TD9 : les étudiants et la ville (correction)

```
> require(ade4)
> #etud.ville<-read.csv("etudiants-ville.csv", sep=';')
> #etud.ville
> #log<-acm.burt(etud.ville, etud.ville)
> log<-read.table("logement.txt")
> # Tableau de Burt avec des trous
> logNA=log[1:11,1:11];
> logNA[1,9:10] = logNA[9:10,1] = logNA[9,4]= logNA[4,9] = logNA[10,10] = NA
> #on enleve les variables occ et anc
> inact=c(1:5, 12:16)
> log1=log[-inact,-inact]
> # On fait une AFC du tableau de Burt.
> coa1<-dudi.coa(log1, scannf=F,nf=20);
> acm1<-coa1
> class(acm1)<-c("acm", "dudi")
> # Les valeurs propres sont le carré des valeurs normales
> acm1$eig<-sqrt(coa1$eig)
> # on cherche à corriger les variances
> acm1$co<-sweep(coa1$co, 2, sqrt(acm1$eig[1:acm1$nf]), "/")
> inert1<-inertia.dudi(acm1,c=T)
> inert1$col.cum<-sweep(inert1$col.cum, 1, rowSums(abs(inert1$col.rel)), "/")*100
> colnames(inert1$col.abs)<-gsub("(%)", "", colnames(inert1$col.abs), fixed=T)
> # les poids des catégories
> poids=as.matrix(round(apply(log,1,sum) / sum(log) * 5 * 100), nrow=nrow(log))
> colnames(poids) = "Poids (%)"
> # variables supplémentaires
> acm.suppl = function(acm, p, tab)
+ {
+  tab = as.matrix(tab)
+  p_eff=apply(tab,1,sum)
+  eff = p_eff / p
+  li=diag(1/p_eff) %*% tab %*% as.matrix(acm$li) %*% diag(1/acm$eig)
+  rownames(li) = rownames(tab)
+  colnames(li)=paste("Comp", seq(1:ncol(acm$li)), sep="")
+  test= diag(sqrt(eff*(sum(eff)-1)/(sum(eff)-eff))) %*% li
+  rownames(test) = rownames(tab)
+  colnames(test)=colnames(li)
+  list(li=li, test=test, eff=eff)
+ }
> suppl1=acm.suppl(acm1, 3, log[inact, -inact])
```

1 Les données

Les données qui suivent sont 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 François Rabelais de Tours. 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 :

- Habitez-vous (variable `occ`) : seul (modalité `seul`), en colocation (`coloc`), en couple (`couple`), avec les parents (`parents`), non réponse (`NR`) ?
- Quel type d'habitation occupez-vous (variable `typ`) : cité universitaire (`cite`), studio (`studio`), appartement (`appart`), chambre chez un particulier (`chambre`), autre (`autre`), non réponse (`NR`) ?
- Si vous vivez en dehors du foyer familial, depuis combien de temps (variable `anc`) : moins de 1 an (`0.1an`), 1 à 3 ans (`1.3ans`), plus de 3 ans (`p3ans`), non applicable (`NA`), non réponse (`NR`) ?
- À quelle distance approximative de l'université vivez-vous (variable `dst`) : moins de 1 km (`0.1km`), 1 à 5 km (`1.5km`), plus de 5 km (`p5km`), non réponse (`NR`) ?
- Quelle est la superficie de votre logement (variable `sur`) : moins de 10 m^2 (`0.10m2`), 10 à 20 m^2 (`10.20m2`), 20 à 30 m^2 (`20.30m2`), plus de 30 m^2 (`p30m2`), non réponse (`NR`) ?

Dans ce qui suit, on représentera les catégories par le nom de la variable suivi du nom de la modalité, comme par exemple `occ.seul` ou `typ.NR`. Non réponse (`NR`) correspond à un défaut des données (réponse oubliée ou non fournie), alors que non applicable (`NA`) est utilisé pour les questions qui n'ont pas de sens pour un individu donné.

Les taux marginaux de réponses aux différentes question (en %) sont reproduit ici

```
> poids[1:5,1,drop=F]   > poids[6:11,1,drop=F]  > poids[12:16,1,drop=F] > poids[17:20,1,drop=F] > poids[21:25,1,drop=F]
```

	Poids (%)		Poids (%)		Poids (%)		Poids (%)		Poids (%)
occ.seul	48	typ.cite	11	anc.0.1an	21	dst.0.1km	27	sur.0.10m2	9
occ.coloc	14	typ.studio	28	anc.1.3ans	25	dst.1.5km	50	sur.10.20m2	18
occ.couple	13	typ.appart	30	anc.p3ans	29	dst.p5km	21	sur.20.30m2	25
occ.parents	23	typ.chambre	5	anc.NA	25	dst.NR	2	sur.p30m2	39
occ.NR	1	typ.autre	20	anc.NR	1			sur.NR	9
		typ.NR	6						

On reproduit ci-dessous la partie du tableau de Burt qui correspond aux variables `occ` et `typ` (le tableau total serait trop grand).

```
> logNA
```

	occ.seul	occ.coloc	occ.couple	occ.parents	occ.NR	typ.cite	typ.studio	typ.appart	typ.chambre	typ.autre	typ.NR
occ.seul	185	0	0	0	0	34	90	40	NA	NA	5
occ.coloc	0	53	0	0	0	5	6	32	2	3	5
occ.couple	0	0	50	0	0	2	10	34	0	3	1
occ.parents	0	0	0	90	0	0	1	9	NA	67	8
occ.NR	0	0	0	0	5	0	1	1	0	0	3
typ.cite	34	5	2	0	0	41	0	0	0	0	0
typ.studio	90	6	10	1	1	0	108	0	0	0	0
typ.appart	40	32	34	9	1	0	0	116	0	0	0
typ.chambre	NA	2	0	NA	0	0	0	0	20	0	0
typ.autre	NA	3	3	67	0	0	0	0	0	NA	0
typ.NR	5	5	1	8	3	0	0	0	0	0	22

Question 1 *7 valeurs sont manquantes (NA) dans le tableau de Burt. Retrouvez les en utilisant les propriétés du tableau.*

On commence par calculer (`typ.chambre`, `occ.parents`) en soustrayant de l'effectif total de `occ.parents` (90) les effectifs connus croisés avec `typ` : on trouve $90 - 1 - 9 - 67 - 8 = 5$.

Pour l'effectif de (`typ.chambre`, `occ.seul`), on fait la somme des effectifs de `typ.chambre` croisé avec les autres modalités de `occ` que l'on compare à l'effectif total de `typ.chambre` (20). On trouve alors $20 - 2 - 5 = 13$.

On peut maintenant calculer l'effectif de (`typ.autre`, `occ.seul`) en regardant les effectifs croisés de `occ.seul`, ce qui donne $185 - 34 - 90 - 40 - 13 - 5 = 3$.

Finalement, l'effectif de `typ.autre` est $3 + 3 + 3 + 67 = 76$.

Question 2 *Est-il exact de dire que « la proportion des gens seuls qui vivent en appartement est plus faible que la proportion des couples qui vivent en appartement » ?*

La proportion de gens seuls qui vivent en appartement est $40/185 = 21\%$. La proportion des couples qui vivent en appartement est $34/50 = 68\%$. L'assertion ci-dessus est donc exacte, même si le nombre de couples vivant en appartement (34) est plus faible que le nombre de personnes seules vivant en appartement (40).

Question 3 *Expliquer pourquoi les catégories `occ.NR`, `anc.NR` et `dst.NR` risquent de poser des problèmes dans une ACM. Quelles pourraient être les solutions envisageables ?*

Ces catégories ont un taux profil marginal très faible ($< 3\%$). Par conséquent, leur contribution à l'inertie totale sera importante. Par exemple, pour occ.NR, on obtient $\frac{1}{5}(1 - 0.0131) = 19.7\%$, ce qui est très important puisqu'aucune catégorie ne peut avoir une contribution supérieure à $1/5 = 0,20$. On a donc un total de 60% de l'inertie totale pour ces trois variables, ce qui ne manquera pas d'influencer les premiers axes principaux (alors qu'elles n'ont pas d'interprétation intéressante).

La solution dans ce cas est normalement de fusionner ces catégories avec d'autres (ce qui veut dire cumuler les effectifs correspondants). La manœuvre est un peu délicate, puisqu'il faut choisir une catégorie d'accueil avec un profil pas trop différent.

Notons qu'il n'est pas possible de traiter une catégorie en élément supplémentaire ; seule une variable peut être gérée comme ça. Il n'est pas possible non plus de retirer les données correspondantes si l'on n'a que le tableau de Burt. Par contre, à partie des données complètes, on peut envisager de retirer les individus de ces catégories..

2 Analyse des correspondances multiples

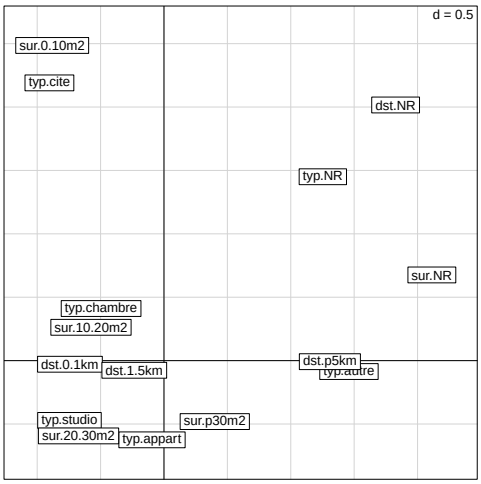
On réalise une analyse des correspondances multiples (ACM) uniquement sur les variables `typ`, `dst` et `sur`, qui décrivent les caractéristiques des logements. Les 6 premières valeurs propres sont

```
> round(acm1$eig[1:6],4)

[1] 0.6990 0.5737 0.5254 0.3892 0.3558 0.3359
```

On donne ci-dessous pour le premier plan factoriel : la projection des catégories actives, les coordonnées de ces catégories sur chaque axe factoriel, leur contribution aux axes (en %), ainsi que la qualité de leur représentation par le plan (en % encore).

```
> s.label(acm1$co)#[-c(6,10,16,20,25),1:2])
```




```
> round(acm1$co[,1:2],2)
```

	Comp1	Comp2
typ.cite	-0.90	2.19
typ.studio	-0.75	-0.47
typ.appart	-0.09	-0.62
typ.chambre	-0.50	0.41
typ.autre	1.46	-0.09
typ.NR	1.25	1.45
dst.0.1km	-0.74	-0.03
dst.1.5km	-0.23	-0.08
dst.p5km	1.31	-0.01
dst.NR	1.83	2.02
sur.0.10m2	-0.88	2.49
sur.10.20m2	-0.58	0.26
sur.20.30m2	-0.67	-0.59
sur.p30m2	0.40	-0.48
sur.NR	2.11	0.68

```
> round(inert1$col.abs[,1:2],1)
```

	Axis1	Axis2
typ.cite	4.2	29.8
typ.studio	7.5	3.6
typ.appart	0.1	6.9
typ.chambre	0.6	0.5
typ.autre	20.1	0.1
typ.NR	4.3	7.0
dst.0.1km	7.1	0.0
dst.1.5km	1.3	0.2
dst.p5km	17.1	0.0
dst.NR	3.7	5.6
sur.0.10m2	3.4	32.8
sur.10.20m2	2.8	0.7
sur.20.30m2	5.4	5.1
sur.p30m2	3.0	5.3
sur.NR	19.5	2.4

```
> round(inert1$col.cum[,1:2],1)
```

	Axis1	Axis1:2
typ.cite	9.8	67.3
typ.studio	21.8	30.6
typ.appart	0.4	17.3
typ.chambre	1.4	2.3
typ.autre	52.7	52.9
typ.NR	9.6	22.4
dst.0.1km	20.2	20.3
dst.1.5km	5.5	6.0
dst.p5km	45.2	45.2
dst.NR	8.0	17.8
sur.0.10m2	7.8	70.0
sur.10.20m2	7.2	8.6
sur.20.30m2	15.0	26.6
sur.p30m2	10.3	25.2
sur.NR	44.9	49.5

Les variables `occ` et `anc`, qui décrivent les occupants des logements, sont utilisées comme variables supplémentaires qualitatives. On calcule les coordonnées et les valeurs test suivantes pour leurs catégories sur les deux premiers axes factoriels

```
> round(suppl1$li[,1:2],2)
```

	Comp1	Comp2
occ.seul	-0.55	0.14
occ.coloc	-0.02	-0.21
occ.couple	-0.06	-0.36
occ.parents	1.13	0.00
occ.NR	0.77	0.61
anc.0.1an	-0.48	0.18
anc.1.3ans	-0.32	-0.02
anc.p3ans	-0.33	-0.20
anc.NA	1.06	0.06
anc.NR	1.60	1.12

```
> round(suppl1$test[,1:2],2)
```

	Comp1	Comp2
occ.seul	-8.61	2.19
occ.coloc	-0.15	-1.59
occ.couple	-0.42	-2.63
occ.parents	11.44	0.02
occ.NR	1.74	1.36
anc.0.1an	-4.55	1.73
anc.1.3ans	-3.30	-0.19
anc.p3ans	-3.77	-2.26
anc.NA	11.00	0.62
anc.NR	2.77	1.94

Question 4 *Calculer la proportion d'inertie expliquée si l'on conserve les deux premiers axes. Commentez la qualité.*

Il y a seulement 3 variables actives (`typ`, `dst` et `sur`). Le nombre de catégories est la somme du nombre de modalités de chaque variable : $6 + 4 + 5 = 15$. Le nombre de valeurs propres que l'on obtiendra est donc $15 - 3 = 12$ et l'inertie totale est $12/3 = 4$.

En conservant deux axes, on a une inertie expliquée égale à $0.70 + 0.57 = 1.27$, soit 31.8%. Il faut donc s'attendre à une qualité de représentation assez faible sur le premier plan. On aurait pu chercher à conserver plus d'axes pour voir si le résultat est interprétable.

Question 5 *Quelles sont les catégories qui définissent les deux premiers axes factoriels ? (on précisera sur quel(s) critère(s) on se fonde).*

On regarde d'abord la contribution de chaque catégorie au premier axe factoriel en la comparant à son poids (on cherche celles pour lesquelles le rapport est supérieur à 1). La contribution d'une catégorie a un axe est est

$$\frac{n_j}{pn}\frac{(a_{jk})^2}{\mu_k},$$

où n_j/pn est le poids de la catégorie, a_{jk} sa coordonnée sur l'axe k et μ_k la valeur propre associée. On choisit de comparer la contribution au double du poids, ce qui revient donc à comparer $|a_{jk}|$ à $\sqrt{2\mu_k}$.

Pour le premier axe, on décide donc de retenir les catégories qui ont une coordonnée supérieure à $\sqrt{2\mu_1} = \sqrt{2 \times 0,6990} = 1,18$. Pour le second axe, on voudrait comparer les coordonnées à $\sqrt{2 \times 0,5737} = 1,07$:

Axe 1

$\ominus$	$\oplus$
	`sur.NR` $(2,11)$
	`dst.NR` $(1,83)$
	`typ.autre` $(1,46)$
	`dst.p5km` $(1,31)$
	`typ.NR` $(1,25)$

Axe 2

$\ominus$	$\oplus$
	`sur.0.10m2` $(2,49)$
	`typ.cite` $(2,19)$
	`dst.NR` $(2,02)$
	`typ.NR` $(1,25)$

Les valeurs absolues des contributions des catégories n'apportent rien de plus ici. On notera que la limite de 2 fois le poids sépare bien les catégories. Il n'y a pas de cas « limite ».

Question 6 *À quoi correspond la qualité de la représentation d'une catégorie par un sous espace vectoriel ? Que peut on dire ici de la qualité de la représentation des catégories par le premier plan factoriel ? Pouvait-on s'y attendre ?*

La qualité de la représentation d'une catégorie par un sous espace vectoriel est le cosinus carré de l'angle entre le vecteur (centre de gravité, point) et sa projection sur ledit sous-espace. La formule correspondante est, pour le premier plan factoriel,

$$\frac{a_{j1}^2 + a_{j2}^2}{\sum_{k=1}^{12} (a_{ik})^2},$$

et les données correspondantes se lisent dans la seconde colonne du dernier tableau. Comme on veut que l'angle soit le plus proche de 0 possible, les points bien représentés doivent avoir une valeur proche de 100. On choisira de ne garder que les valeurs telles que $\cos^2\theta > 0,8$ (qualité > 80), ce qui correspond à $|\cos\theta| > 0,89$, soit un angle $|\theta| < 26°$.

À la consultation du tableau (qui contient déjà les qualités cumulées), il s'avère qu'aucune catégorie ne peut prétendre à cette distinction. Les variables très mal représentées (qualité < 50), forment la grande majorité. Seules `typ.cite`, `typ.autre` et `sur.0.10m2` sont entre les deux. On constate que les variables `NR`, qui déterminaient les axes, sont pourtant mal représentées.

Cela n'est pas étonnant car la qualité de l'ACM est mauvaise : on sait déjà que l'inertie expliquée par le plan représente 32% du total. On ne peut donc pas s'attendre à ce que les catégories soient correctement expliquées. Même si on garde la règle usuelle pour sélectionner les variables ($\mu > 1/p = 0,33$), on est conduit à conserver au moins 6 coordonnées : la décroissance des valeurs propres est très lente. L'inertie expliquée par ces coordonnées est seulement 2.88, soit 72% de l'inertie totale.

Question 7 *Quelles sont les catégories supplémentaires qui sont liées aux axes ? On précisera les conditions sous lesquelles une valeur test permet de conclure.*

Pour qu'une valeur test permette de conclure qu'une variable est liée à un axe il faut
- que cette variable ne soit pas une variable active
- que sont effectif soit au moins 30
- que la valeur test soit supérieure à 2 ou 3

On voit tout de suite que les variables `occ.NR` et `anc.NR` ne sont pas interprétables : leurs effectifs sont trop faibles (pour `anc.NR`, on retrouve que l'effectif est $78/100 * 383 = 3$). Pourtant sur le premier axe `anc.NR` a une valeur test suffisamment grande $(2,76)$. Pour les autres variables, on trouve

Axe 1

$\ominus$	$\oplus$
`occ.seul` $(-8,61)$	`occ.parents` $(11,44)$
`anc.0.1an` $(-4,55)$	`anc.NA` (11)
`anc.p3an` $(-3,77)$	
`anc.1.3an` $(-3,30)$	

Axe 2

$\ominus$	$\oplus$
`occ.couple` $(-2,63)$	`occ.seul` $(2,19)$
`anc.p3ans` $(2,26)$	

On notera que les coordonnées sont modérées par rapport à celles des variables actives et que les valeurs test sont moins significatives sur le second axe.

Question 8 *Comment peut-on interpréter les axes ?*

On peut faire une interprétation rapide : l'axe 1 oppose
- en positif, les étudiants qui vivent loin de l'université dans une maison (`occ.autre`) ; ces étudiants vivent chez leurs parents en général (et donc n'ont pas répondu à la question sur l'ancienneté).
- les étudiants qui vivent seuls, probablement plus proches de l'université.

L'axe 2 décrit oppose
- en positif, les étudiants ceux qui vivent seuls en cité universitaire dans une toute petite surface (`sur.0.10m2`) depuis moins de 3 ans
- en négatif, des étudiants qui vivent en couple depuis plus de 3 ans