

1 Les données

Nous considérons une partie des données issues de l'enquête "Les étudiants et la ville" effectuée en 2001 par des étudiants de sociologie sous la direction de S. Denèfle à l'Université de Tours. Cet exemple est décrit dans : Crucianu M., Asselin de Beauville J-P., Boné R., Méthodes factorielles pour l'analyse des données, Hermès-Lavoisier 2004.

L'analyse porte sur cinq questions en rapport avec le logement étudiant. L'ensemble des individus statistiques est ici un échantillon de 383 étudiants. Les questions sont les suivantes reportées dans la table ci-dessous.

Pour l'ACM, il est quasiment indispensable de regrouper les modalités dont la fréquence est trop faible (inférieure à 5% par exemple) avec d'autres modalités. Aussi, dans les données qui suivent, les modalités "Parents" et "NR1" ont été regroupées pour la variable "mode", de même que "NA" et "NR3" pour la variable "ancienneté" et ">5km" et "NR4" pour la variable "éloignement". Il reste donc 22 modalités distinctes.

2 ACM avec R

Plusieurs packages fournissent des outils permettant de réaliser une analyse factorielle des correspondances. On peut citer :

- dans le package MASS : `mva`
- dans le package ade4 : `dudi.acm`
- dans le package FactoMineR (cf. <http://factominer.free.fr>) : `MCA`

Les résultats les plus complets semblent être ceux fournis par la procédure MCA. Nous allons donc utiliser FactoMineR.

ACM

1. Installer le package FactoMineR à partir du fichier zip.

Question	Réponse possible	Abbréviation
Habitez-vous (variable "mode d'occupation")	Seul Colocataires En couple Avec les parents Non réponse	<i>Seul</i> <i>Coloc</i> <i>Couple</i> <i>Parents</i> <i>NR1</i>
Quel type d'habitation occupez-vous ? (variable "type d'habitation")	Cité universitaire Studio Appartement Chambre chez un particulier Autre Non réponse	<i>Cité</i> <i>Studio</i> <i>Appart</i> <i>Chambre</i> <i>Autre</i> <i>NR2</i>
Si vous vivez en dehors du foyer familial, depuis combien de temps ? (variable "ancienneté")	Moins de 1 an De 1 à 3 ans Plus de 3 ans Non applicable Non réponse	<i>< 1 an</i> <i>1-3 ans</i> <i>> 3 ans</i> <i>NA</i> <i>NR3</i>
A quelle distance approximative de la Fac vivez-vous ? (variable "éloignement")	Moins de 1 km De 1 à 5 km Plus de 5 km Non réponse	<i>< 1 km</i> <i>1 à 5 km</i> <i>> 5 km</i> <i>NR4</i>
Quelle est la superficie de votre logement ? (variable "superficie")	Moins de 10 m ² De 10 à 20 m ² De 20 à 30 m ² plus de 30 m ² Non réponse	<i>< 10 m²</i> <i>10 à 20 m²</i> <i>20 à 30 m²</i> <i>>30 m²</i> <i>NR5</i>

TABLE 1 – Tableau descriptif des données des étudiants en ville

2. Lire les données stockées dans le fichier `Etudiants-Ville-protocole.RData`.
3. Visualiser les distributions des différentes variables et vérifier qu'aucune modalité n'est trop sous représentée.
4. Réaliser une analyse des correspondances multiples en utilisant la fonction `MCA` et produire les graphiques suivants :
 - Projection simultannée des individus et des modalités sur le premier plan factoriel.
 - Projection des variables sur le premier plan factoriel.
 - Projection des modalités sur le premier plan factoriel.
 Certains graphiques sont produits par défaut par la fonction `MCA`, et on obtient les autres en utilisant, par exemple, la fonction `plot.MCA`.
5. Utiliser les graphiques et les sorties de la fonction `MCA` pour interpréter les résultats.

2.1 Comparaison de l'ACM avec l'ACP

1. Comparer les résultats (graphiques) de l'ACM avec les résultats obtenus sur le tableau disjonctif complet par ACP.

3 ACM avec SAS

1. Lire les données à partir du fichier `Etudiants-ville-protocole1.csv`.
2. Utiliser la `proc freq` pour identifier les modalités à faible effectif (par exemple inférieur à 5%). On rappelle qu'on peut le mot clé `_character_` permet de générer implicitement la liste de toutes les variables caractère.
3. Utiliser la `proc corresp` pour réaliser l'analyse des correspondances multiples sur les modalités. On pourra se servir de l'outil d'analyse interactive des données ou de la macro `%gacfx` pour réaliser les graphiques. Il faudrait tenir compte du fait que la macro `%gacfx` lit les coordonnées nécessaire dans une table nommée `resul`.
4. Comment peut-on adapter les commandes de la question précédente pour réaliser l'analyse des correspondances aussi sur les individus? Projeter les individus et les modalités sur le premier plan factoriel.
5. Les variables `Type` et `Superf` sont redondantes. Comment peut-on faire pour pour définir la variable `Superf` en variable supplémentaire? Refaire les graphiques. L'interprétation des axes est-elle rendue plus aisée? Vous semble t'il intéressant de passer d'autres variables en variable supplémentaire?
6. Créer le tableau disjonctif complet en utilisant les instructions suivantes :

```
%let listev = Mode Type Ancien Eloign Superf ;
proc glmmod data=TP4.etudiants outdesign=TP4.ind noprint;
class &listev ;
model ident = &listev /noint ;
run ;
```

```
%let nbc0l = 22;  
%let listmod = Col1--Col&nbc0l ;
```

Puis réaliser l'AFCM sur le tableau disjonctif complet et tracer le graphe présentant la projection correspondante des modalités et des individus sur les 2 premiers plans factoriels. Vous pouvez vous aidez des commandes ci-dessous.

```
proc corresp data=TP4.ind outc=graphe dimens=4;  
var &listmod ;  
id ident ;  
run ;  
%plotit(data=graphe,plotvars=dim2 dim1,labelvar=ident,  
datatype=corresp,options=square) ;
```