Scénario: AFD et exploration de données socio-économiques

Résumé

Analyse de données socio-économiques par ACP puis par Analyse factorielle discriminante avec SAS puis R.

1 Introduction

Ce scénario se focalise sur l'exploration de données socio-économiques observées sur les différents départements regroupés en régions administratives. La question posée est : les régions définies géographiquement sont elles relativement homogènes sur ces aspects socio-économiques. Ce peut être une analyse éclairante que la volonté de supprimer les départements et regrouper des régions. On répond à cette question de façon indirecte et exploratoire, en cherchant à savoir si les variables socio-économiques permettent de discriminer les régions.

Il y a différentes approches du problème de la discrimination selon les hypothèses admises et la méthodologie mise en œuvre. Nous nous limitons ici à l'approche descriptive ou "factorielle". L'approche "décisionnelle" est traité dans le module de data mining. L'objectif est de représenter graphiquement la qualité de discrimination des classes définies par la variables qualitatives à partir des combinaisons linéaires des variables quantitatives explicatives.

2 AFD des départements avec SAS

2.1 Les données

Les données proviennent du Groupe d'Étude et de Réflexion Inter-régional (GERI). Elles décrivent, quatre grands thèmes : la démographie, l'emploi, la fiscalité directe locale, la criminalité, de chacun des départements français métropolitains et la Corse. Les indicateurs ont été observés pendant l'année 1990, ils sont, pour la plupart, des taux calculés relativement à la population totale du département concerné. Voici leur liste :

- identificateur : numéro du département,
- identificateur : code du département,
- identificateur : code de la région,
- URBR : indicateur de concentration de la population mesurant le caractère urbain ou rural d"un département,
- TXCR : taux de croissance de la population sur la période intercensitaire 1982-1990,
- JEUN : part des 0-19 ans dans la population totale,
- AGE : part des plus de 65 ans dans la population totale,
- FE90 : taux de fécondité (pour 1000) égal au nombre de naissances rapportés au nombre de femmes fécondes (15 à 49 ans) en moyenne triennale,
- ETRA : part des étrangers dans la population totale,
- CHOM : taux de chômage,
- CRIM : taux de criminalité : nombre de délits par habitant,
- FISC : produit, en francs constants 1990 et par habitant des quatre taxes directes locales (professionnelle, habitation, foncier bâti, foncier non bâti).

Parts de chaque profession et catégorie socioprofessionnelle (PCS) dans la population active occupée du département :

- AGRI : agriculteurs,
- ARTI : artisans,
- CADR : cadres supérieurs,
- EMPL : employés,
- OUVR : ouvriers,
- PROF : professions intermédiaires,

2.2 Lecture

Charger le fichier depart.dat puis exécuter le programme suivant afin de lire les données et créer la table SAS correspondante :

```
data sasuser.depart;
/* Attention au nom du répertoire */
infile "depart.dat";
input num $ depart $ region $ txcr etra urbr jeun
          age chom agri arti cadr empl ouvr prof
          fisc  crim fe90;
```

```
run;
```

2.3 Analyse en composantes principales

Une première étape descriptive n'a pas conduit à des re-transformations des variables. Celles-ci sont, pour la plupart, déjà des taux (pourquoi ?). Une ACP permet de se faire une première idée sur l'organisation de ces données. Ne pas oublier de charger et exécuter préalablement les macros-commandes.

```
%acp(depart,num,txcr--fe90);
%gacpsx;
%gacpbx;
%gacpvx;
%gacpix;
```

Combien d'axes ? Interpréter ces axes.

2.4 Analyse discriminante

On se propose de mettre en évidence les plus grandes disparités inter-régionales et donc de rechercher les variables ou combinaisons de variables expliquant au mieux le découpage régional. Autre question : les régions administratives sont-elle homogènes d'un point de vue socio-économique. Pour simplifier, nous procédons à des regroupements afin de construire des régions moins nombreuses comprenant des nombres de départements plus semblables. D'autre part, la région "Ile de France", trop particulière et donc trop facile à discriminer, est laissée à part. Elle très influente en définissant à elle seule le premier axe.

2.5 Regroupements

Le programme ci-dessous procède au regroupement des régions. Attention, il est sensible à la casse (majuscules et miniscules) des lettres.

```
data sasuser.depart;
set sasuser.depart;
select;
/* ATTENTION, respecter la casse
     (minuscules et majuscules) */
when(region in"NPC","Pic","HNo","ChA"))
    groupreg="Nd";
when(region in"Als","Lor","FrC")) groupreg="Es";
when(region in"BNo","Bre","PaL")) groupreg="Ws";
when(region in"Cen","Bou")) groupreg="CN";
when(region in"PoC","Lim")) groupreg="CW";
when(region in"Auv","RhA")) groupreg="CE";
when(region in"Aqu","MiP")) groupreg="SW";
when(region in"LaR","PAC","Cor")) groupreg="SE";
otherwise delete;
end;
run;
```

Charger et exécuter les macros commandes de l'analyse discriminante.

```
%afd(depart,num,groupreg,txcr--fe90,7);
```

Retrouver dans les résultats de la fenêtre `output` les différents tableaux issus du calcul de l'AFD et nécessaires à l'interprétation ou au tracé des graphiques : matrice des corrélations, des corrélations expliquées, valeurs propres, corrélations variables×facteurs (BCS), les coordonnées des barycentres dans la nouvelle base, celles des individus.

```
%gafdvx;
%gafdix;
```

Interpréter les axes factoriels discriminants. Que pouvez vous dire sur la cohésion régionale ?

3 AFD avec R

L'analyse factorielle discriminante n'est pas directement accessible dans une librairie, elle nécessite quelques calculs pour obtenir une représentation graphique satisfaisante. C'est l'objet de cette section.

3.1 Importation des données

Comme précédemment, il faudrait adapter le fichier pour une importation facile dans R en ajoutant une première ligne contenant le nom des variables. Il

est plus simple, et utile à savoir, de transférer directement les données de SAS vers R. Cela évite de reprogrammer en R le regroupement des départements.

```
/* Exportation d'un fichier SAS en format .csv */
proc export data=sasuser.dat.depart
     outfile= "depart.csv"
     DBMS=CSV REPLACE;
run;
```

Rechercher ou SAS a "rangé" le ficher `depart.csv`, sans doute dans le répertoire duquel l'exécution a été lancé. Éventuellement le déplacer dans le répertoire courant de R.

```
depart=read.csv("depart.csv")
# Vérifier que tout s'est bien passé
summary(depart)
# rectification de la première colonne
depart[,1]=as.character(depart[,1])
```

3.2 AFD avec la fonction lda de R

La fonction `lda` (*linear discriminant analysis*) de R calcule bien les facteurs discriminants et les coordonnées des individus mais, cela manque, pas celles des barycentres des classes. Ceci nécessite quelques calculs complémentaires pour l'obtention d'un graphique lisible.

```
library(MASS)
# calcul de l'afd
dep.afd=lda(depart[,4:18],depart$groupreg)
print(dep.afd)
plot(dep.afd)
# calcul des barycentres
dep.pred=predict(dep.afd,data=depart[,4:18])
m=matrix(rep(0,15),nrow=8,ncol=2)
for (i in 1:8){
 for (j in 1:2){
 m[i,j]=mean(dep.pred$x[
    unclass(depart$groupreg)==i,j])}}
# graphe dans les axes 1 et 2
color=as.integer(depart$groupreg)
plot(dep.pred$x[,1],dep.pred$x[,2],
    bg=color,pch=21)
abline(0,0,h=0); abline(0,0,v=0)
text(m[,1],m[,2],labels=
    levels(depart$groupreg),cex=1,col=1:8)
```