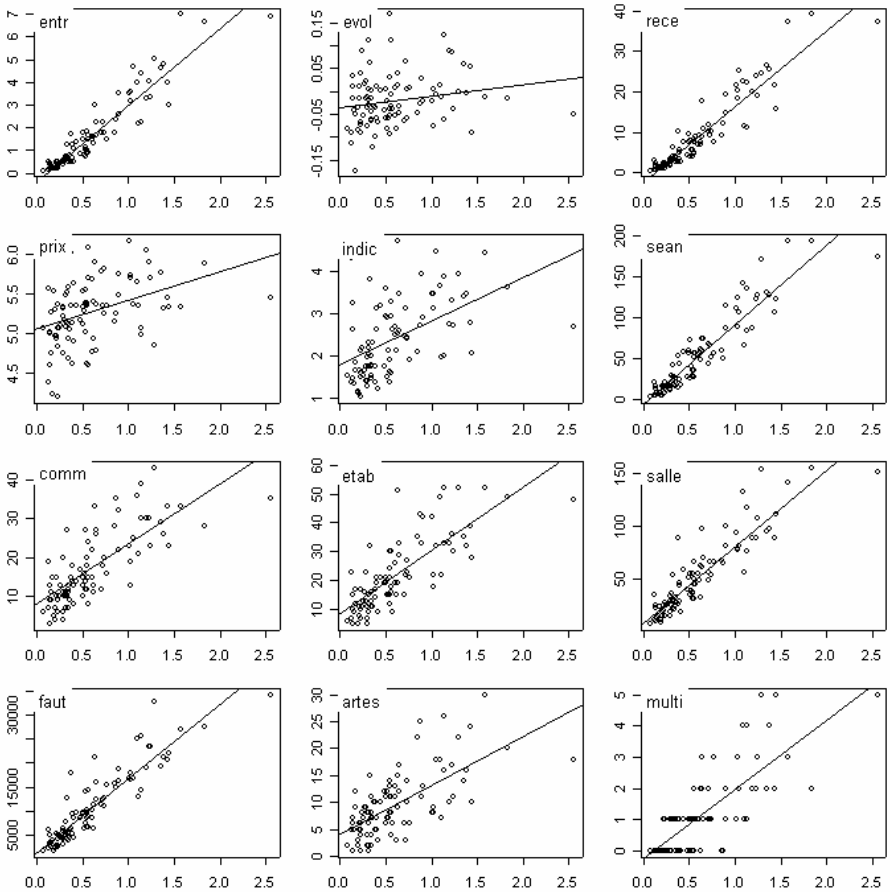


Identifier les concentrations de populations.

4. Etablir les données définitives

4.1. Modifications initiales

Décrire toutes les opérations effectuées sur le jeu de données originales, à la fois pour respecter les auteurs des données d'origine et pour permettre au lecteur de refaire les opérations.



La variable `evol` n'a rien à voir avec le reste et n'a de sens que dans l'étude de l'évolution du secteur. Elle ne nous intéresse pas ici. Affirmer qu'on cherche une solution à une question est essentielle à l'analyse statistique. Ce n'est pas parce qu'une colonne est là qu'on doit forcément s'en servir. La variable prix est un rapport. Nous savons faire cela dans R si nous le désirons :

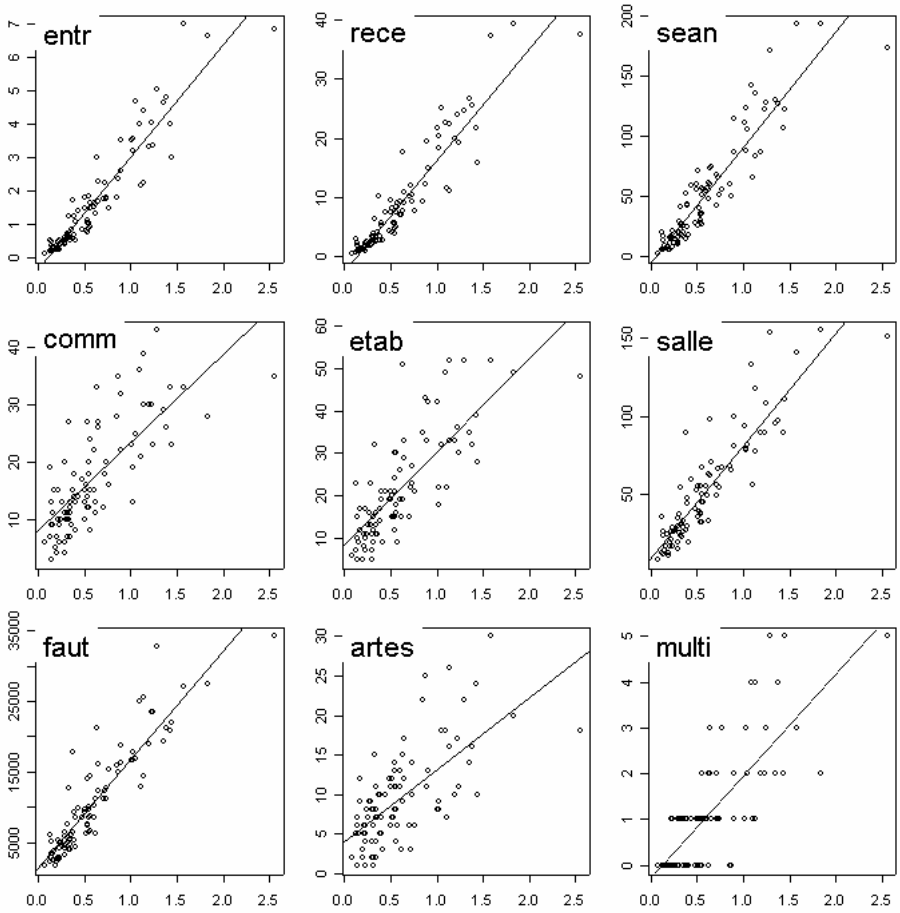
```
cor(cnc1$prix,cnc1$rece/cnc1$entr)
[1] 1
```

La variable prix n'a donc rien à faire ici. Même remarque sur l'indice de fréquentation : c'est un rapport (entrées par habitants) :

```
cor(cnc1$indic,cnc1$entr/cnc1$popu)
[1] 1
```

Enlever le superflu.

```
sco.quant(cnc2$popu,cnc2[,-1],abl=1,csub=3)
```



4.2. Questions techniques

Faut-il maintenant diviser toutes les variables par le nombre d'habitants ? Cela signifie que les variables observées suivent un modèle sans terme constant. Est-ce vrai ?

```
apply(cnc2[,-1],2,function(x) {summary(lm(x~cnc2$pop))})
```

```
$entr
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.3901     0.0975    -4.0  0.00013
cnc2$pop       3.3842     0.1307    25.9  < 2e-16

$rece
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.402      0.552   -4.35  3.6e-05
cnc2$pop      18.830      0.740   25.43  < 2e-16

$sean
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    -5.35       3.09   -1.73    0.087
cnc2$pop       96.29       4.15   23.23   <2e-16

$comm
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     8.05       1.17    6.88  7.3e-10
cnc2$pop       15.42       1.57    9.84  5.5e-16

$etab
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     8.67       1.57    5.52  3.1e-07
cnc2$pop       21.84       2.10   10.39  < 2e-16

$salle
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     8.76       2.73     3.2   0.0019
cnc2$pop       72.03       3.66    19.7   <2e-16

$faut
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     1311        549    2.39    0.019
cnc2$pop       15459        736   21.02   <2e-16

$artes
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.076      0.783    5.21  1.2e-06
cnc2$pop       9.120      1.049    8.70  1.4e-13

$multi
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.263      0.141   -1.86    0.066
cnc2$pop       2.218      0.190   11.70   <2e-16
```

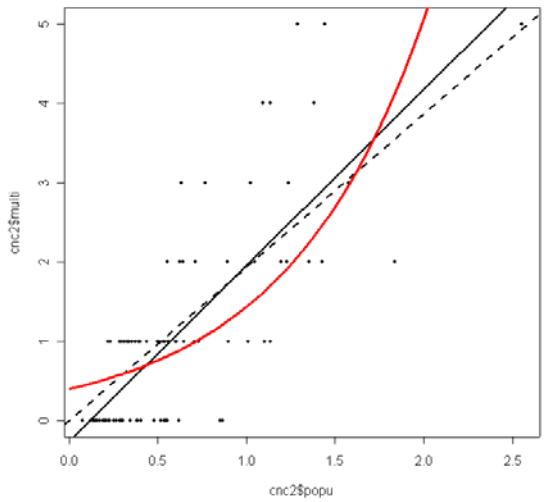
Le cas le plus fréquent est que l'ordonnée à l'origine n'est pas nulle. On prendra donc les résidus des modèles linéaires.

```
cnc3=apply(cnc2[,-1],2,function(x) residuals(lm(x~cnc2$popu)))
cnc3=as.data.frame(cnc3)
row.names(cnc3)=row.names(cnc2)
```

Les variables qui figurent dans ce tableau ne sont plus perturbées par les variations de densités de populations. On peut également se poser des problèmes techniques plus raffinées. La variable nombre de multiplexes actifs est très discrète. Sa variance est plutôt de type poissonien.

4.3. Essais partiels

On peut montrer qu'une idée est restée sans suite, parce que les données ne le supportaient pas, que les résultats n'étaient pas convaincants, que c'était plus difficile que prévu …

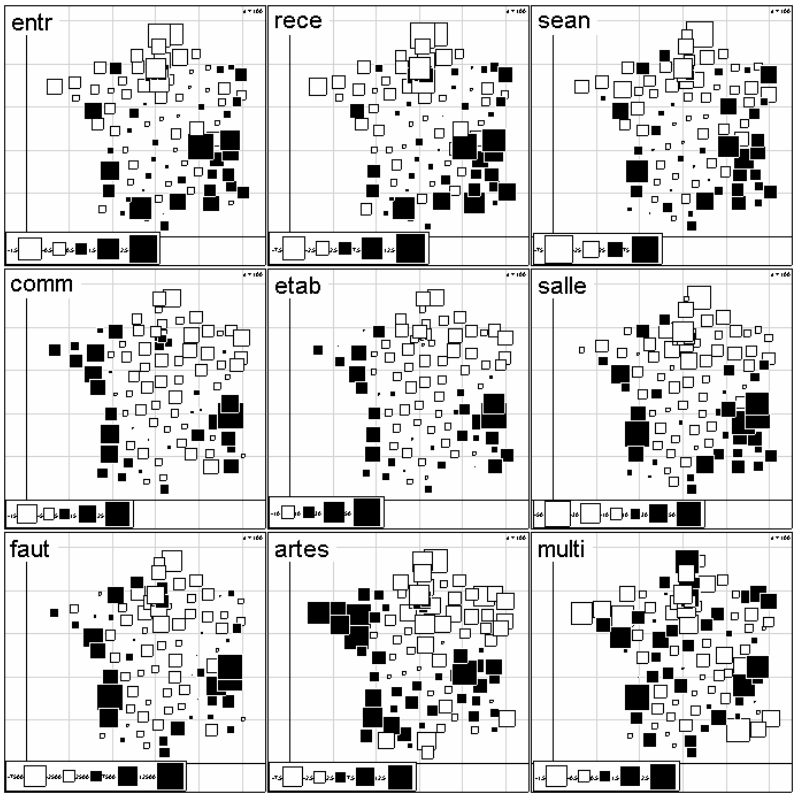



```
plot(cnc2$popu,cnc2$multi,pch=20)
abline(lm(cnc2$multi~cnc2$popu),lwd=2)
abline(lm(cnc2$multi~-1+cnc2$popu),lwd=2,lty=2)
x0=seq(0,2.5,le=50)
y0=predict(glm(multi~popu,data=cnc2,family="poisson"),
      newd=list(popu=x0),type="response")
lines(x0,y0,lwd=3,col="red")
```

Etudier le modèle et détecter le point pivot :

```
glm(multi~popu,data=cnc2,family="poisson"))
```

5. Expression graphique

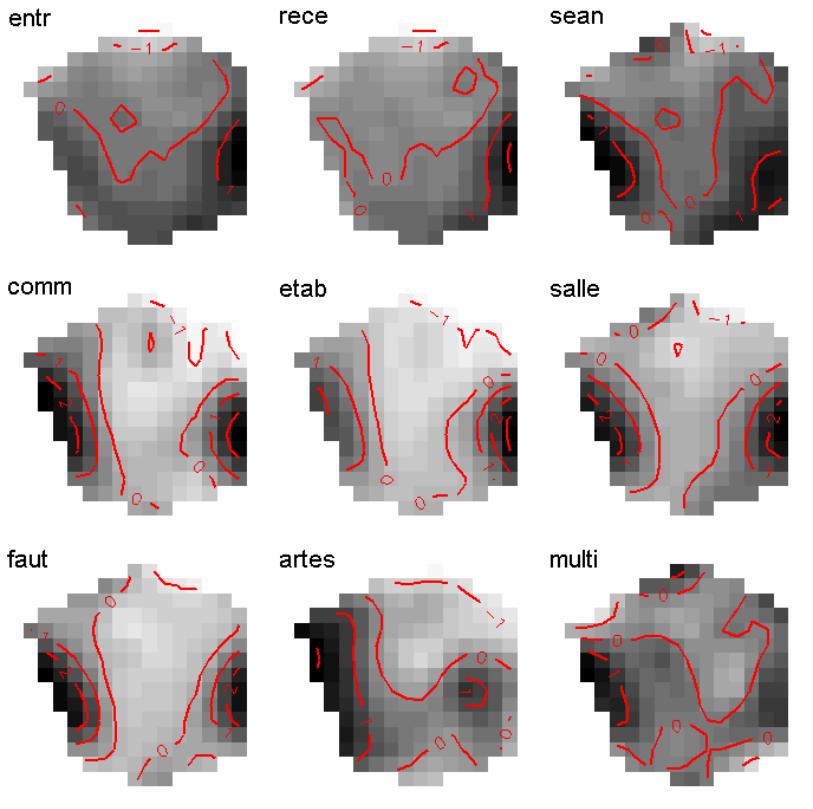


Cartographier les variables par carrés :

```
par(mfrow=c(3,3))
for(k in 1:9) s.value(xy,cnc3[,k],sub=names(cnc3)[k],csub=3,csi=1.5)
```

Cartographier les variables par courbes de niveaux :

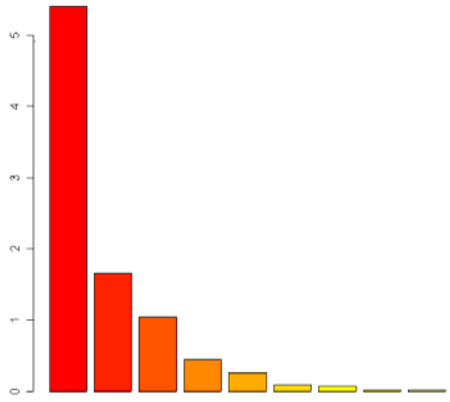
```
for(k in 1:9) s.image(xy,cnc3[,k],sub=names(cnc3)[k],csub=3,span=0.3)
```



La composante régionale est indiscutable.

Synthétiser :

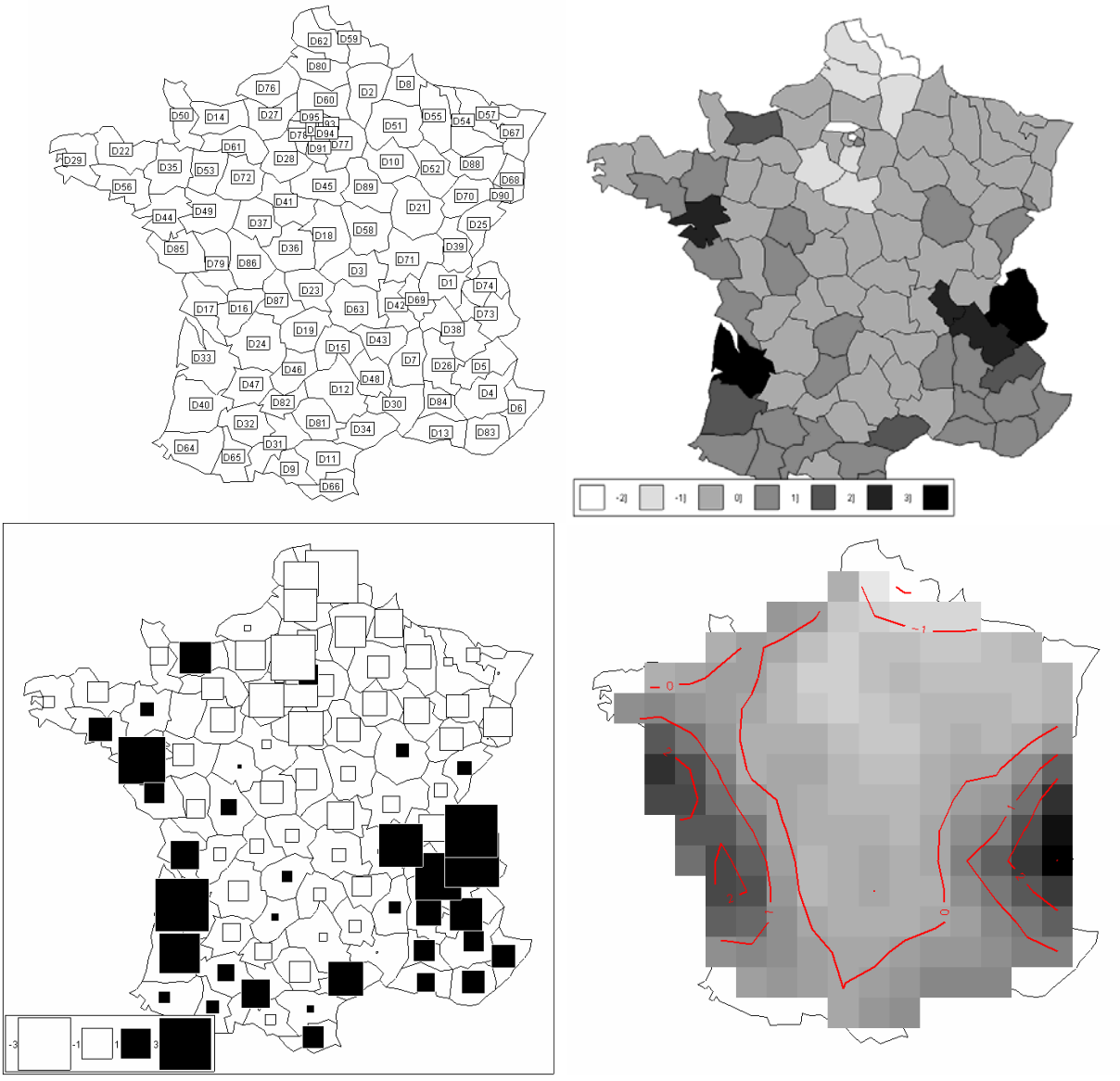
```
dudi3=dudi.pca(cnc3)
Select the number of axes: 3
```



Cartographier la première coordonnée factorielle :

Préparer un nouveau moyen graphique :

```
area=elec88$area
names(area)
[1] "V1" "V2" "V3"
area=area[area$V1!="D75",]
area$V1=as.factor(as.character(area$V1))
area.plot(area,lab=unique(area$V1),clab=0.75)
area.plot(area,val=dudi3$l1[,1])
area.plot(area)
xy=elec88$xy[row.names(elec88$xy)!="D75",]
s.value(xy,dudi3$l1[,1],csi=1.5,add.plot=T)
s.image(xy,dudi3$l1[,1],contour=elec88$contour,span=0.20)
```



6. Le raisonnement statistique

Le but est d'extraire de l'information, rien que de l'information (se contenter de la première pression à froid, il n'est pas nécessaire de faire de l'huile de moteur avec les noyaux !). Expliquer sa démarche est le plus important.

6.1. Retour aux données

Il est éclairant de retourner directement aux données brutes pour vérifier le caractère vraisemblable des observations faites. D'après l'analyse, la corrélation entre les variables persiste après la correction et reste importante, toute chose égale par ailleurs. On distingue 4 départements de la façade atlantique, 5 départements du Nord et 5 départements de la région Rhône-Alpes :

```
w=cnc2/cnc2$popu

apply(w[c(17,33,40,84),],2,mean)/ apply(w[c(62,59,80,02,08),],2,mean)
 popu  entr  rece  sean  comm  etab salle  faut artes multi
1.000 1.664 1.585 1.549 1.304 1.472 1.450 1.493 1.022 2.178

apply(w[c(69,38,73,74,05),],2,mean)/ apply(w[c(62,59,80,02,08),],2,mean)
 popu  entr  rece  sean  comm  etab salle  faut artes multi
1.000 1.607 1.657 1.393 2.425 2.605 1.838 1.871 1.421 1.286
```

L'offre et la consommation de cinéma est bien de 50 à 100 % plus forte dans les deux régions Sud que dans le Nord. Comme Paris ville joue un rôle particulier en rapport avec les départements de la couronne, la véritable échelle du phénomène est peut-être régionale.

6.2. Déplacer la question

C'était techniquement plus utile mais moins pertinent de travailler sur les départements. Le fichier

http://www.cnc.fr/cncinfo/288/tableau11.pdf

est accessible de la même manière. Le titre de l'étude était "La géographie du cinéma".

Equipement et résultats d'exploitation en 2002 par région(1)

Région	Population (millions)(2)	Entrées (millions)	Evolution des entrées 2002/2001	Recettes guichets (M€)	Recette moyenne par entrée (€)	Indice de fréquentation	Séances (milliers)	Communes équipées	Etab. actifs	Salles actives	Fauteuils	Etab. Art et Essai	Multipl. actifs
ALSACE	1,73	5,316	-1,9%	30,453	5,73	3,07	167	25	33	127	27 813	14	4
AQUITAINE	2,91	9,002	+0,6%	44,303	4,92	3,10	314	116	136	317	65 809	68	7
AUVERGNE	1,31	2,544	-5,2%	14,186	5,58	1,94	90	41	50	107	17 830	25	1
BASSE-NORMANDIE	1,42	3,502	-4,0%	17,303	4,94	2,46	115	55	63	129	29 169	32	1
BOURGOGNE	1,61	3,294	-5,9%	18,005	5,47	2,05	141	47	60	148	27 306	28	2
BRETAGNE	2,91	6,833	-4,7%	36,021	5,27	2,35	199	109	128	232	51 730	78	2
CENTRE	2,44	5,320	-5,1%	27,208	5,11	2,18	185	54	68	166	34 037	27	4
CHAMPAGNE-ARDENNE	1,34	2,653	-3,7%	15,707	5,92	1,98	95	22	26	89	18 367	14	2
CORSE	0,26	0,351	-3,4%	2,135	6,08	1,35	13	17	22	32	8 209	4	0
FRANCHE-COMTE	1,12	2,447	-4,7%	12,273	5,01	2,19	90	37	49	120	22 282	17	2
HAUTE-NORMANDIE	1,78	4,387	+0,3%	24,678	5,63	2,46	163	35	42	140	29 971	19	4
ILE-DE-FRANCE	10,95	55,596	+0,6%	332,885	5,99	5,08	1 422	200	322	960	203 814	142	19
LANGUEDOC-ROUSSILLON	2,30	7,126	+3,4%	36,850	5,17	3,10	254	70	94	252	48 725	28	6
LIMOUSIN	0,71	1,490	-5,6%	7,304	4,90	2,10	53	25	29	73	13 386	20	2
LORRAINE	2,31	6,148	-3,0%	34,161	5,56	2,66	179	57	66	174	38 347	20	5
MIDI-PYRENEES	2,55	7,306	-1,4%	37,727	5,16	2,86	197	106	125	231	44 228	69	3
NORD-PAS-DE-CALAIS	4,00	9,843	-6,2%	53,362	5,42	2,46	297	58	76	262	56 283	28	10
PAYS DE LA LOIRE	3,22	8,803	-3,4%	46,537	5,29	2,73	279	109	133	295	64 562	63	8
PICARDIE	1,86	3,077	+1,9%	16,809	5,46	1,66	111	45	54	137	29 409	18	3
POITOU-CHARENTES	1,64	3,896	-3,0%	18,793	4,82	2,38	144	62	78	171	33 636	36	3
PROVENCE-ALPES-COTE D'AZUR	4,51	15,538	-2,0%	90,639	5,83	3,45	490	128	190	440	79 845	64	5
RHONE-ALPES	5,65	19,988	-1,0%	110,525	5,53	3,54	626	214	302	655	125 764	118	13
FRANCE	58,52	184,461	-1,5%	1 027,865	5,57	3,15	5 625	1 632	2 146	5 257	1 070 522	932	106

(1) Données provisoires - (2) INSEE - Recensement 1999.

L'objet elec88 donne le fond de cartes par départements. Le classement des départements par régions est utile. Implanter le facteur qui donne la région par départements :

```
                      dep                    reg
D1                    Ain             Rhône-Alpes
D2                  Aisne                Picardie
D3                 Allier               Auvergne
D4  Alpes de Hautes Provence Provence-Alpes-Côted'Azur
D5              Hautes Alpes Provence-Alpes-Côted'Azur
D6           Alpes Maritimes Provence-Alpes-Côted'Azur
D7                   Ardèche             Rhône-Alpes
```

```
D8                  Ardennes         Champagne-Ardenne
D9                    Ariège             Midi-Pyrénées
D10                     Aube         Champagne-Ardenne
D11                     Aude      Languedoc-Roussillon
...
D86                   Vienne          Poitou-Charentes
D87             Haute Vienne                 Limousin
D88                   Vosges                 Lorraine
D89                    Yonne                Bourgogne
D90    Territoire de Belfort            Franche-Comté
D91                  Essonne            Ile-de-France
D92          Hauts de Seine             Ile-de-France
D93          Seine St Denis             Ile-de-France
D94            Val de Marne             Ile-de-France
D95              Val d'Oise             Ile-de-France
```

```
apply(cnc,2,function(x) tapply(x,reg.dep,sum))[,c(1,2,4,7:13)]
```

```
                          popu   entr    rece sean comm etab salle   faut artes multi
Alsace                   1.734  5.316  30.453  167   25   33   127  27813    14     4
Aquitaine                2.907  9.002  44.303  314  116  136   317  65809    68     7
Auvergne                 1.309  2.543  14.186   89   41   50   107  17830    25     1
Basse-Normandie          1.421  3.502  17.303  114   55   63   129  29169    32     1
Bourgogne                1.610  3.295  18.006  141   47   60   148  27306    28     2
Bretagne                 2.906  6.833  36.021  200  109  128   232  51730    78     2
Centre                   2.440  5.320  27.208  184   54   68   166  34037    27     4
Champagne-Ardenne        1.342  2.652  15.707   95   22   26    89  18367    14     2
Franche-Comté            1.117  2.448  12.273   90   37   49   120  22282    18     1
Haute-Normandie          1.780  4.387  24.678  163   35   42   140  29971    19     4
Ile-de-France           10.951 55.597 332.885 1421  200  322   960 203814   142    19
Languedoc-Roussillon     2.296  7.125  36.850  253   70   94   252  48725    28     6
Limousin                 0.711  1.490   7.305   54   25   29    73  13386    20     2
Lorraine                 2.310  6.148  34.160  179   57   66   174  38347    20     5
Midi-Pyrénées            2.550  7.305  37.728  196  106  125   231  44228    69     3
Nord-Pas-de-Calais       3.997  9.844  53.362  297   58   76   262  56283    28    10
Pays de la Loire         3.222  8.802  46.538  279  109  133   295  64562    63     8
Picardie                 1.858  3.076  16.809  111   45   54   137  29409    18     3
Poitou-Charentes         1.640  3.895  18.794  144   62   78   171  33636    36     3
Provence-Alpes-Côted'Azur 4.506 15.539  90.640  489  128  190   440  79845    64     5
Rhône-Alpes              5.646 19.989 110.525  625  214  302   655 125764   118    13
```

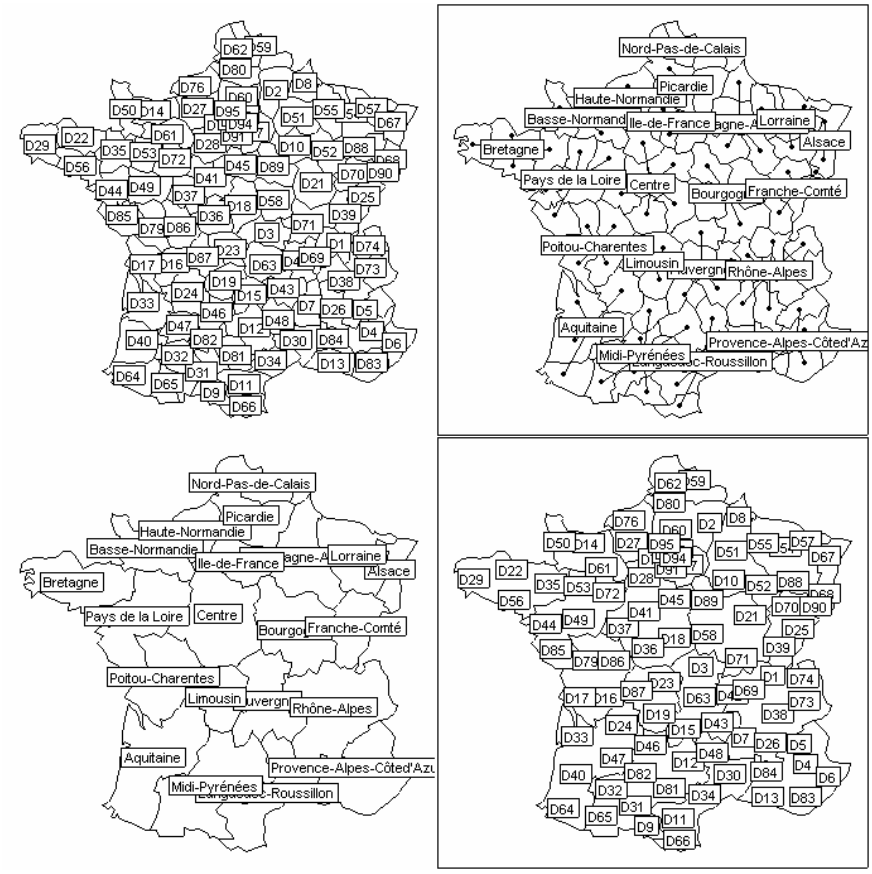
6.3. Ecrire une fonction

C'est une autre manière de travailler. Les données deviennent alors source d'illustrations pour une nouvelle fonction. Si elec88$area contient sous forme de data.frame facteur-x-y les contours des départements on peut chercher à avoir automatiquement les contours par régions :

```
"area.util.class" <- function (area, fac)
{
    if (nlevels(area[, 1] != length(fac)))
        stop("non convenient matching")
    lreg <- split(as.character(unique(area[, 1])), fac)
    "contour2poly" <- function(x) {
        a = paste(x[, 1], x[, 2], sep = "_")
        b = paste(x[, 3], x[, 4], sep = "_")
        a = cbind(a, b)
        points = a[1, 1]
        curr = a[1, 1]
        rowcur = 1
        colcur = 1
        npts = nrow(x)
        for (k in (1:(npts - 2))) {
            colnew = 3 - colcur
            curnew = a[rowcur, colnew]
            points = c(points, curnew)
            a <- a[-rowcur, ]
            coo = which(a == curnew, arr = TRUE)
            rowcur = coo[1, 1]
```

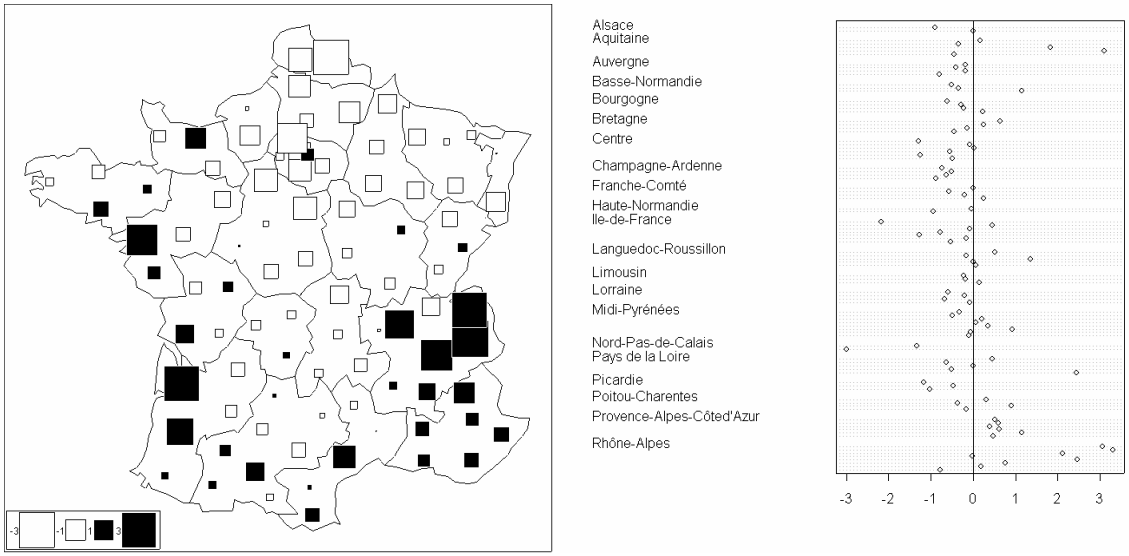
```
            colcur = coo[1, 2]
            curr = a[rowcur, colcur]
        }
        colnew = 3 - colcur
        curnew = a[rowcur, colnew]
        points = c(points, curnew)
        return(matrix(as.numeric(unlist(strsplit(points, "_"))),
            ncol = 2, byr = TRUE))
    }
    "souscontour" <- function(k) {
        sel = unlist(lapply(lreg[[k]], function(x) which(area[,
            1] == x)))
        area.sel = area[sel, ]
        area.sel[, 1] = as.factor(as.character(area.sel[, 1]))
        w = area.util.contour(area.sel)
        w = contour2poly(w)
        w = cbind(rep(k, nrow(w)), w)
        return(w)
    }
    lcontour <- lapply(1:nlevels(fac), souscontour)
    w = lcontour[[1]]
    for (k in 2:length(lcontour)) w <- rbind.data.frame(w, lcontour[[k]])
    w[, 1] <- as.factor(levels(fac)[w[, 1]])
    return(w)
}
```

```
area.dep = elec88$area
area.reg = area.util.class(area.dep,reg.dep)
par(mfrow=c(2,2))
area.plot(area.dep,clab=1)
area.plot(area.dep,clab=0)
s.class(elec88$xy,reg.dep,cell=0,add.p=T)

area.plot(area.reg,clab=1)
area.plot(area.reg,clab=0)
s.label(elec88$xy,clab=1,add.p=T)
```



```
area.plot(area.reg)
xy=elec88$xy[row.names(elec88$xy)!="D75",]
s.value(xy,dudi3$l1[,1],csi=1,add.plot=T)

fac1=reg.dep[row.names(elec88$tab)!="D75"]
dotchart(dudi3$l1[,1],gr=fac1)
abline(v=0)
```

Dans cet exemple, l'essentiel du travail a consisté à observer que la librairie ade4 ne contenait aucun moyen de passer directement des départements aux régions, plus généralement de passer d'un découpage en unités surfaciques à un autre moins fin. L'écriture et l'illustration de la fonction devient alors l'essentiel du travail personnel.



Composante spatiale, sans doute, déterminisme régional, c'est moins sûr.

7. Eléments pour une grille de lecture

La première vertu d'un compte-rendu peut être la transparence, l'honnêteté, le savoir faire, l'originalité, la pédagogie, … Dans tous les cas, il devra être correctement rédigé : c'est la marque du respect du lecteur.

Cinq questions sur la présentation :

Le rapport a-t-il un titre qui indique le contenu et l'intention ? "Analyse de données" "Devoir de Statistique", "Analyse multivariée" ou "Analyse d'un jeu de données" ne sont pas des titres. "Etude du sommeil chez différents mammifères" en est un.

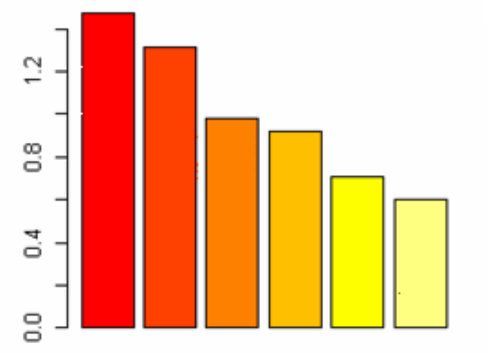
Le rapport a-t-il un plan et une table des matières ? Même pour quatre pages, on peut faire une introduction, deux parties et une conclusion et l'indiquer sous le titre. Eviter les déséquilibres majeurs du style "Origine des données (p. 1) – Résultats (p. 2) – Conclusion (p.17) – Références (p.17)".

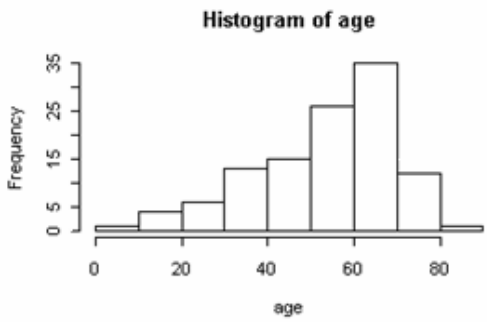
Le rapport utilise-t-il des styles minimaux pour le texte, les lignes de commandes et les légendes des figures ? A proscrire : un tableau de résultat en Times 12 avec 8 chiffres significatifs qui fait désordre ou encore cette matrice de corrélation (sur plusieurs pages) :

```
> round(cor(soupe),dig=3)
       tendr  soup   mou   sec    fibre   grainu sable  homog elast   sale   doux  bouil   roti  aroma  cuisi   gras
tendr  1.000
soup   0.836  1.000
mou    0.838  0.765  1.000
sec   -0.740 -0.723 -0.522 1.0000
fibre -0.303 -0.017 -0.183 0.3639 1.0000
grainu -0.357 -0.417 -0.233 0.8128 0.1154 1.0000
sable -0.428 -0.583 -0.239 0.8525 0.0244 0.8801 1.000
homog  0.076 -0.031  0.140 0.2274 0.1570 0.1751 0.437 1.000
elast  0.012  0.149 -0.059 -0.1048 0.3722 -0.0402 -0.296 -0.288 1.000
sale   0.447  0.636  0.348 -0.5103 0.3678 -0.3680 -0.652 -0.108 0.614 1.000
doux   0.433  0.458  0.602 -0.1760 -0.0014 -0.0878 0.123 0.661 -0.298 -0.058 1.0000
bouil -0.014  0.162  0.314  0.1980 0.2730 0.2184 0.074 -0.278 0.191 0.095 -0.0373 1.000
roti   0.405  0.658  0.388 -0.3877 0.3768 -0.3139 -0.371 0.248 -0.011 0.511 0.4939 -0.057 1.000
aroma  0.614  0.731  0.493 -0.7035 0.1009 -0.5730 -0.615 0.218 0.282 0.717 0.4071 -0.284 0.728 1.0000
cuisi  0.667  0.788  0.580 -0.7746 0.0271 -0.6483 -0.662 0.163 0.147 0.655 0.4489 -0.227 0.747 0.9723 1.0000
gras   0.582  0.750  0.636 -0.4273 0.3519 -0.2490 -0.379 0.232 0.452 0.797 0.4081  0.242 0.612 0.7808 0.7566 1.000
rance  0.144 -0.052  0.236  0.2310 -0.1113 0.4627 0.495 0.216 -0.251 -0.265 0.1143 0.423 -0.064 -0.3094 -0.2544 0.019
```

La rédaction est-elle convenable ? On peut ne pas avoir un style admirable mais on évitera de se croire sur un forum *(... je voulai savoir si vous auriez pas une idee de prog facil qui pourai etre utile car g beau chercher je trouve que des prog assez dur …)*.

Les graphiques sont-ils lisibles ? Penser que le "prof" est âgé et que la vignette microscopique heurte sa presbytie de même que le remplissage pleine page de figures insignifiantes heurte sa confiance en l'intelligence commune.





Cinq questions sur les données :

Les données sont-elles originales et en quoi ? On peut s'intéresser à une foule de sujets mais on évitera le forum statlib ou les data des librairies réputées. On peut prendre la base de données de son club de belote, les carnets de chasse de son père, ou un parmi les milliers de tableaux disponibles sur le web. Eviter simplement de dire "je sais que ces données sont sans intérêt" ou "vous excuserez le caractère bien connu de ce tableau".

Les sources sont-elles correctement citées ? On pourra utiliser tout type d'information avec la citation correcte des sources. Si vous copiez-collez une introduction admirable de précision, de concision et d'humour sans citation, il vaudra mieux que le reste du rapport soit de même qualité. Pensez à Google ! Exemple :

1 Présentation

1.1 Contexte

Le whisky écossais (scotch whisky) est le résultat de la distillation de céréales fermentées sous l'action des levures, et d'eau pure. Le whisky écossais est distillé depuis plusieurs centaines d'années, à l'origine pour une production et une consommation purement personnelle. Chacun avait son propre petit alambic pour jouir des excédants de sa production d'orge de la ferme.

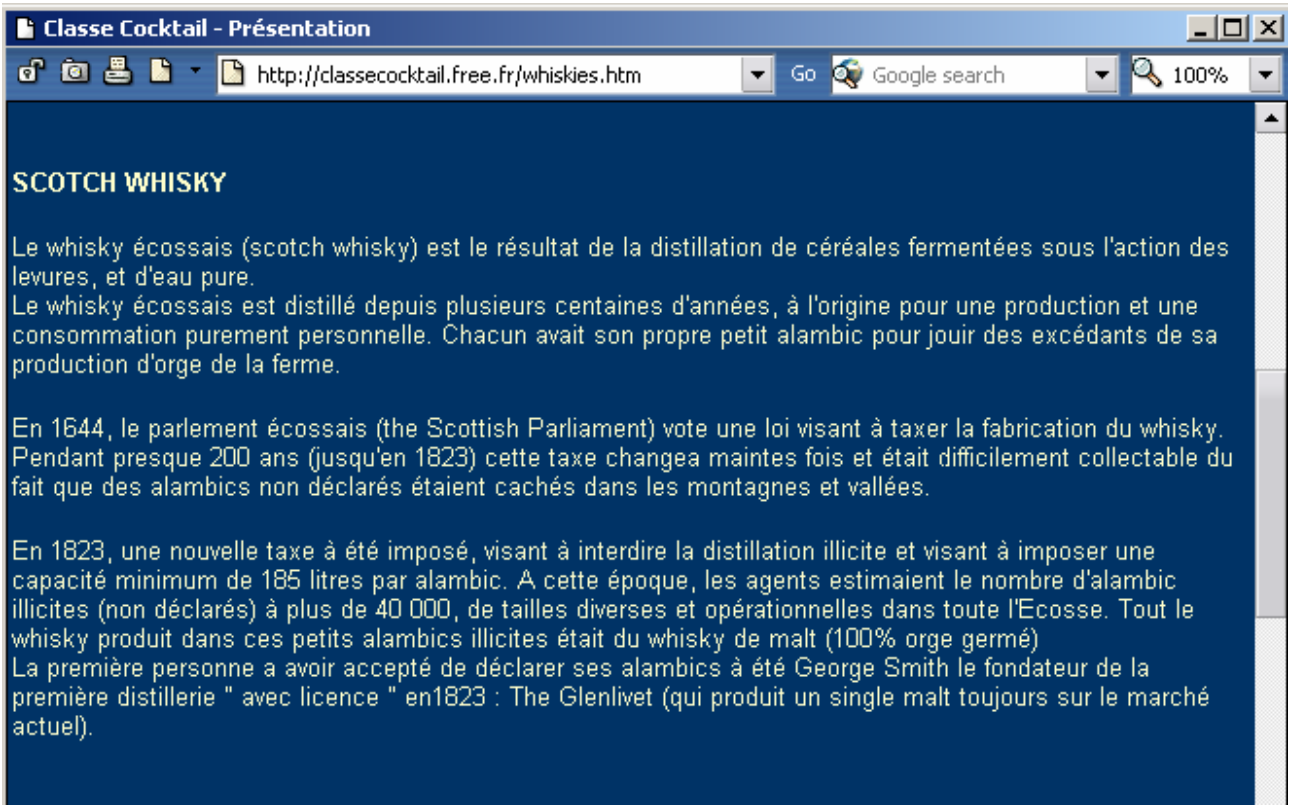
En 1644, le parlement écossais (the Scottish Parliament) vote une loi visant à taxer la fabrication du whisky. Pendant presque 200 ans (jusqu'en 1823) cette taxe changea maintes fois et était difficilement collectable du fait que des alambics non déclarés étaient cachés dans les montagnes et vallées.

En 1823, une nouvelle taxe à été imposé, visant à interdire la distillation illicite et visant à imposer une capacité minimum de 185 litres par alambic. A cette époque, les agents estimaient le nombre d'alambic illicites (non déclarés) à plus de 40 000, de tailles diverses et opérationnelles dans toute l'Ecosse. Tout le whisky produit dans ces petits alambics illicites était du whisky de malt (100% orge germé) La première personne a avoir accepté de déclarer ses alambics à été George Smith le fondateur de la première distillerie " avec licence " en1823 : The Glenlivet (qui produit un single malt toujours sur le marché actuel).

C'est bien, mais ça se gâte très vite :

Depuis des centaines de distillerie, on parsemé toute l'écosse, chacune avec un whisky unique en goût. Je vais travailler ici sur un jeu de données carractérisant 109 whisky écossais. Essayer de carractériser différent groupe de whisky en utilisant des méthodes statistiques.

Donc :



Sans citation ! Sans commentaire ! Carton rouge.

Les données traitées sont-elles accessibles pour un contrôle ? Peut-on en discuter publiquement ? A éviter :

> Les données sont contenues dans le fichier *dipteres.txt*. Lorsque l'on fait, sous R, load(« dipteres.txt »), on a ensuite accès à la liste *dipteres*, composée de plusieurs data.frame. *initData* contient le tableau initial des données, *names* contient les 31 noms des familles, *temps* est la moyenne des trois relevés, *sem1*, *sem2* et *sem3* correspondent aux trois relevés, et *temperature* contient les températures minimales et maximales de chaque semaine de relevé, pour chaque strate.
>
> Ces donnés ont été transmises sous promesse de non diffusion et non publication.

Les données posent-elles un problème ? Lequel ? On peut avoir une question sémantique, technique, mathématique, pédagogique. Il faut annoncer la couleur.

La quantité a peu à voir dans l'affaire, mais à partir d'un certain niveau, l'espoir de faire autre chose que de l'illustration s'amenuise fortement. Il n'y a alors qu'un tableau qui exprime que les données sont traitées ou n'ont pas à l'être, ou encore sont publiées pour être lues directement :

I) Le jeu de données

1 .1) Les sources

Il s'agit d'un jeu de données obtenu sur le site de l'INSEE (http://www.insee.fr/fr/ffc/tef/tef06.pdf)

Activité de la population dans quelques pays européens en 2002 [2]

	Population active occupée millions	Taux d'emploi* %	Taux d'emploi des 55-64 ans %	Taux d'emploi féminin* %
Allemagne	36,3	65,4	38,4	58,8
Autriche	3,7	68,2	28,1	61,1
Belgique	4,1	59,7	25,8	51,1
Danemark	2,7	76,4	57,3	72,6
Espagne	16,2	58,4	39,8	44,0
Finlande	2,4	69,1	47,8	67,3
France	**23,9**	**62,3**	**33,8**	**55,5**
Grèce	3,9	56,9	33,2	42,7
Irlande	1,8	65,0	48,0	55,2
Italie	21,8	55,4	28,6	41,9
Luxembourg	0,2	63,6	27,9	51,5
Pays-Bas	8,2	74,5	42,0	65,9
Portugal	5,1	68,6	51,4	61,2
Royaume-Uni	28,3	71,5	53,3	65,3
Suède	4,3	74,0	68,3	72,5
UE à 15	**163,0**	**64,2**	**39,8**	**55,5**
Hongrie	3,8	56,5	25,9	49,9
Pologne	13,8	51,7	26,6	46,7
Rep. Tchèque	4,8	65,6	40,4	57,2

* Proportion de personnes ayant un emploi dans la population de 15 à 64 ans.

Une question posée, voilà qui est plus prometteur :

> Grâce à ces données, on va pouvoir répondre à plusieurs questions qui se posent : tout d'abord, est-ce que le rang en HDI dépend fortement ou au contraire peu du PNB et beaucoup des autres critères de développement ? est-ce qu'il y a des pays qui sont bien développés matériellement mais peu au vu des autres critères, ou au contraire des pays pauvres mais ayant un niveau d'éducation et d'espérance de vie grand au vu de leur situation matérielle ? Peut-on dégager des groupes parmi ces tendances, et comment expliquer ces groupes (cela ne peut pas s'expliquer directement par les données elles-mêmes bien sûr, mais par les liens entre les pays concernés...) ? Quels sont les pays où les deux types de facteurs s'équilibrent bien ?
> On peut ensuite voir l'influence géographique, l'influence du régime politique peut-être, ou encore d'une culture particulière...et ainsi voir que la notion de développement est assez relative, si les pays considérés habituellement comme développés ne le sont pas beaucoup au niveau social en comparaison avec leurs richesses.

Ou encore cette introduction, véritable cas d'école :

> ## 1-Introduction
>
> **Etant moi-même joueur de golf, j'ai décidé de m'intéresser aux statistiques de l'année 2002 des 50 meilleurs joueurs du PGA tour. Le PGA tour se déroule aux Etats-Unis et est considéré comme le meilleur circuit au monde. Tous les champions actuels s'y côtoient, on peut ainsi y voir évoluer le désormais très célèbre Tiger Woods et d'autres champions bien connus des initiés comme Phil Mickelson ou Ernie Els.**
>
> **Le site www.pgatour.com regroupe toutes les statistiques de ces joueurs de façon très détaillée et a donc été ma source de renseignements.**
>
> **Dans cette analyse , je vais essayer de « décrypter » toutes ces données diverses avec les outils que je possède. Ainsi je pourrais peut-être faire ressortir des faits intéressants pouvant me permettre de mieux comprendre le golf.**
>
> ## 2-Termes techniques et définitions
>
> **Pour les non-initiés, commençons par donner quelques définitions nécessaires à la compréhension de l'analyse (le schéma de la page suivante montre la localisation des différentes aires de jeu) :**

Il faut bien, un jour ou l'autre arrêter de faire des exercices pour faire plaisir à l'enseignant :

1 Introduction

Dans le cadre du cours de statistiques et d'analyse de données, nous allons étudier un jeu de données. Le jeu choisi concerne le taux d'homicide à Détroit de 1961 à 1973. Plusieurs facteurs ont été mesurés chaque année. Ainsi chaque ligne du tableau de données représente une année. Les différents facteurs sont décrits ci-dessous. Il sera possible d'émettre des hypothèses sur ces données. Nous pourrons alors voir si elles se vérifient ou si au contraire elles sont contredites par l'analyse statistique.

Les données ont-elles des qualités techniques acceptables ? On ne peut pas donner la liste des erreurs potentielles. Simplement, un exemple :

Mes données proviennent du Quid 2002 et concernent les croyances religieuses de la population mondiale. Elles sont résumées dans un tableau à double entrée. La planète est divisée en 6 grandes régions : l'Afrique, l'Amérique du Nord, l'Amérique latine, l'Europe et l'Océanie.

Les croyances religieuses, quant à elles, sont divisées en 19 ou 23 catégories, selon qu'on regroupe tous les chrétiens dans un seul et même groupe ou qu'on les subdivise en 5 grandes classes : anglicans, catholiques romains, protestants, orthodoxes et autres chrétiens. On distingue parmi les individus d'une part les non-croyants (athées et agnostiques) et d'autre part les croyants. Ceux-là sont très subdivisés entre les différentes grandes religions du monde, les religions nouvelles (fondées à partir de 1800 mais dont la majorité l'ont été à partir de 1945) et les religions diverses. Ce tableau permet donc de diviser la population mondiale selon deux variables : la géographie, qui contient 6 modalités, et la religion, qui en contient 23.

On a donc un tableau de dimension 19x6. Les effectifs sont en milliers d'individus.

Voici le tableau considéré :

	Afrique	Am Nord	Am lat	Asie	Europe	Océanie
Agnostiques	4877	28201	1824	602992	107478	3268
Athées	411	1628	2714	121467	23140	360
Animistes	94934	434	1266	127260	1264	263
Baha'is	1694	770	850	3382	128	108
Bouddhistes	132	2637	635	351043	1533	290
Rlg chinoises trad	32	844	190	380250	253	63
Chrétiens	***351276***	***258770***	***473713***	***306401***	***559212***	***24809***
Confucianistes	0	0	0	6219	11	23
Hindouistes	2312	1308	761	792897	1401	349
Jaïnistes	65	7	0	4079	0	0
Juifs	212	6015	1133	4323	2534	96
Mandéens	***0***	***0***	***0***	***38***	***0***	***0***
Musulmans	310529	4389	1646	807034	31219	292
Shintoïstes	0	56	7	2715	0	0
Sikhs	52	514	0	22015	238	18
Spiritistes	3	149	11894	0	131	7
Zoroastriens	1	76	0	2407	1	1
Nouvelles rlg	28	813	613	99734	156	62
Rlg diverses	65	591	96	23	235	9

La ligne 7 représente 33 % des individus, la ligne 12 moins de 7 millionièmes. Peut-on considérer que l'AFC du tableau comparera des modalités comparables ? Pourquoi six régions ? Comment ces chiffres sont-ils obtenus ? Quelle réalité recouvrent-ils ?

Cinq questions sur les méthodes :

Sont-elles basiques ou évoluées, simples ou complexes, variées ou limitées ?

Sont-elles adaptées aux contraintes ? Ne faites pas semblant de découvrir de l'information avec une ACP sur trois variables.

Sont-elles adaptées aux objectifs ? Une ACP à trois variables est justifiée au titre de l'illustration d'une idée.

Sont-elles en interaction avec les données ? Le traitement est-il un processus ou une simple recette ?

L'exécution technique est-elle maîtrisée ? Eviter l'image de l'exercice de commande qui n'amuse personne. Reconnaître les difficultés peut être la marque des bons travaux :

I/ Contexte

Les données que j'ai décidé d'étudier correspondent aux notes de gymnastes au championnat de France UFOLEP, le 1er et 2 juin 2002 à Elancourt.

La gymnastique est une discipline où :'on retrouve (chez les féminines) quatre agrès : le sol, le saut de cheval, la poutre et les barres asymétriques.

La compétition regroupe deux fédérations : l'UFOLEP (Union Française des Œuvres Laïques d'Education Physique) et la FFG (Fédération Française de Gymnastique). Certaines filles sont en double appartenance, c 'est à dire qu'elles sont affiliées aux deux fédérations :

...

Conclusion

L'étude de ce jeu de données a présenté pour moi de nombreux intérêts.

D'une part, je me suis rendue compte de la difficulté d'analyser un jeu de données **même en connaissant relativement bien le sujet**. En effet, toutes difficultés statistiques mises à part, il n'est pas évident de donner un sens aux données. J'ai eu beaucoup de difficultés à donner un sens au deuxième facteur de l'ACP (sans parler du troisième). Peut-petre me manquait-il des informations... la taille des gymnastes aurait pu être une information intéressante qui aurait pu expliquer certains résultats...

L'analyse de ce jeu de données m'a aussi permis de comprendre certaines choses d'un point de vu strictement statistique (l'utilisation de certaines fonctions graphiques, les tests non paramétriques....)

Cinq questions sur l'ensemble du rapport :

Existe-t-il un lien cohérent entre introduction et conclusion ?

Les données ont-elles encore un intérêt après l'analyse ? L'analyse peut montrer qu'on a eu tort d'attaquer cette question, l'essentiel est que l'analyse apporte réellement quelque chose.

Quelle est l'impression générale au niveau technique ?

Quelle est l'impression générale au niveau de la démarche ?

Quelle est l'impression générale au niveau de la communication ?

En bref, il y a de multiples manières de faire un bon rapport et quelques unes pour en faire un mauvais. La statistique regroupe les techniques d'extraction de l'information des sources numériques. La question de fond est donc **qu'a-t-on appris de l'analyse des données** ? Qu'il s'agisse de production d'un troupeau de vaches laitières, de manières d'être un bon joueur de golf ou de biais d'usage des tétranucléotides chez les virus, si l'usage d'une technique quelconque a rendu visible une information plus ou moins cachée, le contrat est rempli. Si le seul résultat obtenu est une trivialité (du genre il fait constamment plus chaud dans le sud), la statistique est une ânerie inutile. Mais non !