# Variable selection to construct indicators of quality of life for data structured in groups

Marie Chavent[a,b], Vanessa Kuentz-Simonet[c], Amaury Labenne[c], Jérôme Saracco[a,b]

[a]Univ. Bordeaux, IMB, UMR 5251

[b]INRIA Bordeaux Sud-Ouest, CQFD

[c]IRSTEA, UR ADBX

COMPSTAT, August 21, 2014



# Introduction

- Measure of quality of life can be made with two different and complementary approaches:
  - Citizen Survey: measuring levels of life satisfaction.
  - Analysis of national databases: creation of composite indicators of living conditions at the municipal scale

# Introduction

- Measure of quality of life can be made with two different and complementary approaches:
  - Citizen Survey: measuring levels of life satisfaction.
  - Analysis of national databases: creation of composite indicators of living conditions at the municipal scale

- Need to create composite indicators to summarize the information brings by national data:
  - Factor analysis methods for data structured in groups
  - Principal components = Linear combination of variables = COMPOSITE INDICATORS

# Introduction

- Measure of quality of life can be made with two different and complementary approaches:
  - Citizen Survey: measuring levels of life satisfaction.
  - Analysis of national databases: creation of composite indicators of living conditions at the municipal scale

- Need to create composite indicators to summarize the information brings by national data:
  - Factor analysis methods for data structured in groups
  - Principal components = Linear combination of variables = COMPOSITE INDICATORS

- These methods raise several questions:
  - How many principal components to keep?
  - Can we, by selecting a lower number of variables get indicators that are highly correlated with indicators calculated on all variables? (facilitate interpretation)

# Data presentation

## Data structured in groups:

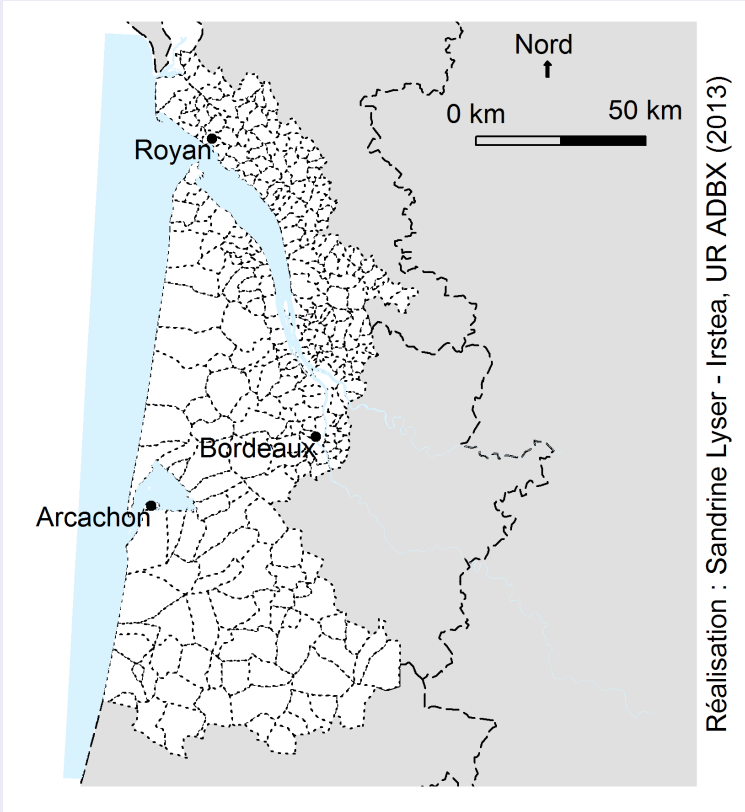
- The analysis focus on n=303 cities described by p=44 variables (numerical and categorical).
- The p=44 variables can be divided into G=5 groups of variables :
  - Economic conditions,
  - Living conditions,
  - Family situations,
  - Services access,
  - Natural Environment.

# Data presentation

## Data structured in groups:

- The analysis focus on n=303 cities described by p=44 variables (numerical and categorical).
- The p=44 variables can be divided into G=5 groups of variables :
  - Economic conditions,
  - Living conditions,
  - Family situations,
  - Services access,
  - Natural Environment.

## Map of the studied area:



# Outline

1. The MFAmix method to create composite indicators
2. Choice of the number of principal components according to the stability
3. Select the most important variables in the creation of the indicators

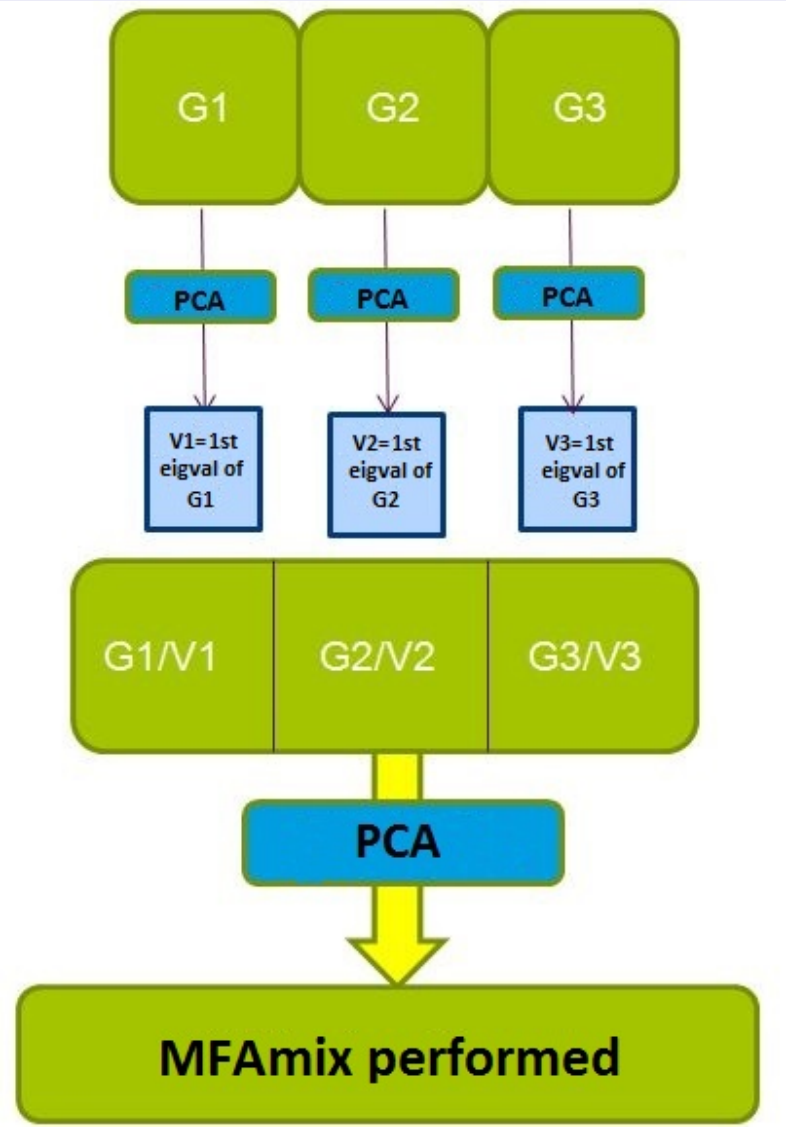
# The MFAmix method (1/5)

## The multiple factor analysis

- Multiple factor analysis (MFA) for numerical variables: Escofier and Pages (1983).
- MFA for numerical or categorical variables by groups: Pages J (2002).
- Une extension de l'analyse factorielle multiple pour des groupes de variables mixtes : MFAmix. Chavent et al. (2013) 2emes rencontres R, Lyon
  $\longrightarrow$ Based on PCAmix (Chavent et al. 2012): Factor analysis of mixed data type: Mix between PCA and MCA
  $\longrightarrow$ Package `PCAmixdata`

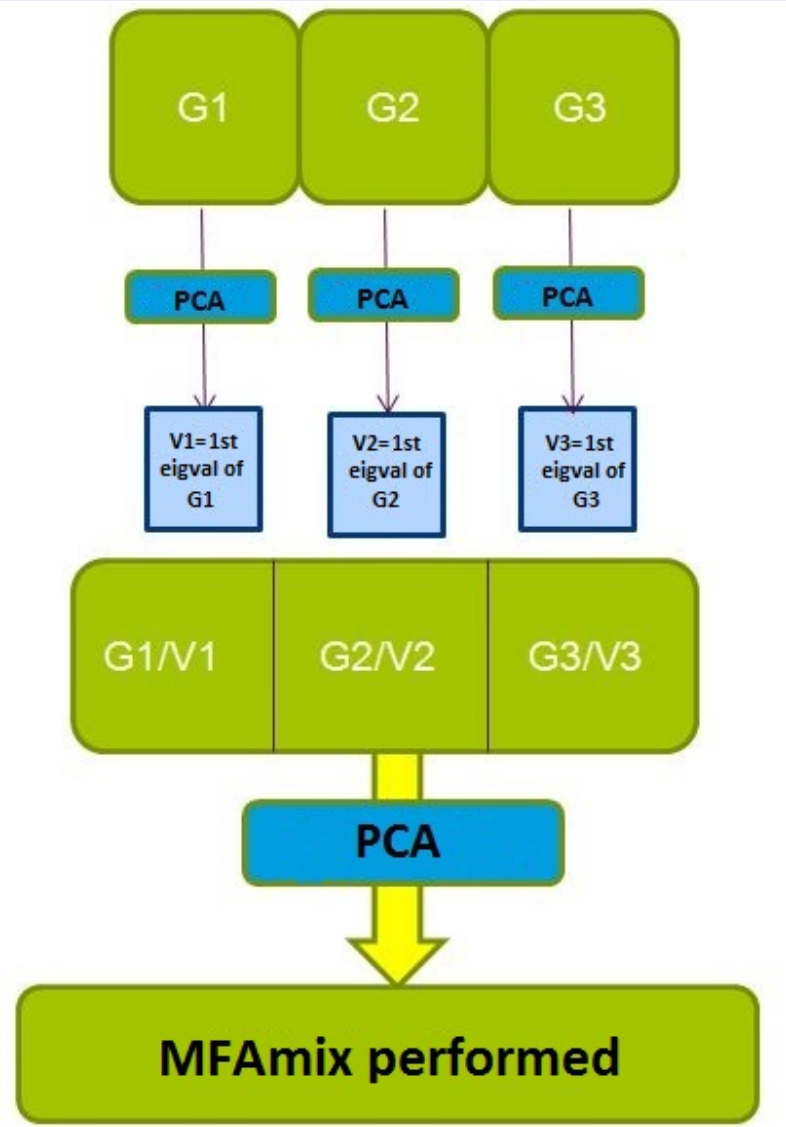
# The MFAmix method (2/5)

## Explanatory schema



# The MFAmix method (2/5)

## Explanatory schema




## Principle method:

The MFAmix method is based on the generalized singular value decomposition (GSVD) of $\mathbf{Z}$ (the raw data matrix previously recoded) with metrics $\mathbf{D}$ for individuals and $\mathbf{M}$ for variables. It gives :

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

- $\mathbf{\Lambda} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$ the singular value matrix of $\mathbf{ZDZ}^t\mathbf{M}$ and $\mathbf{Z}^t\mathbf{DZM}$ where $r$ is the rank of $\mathbf{Z}$ ;
- $\mathbf{U}$ the $n \times r$ eigenvector matrix of $\mathbf{ZMZ}^t\mathbf{D}$ and $\mathbf{U}^t\mathbf{DU} = \mathbb{I}_r$ ;
- $\mathbf{V}$ the $p \times r$ eigenvector matrix of $\mathbf{Z}^t\mathbf{DZM}$ and $\mathbf{V}^t\mathbf{MV} = \mathbb{I}_r$.

# The MFAmix method (3/5)

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

```
print(res.MFA)

## **Results of the Multiple Factor Analysis for mixed data (MFAmix)**
## The analysis was performed on 303 individuals, described by 44 variables
## *Results are available in the following objects :
##
##    name                  description
## 1  "$eig"                "eigenvalues"
## 2  "$eig.separate"       "eigenvalues of the separate analyses"
## 3  "$separate.analyses" "separate analyses for each group of variables"
## 4  "$group"              "results for all the groups"
## 5  "$partial.axes"       "results for the partial axes"
## 6  "$ind"                "results for the individuals"
## 7  "$ind.partial"        "results for the partial individuals"
## 8  "$quanti"             "results for the quantitative variables"
## 9  "$levels"             "results for the levels of the qualitative variables"
## 10 "$quali"              "results for the qualitative variables"
## 11 "$sqload"             "squared loadings"
## 12 "$global.pca"         "results for the global PCA"
```

# The MFAmix method (4/5)

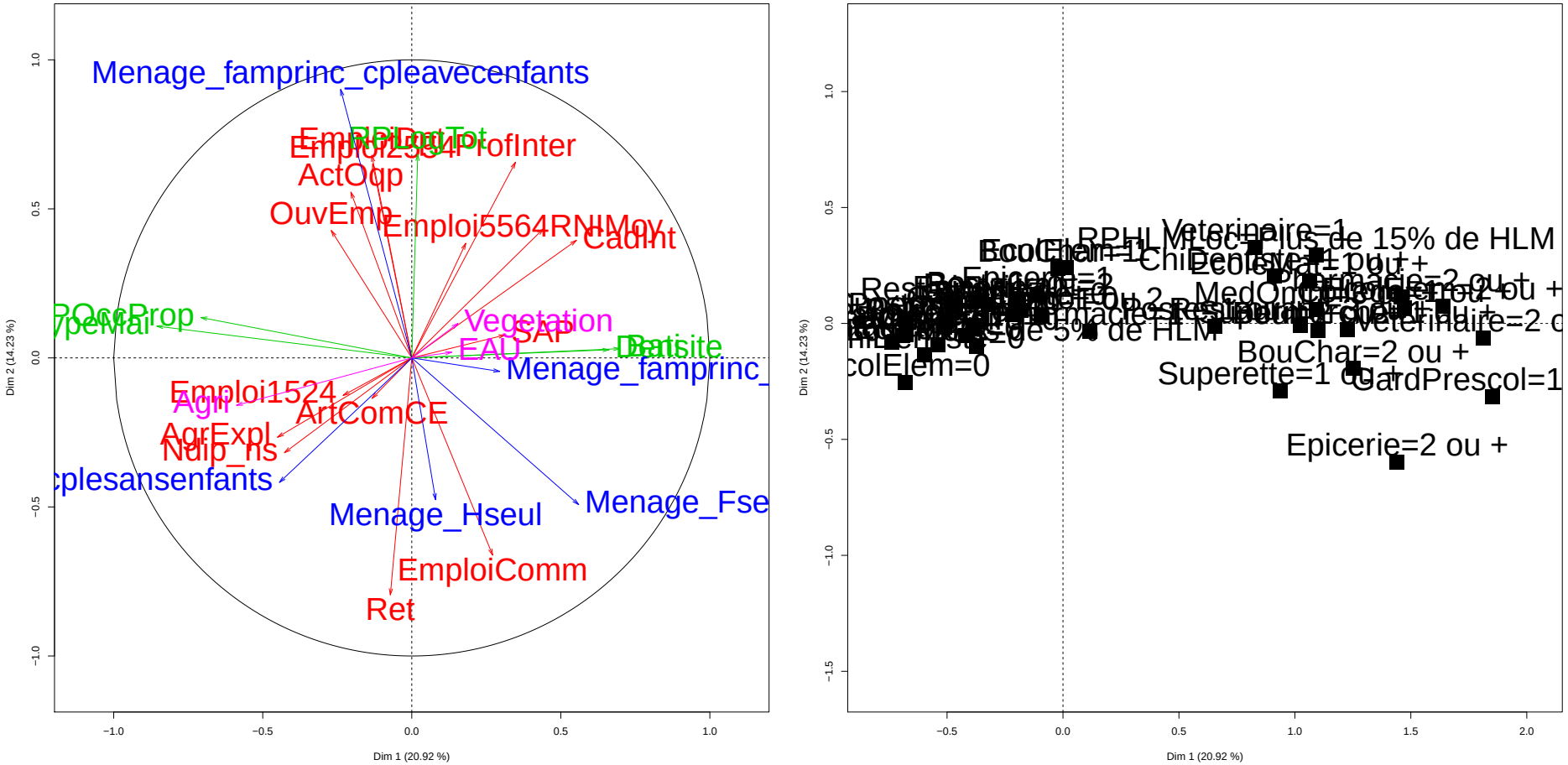



Figure : Correlation circle and map of levels of categorical variables

$\longrightarrow$ Lot of variables, hard to interpret

# The MFAmix method (5/5)

## Conclusions about MFAmix

- The MFAmix method allowed to create composite indicators (the principal components).
- How many principal components (PC) to keep?
- These PC are linear combinations of 44 variables. It is necessary to reduce the number of variables to facilitate the interpretation.

# Choice of the number of components to keep (1/3)

- The choice of the number of components ($q$) to keep is a problem very often approached in factor analysis
- Several methods exist in the literature:
  - Besse, P. (1992). PCA stability and choice of dimensionality. *Statistics and Probability Letters*, **13**, 405-410.
  - Josse, J., Husson, F. (2012). Selecting the number of components in PCA using cross-validation approximations. *Computational Statistics and Data Analysis*, **56**, pp. 1869-1879.
- Here we chose to use the method proposed by Besse using a bootstrap approach.

# Choice of the number of components to keep (2/3)

We define $\mathbf{P}_q$, the projection matrix $\mathbf{M}$-orthogonal of rows of $\mathbf{Z}$ on $E_q = Im(\mathbf{V}_q)$ ($q$ is the number of PC chosen) as follow:

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

The loss function based on the euclidean distance between two orthogonal projectors is given by:

$$\mathcal{L}_q = \frac{1}{2} \|\mathbf{P}_q - \mathbf{P}_q{}^\star\|^2 = q - Tr(\mathbf{P}_q \mathbf{P}_q{}^\star).$$

Then, the risk is defined as the expectation of the loss function:

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

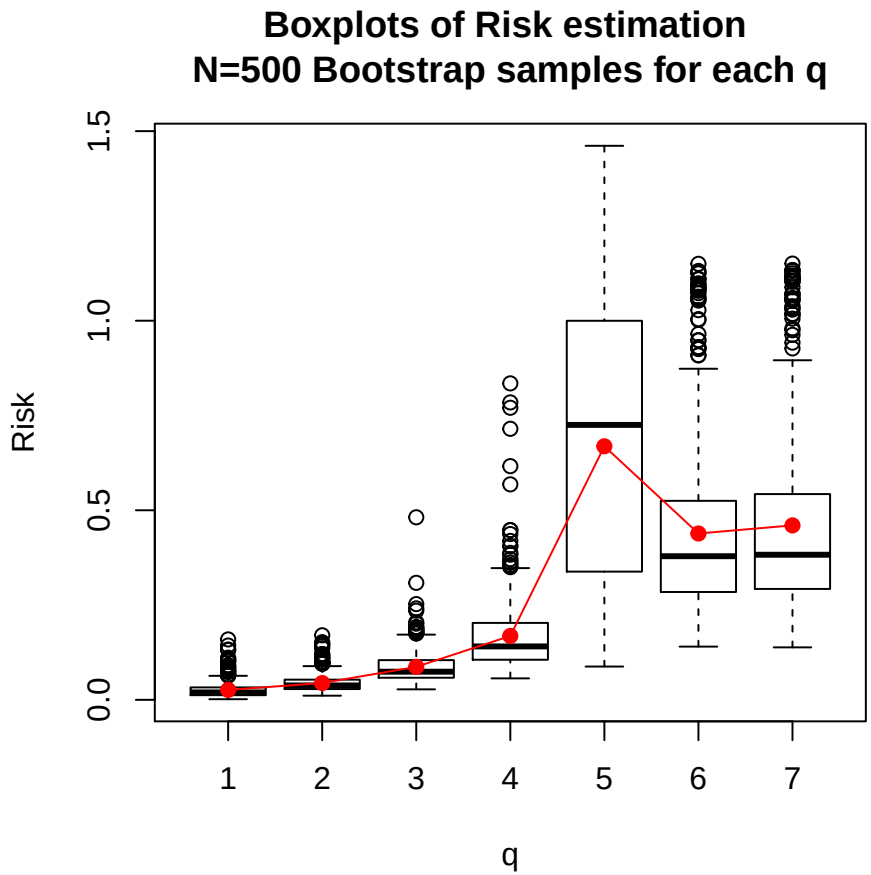
The idea is to estimate $R_q$ by a bootstrap estimator:

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

Where $B$ is the number of bootstrap samples, $P_q^{\star b}$ is the projection matrix obtained with MFAmix on the $b^{th}$ bootstrap sample and $P_q^{\star (.)} = \frac{1}{B} \sum_{b=1}^{B} P_q^{\star b}$.

# Choice of the number of components to keep (3/3)

```
library(boot)
res.stab<-stability.CP.bootstrap(res.mfa=res.MFA,i,q.max=7,NB=500,graph=TRUE)
```




We chose to keep the $q = 3$ first principal components.

# Reduction of the number of variables (1/6)

Goal :

Select a subset of $p^\star < p$ variables such as the PC (denoted $\mathbf{F}^\star$) calculated on this subset are as close as possible to the benchmark PC (calculated on all variables and denoted $\mathbf{F}$ ) in the sense of a measure of connection.

- Method "Closest Submodel Selection" (CSS) developed by Coudret and al. (2014) within the framework of SIR regression.
- Required to define a measure of connection between two groups of 3 variables (the PC $\mathbf{F}$ and $\mathbf{F}^\star$)

# Reduction of the number of variables (2/6)

We denote:

- $\mathbf{F} = \mathbf{U}\mathbf{\Lambda}$ the PC obtained with MFAmix on the $p$ variables,
- $\mathbf{F}^\star = \mathbf{U}^\star\mathbf{\Lambda}^\star$ the PC obtained with MFAmix on the $p^\star$ variables.

The measure of connection between $\mathbf{F}$ and $\mathbf{F}^\star$ is defined as follow:

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

Where:

- $P_F = \mathbf{F}\mathbf{F}^T\mathbf{D}$ is the $\mathbf{D}$-orthogonal projection matrix on $\mathbf{F}$,
- $P_F^\star = \mathbf{F}^\star\mathbf{F}^{\star T}\mathbf{D}$ is the $\mathbf{D}$-orthogonal projection matrix on $\mathbf{F}^\star$.

This measure of connection will allow us, with the CSS method, to choose a subset of variable in order to obtain PC $\mathbf{F}^\star$ the most linked to the benchmark PC $\mathbf{F}$.

# Reduction of the number of variables (3/6)

## The CSS method in MFA

- Step 1: Choose $N_0$, the number of subspace to evaluate and $\zeta$ the percentage of subspace kept among the $N_0$ evaluated.
- Step 2: For $a = 1 \dots N_0$, repeat:
  - Step 2.1: Randomly select $p_0$ variables among the $p$ and construct the matrix $X^{(a)}$ containing the selected variables.
  - Step 2.2: Perform MFAmix on $X^{(a)}$ and calculate $\mathcal{D}(\mathbf{FF}^{(a)})$.
- Step 3: Keep the $N_1 = \zeta N_0$ subspace with the best measure of connection.
- Step 4: Count the number of times when appears every variable in the $N_1$ best subspaces. Afterward, these variables are kept to perform MFAmix.

# Reduction of the number of variables (4/6)

```
reduc.CSS <- reduc.MFAmix.CSS(object = res.MFA, N0 = 20000, p0 = 20, zeta = 5/100, ndim = 3,
    q = 3, graph = FALSE, p.fixed = NULL)
```

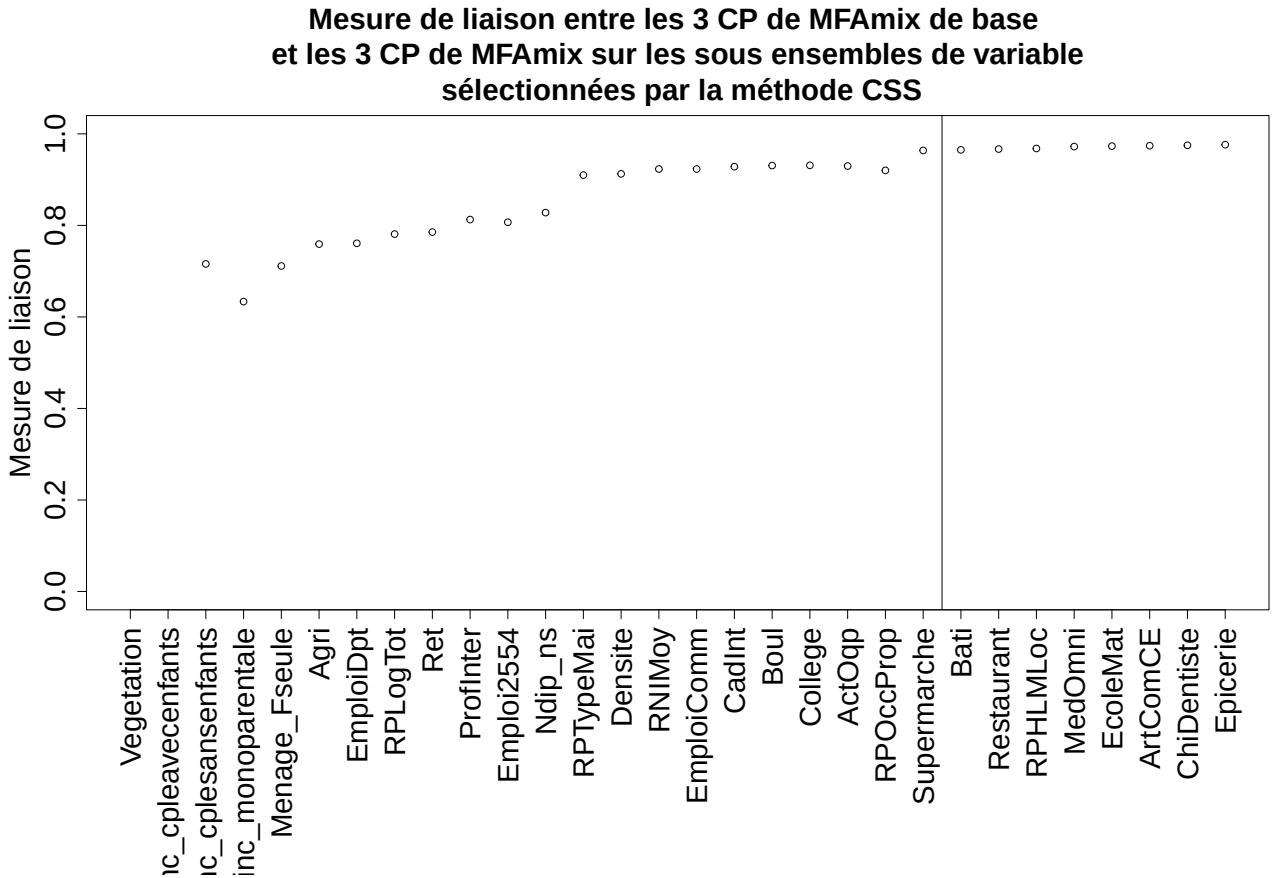
We show here (in decreasing order) the number of times when every variable appeared in the $N_1$ best subspaces.

```
reduc.CSS$var.chosen[1:10]

##                       Vegetation Menage_famprinc_cpleavecenfants
##                              981                              818
## Menage_famprinc_cplesansenfants   Menage_famprinc_monoparentale
##                              796                              729
##                     Menage_Fseule                             Agri
##                              561                              536
##                        EmploiDpt                         RPLogTot
##                              524                              523
##                              Ret                        ProfInter
##                              515                              498
```

# Reduction of the number of variables (5/6)

We will look more specifically what is the measure of connection according to the subset of selected variables.




Measure of connection between the PC according the chosen subset of variables (sorted by descending order of appearance in CSS).

# Reduction of the number of variables (6/6)

Here, we are going to reperform MFAmix on the subset of 20 variables which appeared most of the time in the CSS method then briefly interpret the results.

```
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.level =
```

# Reduction of the number of variables (6/6)

Here, we are going to reperform MFAmix on the subset of 20 variables which appeared most of the time in the CSS method then briefly interpret the results.

```
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.level =
```

We look at the correlations (2 by 2) between PC from the MFAmix benchmark and PC obtained from the CSS method.

```
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 0.969525 -0.13252  0.0116057
## dim 2 0.133680  0.97539 -0.0008336
## dim 3 0.007293 -0.02094  0.9283049
```

We see that PC are highly correlated, so we can perform MFAmix on the 20 selected variables without losing too much information.

# Basic interpretation of the composite indicators (1/5)

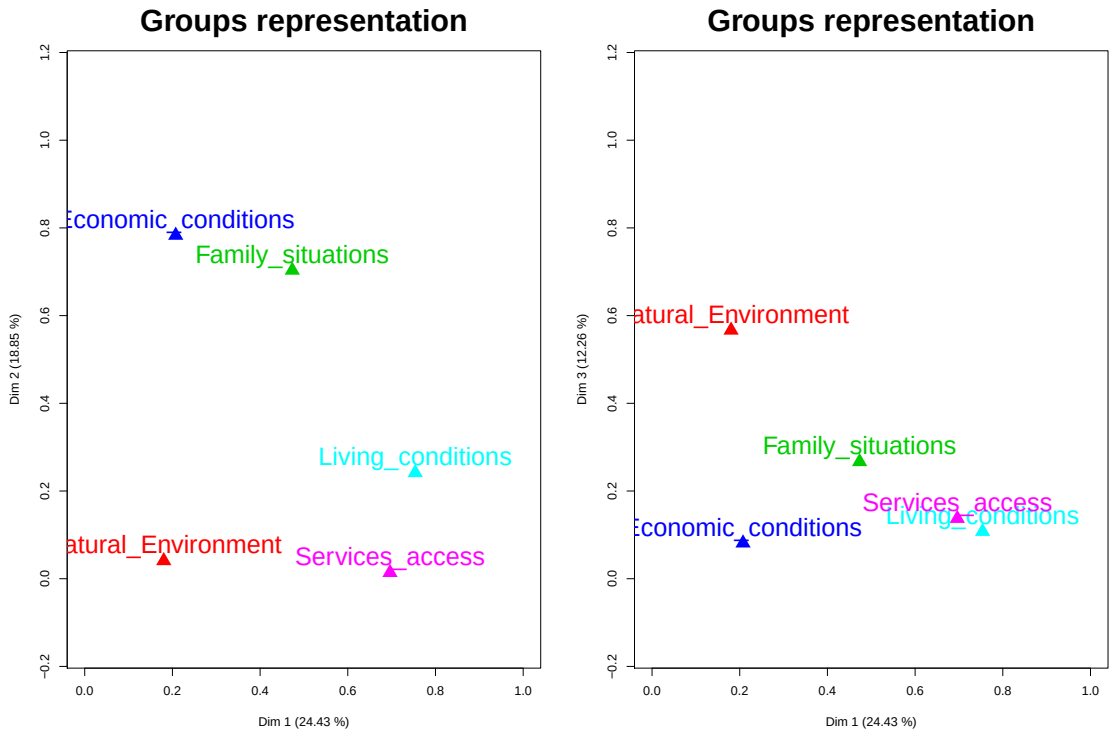



Figure : Representation of groups on the two first factor plans

Axis 1: Living conditions and Services
Axis 2: Economic conditions and family situations
Axis 3: Environment

# Basic interpretation of the composite indicators (2/5)

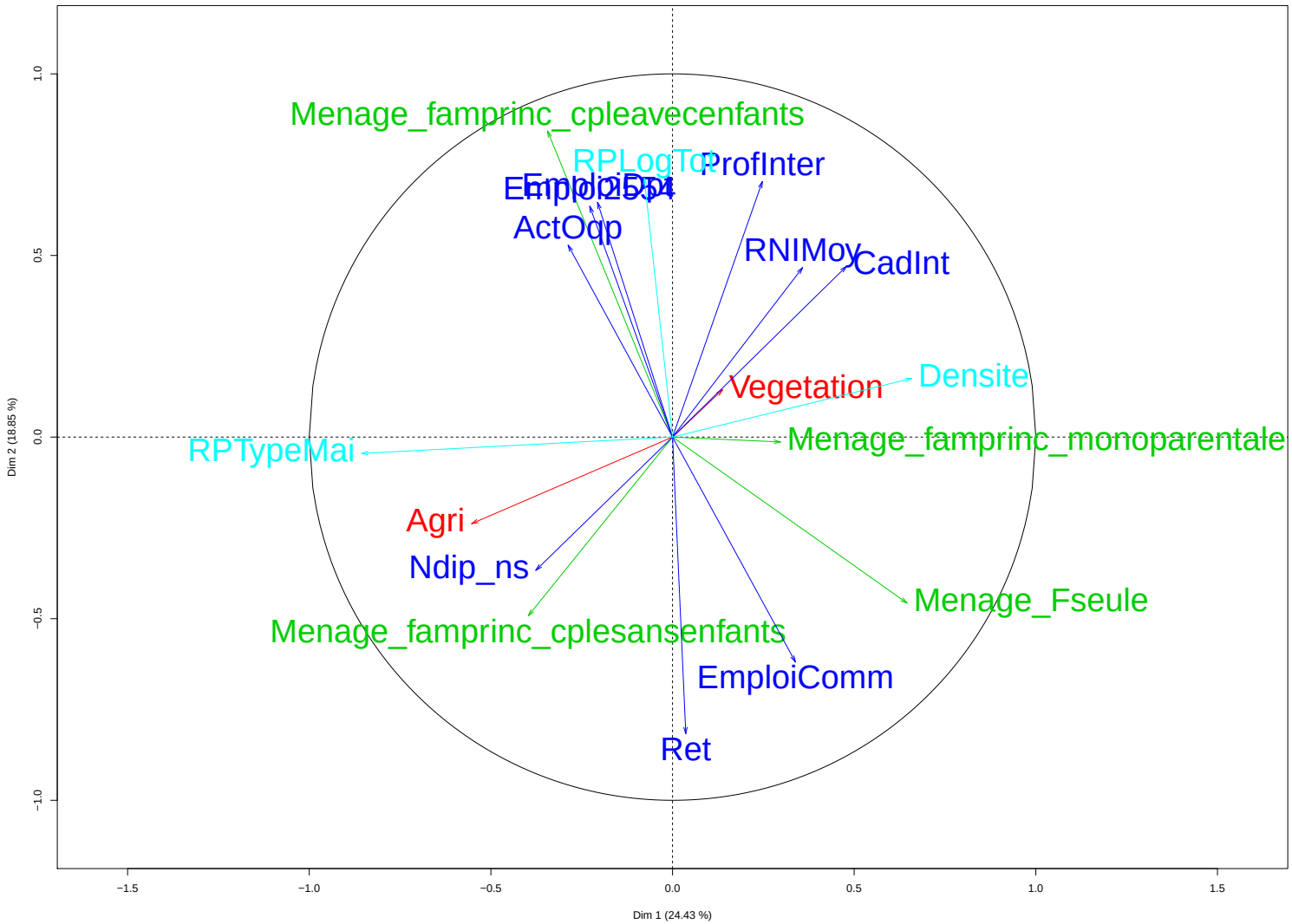



Figure : Correlation circle of numerical variables on the plan (1-2)

# Basic interpretation of the composite indicators (3/5)

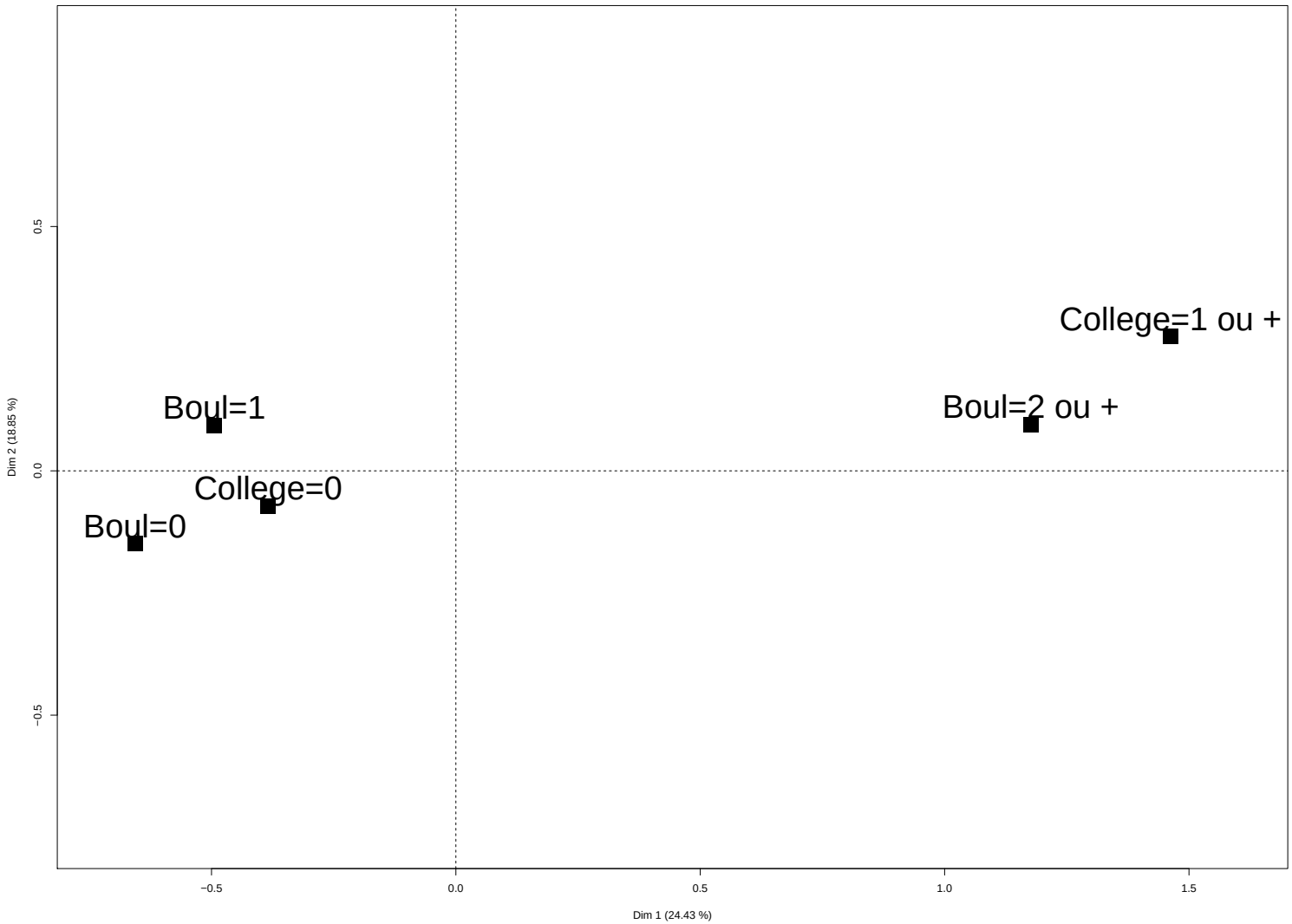



Figure : Representation of levels of categorical variables on the plan (1-2)

# Basic interpretation of the composite indicators (4/5)

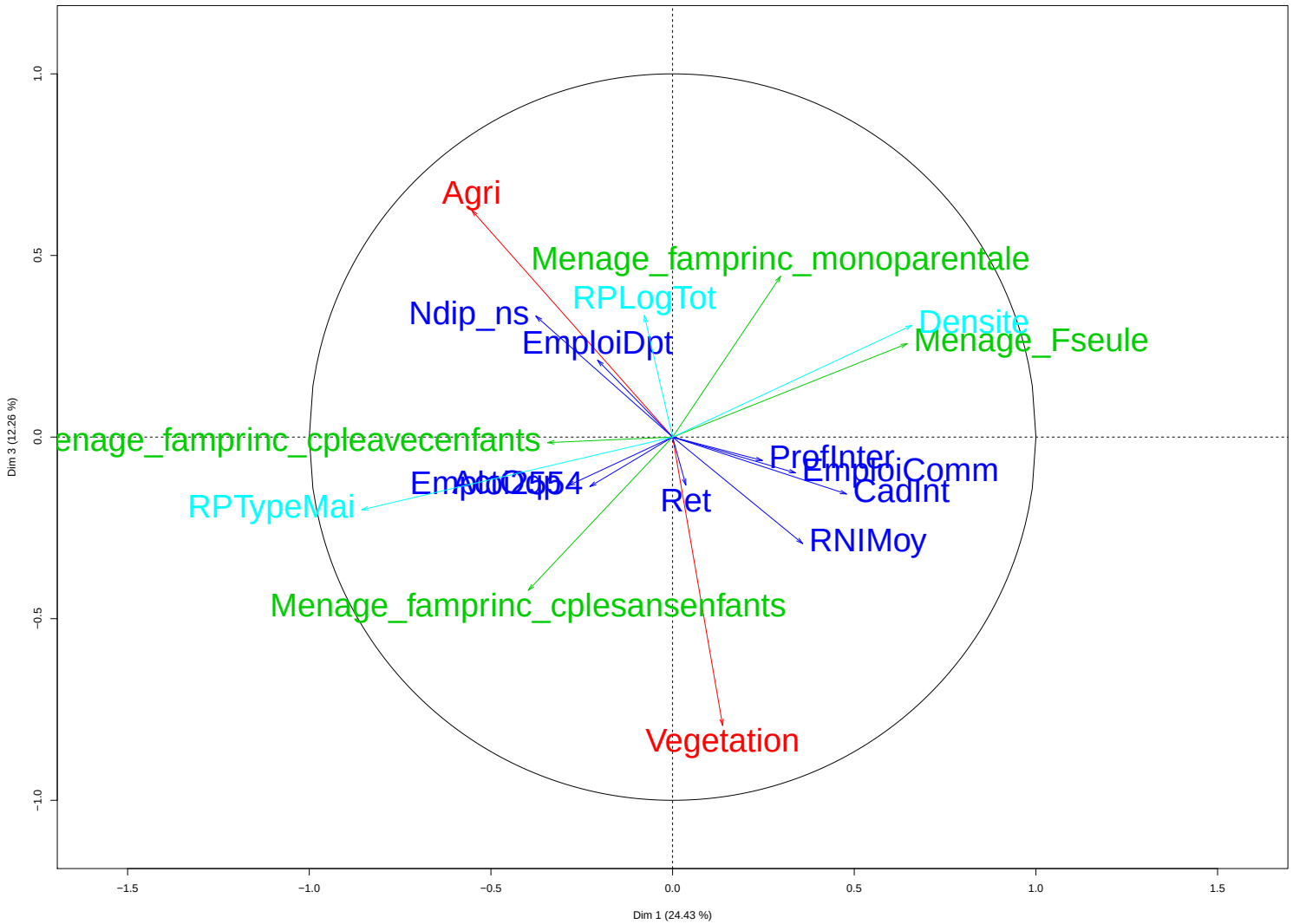



Figure : Correlation circle of numerical variables on the plan (1,3)

# Basic interpretation of the composite indicators (5/5)

- **The indicator 1** mainly characterize access to services and housing conditions.
  - $\oplus$: High population density.
  - $\ominus$: Low density, high proportion of houses instead of apartments.
- **The indicator 2** mainly characterize employment conditions.
  - $\oplus$: Greater proportion of retired people.
  - $\ominus$: Greater proportion of family with children.
- **The indicator 3** contrast agricultural territories with forest territories.
  - $\oplus$: Mainly agricultural territories.
  - $\ominus$: Mainly forest territories.

# Conclusion

- The MFAmix method allowed to perform MFA when data are mixed within groups. This method can be usefull to create composite indicators.
- The use of a method of selection of variables allowed us to create new indicators more easily understable and interpretable.
- Fonctions MFAmix and PCAmix are available in the Package `PCAmixdata` available on CRAN.

# Thank you for your attention

## Some references


1. Noll, Heinz-Herbert (2006). Towards a European System of Social Indicators: Theoretical Framework and System Architecture. *Social Indicators Research*.
2. Escofier B et Pagès J (1983), Méthode pour l'analyse de plusieurs groupes de variables. Application à la caractérisation des vins rouges du Val de Loire, *Revue de statistique appliquée*, 31(2) : 43-59.
3. Chavent, M., Kuentz-Simonet, V., Labenne, A.,Saracco, J. (2013).Une extension de l'analyse factorielle multiple pour des groupes de variables mixtes : MFAmix. *45èmes Journées de la Statistique, Toulouse*.
4. Besse, P. (1992). PCA stability and choice of dimensionality. *Statistics and Probability Letters*, **13**, 405-410.
5. Josse, J., Husson, F. (2012). Selecting the number of components in PCA using cross-validation approximations. *Computational Statistics and Data Analysis*, **56**, pp. 1869-1879.
6. Coudret, R., Liquet, B., Saracco, J. (2014). Comparison of sliced inverse regeression method approaches for undeterminated cases.*Journal de la Société française de Statistique*, in press.