1 Objectif

Description du test de sphéricité de Bartlett et de l'indice KMO (Kaiser – Mayer – Olkin).

L'analyse en composantes principales (ACP) est une technique exploratoire très populaire. Il y a différentes manières de l'appréhender, en voici une très simplifiée : « partant d'une base de données à ''n'' observations et ''p'' variables, toute quantitatives, on cherche à résumer l'information disponible à l'aide de quelques variables synthétiques qu'on appelle facteurs ». Leur nombre n'est pas défini à l'avance, sa détermination est d'ailleurs un enjeu fort dans le processus. Généralement, on en prend au moins deux afin de disposer d'une représentation graphique des individus et des variables dans le plan.

Nous avons présenté maintes fois l'ACP auparavant, tant pour le logiciel R que pour Tanagra[1]. Dans ce tutoriel, nous décrivons deux indicateurs de qualité de l'analyse qui sont directement proposés dans des logiciels commerciaux célèbres (SPSS et SAS pour ne pas les citer), mais que l'on retrouve peu ou prou dans les logiciels libres. On notera qu'ils ne sont pas non plus repris dans les ouvrages qui font référence en français. Il s'agit du test de sphéricité de Bartlett et de l'indice KMO (Kaiser – Mayer – Olkin) mesurant l'adéquation de l'échantillon. Plusieurs internautes m'ayant posé la question sur la manière des le obtenir sous R, je me suis dit qu'il y a avait là quelque chose à faire.

Dans ce qui suit, nous présentons succinctement les formules de calcul, nous leur associons un programme écrit en R, et nous montrons leur mise en œuvre sur un fichier de données. Nous comparons nos sorties avec celles du logiciel SAS.

2 Données – Analyse en composantes principales

Trouver la description d'une méthode est une chose, la décrypter pour la programmer correctement en est une autre. Une bonne manière de vérifier si notre implémentation est correcte est de comparer, à données égales, nos résultats avec ceux d'un logiciel reconnu. Dans ce tutoriel, nous avons récupéré les formules sur le site de documentation de SPSS, et nous avons essayé de reproduire les résultats de SAS à partir d'un exemple décrit sur leur site web.

Le fichier « socioeconomics.xls » est composé de ''n = 12'' observations (secteurs de recensement de la région métropolitaine de Los Angeles) décrites par ''p = 5'' variables socio économiques (POPULATION, nombre d'années d'études – SCHOOL, taux d'activité – EMPLOYMENT, divers services professionnels – SERVICES, médiane de la valeur des habitations – HOUSEVALUE).

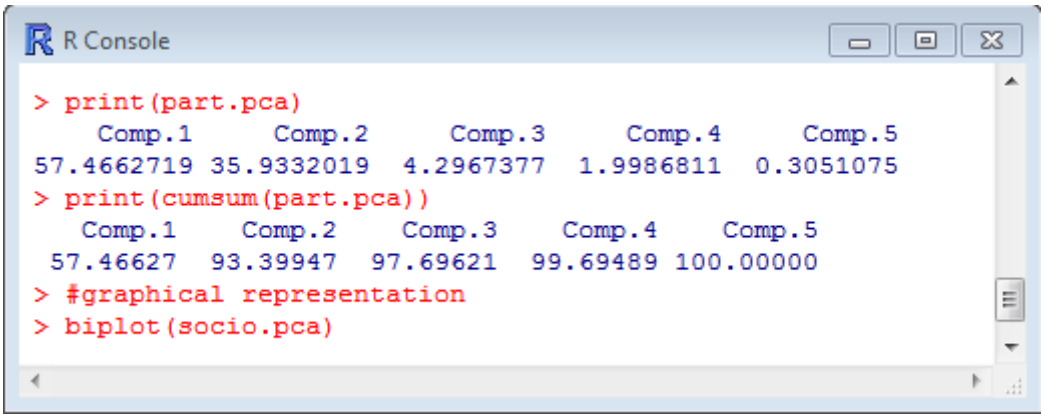
Nous avons utilisé le code R suivant pour charger les données et lancer l'ACP.

```
#importation des données – fichier Excel (.xls)
library(xlsx)
socio.data <- read.xlsx(file="socioeconomics.xls", header=T, sheetIndex=1)
#réalisation de l'ACP via princomp
```

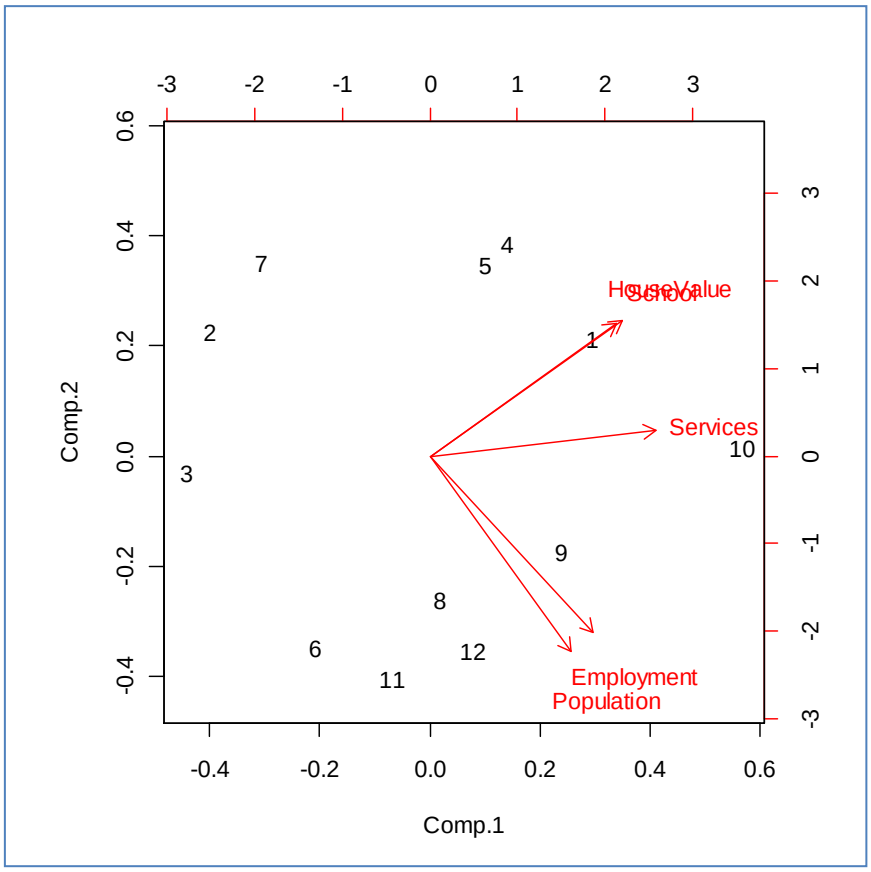
[1] Entres autres : « ACP – Description de véhicules », http://tutoriels-data-mining.blogspot.fr/2008/03/acp-description-de-vhicules.html ; « Analyse en Composantes principales avec R », http://tutoriels-data-mining.blogspot.fr/2009/05/analyse-en-composantes-principales-avec.html; etc.

```
socio.pca <- princomp(socio.data, cor=T)
#proportion de variance expliquée par les facteurs
part.pca <- socio.pca$sdev^2/sum(socio.pca$sdev^2)*100
print(part.pca)
print(cumsum(part.pca))
#représentation graphique
biplot(socio.pca)
```

Nous obtenons :



Avec la représentation « biplot » :



Les deux premiers axes traduisent 93.4% de l'information disponible. On peut considérer que la représentation dans le premier plan factoriel est assez fidèle du positionnement relatif des objets.

A titre de comparaison, voici les sorties la procédure FACTOR de SAS sur les mêmes données.

Eigenvalues of the Correlation Matrix: Total = 5 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.87331359	1.07665350	0.5747	0.5747
2	1.79666009	1.58182321	0.3593	0.9340
3	0.21483689	0.11490283	0.0430	0.9770
4	0.09993405	0.08467868	0.0200	0.9969
5	0.01525537		0.0031	1.0000

3 Test de sphéricité de Bartlett

3.1 Calculer les corrélations croisées

L'idée sous-jacente à ces indicateurs est la suivante : est-ce qu'il est possible d'obtenir un bon résumé ? En effet, on peut considérer l'ACP comme **une compression de l'information**. Elle **n'est possible que si les données présentent une certaine redondance**. Si les variables sont parfaitement corrélées, un seul axe factoriel suffit, il restituera 100% de l'information disponible. A l'inverse, si elles sont deux à deux orthogonales, a fortiori si elles sont deux à deux indépendantes, le nombre adéquat de facteurs à retenir est égal au nombre de variables. Dans ce dernier cas, la matrice de corrélation – impliquée dans le calcul de la solution – est la matrice unité (ou matrice identité).

Une première stratégie consiste simplement à inspecter la matrice de corrélation. Les éléments hors diagonale prennent des valeurs faibles (en valeur absolue) lorsque les variables sont peu liées entres elles. Il est dès lors illusoire d'espérer obtenir un résumé efficace en un nombre de facteurs réduit.

Nous calculons la matrice de corrélation avec le code suivant sous R :

```
#matrice de corrélation
R <- cor(socio.data)
print(R)
```

Nous obtenons.

	Population	School	Employment	Services	HouseValue
Population	1.00000000	0.00975059	0.9724483	0.4388708	0.02241157
School	0.00975059	1.00000000	0.1542838	0.6914082	0.86307009
Employment	0.97244826	0.15428378	1.0000000	0.5147184	0.12192599
Services	0.43887083	0.69140824	0.5147184	1.0000000	0.77765425
HouseValue	0.02241157	0.86307009	0.1219260	0.7776543	1.00000000

Figure 1 - Matrice de corrélation

Certaines variables sont fortement corrélées (Population et Employment : 0.97, School et HouseValue : 0.86), d'autres moins. Il est difficile de donner une valeur seuil à partir de laquelle il faut décider que des liaisons exploitables existent. Il s'agit avant tout d'avoir une « impression

générale de corrélation » en consultant la matrice. Certains auteurs[2] indiquent par exemple que lorsque les corrélations croisées sont toutes supérieures à 0.3 en valeur absolue, l'analyse en composantes principales est appropriée. Mais cette approche est très empirique. De plus, elle devient rapidement impraticable lorsque le nombre de variables augmente.

3.2 Test de Bartlett

Le test de sphéricité de Bartlett propose une mesure globale en s'appuyant sur une démarche statistique. Il vise à détecter dans quelle mesure la matrice de corrélation $\mathbf{R=(r_{ij})_{(p \times p)}}$ calculée sur nos données (matrice observée) diverge significativement de la matrice unité (matrice théorique sous hypothèse nulle H0). Essayer de procéder à un résumé est illusoire lorsque l'hypothèse nulle n'est pas démentie par les données. En revanche, il est possible de compresser l'information, jusqu'à quel point on ne le sait pas, en un nombre plus réduit de facteurs lorsque l'hypothèse nulle est rejetée. Cela ne veut pas dire pour autant que nous allons trouver des informations « intéressantes »[3] dans notre ACP.

Pour mesurer le lien entre les variables, nous calculons le déterminant |R| de la matrice de corrélation. Sous H0, |R| = 1 ; s'il y a des colinéarités parfaites, nous aurions |R| = 0. Ici également, fixer des valeurs seuils est difficile. Généralement, lorsque |R| est inférieur à 0.00001, on considère qu'il y a de très fortes redondances dans les données[4] c.-à-d. elles ne recèlent qu'un seul type d'information. Le résultat sera d'une très grande trivialité (ex. les personnes de grande taille sont plus lourdes, courent plus vite, sautent plus haut et ont des grands pieds). A l'inverse, lorsque |R| *se rapproche* de 1, l'ACP ne servira pas à grand-chose car les variables sont quasiment orthogonales deux à deux.

Le test de Bartlett vise justement à vérifier si l'on s'écarte significativement de cette situation de référence |R| = 1. La statistique de test s'écrit :

$$\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right) \times \ln|R|$$

Sous H0, elle suit une loi du χ^2 à [p x (p-1) / 2] degrés de liberté.

Voyons ce qu'il en est sur nos données, nous avons utilisé le code suivant :

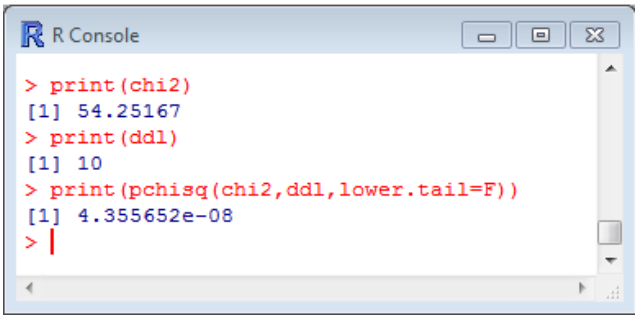
```
#calculer la statistique de Bartlett
n <- nrow(socio.data)
p <- ncol(socio.data)
chi2 <- -(n-1-(2*p+5)/6)*log(det(R))
ddl <- p*(p-1)/2
print(chi2)
print(ddl)
print(pchisq(chi2,ddl,lower.tail=F))
```

[2] B. Williams, A. Onsman, T. Brown, « Exploratoyrfactor analysis: a five-step guide for novices », Journal of Emergency Primary Health Care (JEPHC), Vol. 8, Issue 3, 2010 (http://www.jephc.com/uploads/9900399BW.pdf).

[3] « Intéressant » est très subjectif. L'expert du domaine joue un rôle essentiel dans l'interprétation des résultats.

[4] Jacques Baillargeon, http://www.uqtr.ca/cours/srp-6020/acp/acp.pdf

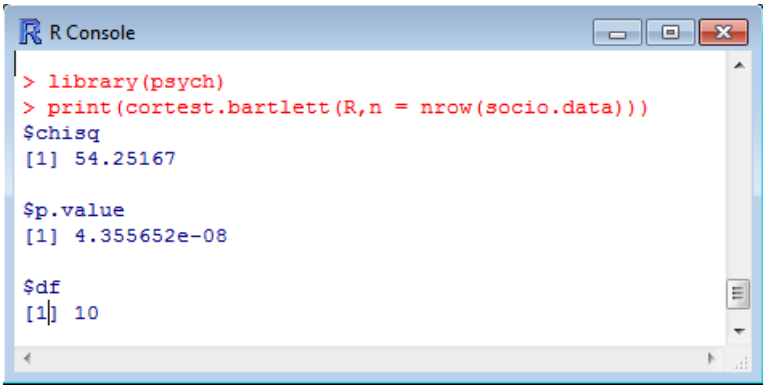
Nous obtenons :



La matrice de corrélation s'écart significativement de la matrice unité pour un test à 5% (p-value = $4.35 \times 10^{-8} < 0.05$). Il est possible d'initier une ACP sur ces données.

Il faut quand même prendre avec énormément de prudence ce test. Il a tendance à considérer que toute configuration est significative à mesure que la taille de l'échantillon « n » augmente. Ce n'est guère étonnant, « n » intervient dans la statistique mais pas dans le calcul des degrés de liberté. De fait, on recommande[5] généralement de ne l'utiliser que lorsque le ratio n:p (nombre d'observations sur nombre de variables) est (approximativement) inférieur à 5 c.-à-d. nous avons moins de 5 observations par variable.

Après coup, je me suis rendu compte que le test de Bartlett est implémenté dans le package « psych ». Le résultat est exactement le même bien évidemment.



4 Indice KMO

L'indice KMO participe de la même idée : est-ce qu'il est possible de trouver une factorisation intéressante des données ? Pour y répondre, il s'appuie sur une stratégie différente.

Le point de départ est toujours la matrice de corrélation. On sait que les variables sont plus ou moins liées dans la base. La corrélation brute entre deux variables est influencée par les (p-2) autres. Nous utilisons la corrélation partielle pour mesurer la relation (nette) entre deux variables en retranchant l'influence des autres[6]. L'indice cherche alors à confronter la corrélation brute avec la corrélation partielle. Si la seconde est nettement plus faible (en valeur absolue), cela veut dire que la liaison est

[5] P. Dugard, J. Todman, H. Staines, « Approaching multivariate analysis – A practical introduction », Psychology Press, 2010 (http://www.psypress.com/multivariate-analysis/medical-examples/chapter08/med_factor_EFA.pdf).

[6] http://en.wikipedia.org/wiki/Partial_correlation

effectivement déterminée par les autres variables. Cela accrédite l'idée de redondance, et donc la possibilité de mettre en place une réduction efficace de l'information. A contrario, si la seconde est équivalente, voire plus élevée, en valeur absolue, cela veut dire qu'il y a une relation directe entre les deux variables. Elle sera difficilement prise en compte par l'ACP. Dans les faits, ces deux variables détermineront souvent un axe factoriel à elles seules.

4.1 Matrice des corrélations partielles

Les corrélations partielles peuvent être déduites de la matrice de corrélation brute. Nous inversons cette dernière, nous obtenons la matrice $R^{-1} = (v_{ij})$. La matrice de corrélation partielle $A = (a_{ij})$ est formée à l'aide de la formule suivante :

$$a_{ij} = -\frac{v_{ij}}{\sqrt{v_{ii} \times v_{jj}}}$$

En programme R pour nos données, cela donne[7]

```
#inverse of the correlation matrix
invR <- solve(R)
#partial correlation matrix (-1 * spss anti-image matrix, unless the diagonal)
A <- matrix(1,nrow(invR),ncol(invR))
for (i in 1:nrow(invR)){
  for (j in (i+1):ncol(invR)){
    #above the diagonal
    A[i,j] <- -invR[i,j]/sqrt(invR[i,i]*invR[j,j])
    #below the diagonal
    A[j,i] <- A[i,j]
  }
}
colnames(A) <- colnames(socio.data)
rownames(A) <- colnames(socio.data)
print(A)
```

Nous obtenons :

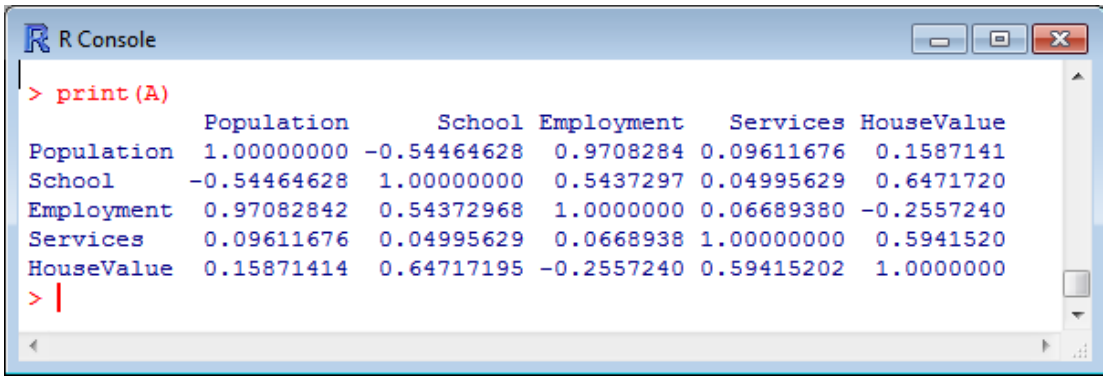


Figure 2 - Matrice des corrélations partielles

[7] Je l'ai fait de la manière la plus simple possible pour plus de clarté. J'imagine qu'on peut le faire sans une double boucle. Les puristes pourront s'y essayer.

On notera par exemple que la liaison entre Population et Employment n'est absolument pas influencée par les autres variables. La corrélation partielle est quasi identique à la corrélation brute.

Voici la matrice fournie par SAS, les valeurs sont les mêmes :

Partial Correlations Controlling all other Variables					
	Population	School	Employment	Services	HouseValue
Population	1.00000	-0.54465	0.97083	0.09612	0.15871
School	-0.54465	1.00000	0.54373	0.04996	0.64717
Employment	0.97083	0.54373	1.00000	0.06689	-0.25572
Services	0.09612	0.04996	0.06689	1.00000	0.59415
HouseValue	0.15871	0.64717	-0.25572	0.59415	1.00000

Sous SPSS, on parle de « **matrice anti-image** ». A la différence que SPSS place sur la diagonale l'indice KMO par variable qui nous présenterons plus loin.

4.2 Indice KMO global

Il ne nous reste plus qu'à calculer l'indice KMO global

$$KMO = \frac{\sum_i \sum_{j \neq i} r_{ij}^2}{\sum_i \sum_{j \neq i} r_{ij}^2 + \sum_i \sum_{j \neq i} a_{ij}^2}$$

L'indice KMO varie entre 0 et 1. S'il est proche de 0, les corrélations partielles sont identiques aux corrélations brutes. Dans ce cas, une compression efficace n'est pas possible. Les variables sont deux à deux orthogonales. S'il est proche de 1, nous aurons un excellent résumé de l'information sur les premiers axes factoriels.

On nous donne parfois ici et là des grilles de lecture : « mauvais » en dessous de 0.5, « bon » entre 0.8 et 0.9 (document SAS) ; ou encore, « inacceptable » en dessous de 0.5, « médiocre » entre 0.5 et 0.6, « moyen » entre 0.6 et 0.7, « bien » entre 0.7 et 0.8, « très bien » entre 0.8 et 0.9, et « excellent » au dessus de 0.9 (http://peoplelearn.homestead.com/Topic20-FACTORanalysis3a.html ; http://www.uqtr.ca/cours/srp-6020/acp/acp.pdf). Pourquoi pas. N'oublions pas simplement qu'il s'agit de mesurer la compressibilité de l'information ici. Un indice « excellent » ne veut pas dire qu'on va tirer des résultats extraordinairement exploitables à partir des données.

Sous R, la formule devient :

```
kmo.num <- sum(R^2) - sum(diag(R^2))
kmo.denom <- kmo.num + (sum(A^2) - sum(diag(A^2)))
kmo <- kmo.num/kmo.denom
print(kmo)
```

Nous obtenons la valeur **KMO = 0.5753676**, exactement comme sous SAS :

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.57536759				
Population	School	Employment	Services	HouseValue
0.47207897	0.55158839	0.48851137	0.80664365	0.61281377

Figure 3 - Calcul des KMO sous SAS

Avec un KMO = 0.575, notre échantillon est plutôt « *médiocre* ». Dans la documentation SAS, on nous conseille d'ajouter de nouvelles variables dans la base pour préciser notre analyse. En effet, toujours selon SAS, une règle usuelle serait d'en avoir au moins 3 par facteurs retenus. Nous avons p = 5 variables ici, avec k = 2 facteurs sélectionnés. Les résultats ne sont pas très « fiables ».

Remarque : L'indice KMO est parfois appelé MSA (Measure of Sampling Adequacy) dans les logiciels anglo-saxons. Il faudrait remonter à l'article originel pour en comprendre la signification. Mais il est clair que la traduction française « mesure d'adéquation de l'échantillon » n'est pas très heureuse. On ne saurait dire en quoi les données seraient adéquates ou non. Il faudrait plutôt utiliser « mesurer de compressibilité des données » ou quelque chose de ce genre, en tous les cas une appellation qui traduit la faculté à résumer efficacement l'information qu'elles recèlent.

4.3 Indice KMO par variable

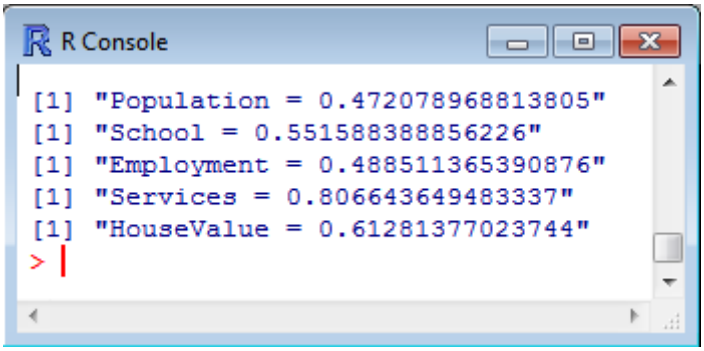
Il est possible de calculer un indice KMO par variable pour repérer les fauteurs de trouble. Il est défini par la formule suivante :

$$KMO_j = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} a_{ij}^2}$$

Que nous codons en R comme suit :

```
#KMO per variable (diagonal of the spss anti-image matrix)
for (j in 1:ncol(socio.data)){
 kmo_j.num <- sum(R[,j]^2) - R[j,j]^2
 kmo_j.denom <- kmo_j.num + (sum(A[,j]^2) - A[j,j]^2)
 kmo_j <- kmo_j.num/kmo_j.denom
 print(paste(colnames(socio.data)[j],"=",kmo_j))
}
```

R nous fournit à la sortie (voir aussi Figure 3) :



Population (0.472) et Employment (0.488) semblent être les variables « problématiques ». Mais dans quel sens doit-on le prendre ? La réponse est dans la matrice de corrélation (Figure 1). Ces deux variables sont fortement liées entres elles (r = 0.97), mais l'une comme l'autre le sont peu avec les 3 autres. Constatation confirmée par la corrélation partielle de 0.97 (Figure 3).

Ce n'est pas vraiment un problème en réalité. Ces deux variables – qui font cavaliers seuls en quelque sorte – sont celles qui pèsent le plus sur le second facteur, les 3 autres déterminent le premier.

Les sorties de l'analyse en composantes principales de Tanagra (corrélation variables – facteurs et COS^2), sous une forme différente, permettent de retrouver cette conformation des données.

Factor Loadings [Communality Estimates]

Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
Population	0.5810	*34 % (34 %)*	0.8064	*65 % (99 %)*
School	0.7670	*59 % (59 %)*	-0.5448	*30 % (89 %)*
Employment	0.6724	*45 % (45 %)*	0.7260	*53 % (98 %)*
Services	0.9324	*87 % (87 %)*	-0.1043	*1 % (88 %)*
HouseValue	0.7912	*63 % (63 %)*	-0.5582	*31 % (94 %)*
Var. Expl.	2.8733	*57 % (57 %)*	1.7967	*36 % (93 %)*

De la même manière, en procédant à une classification des variables (composant VARHCA de Tanagra), nous avons (School, HouseValue, Services) d'un côté, (Population, Employment) de l'autre pour une subdivision en deux groupes

Cluster summary

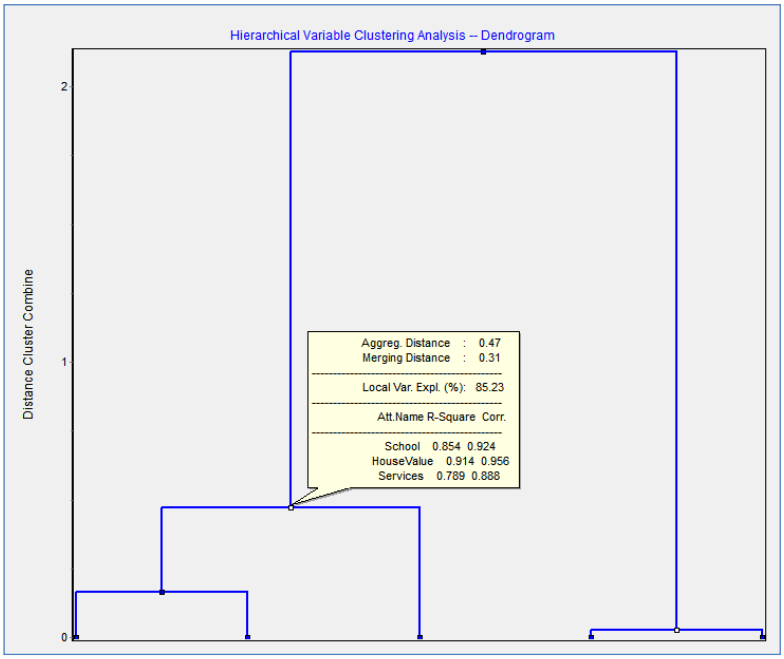
Cluster	# Members	Variation Explained	Proportion Explained
1	3	2.5569	0.8523
2	2	1.9724	0.9862
Total		4.5293	0.9059

Cluster members and R-square values

Cluster	Members	Own Cluster	Next Closest	1-R^2 ratio
1	School	0.8544	0.0068	0.1466
	HouseValue	0.9139	0.0053	0.0866
	Services	0.7886	0.2305	0.2748
2	Population	0.9862	0.0270	0.0142
	Employment	0.9862	0.0785	0.0149

Cluster correlations -- Structure

Attribute	# membership	Cluster 1	Cluster 2
Population	1	0.1643	0.9931
School	1	0.9243	0.0826
Employment	1	0.2801	0.9931
Services	1	0.8880	0.4801
HouseValue	1	0.9560	0.0727



On est toujours rassuré quand plusieurs approches donnent des résultats convergents.

5 Conclusion

Je ne suis pas sûr que ces indicateurs (test de Bartlett et indice KMO) soient si importants que cela. Mais comme c'est une question qui revenait assez souvent chez les internautes, perplexes devant des sorties de logiciels dont ils comprenaient mal la portée, il me semblait intéressant de mieux les expliciter et de monter leur implémentation sous R. Ils seront peut être intégrés un jour dans Tanagra. Mon hésitation vient du fait que leur calcul intègre une contrainte forte qui ne se justifie pas réellement dans le processus d'analyse en composante principale en tant que tel : la matrice de corrélation doit être inversible, de déterminant non nul.