Une nouvelle méthode statistique pour la construction d'indicateurs composites de qualité de vie à l'échelle communale

Marie Chavent, Vanessa Kuentz-Simonet, Amaury Labenne, Jérôme Saracco

3èmes Rencontres R, Montpellier, 27 juin 2014

Introduction

- Mesure de la qualité de vie via deux approches différentes :
  - Enquête auprès des citoyens : mesure des niveaux de satisfaction de la vie
  - Analyse des bases de données nationales : Création d'indices composites de condition de vie à l'échelle communale

- Nécessité de créer des indices composites résumant au mieux l'information des données nationales
  - Méthode d'analyse factorielle pour variables structurées en groupes
  - Composantes principales = Combinaison linéaire des variables = INDICE COMPOSITE

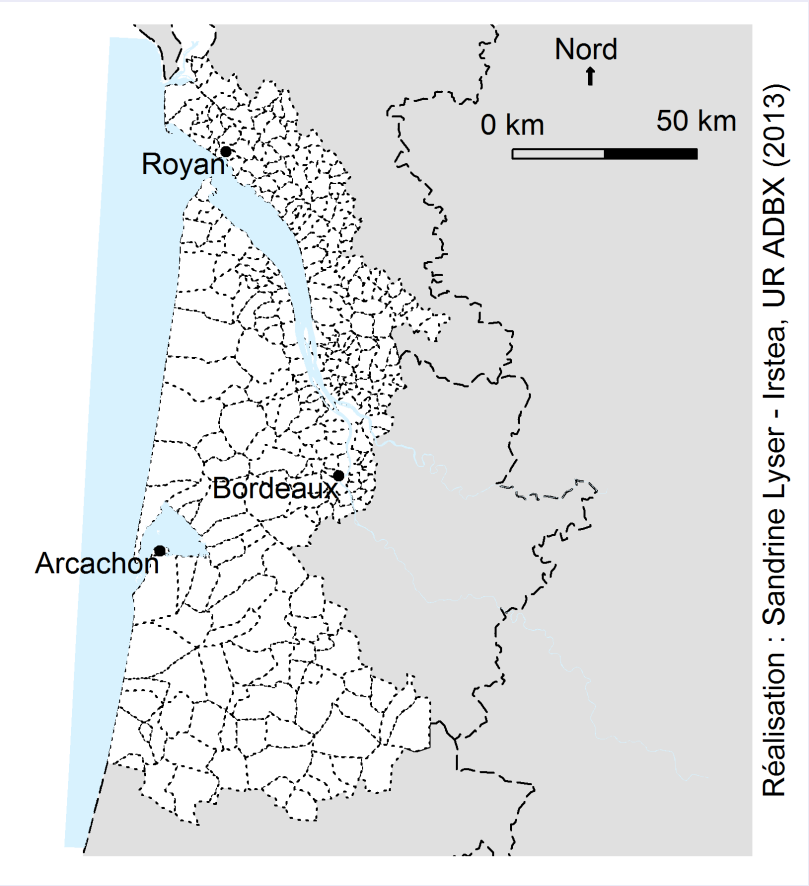
- Ces méthodes soulèvent plusieurs questions :
  - Combien de composantes principales retenir ?
  - Peut on obtenir des indices sur un nombre restreint de variables et qui soient fortement corrélés aux indices calculés sur toutes les variables : Facilité d'interprétation

Présentation des données

Les données réparties en groupes :

- L'analyse porte sur n=303 communes décrites par p=44 variables quantitatives et qualitatives.
- Les p=44 variables peuvent être réparties en G=5 groupes de variables :
  - Economic conditions,
  - Living conditions,
  - Family situations,
  - Services access,
  - Natural Environment.

Carte du territoire étudié :



Données mixtes structurées en groupe $\longrightarrow$ MFAmix : analyse factorielle multiple de données mixtes *(Chavent, M., Kuentz-Simonet, V., Labenne, A.,Saracco, J., Rambonilaza, T. (2013).Une extension de l'analyse factorielle multiple pour des groupes de variables mixtes : MFAmix. 2èmes rencontres R, Montpellier)*

Sommaire

La méthode MFAmix (1/7)

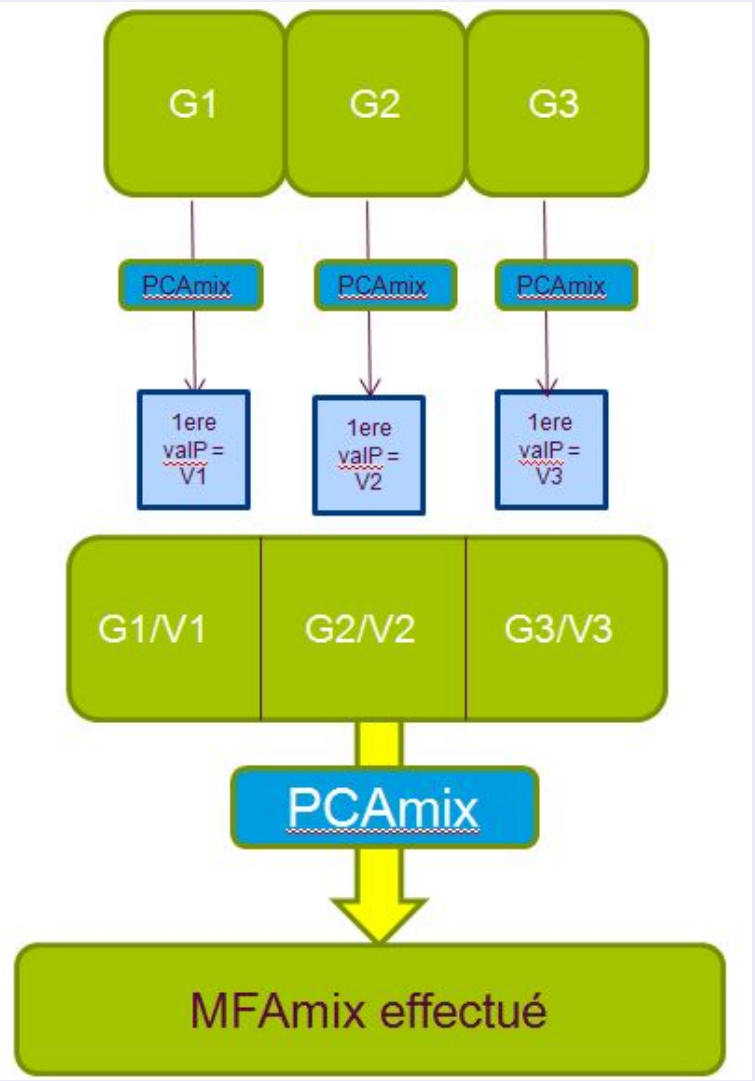
Principe de la méthode :

La méthode MFAmix est basée sur la DVSG de $\mathbf{Z}$ (la matrice des données brutes précédemment recodées) avec les métriques $\mathbf{D}$ pour les individus et $\mathbf{M}$ pour les variables. On a ainsi :

$$\mathbf{Z} = \mathbf{U\Lambda V}^t, \text{ avec :}$$

- $\mathbf{\Lambda} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$ la matrice des valeurs singulières de $\mathbf{ZDZ}^t\mathbf{M}$ et $\mathbf{Z}^t\mathbf{DZM}$ où r est le rang de $\mathbf{Z}$;
- $\mathbf{U}$ la matrice $n \times r$ des vecteurs propres de $\mathbf{ZMZ}^t\mathbf{D}$ et $\mathbf{U}^t\mathbf{DU} = \mathbb{I}_r$;
- $\mathbf{V}$ la matrice $p \times r$ des vecteurs propres de $\mathbf{Z}^t\mathbf{DZM}$ et $\mathbf{V}^t\mathbf{MV} = \mathbb{I}_r$.

Schéma explicatif



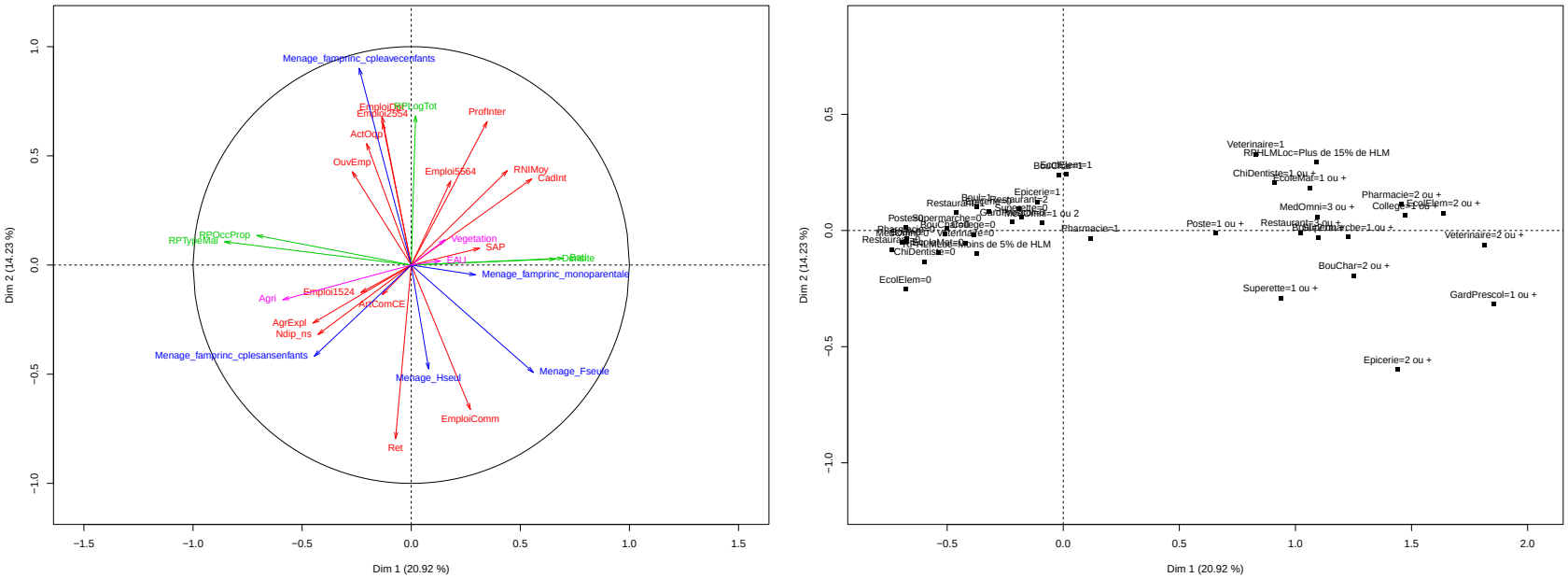
La méthode MFAmix (2/7)

```
res.MFA<-MFAmix(data=data_littoral,
                group=vect.group,name.group=name.group,ndim=10,rename.categ=TRUE)

## **Results of the Multiple Factor Analysis for mixed data (MFAmix)**
## The analysis was performed on individuals, described by variables
## *Results are available in the following objects :
##
##   name                  description
## 1 "$eig"                "eigenvalues"
## 2 "$separate.analyses"  "separate analyses for each group of variables"
## 3 "$group"              "results for all the groups"
## 4 "$partial.axes"       "results for the partial axes"
## 5 "$ind"                "results for the individuals"
## 6 "$quanti.var"         "results for the quantitatives variables"
## 7 "$quali.var"          "results for the categorials variables"
## 8 "$global.pca"         "results for the global PCA"
## 9 "$recap.eig.separate" "ndim first eigenvalues of the separate analyses"
```

La méthode MFAmix (3/7)

```
par(mfrow=c(1,2))
plot(res.MFA,choice="var",habillage="group",cex=0.9,leg=FALSE,axes=c(1,2))
plot(res.MFA,choice="ind",invisible="ind",habillage="group",cex=0.9,leg=FALSE,axes=c(1,2))
```



Beaucoup de variables, Difficulté d'interprétation

La méthode MFAmix (7/7)

Conclusions sur MFAmix

- La méthode MFAmix a permis de créer des indices composites (les composantes principales).
- Combien de composantes principales retenir ?
- Ces CP sont des combinaisons linéaires des 44 variables. Il est nécessaire de restreindre le nombre de variables pour faciliter l'interprétation.

Choix du nombre de CP à retenir (1/2)

On définit la matrice de projection **M**-orthogonale des lignes de **Z** sur $E_q = Im(\mathbf{V}_q)$, comme suit :

$$\widehat{\mathbf{P}_q} = \mathbf{V}_q \mathbf{V}_q^t \mathbf{M}$$

La fonction de perte reposant sur la distance euclidienne entre deux projecteurs orthogonaux est donnée par :

$$\mathcal{L}_q = \mathcal{Q}(E_q, \widehat{E_q}) = \frac{1}{2}\|\mathbf{P}_q - \widehat{\mathbf{P}_q}\|_2^2 = q - Tr(\mathbf{P}_q \widehat{\mathbf{P}_q}).$$

Finalement, le risque est défini comme l'espérance de la fonction de perte:

$$R_q = E[\mathcal{L}_q].$$

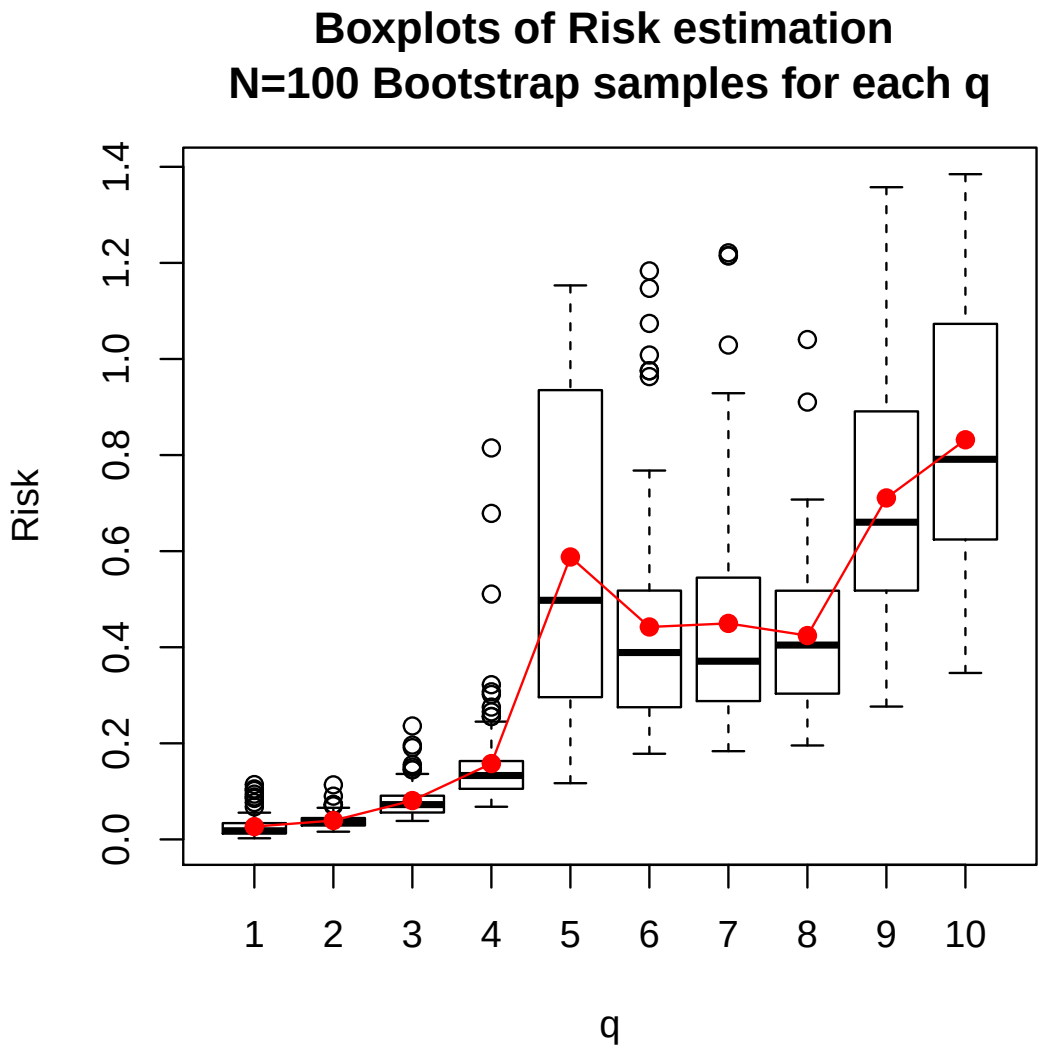
L'idée est d'estimer R_q par un estimateur bootstrap :

$$\widehat{R_{Bq}} = \frac{1}{B}\sum_{b=1}^{B}\Big(q - Tr(P_q^{\star b} P_q)\Big) = q - Tr(P_q^{\star(.)} P_q).$$

Où B est le nombre d'échantillons bootstrap, $P_q^{\star b}$ est la matrice de projection obtenue avec MFAmix sur le b-ème échantillon et $P_q^{\star(.)} = \frac{1}{B}\sum_{b=1}^{B} P_q^{\star b}$.

Choix du nombre de CP à retenir (2/2)

```
library(boot)
res.stab<-stability.CP.bootstrap(res.mfa=res.MFA,i,q.max=8,NB=100,graph=TRUE)
```



Reduction du nombre de variables (1/5)

Nous avons fixé le nombre de CP à retenir à $q = 3$.
Nous allons maintenant chercher si en réalisant la méthode MFAmix sur un nombre de variables $p^\star$, inférieur aux $p = 44$ variables initiales nous obtenons des CP qui soient le plus liées possibles aux CP de bases.

Pour cela il est nécessaire de définir une mesure de liaison (equivalente à la corrélation) entre deux groupes de 3 variables (ici les CP).
On note $\mathbf{F} = \mathbf{U\Lambda}$ les CP obtenues grâce à MFAmix sur les p variables et $\mathbf{F}^\star = \mathbf{U}^\star\mathbf{\Lambda}^\star$ les CP obtenues grâce à MFAmix réalisé sur les $p^\star$ variables.
La mesure de liaison entre $\mathbf{F}$ et $\mathbf{F}^\star$ est défini comme suit :

$$\mathcal{D}(\mathbf{F}\mathbf{F}^\star) = \frac{1}{q} Tr(P_F P_{F^\star})$$

Où :

- $P_F = \mathbf{U}\mathbf{U}^T\mathbf{D}$ est la matrice de projection $\mathbf{D}$-orthogonale sur $\mathbf{F}$,
- $P_F^\star = \mathbf{U}^\star\mathbf{U}^{\star T}\mathbf{D}$ est la matrice de projection $\mathbf{D}$-orthogonale sur $\mathbf{F}^\star$.

Cette mesure de liaison (distance entre deux sous espaces) va nous permettre via la méthode Closest Submodel Selection (CSS) de choisir le meilleur sous ensemble de variable tel que MFAmix réalisé sur celui ci donne des CP $\mathbf{F}^\star$ le plus proche possible des CP $\mathbf{F}$ obtenues grace à MFAmix sur les p variables initiales.

Reduction du nombre de variables (2/5)

La méthode CSS

L'idée de la méthode est de trouver des sous ensembles de p_0 variables ($p_0 < p$) sur lesquels on réalise MFAmix et de sélectionne les meilleurs sous espaces (au sens de la distance définit précedemment). Les variables qui apparaissent le plus souvent dans les meilleurs sous espaces seront retenues comme les plus importantes.

- Step 1 : Choisir N_0, le nombre de sous espaces a évaluer et ζ le pourcentage de sous espaces retenues parmis les N_0 évalués. On pose $a = 1$
- Step 2 : Sélectionner aléatoirement p_0 variables parmis les p et construire la matrice $X^{(a)}$ contenant les variables sélectionnées.
- Step 3 : Réaliser MFAmix sur $X^{(a)}$ et calculer $\mathcal{D}(\mathbf{FF}^{(a)})$. Poser $a = a + 1$. Répéter les étapes 2 et 3 N_0 fois.
- Step 4 : Retenir les $N_1 = \zeta N_0$ sous espaces ayant la plus grande mesure de liaisoon avec $\mathbf{F}$.
- Step 5 : Compter le nombre de fois ou apparait chaque variables dans les N_1 meilleurs sous espaces. Ces variables sont retenues pour réaliser MFAmix par la suite.

Reduction du nombre de variables (3/5)

```
reduc.CSS <- reduc.MFAmix.CSS(object = res.MFA, N0 = 25000, p0 = 20, zeta = 5/100,
    ndim = 3, q = 3, graph = FALSE, p.min.group = c(0, 0, 0, 0, 0))
```

On affiche ici le nombre de fois ou chaque variable est apparu dans les N_1 meilleurs modèles

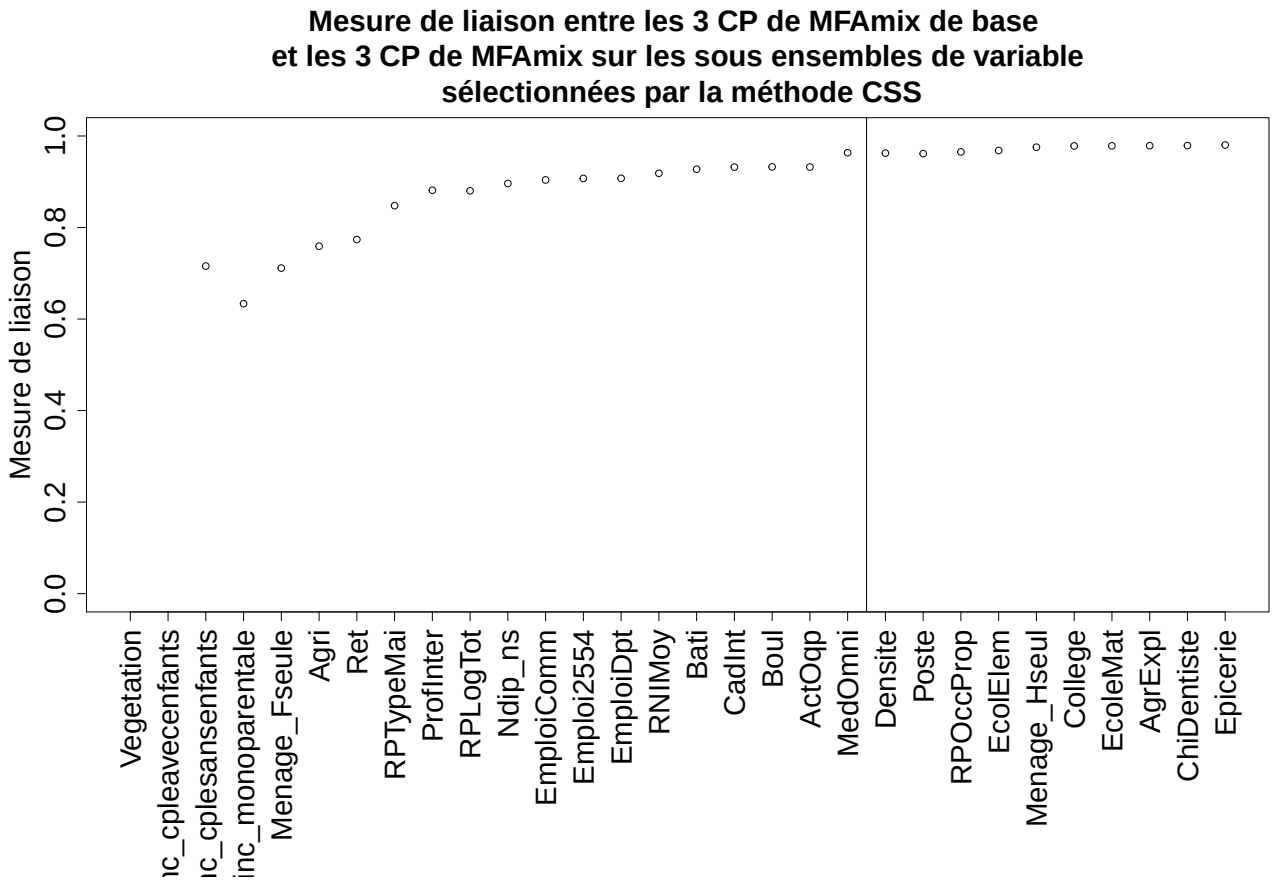
```
reduc.CSS$var.chosen[1:20]
```

```
##                       Vegetation Menage_famprinc_cpleavecenfants
##                             1221                             996
## Menage_famprinc_cplesansenfants   Menage_famprinc_monoparentale
##                              993                             922
##                     Menage_Fseule                            Agri
##                              709                             658
##                              Ret                         RPTypeMai
##                              653                             648
##                        ProfInter                         RPLogTot
##                              638                             624
##                          Ndip_ns                        EmploiComm
##                              623                             596
##                       Emploi2554                        EmploiDpt
##                              594                             584
##                           RNIMoy                            Bati
##                              572                             542
##                           CadInt                            Boul
##                              541                             534
##                           ActOqp                         MedOmni
##                              526                             519
```

Reduction du nombre de variables (4/5)

On va regarder plus précisément quelle est la mesure de liaison en fonction du sous ensemble de variable sélectionné

```
CSS.chosen <- MFAmix.reduc.choice(obj = reduc.CSS, res.mfa.global = res.MFA,
    nb.var = 20, q = 3, data.base = data_littoral)
```



Reduction du nombre de variables (5/5)

Ici, on va relancer MFAmix sur le sous ensemble de variables choisis puis interpréter briévement les résultats

```
base.CSS <- data_littoral[, names(CSS.chosen[[2]][1:20])]
MFAmix.CSS <- MFAmix.sub.var(object = res.MFA, data.sub.var = base.CSS, ndim = 3,
    rename.categ = TRUE)
```

On regarde les corrélations (2 à 2) entres les CP de MFAmix de base et les CP obtenues grâce à la méthode CSS.

```
indice.base<-res.MFA$ind$coord[,1:3]
indice.CSS<-MFAmix.CSS$ind$coord[,1:3]
cor(indice.CSS,indice.base)

##           dim 1    dim 2     dim 3
## dim 1 0.960501 -0.13914  0.002414
## dim 2 0.133084  0.97429 -0.021067
## dim 3 0.008041  0.01013  0.981221
```

Interprétation sommaire des indices synthétiques

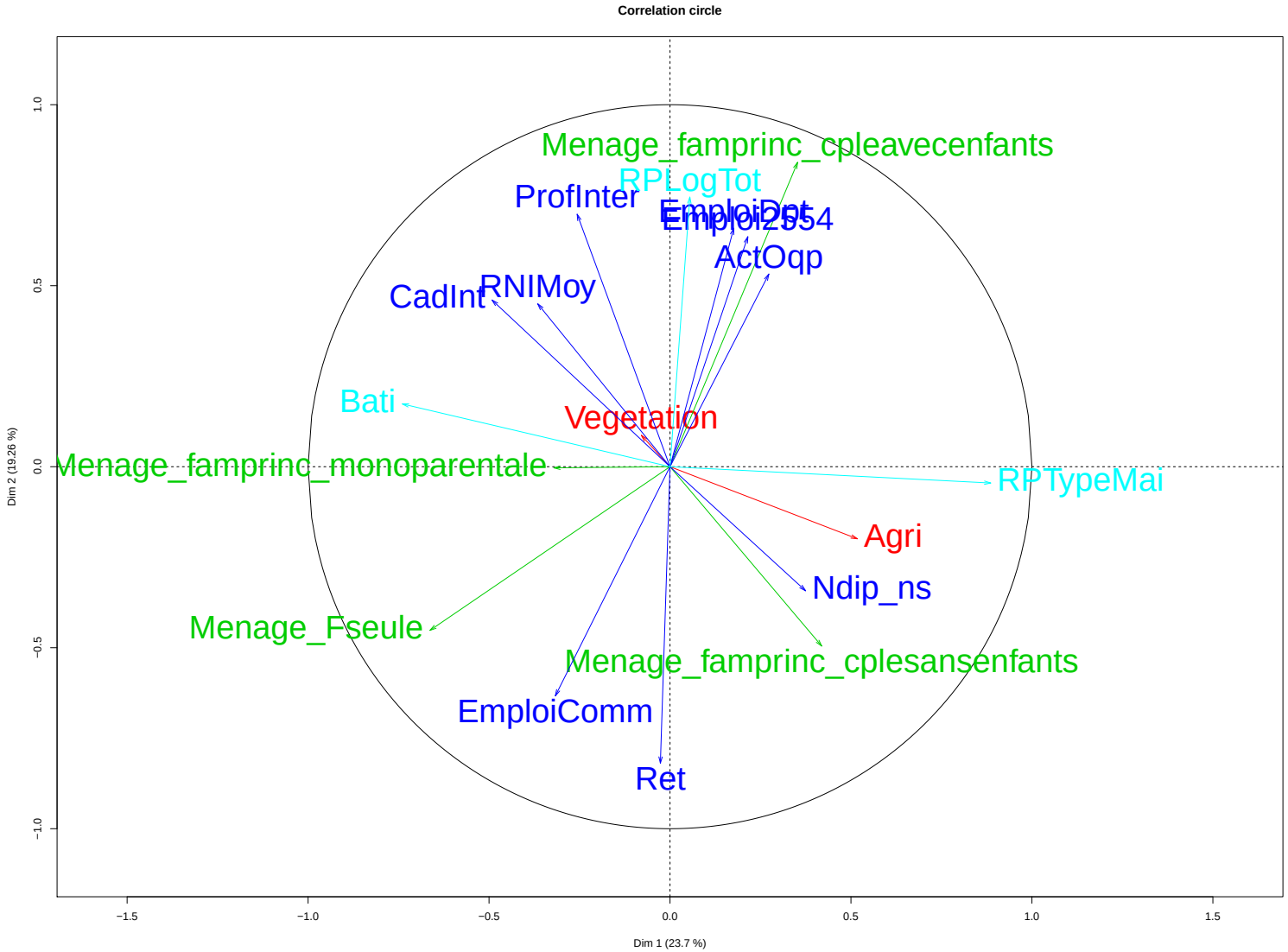


Figure : Cercle des corrélation des variables quantitatives sur le plan (1,2)

Interprétation sommaire des indices synthétiques

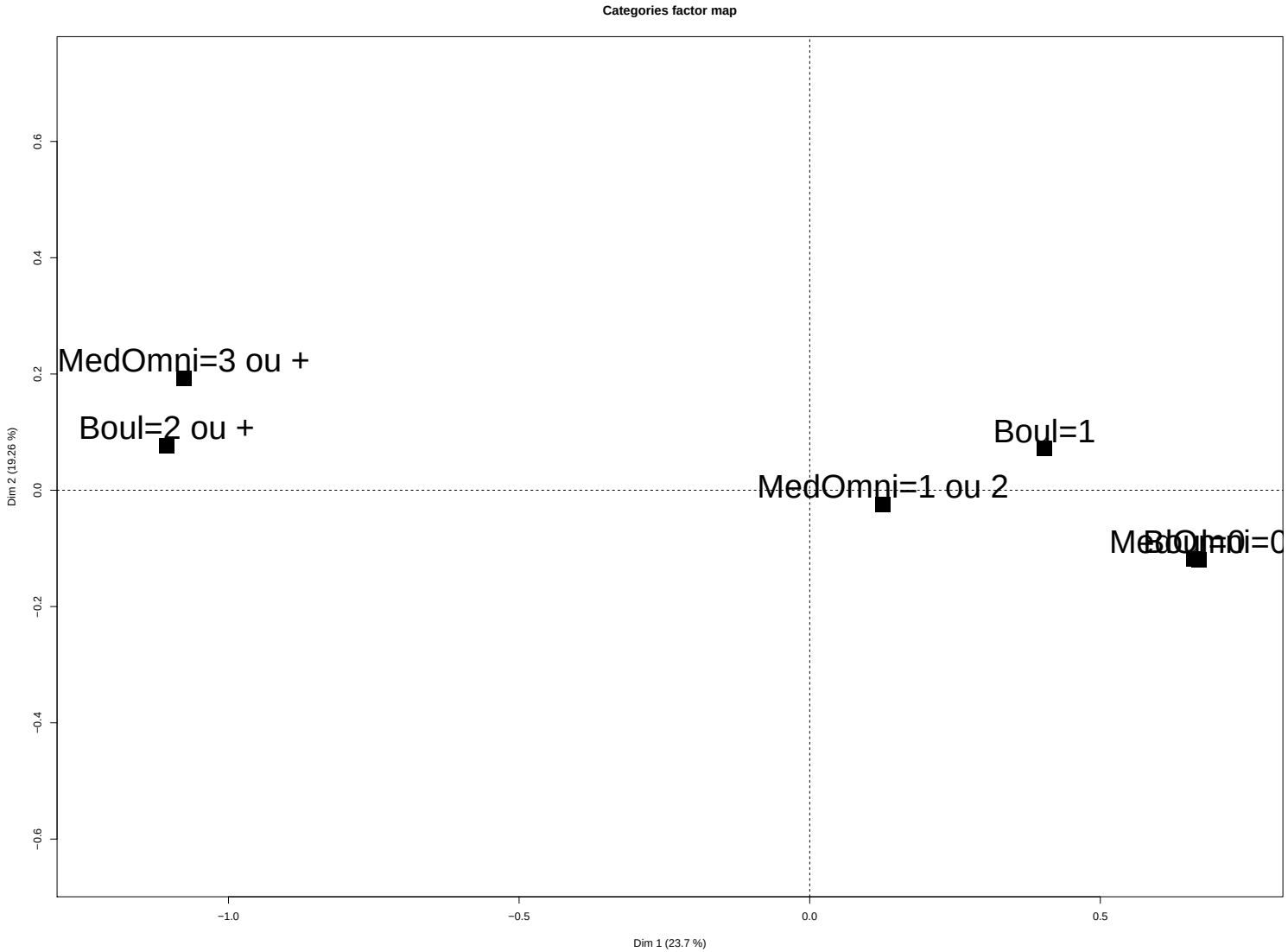


Figure : Cercle des corrélation des variables quantitatives sur le plan (1,3)

Interprétation sommaire des indices synthétiques

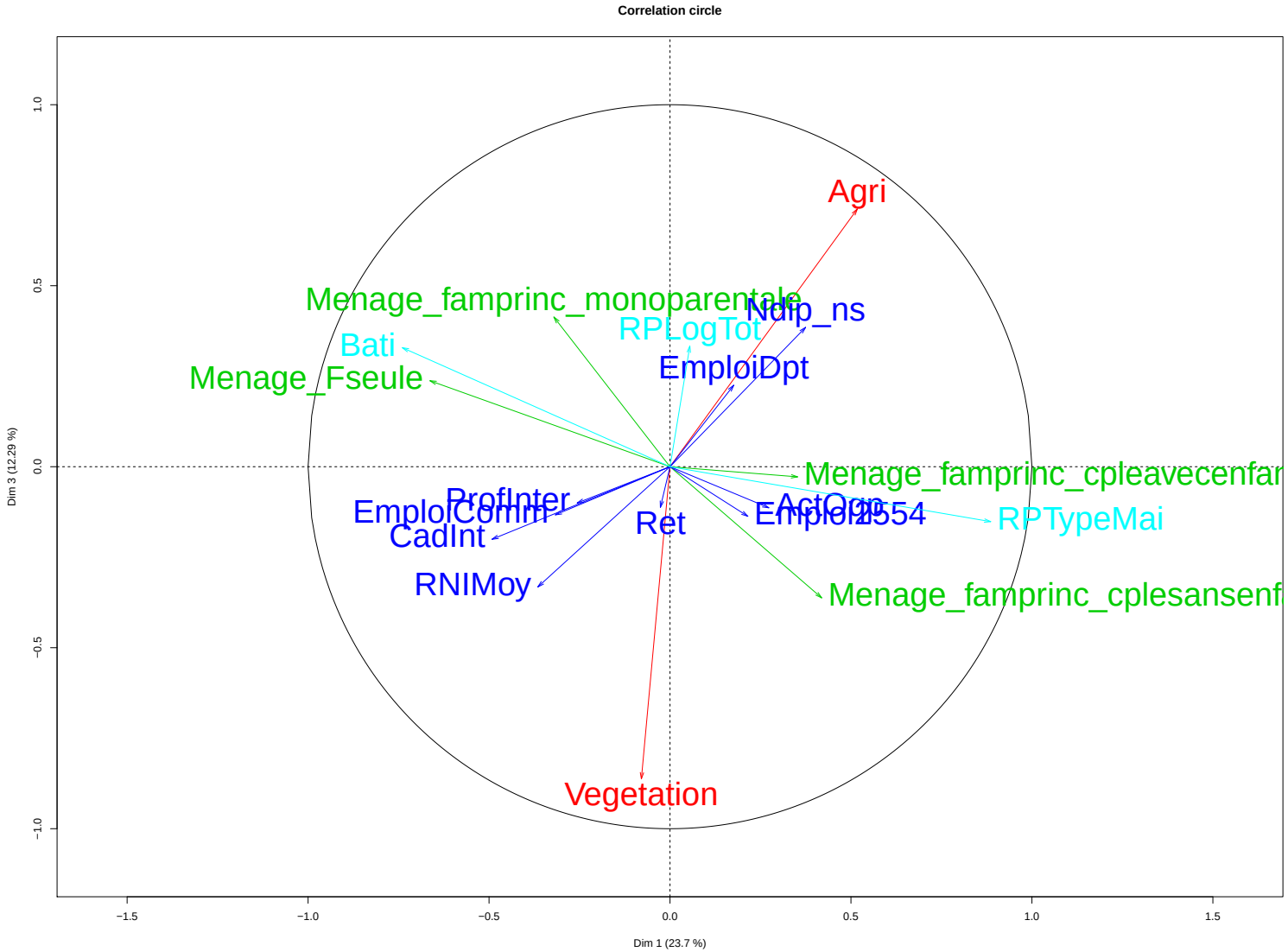


Figure : Représentation des groupes sur les deux premiers plans factoriels

Interprétation sommaire des indices synthétiques

- **L'indice 1** oppose les communes habitées par des couples avec enfants qui ont tendance à travailleur à l'extérieur aux communes habitées par des personnes y travaillant et habitées par des retraités.
- **L'indice 2** oppose les communes avec une forte proportion de maisons aux communes avec une plus grande proportion d'appartements plus souvent occupées par des femmes seules.
- **L'indice 3** oppose les communes de type agricole aux communes plus végétalisée (territoires forestiers).

Conclusion

- Méthode MFAmix semble pertinente pour la création d'indices synthétiques lorsque les données sont structurées en groupe.
- Il existe d'autres méthodes plus robustes que l'éboulis des valeurs propres pour choisir le nombre de CP à conserver.
- Il est possible de créer des indices fortement corrélés aux indices de base calculés sur un nombre restreint de variables.