FactoMineR

An R package for exploratory data analysis for teaching and research

François Husson, Julie Josse & Sébastien Lê

AGROcampus RENNES

FactoMineR

Why FactoMineR ?

- To make exploratory multivariate data analysis with a free software R
- The possibility to propose new methods (taking into account different structure on the data)
- To have a package user friendly and oriented to practitioner (a very easy GUI)

AGROcampus RENNES

FactoMineR

1 – The classical methods

- Methods implemented are similar in their main objective: to sum up and simplify the data by reducing the dimensionality of the dataset

- Continuous variables: Principal Components Analysis
- Contingency table: Correspondence Analysis
- Categorical variables: Multiple Correspondence Analysis
- Continuous and categorical variables: Mixed Data Analysis

PCA Example

Data : performances of 41 athletes during two meetings of decathlon

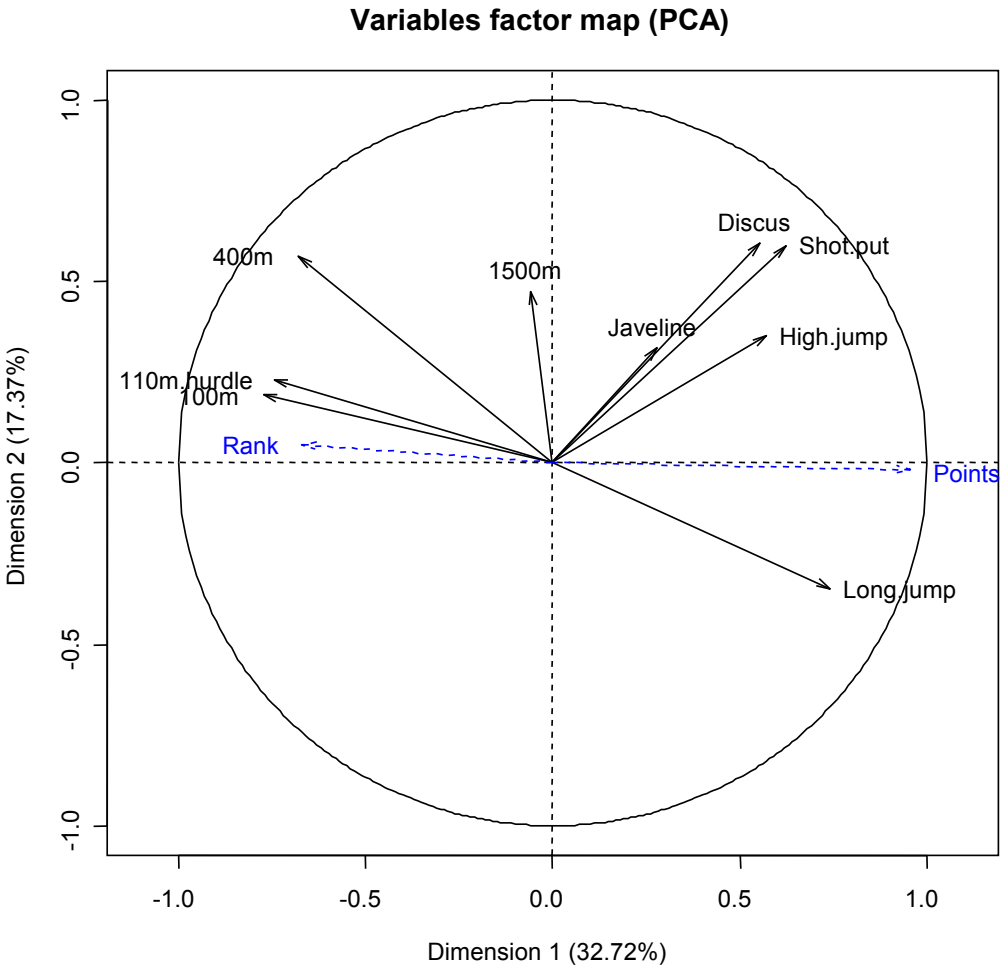
	100m	Long.jump	Shot.put	High.jump	400m	110m.hurdle	Discus	Pole.vault	Javeline	1500m	Rank	Points	Competition
SEBRLE	11.04	7.58	14.83	2.07	49.81	14.69	43.75	5.02	63.19	291.70	1	8217	Decastar
CLAY	10.76	7.40	14.26	1.86	49.37	14.05	50.72	4.92	60.15	301.50	2	8122	Decastar
KARPOV	11.02	7.30	14.77	2.04	48.37	14.09	48.95	4.92	50.31	300.20	3	8099	Decastar
BERNARD	11.02	7.23	14.25	1.92	48.93	14.99	40.87	5.32	62.77	280.10	4	8067	Decastar
YURKOV	11.34	7.09	15.19	2.10	50.42	15.31	46.26	4.72	63.44	276.40	5	8036	Decastar
Sebrle	10.85	7.84	16.36	2.12	48.36	14.05	48.72	5.00	70.52	280.01	1	8893	OlympicG
Clay	10.44	7.96	15.23	2.06	49.19	14.13	50.11	4.90	69.71	282.00	2	8820	OlympicG
Karpov	10.50	7.81	15.93	2.09	46.81	13.97	51.65	4.60	55.54	278.11	3	8725	OlympicG
Macey	10.89	7.47	15.73	2.15	48.97	14.56	48.34	4.40	58.46	265.42	4	8414	OlympicG
Warners	10.62	7.74	14.48	1.97	47.97	14.01	43.73	4.90	55.39	278.05	5	8343	OlympicG

AGROcampus RENNES

FACTOMINER

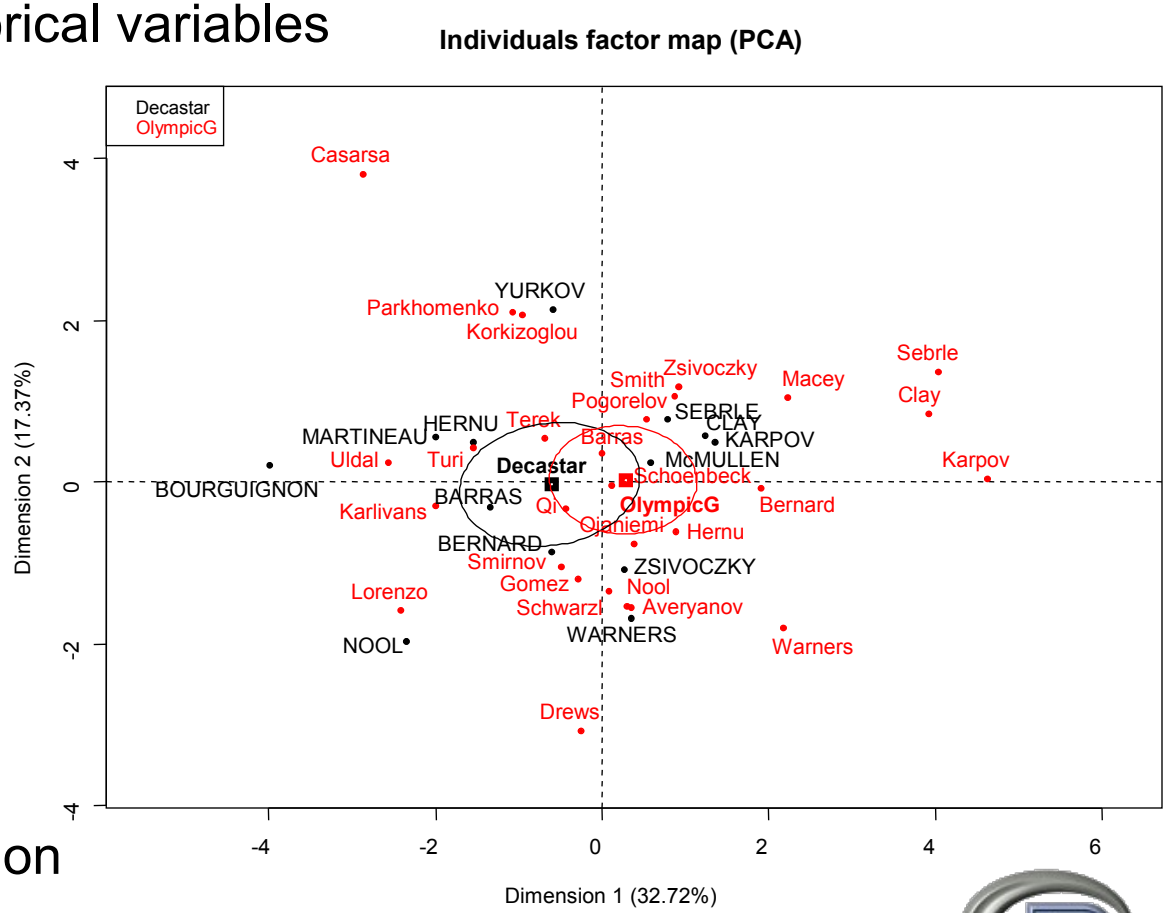
PCA example

- Introduction of supplementary information:
  - supplementary continuous variables
- Graphs enriched by :
  - representing the variables according to their quality of representation
- Indicators:
  - contribution
  - quality of representation



PCA example

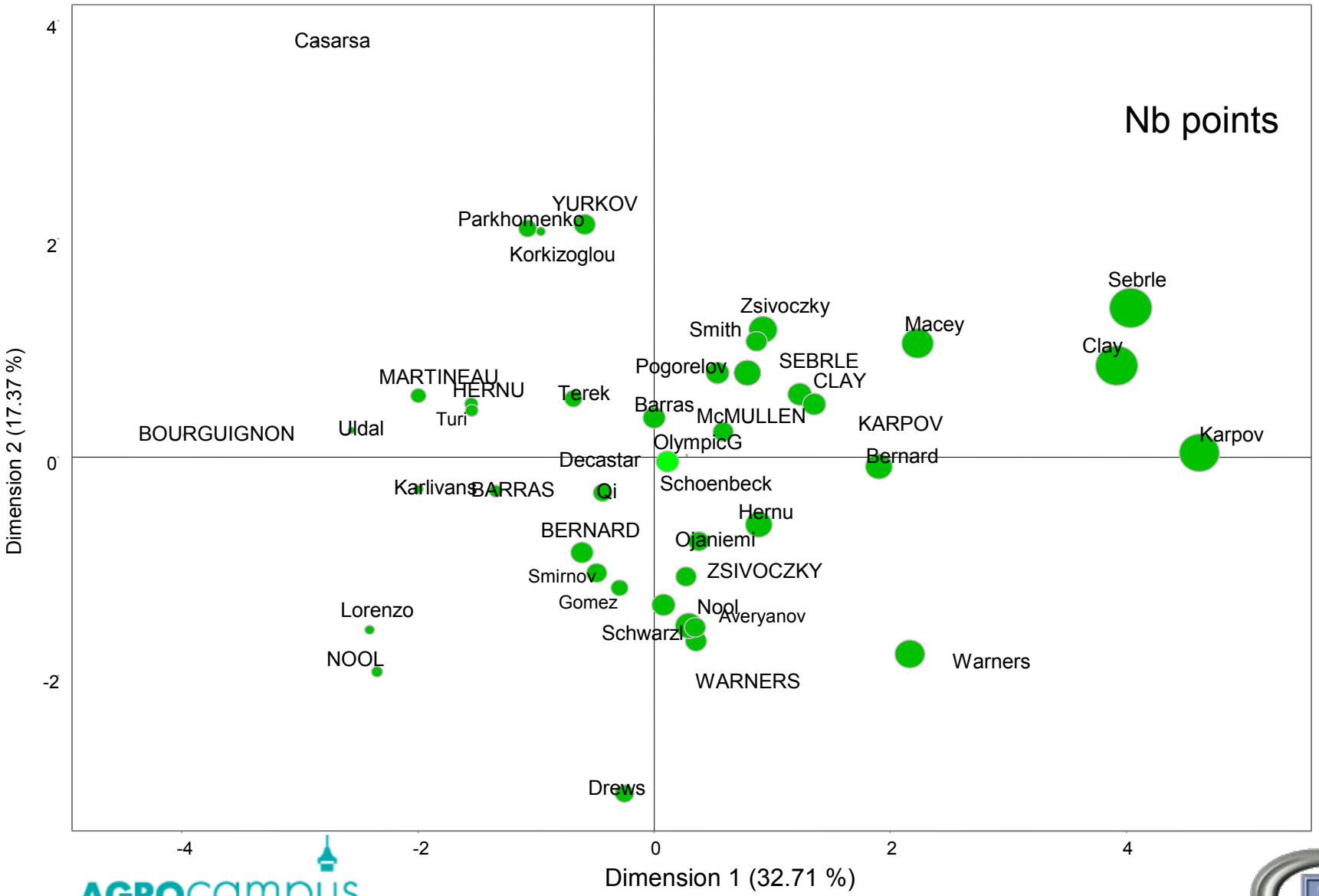
- Introduction of supplementary information:
  - supplementary individuals
  - supplementary categorical variables
- Graphs enriched by:
  - coloring according to supplementary information
  - confidence ellipses around the categories
- Indicators:
  - contribution
  - quality of representation



AGROcampus RENNES

FACTOMINER

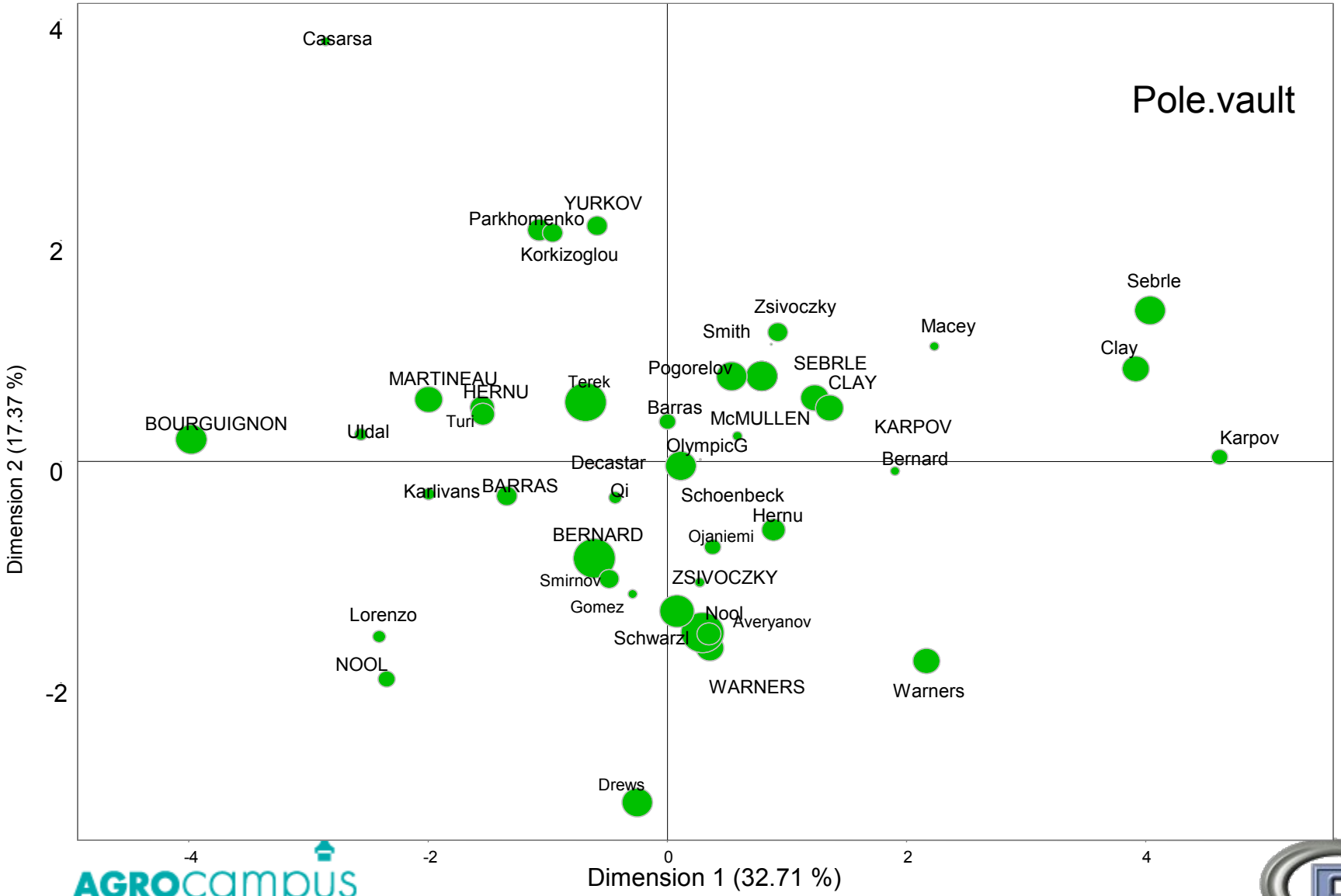
PCA example



AGROcampus RENNES

FACTOMINER

PCA example



AGROcampus RENNES

FACTOMINER

Description of the dimensions

- **By the quantitative variables:**
  - The correlation between each variable and the coordinate of the individuals on the axis *s* is calculated
  - The correlation coefficients are sorted
  - Only the significant correlations are given

Best variable to describe the 1st dimension →

```
$Dim.1
$Dim.1$quanti
              Dim.1
Points         0.96
Long.jump      0.74
Shot.put       0.62
Rank          -0.67
400m          -0.68
110m.hurdle   -0.75
100m          -0.77
```

```
$Dim.2
$Dim.2$quanti
              Dim.2
Discus         0.61
Shot.put       0.60
```

Significant level = 0.05

AGROcampus RENNES

FACTOMINER R

Description of the dimensions

➢ By the qualitative variables:

- Perform a one-way analysis of variance with the coordinates of the individuals on the axis explained by the qualitative variable

```
$Dim.1$quali
               P-value
Competition      0.155

$Dim.1$category
          Estimate P-value
OlympicG   0.4393   0.155
Decastar  -0.4393   0.155
```

Significant level = 0.2

- A *F*-test by variable

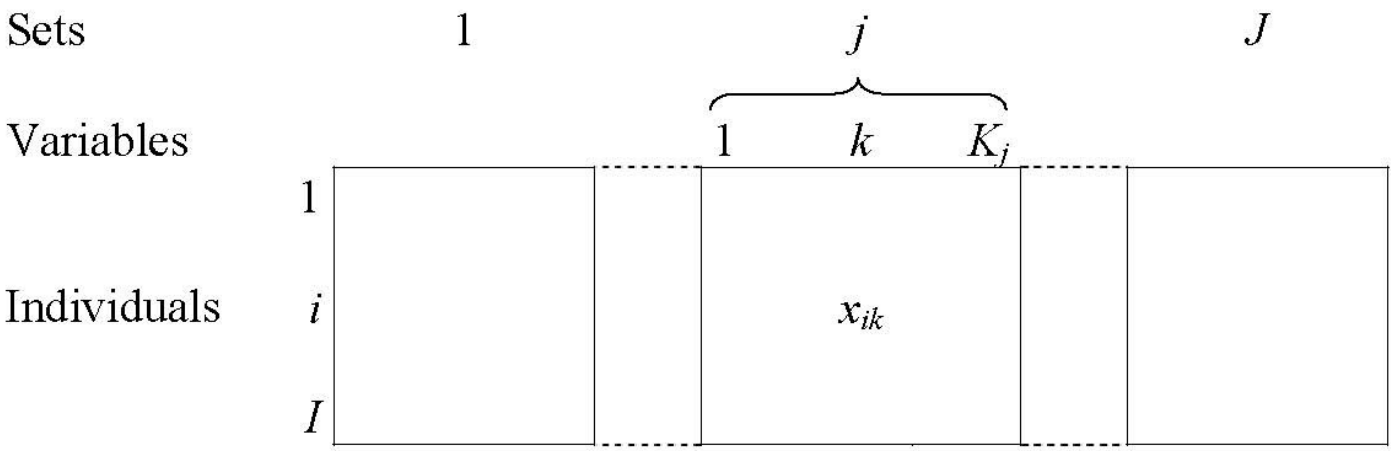
- For each category, a student *T*-test to compare the average of the category with the general mean

2 – Structure on the data

Different structure on the data are proposed:

- a partition on the variables: several sets of variables are simultaneously studied: Multiple Factor Analysis, Generalized Procrustes Analysis
- a hierarchy on the variables: variables are grouped and subgrouped (like in questionnaires structured in topics and subtopics): Hierarchical Multiple Factor Analysis
- a partition on the individuals: several sets of individuals described by the same variables: Dual Multiple Factor Analysis

Groups of variables (MFA)



Groups of variables are quantitative and/ or qualitative

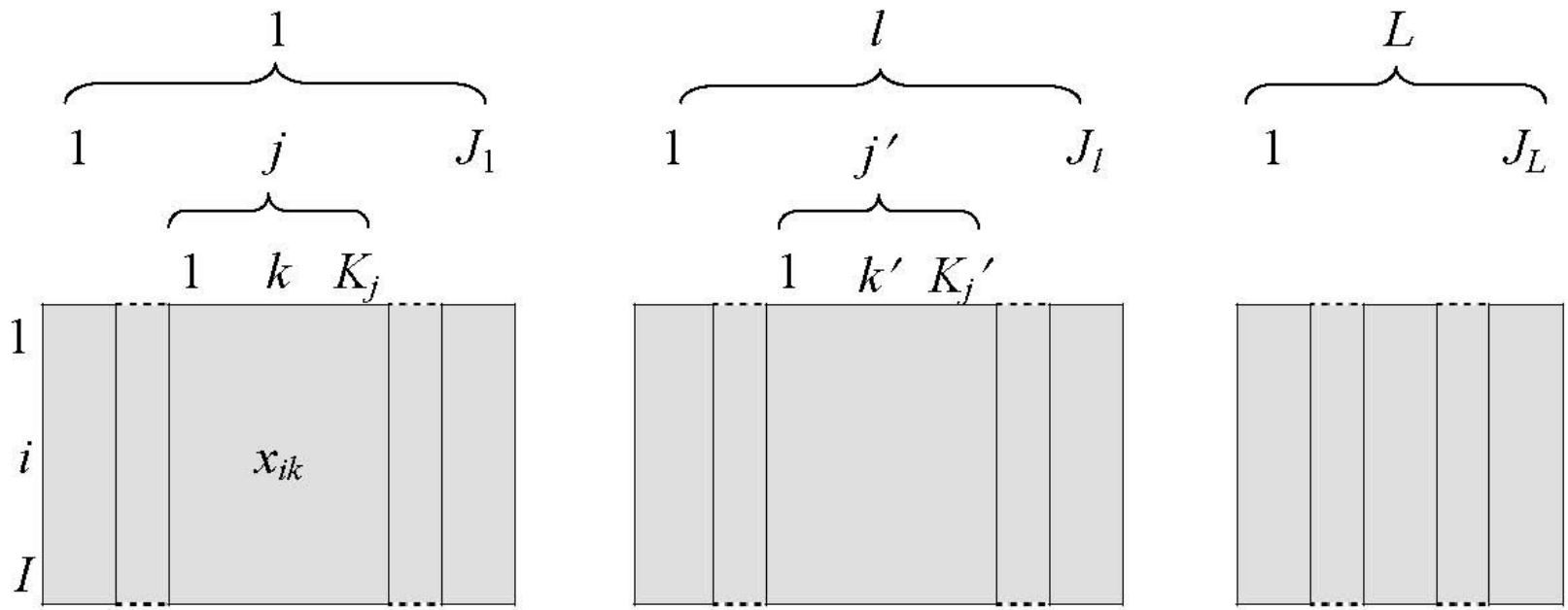
Objectives : - study the link between the sets of variables
- balance the influence of each group of variables
- give the classical graphs but also specific graphs:
groups of variables - partial representation

Examples : - Genomic: DNA, protein
- Sensory analysis: sensorial, physico-chemical
- Comparison of coding (quantitative / qualitative)

AGROcampus RENNES

FACTOMINER

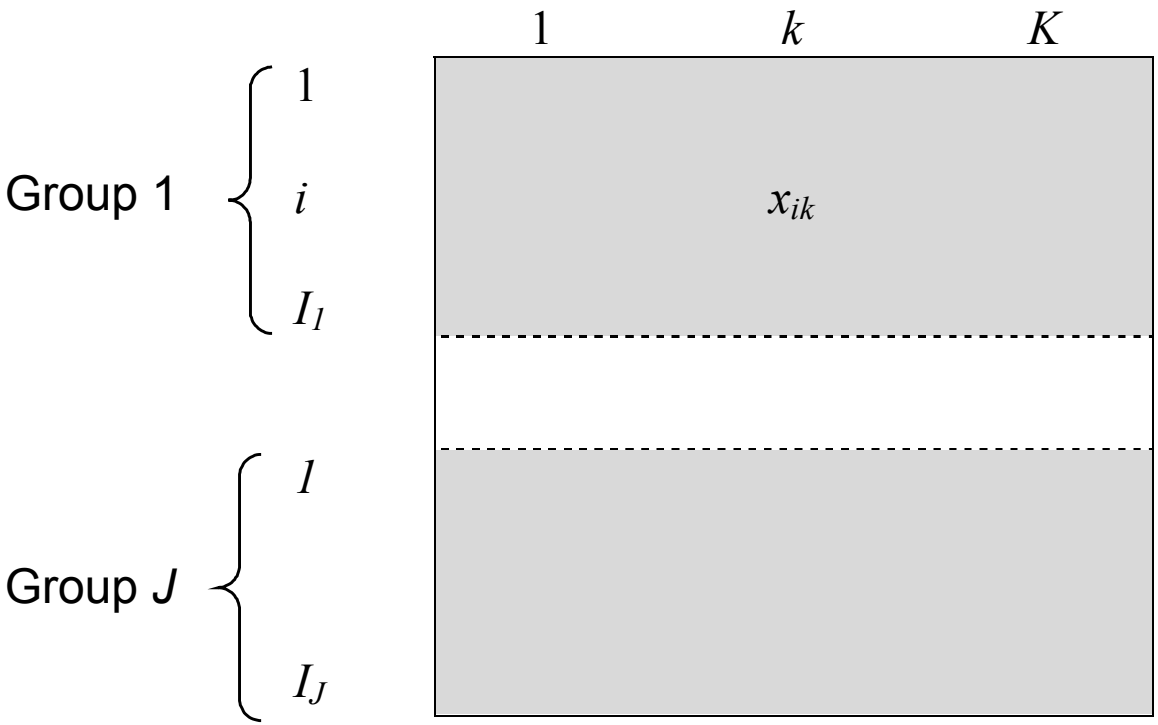
Hierarchy on the variables (HMFA)



Two levels for the hierarchy: the first one contains *L* groups,
each *l* group contains J_l subgroups, and each subgroup have K_j variables

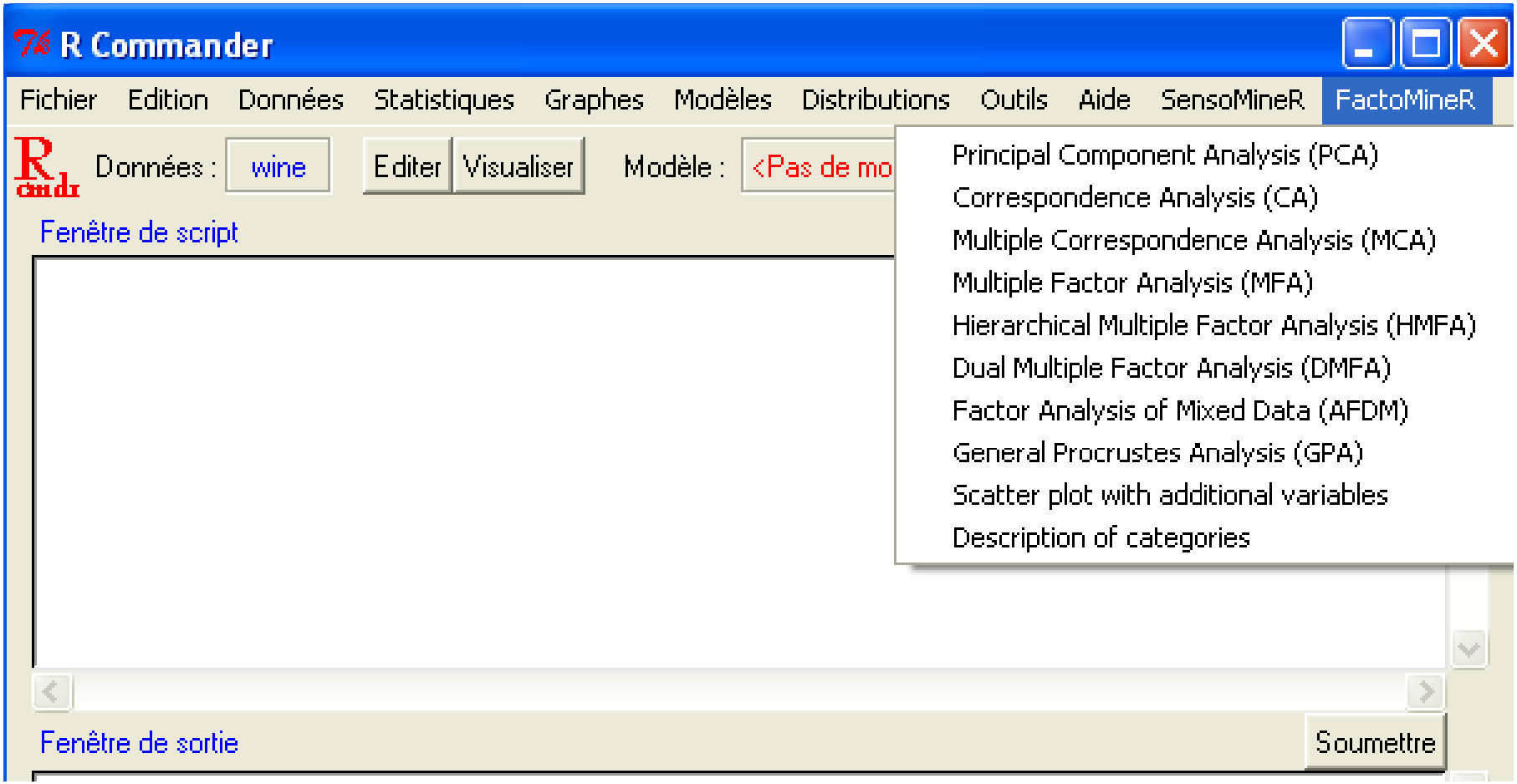
Objective: to balance the groups and the subgroups of variables

Partition on the individuals (DMFA)



Objective: to compare the covariance matrices

3 – Graphical User Interface



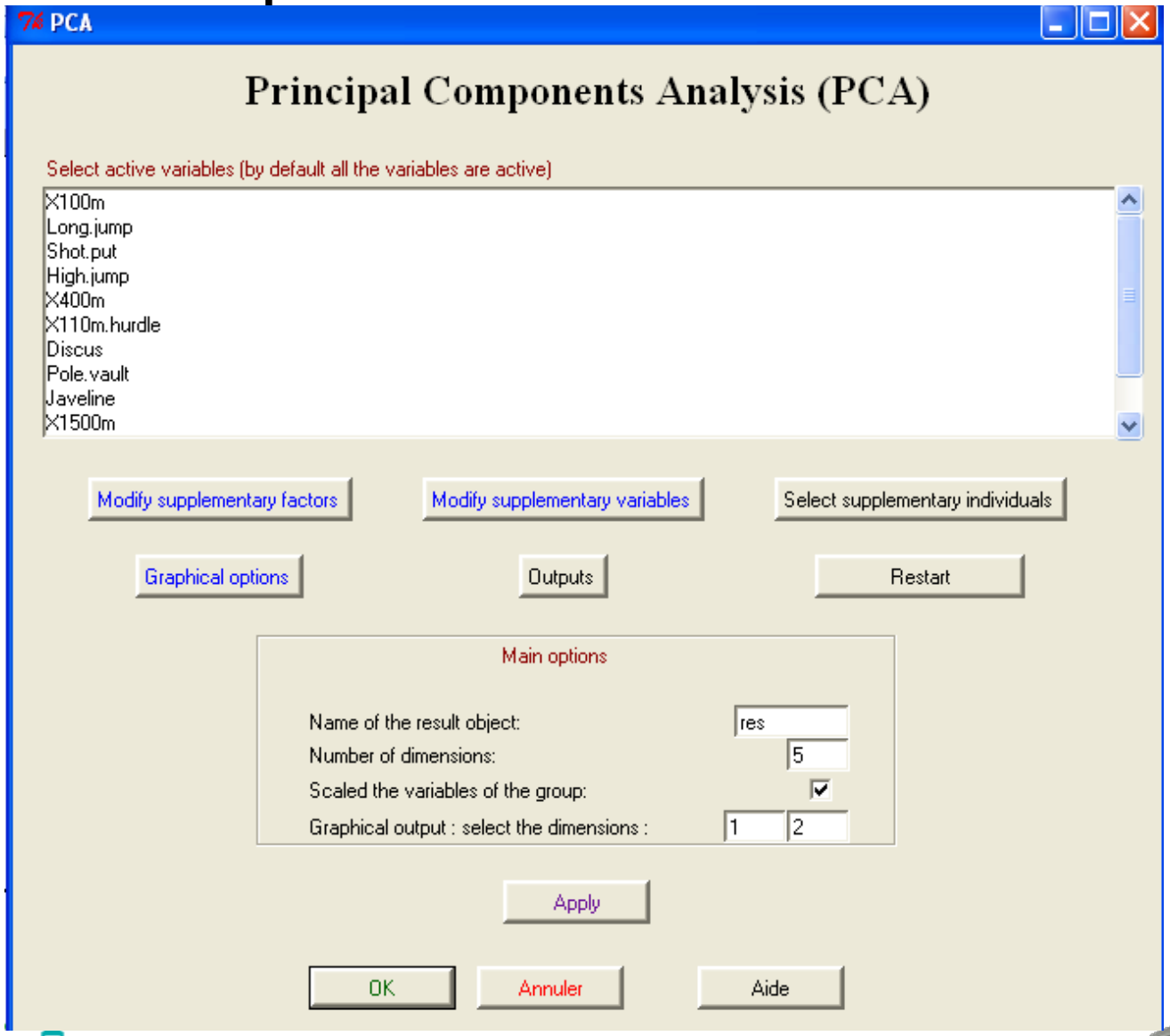
Menu of the FactoMineR GUI

AGROcampus RENNES

FACTOMINER

3 – Graphical User Interface

Main window of the PCA

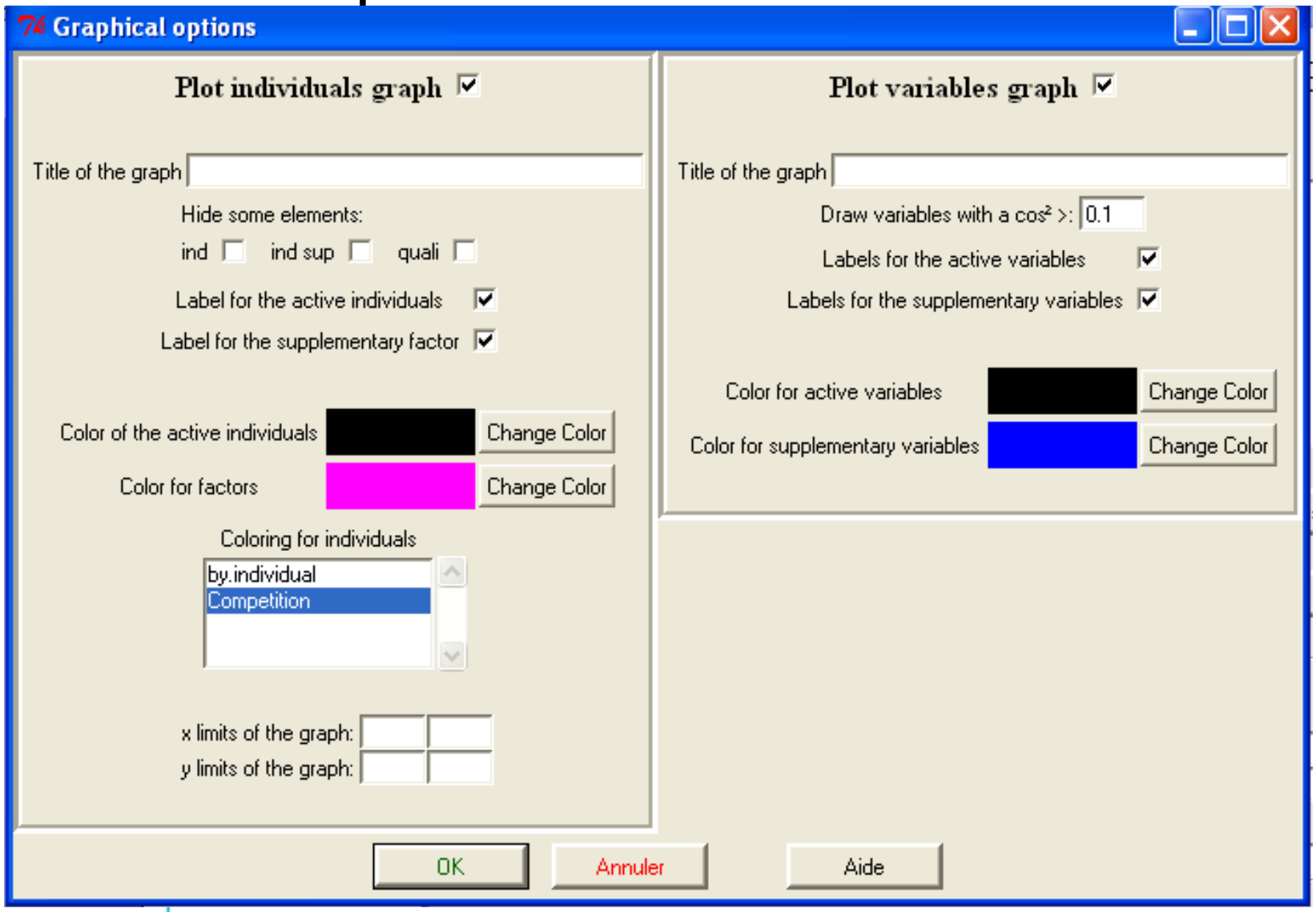


AGROcampus RENNES

FACTOMINER

3 – Graphical User Interface

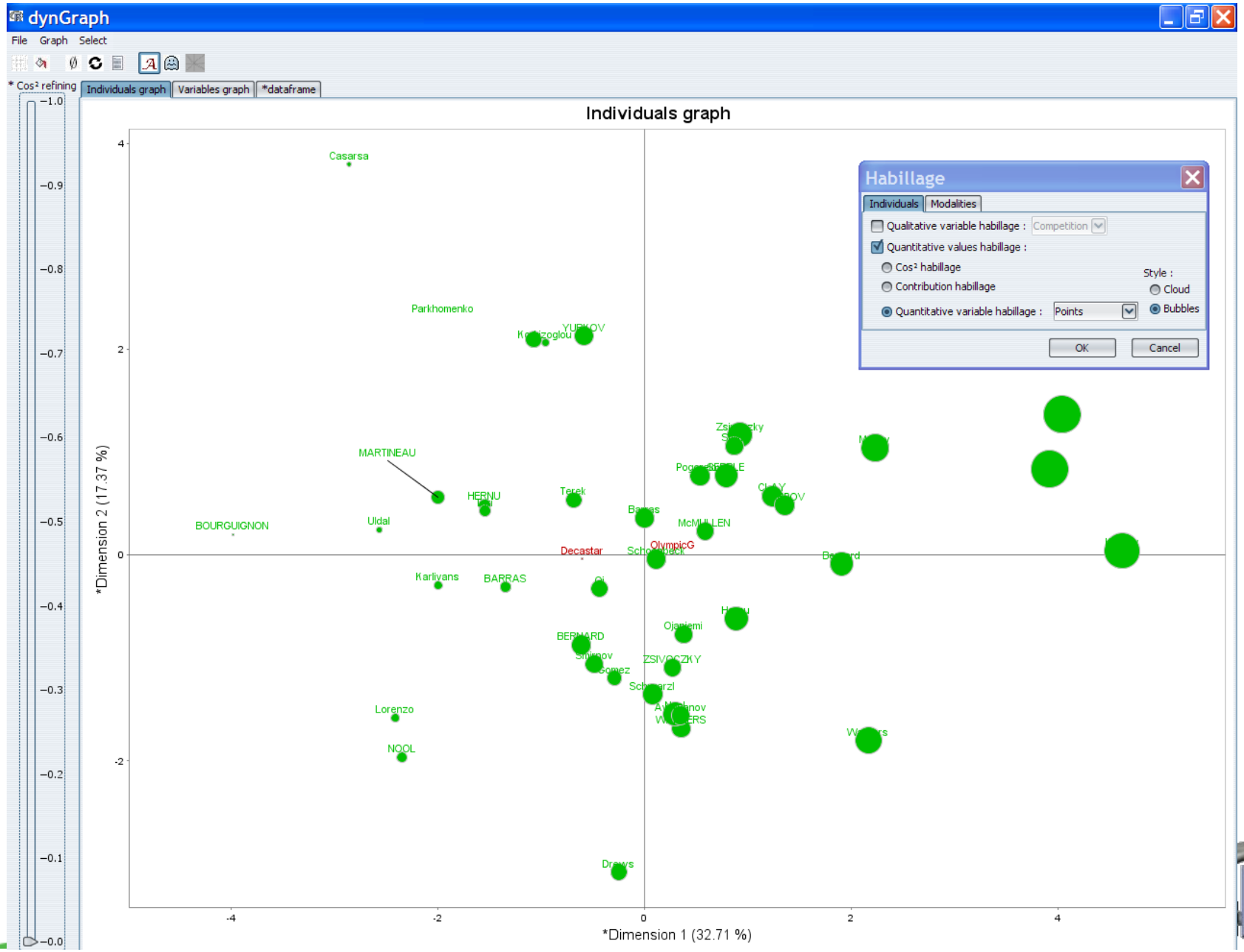
Graphical options



AGROcampus RENNES

FACTOMINER

3 – Graphical User Interface



4 – Conclusion

For researchers, practitioners and students: with classical and advanced methods

The FactoMineR package is available on the CRAN

The GUI can be simply loaded:

```
source("http://factominer.free.fr/install-facto.r")
```

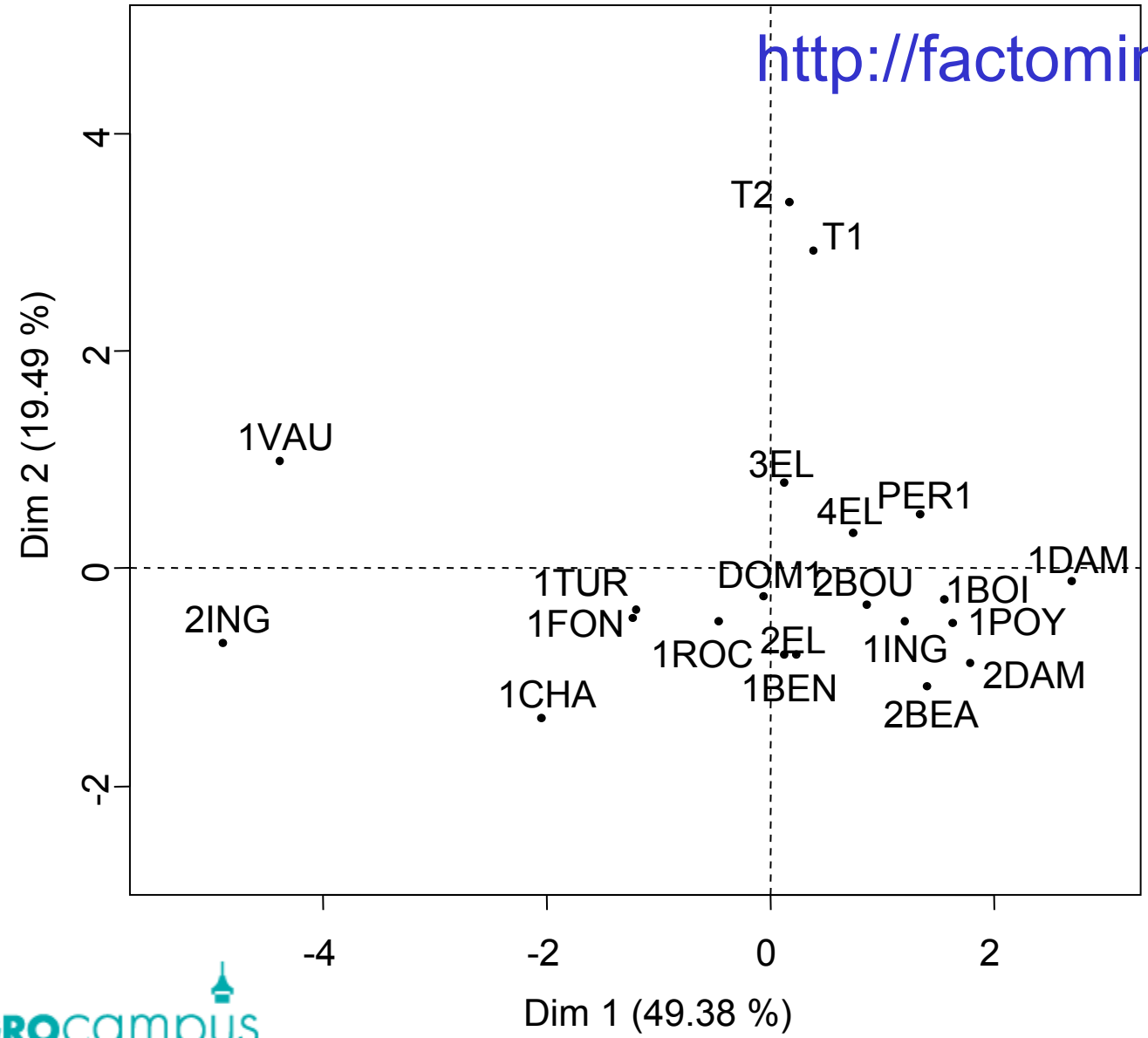
A website is dedicated to this package: http://factominer.free.fr

Future: dynamical graphs
Perspective: UseR!2008 (2 tutorials), UseR!2009 at Rennes

AGROcampus RENNES

FACTOMINER

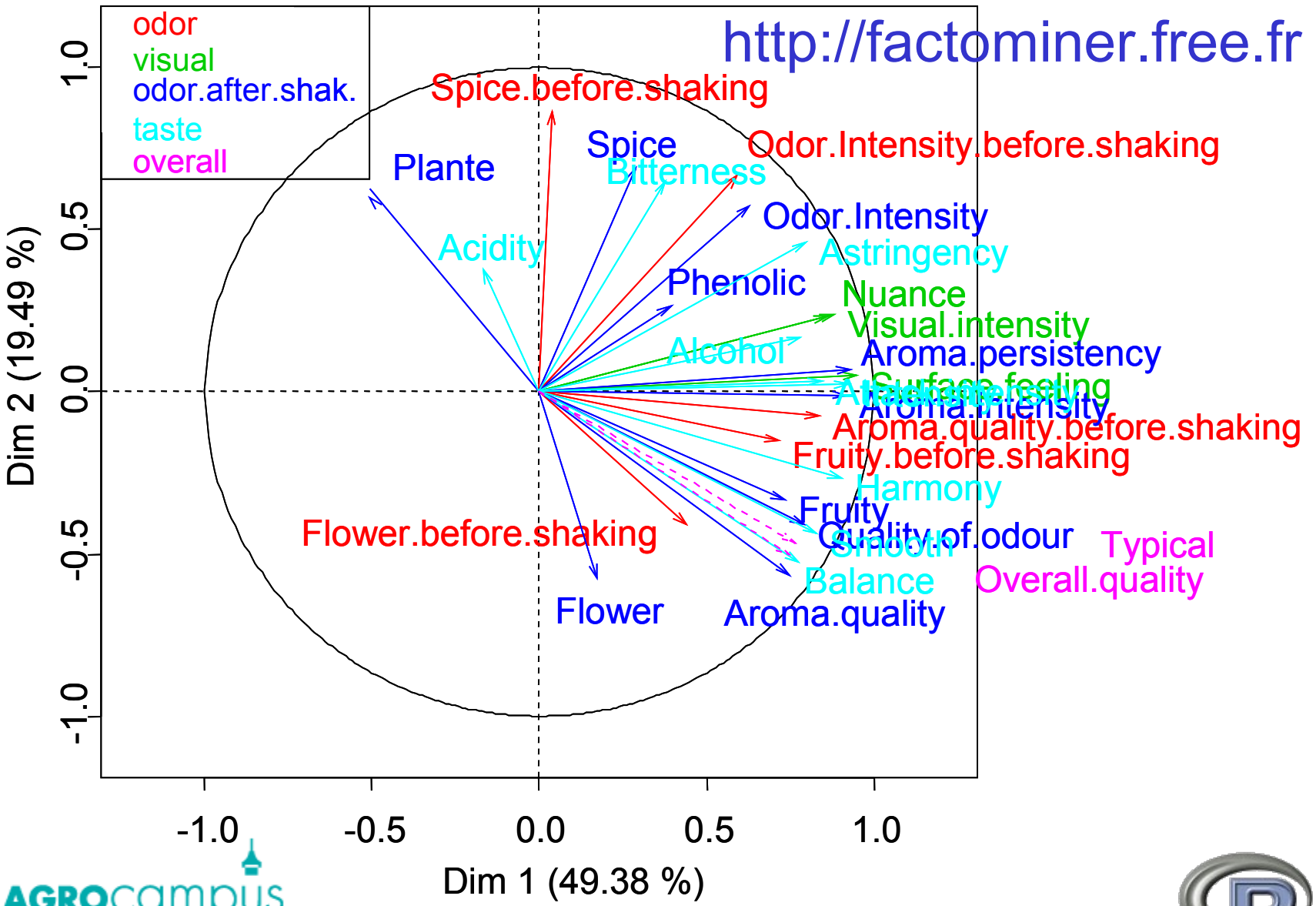
MFA example: representation of the individuals



AGROcampus RENNES

FACTOMINER

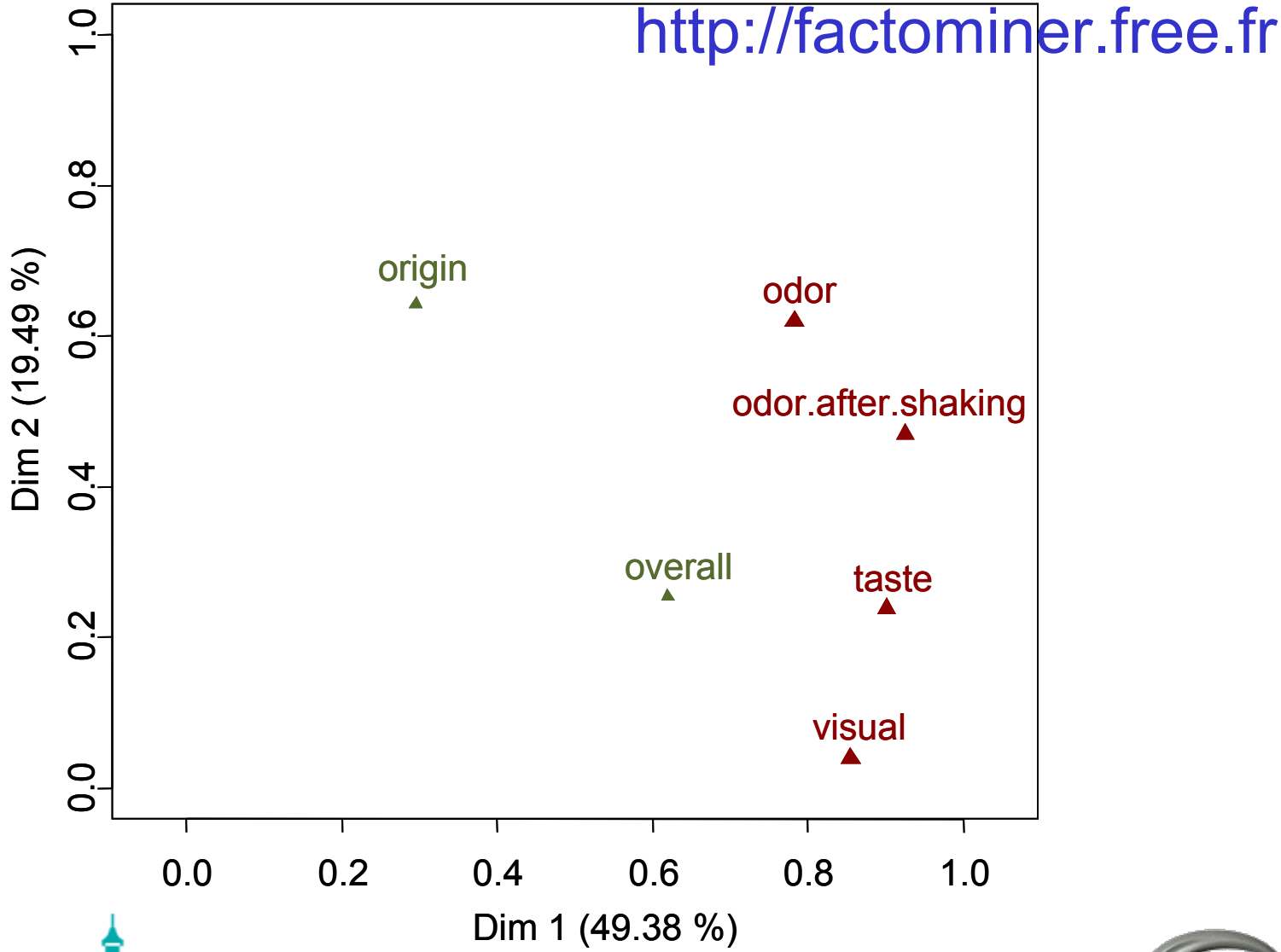
MFA example: representation of the variables



AGROcampus RENNES

FACTOMINER

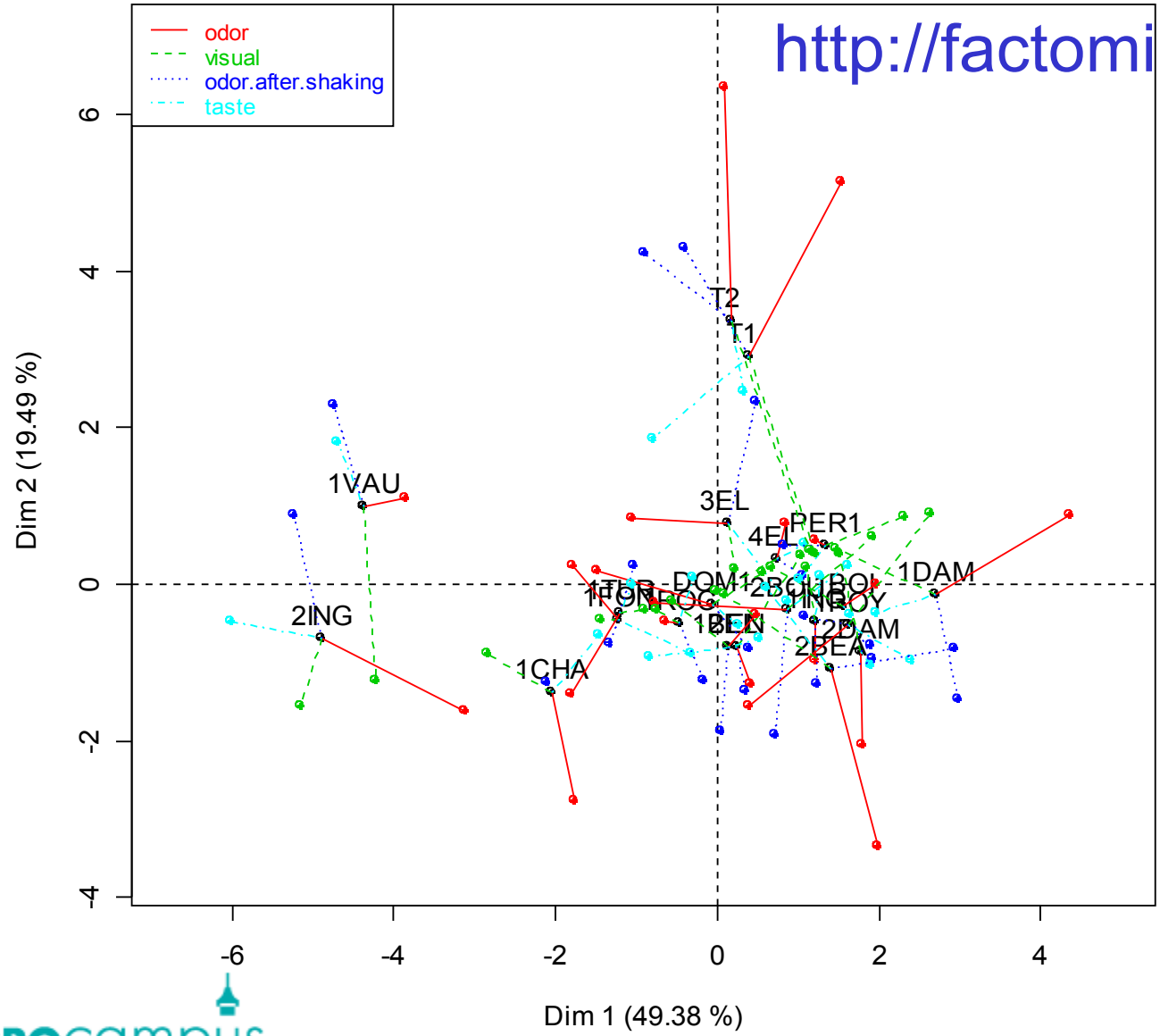
MFA example: representation of the groups



AGROcampus RENNES

FACTOMINER

MFA example: representation of the partial points

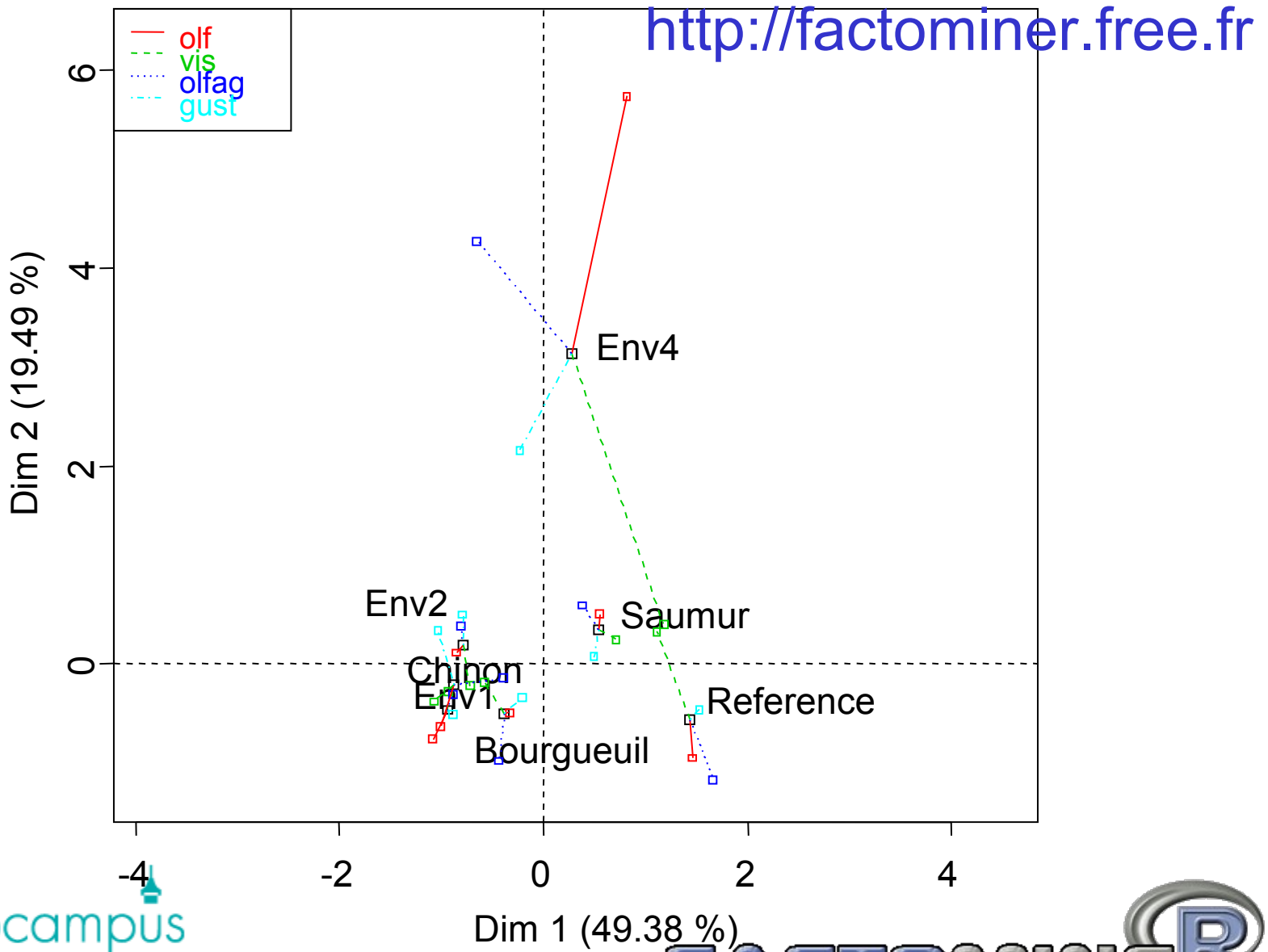


http://factominer.free.fr

AGROcampus RENNES

FACTOMINER

MFA example: representation of the partial points

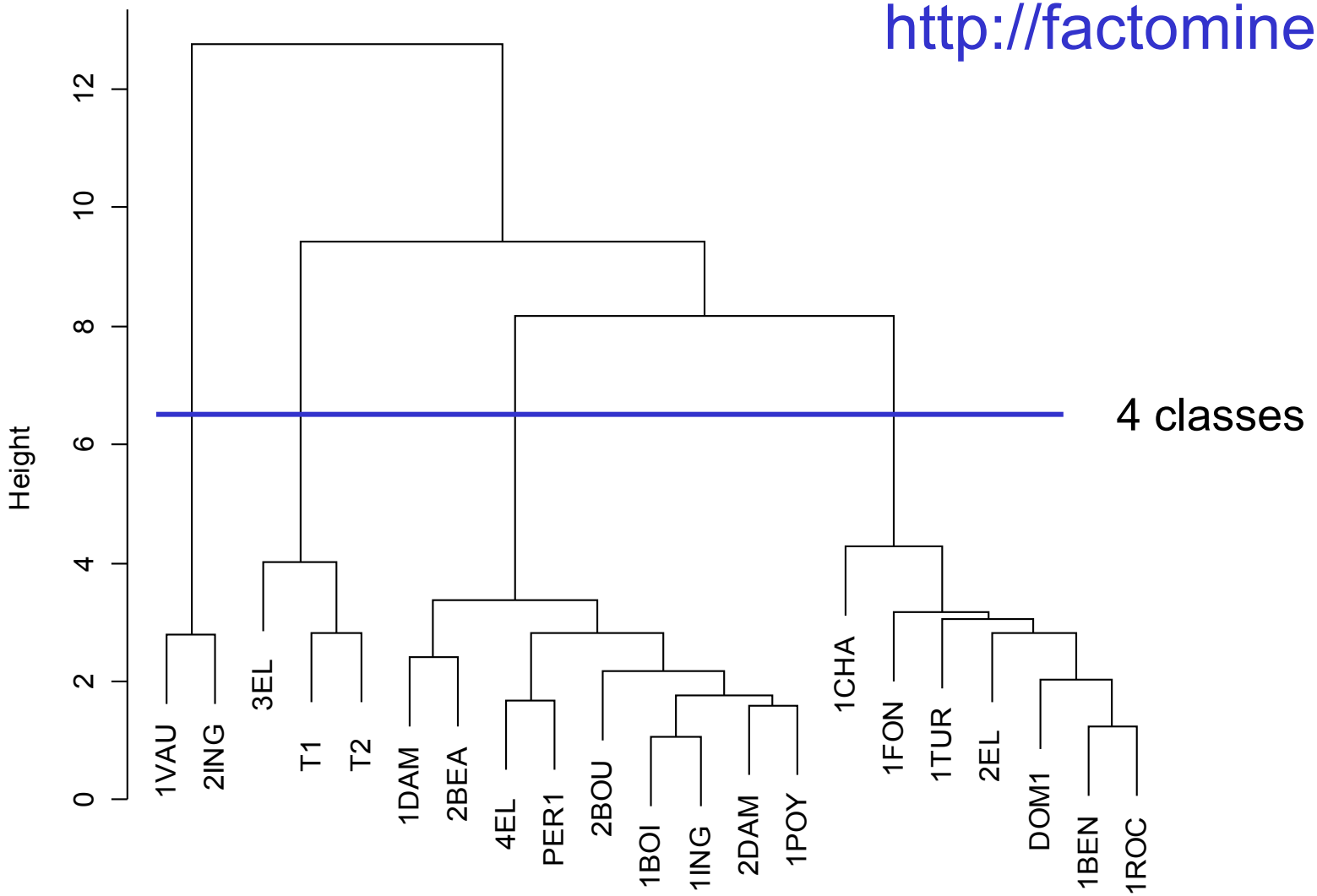


AGROcampus RENNES

FACTOMINER

Unsupervised classification

http://factominer.free.fr

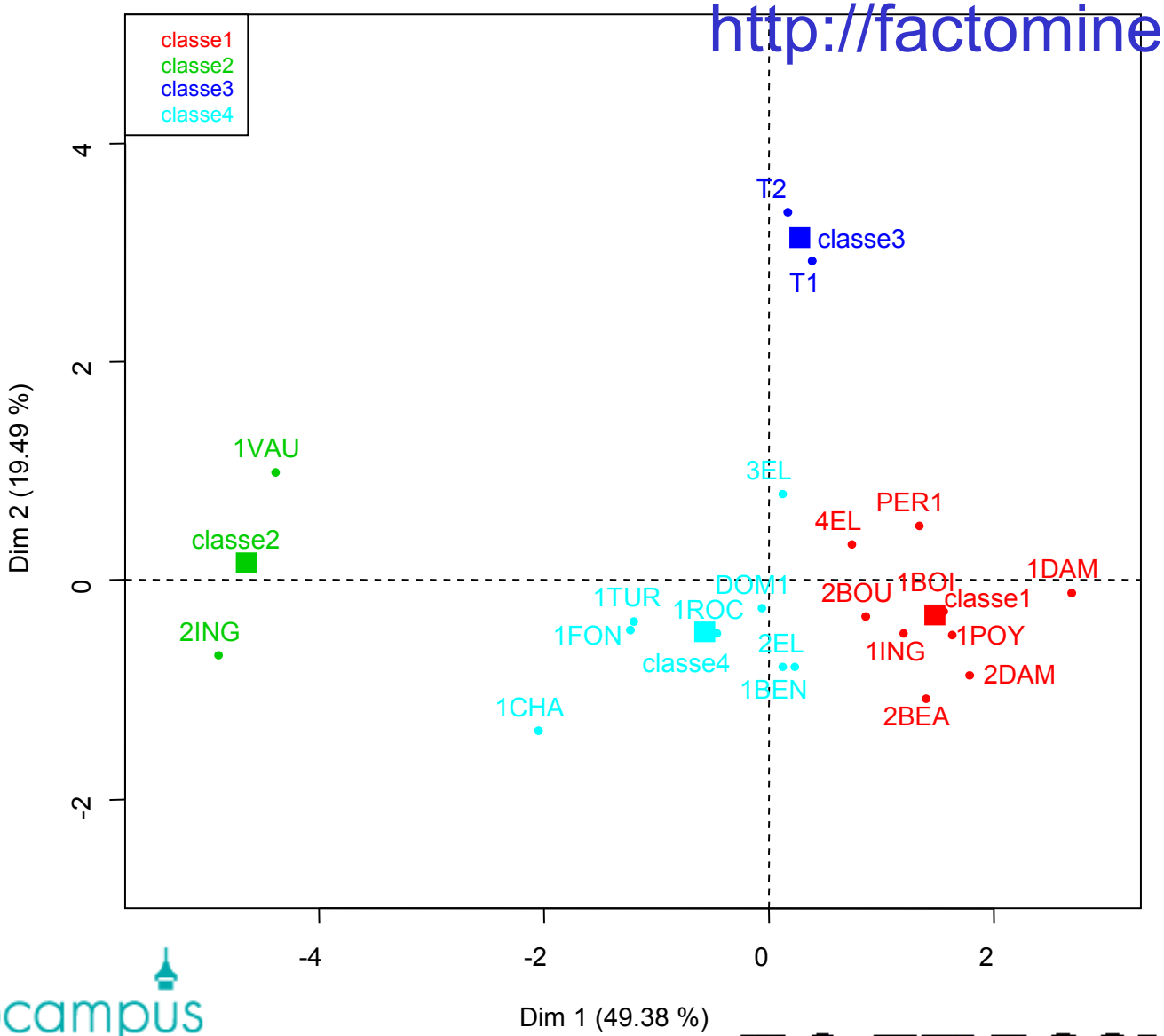


AGROcampus RENNES

FACTOMINER

MFA example: representation of the individuals

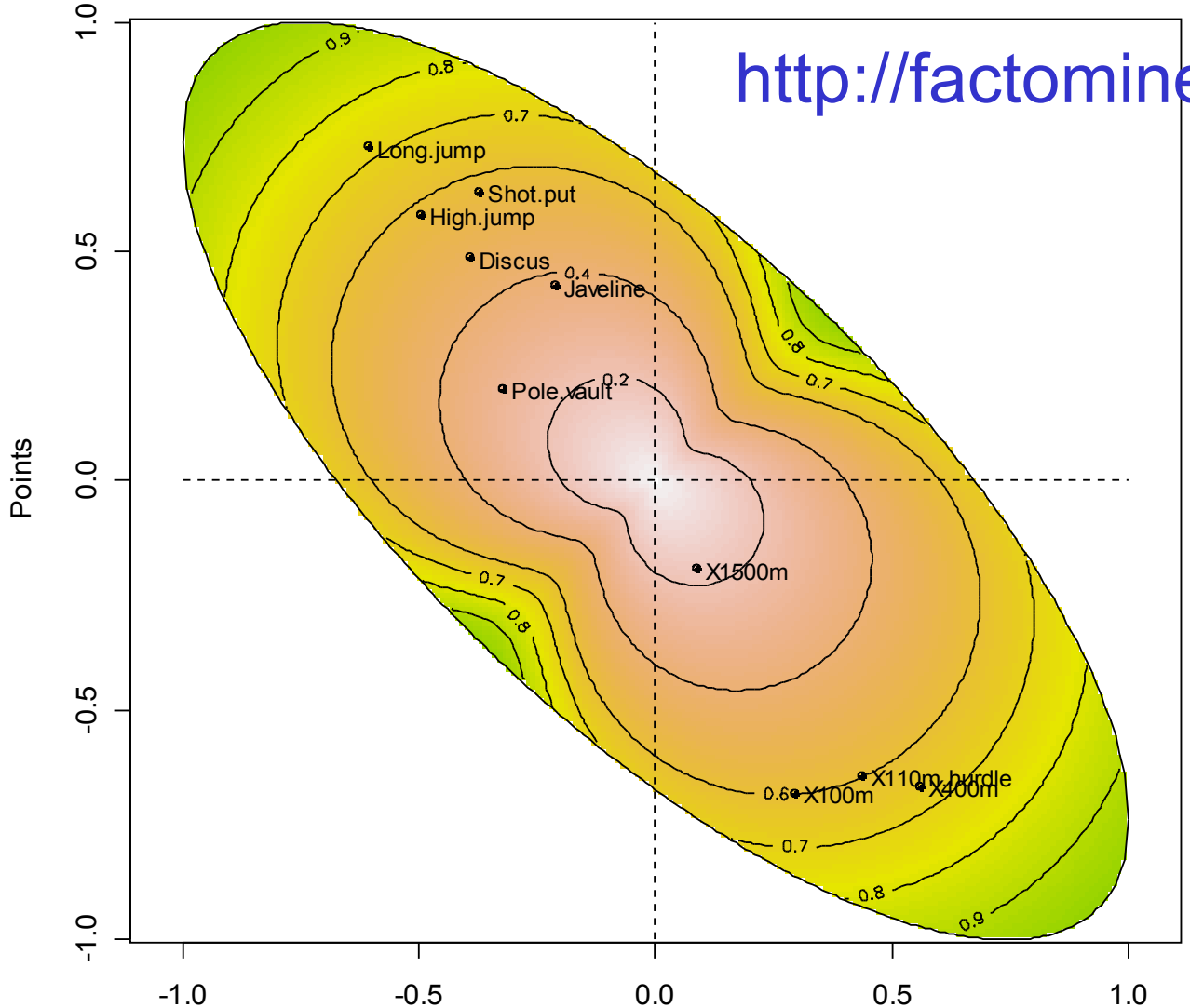
http://factominer.free.fr



AGROcampus RENNES

FACTOMINER

Prefmap-PLS graph between Rank and Points

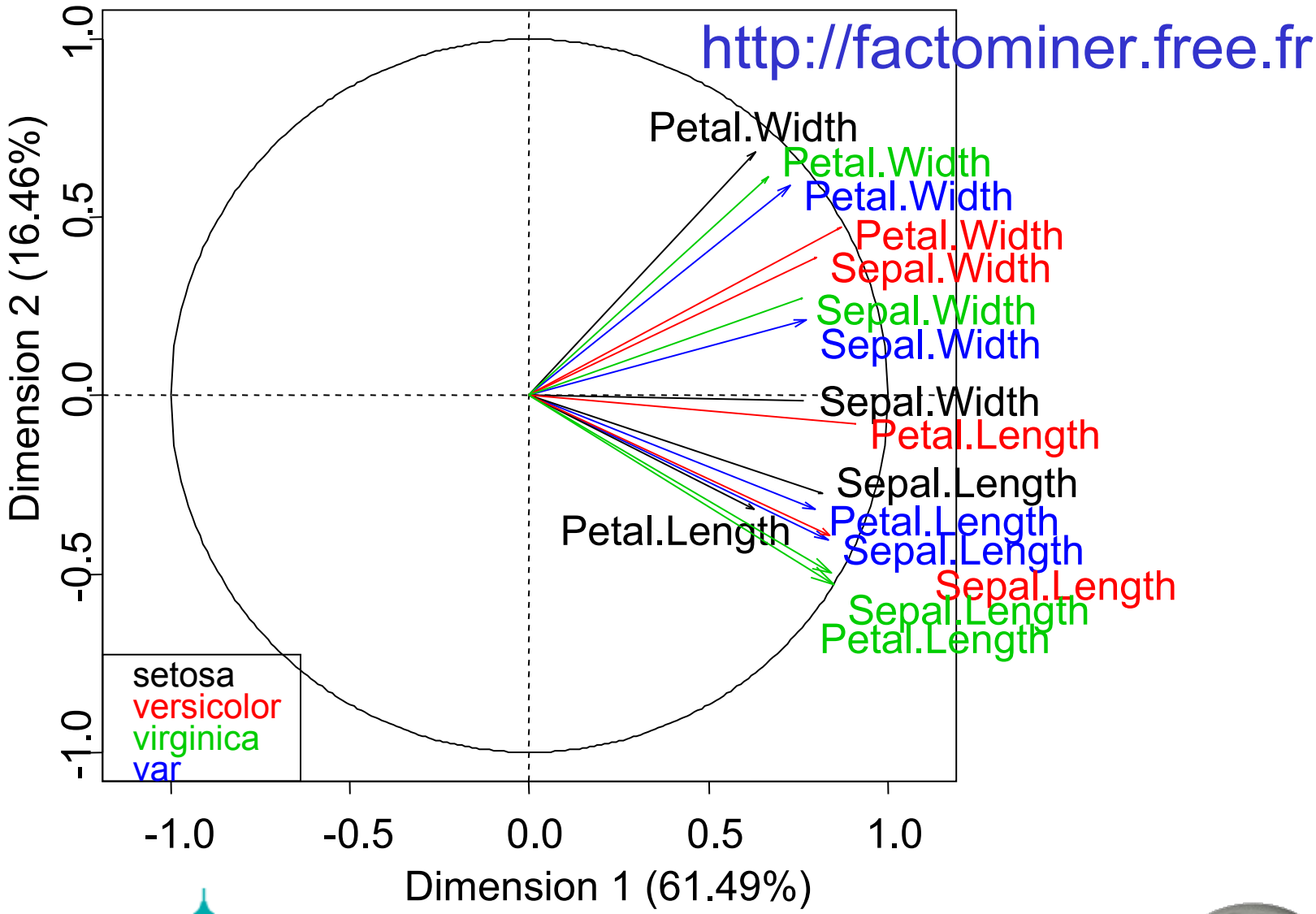


Correlation between Rank and Points : -0.7392

AGROcampus RENNES

FACTOMINER

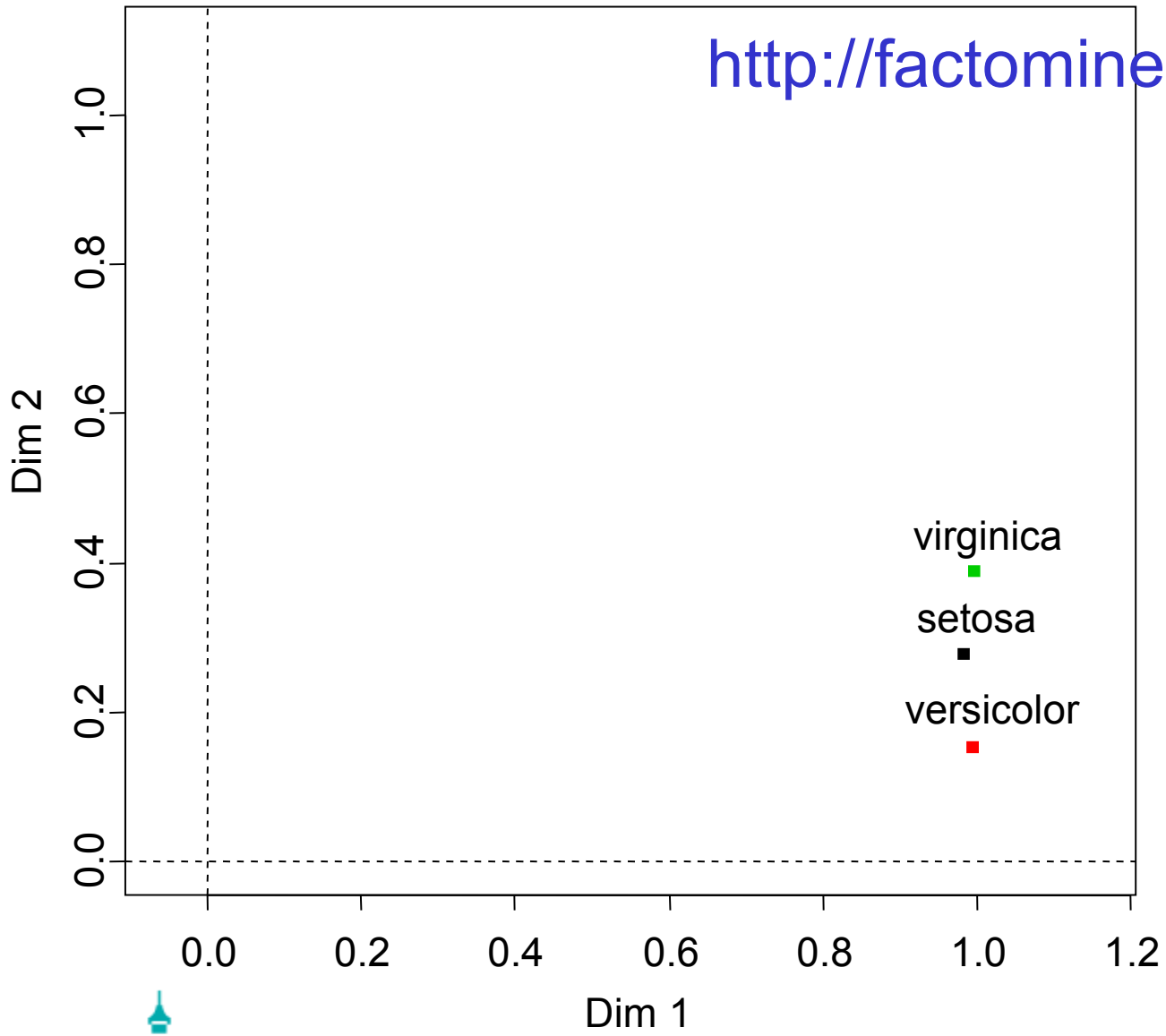
MFA example: representation of the variables



AGROcampus RENNES

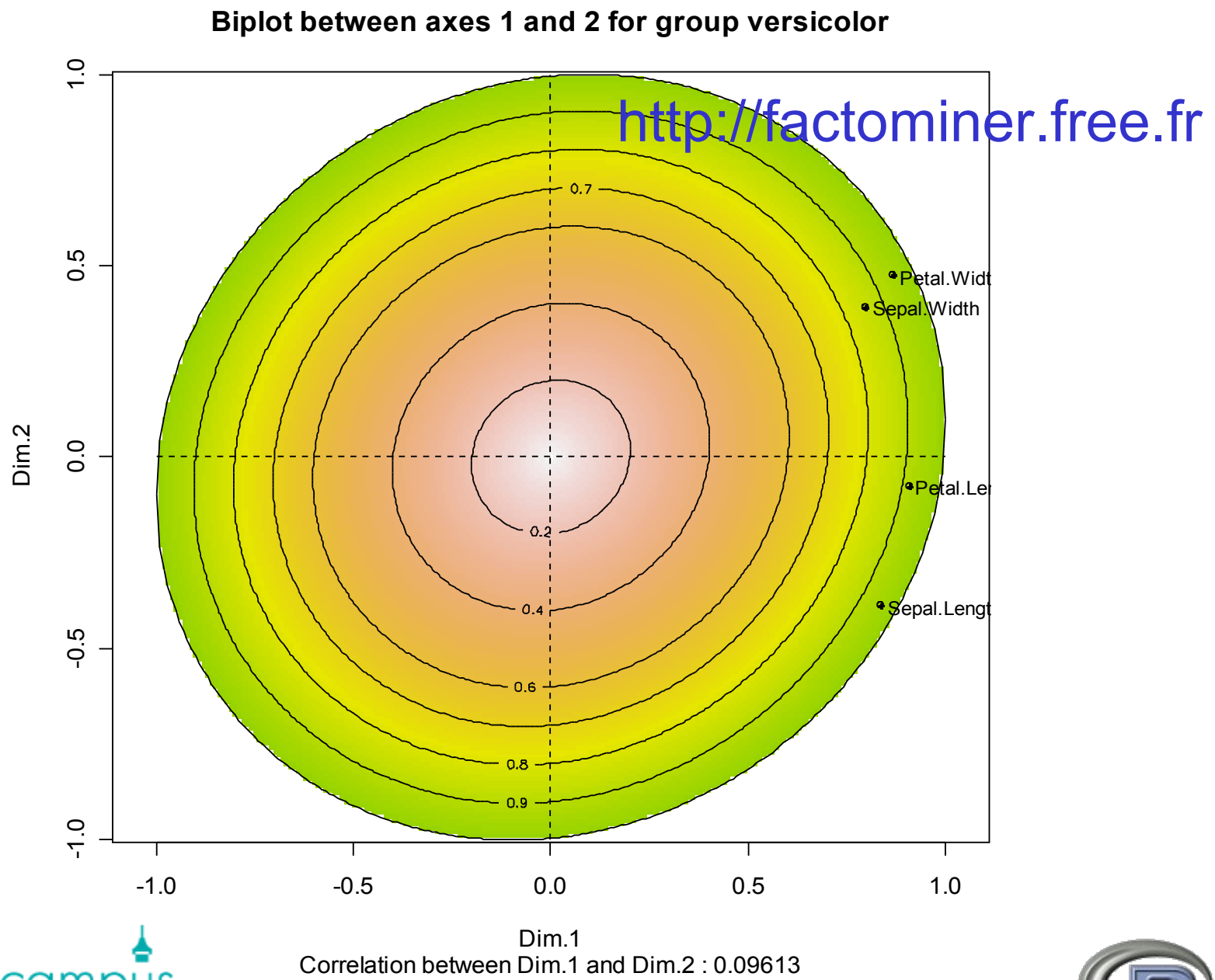
FACTOMINER

Projection of the groups



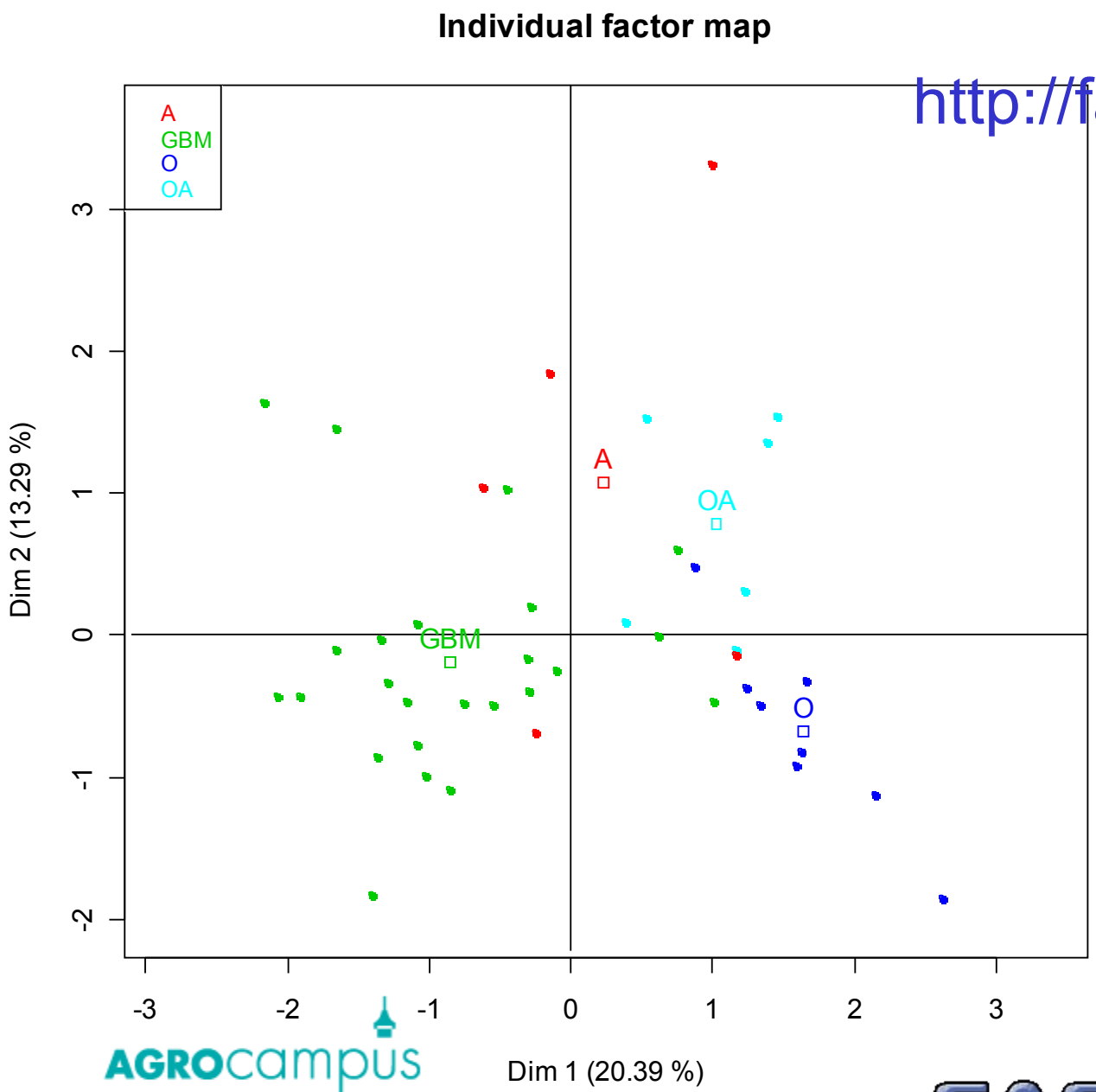
AGROcampus RENNES

FACTOMINER



AGROcampus RENNES

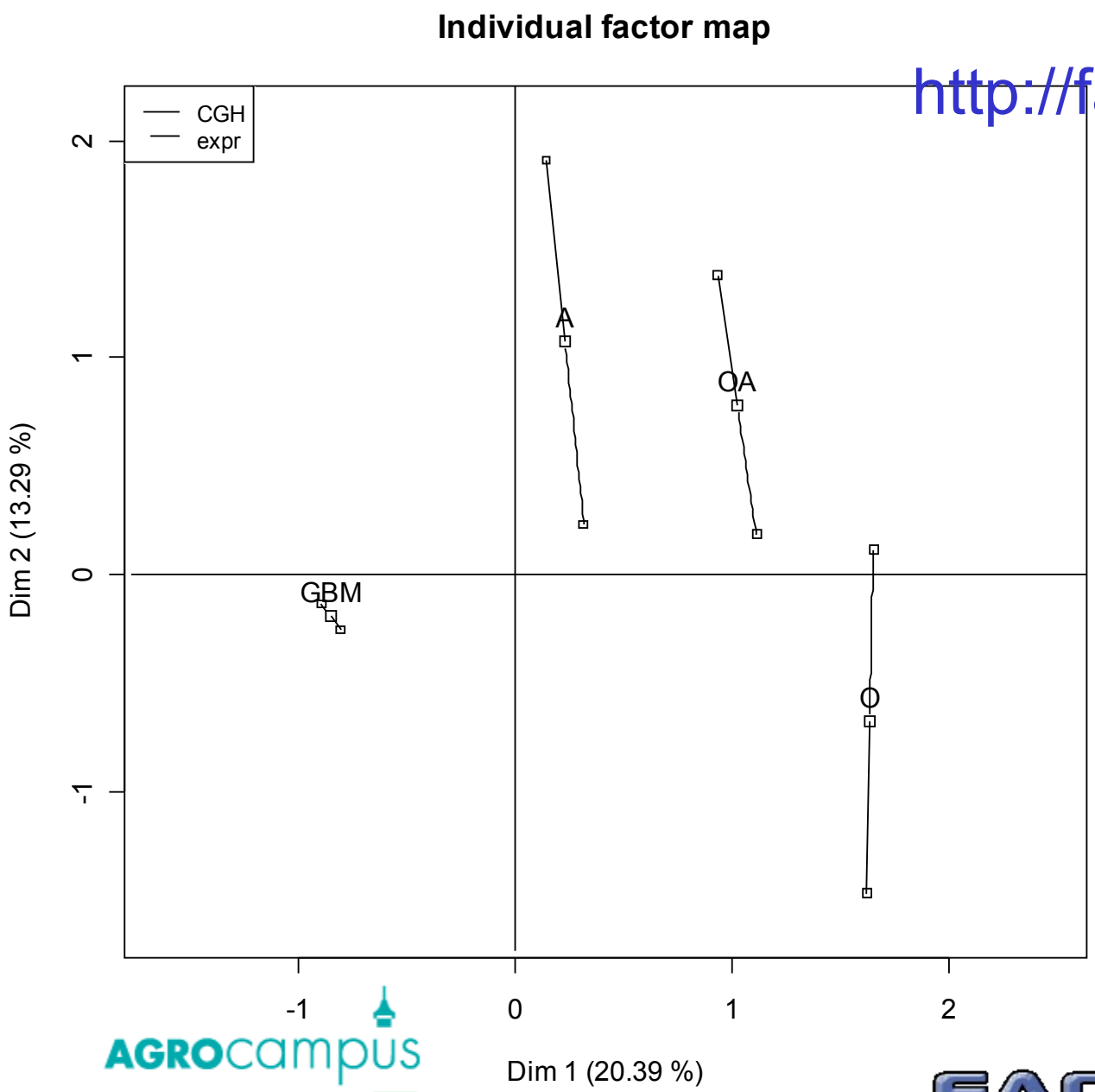
FACTOMINER



http://factominer.free.fr

AGROcampus RENNES

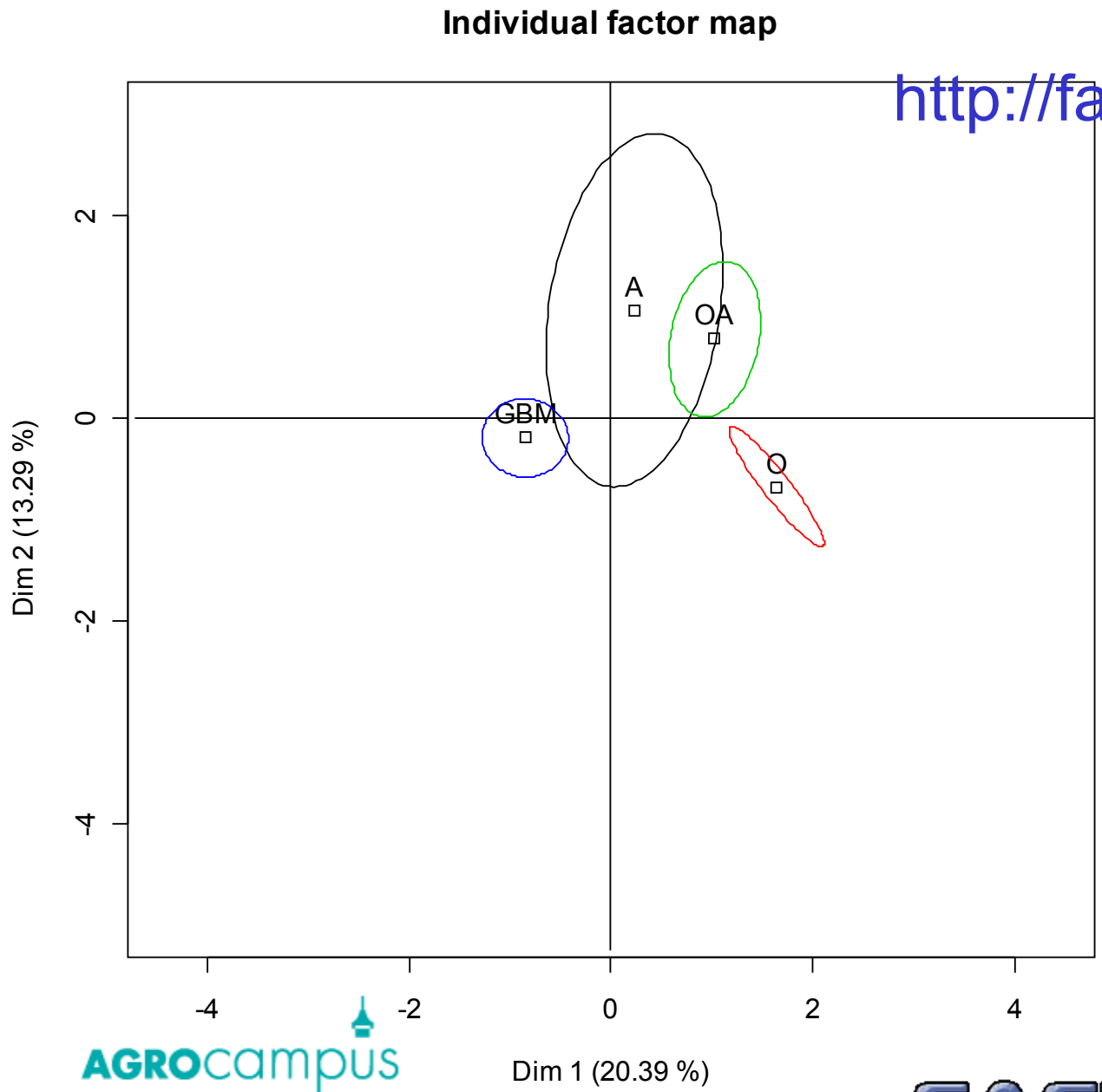
FACTOMINER



http://factominer.free.fr

AGROcampus RENNES

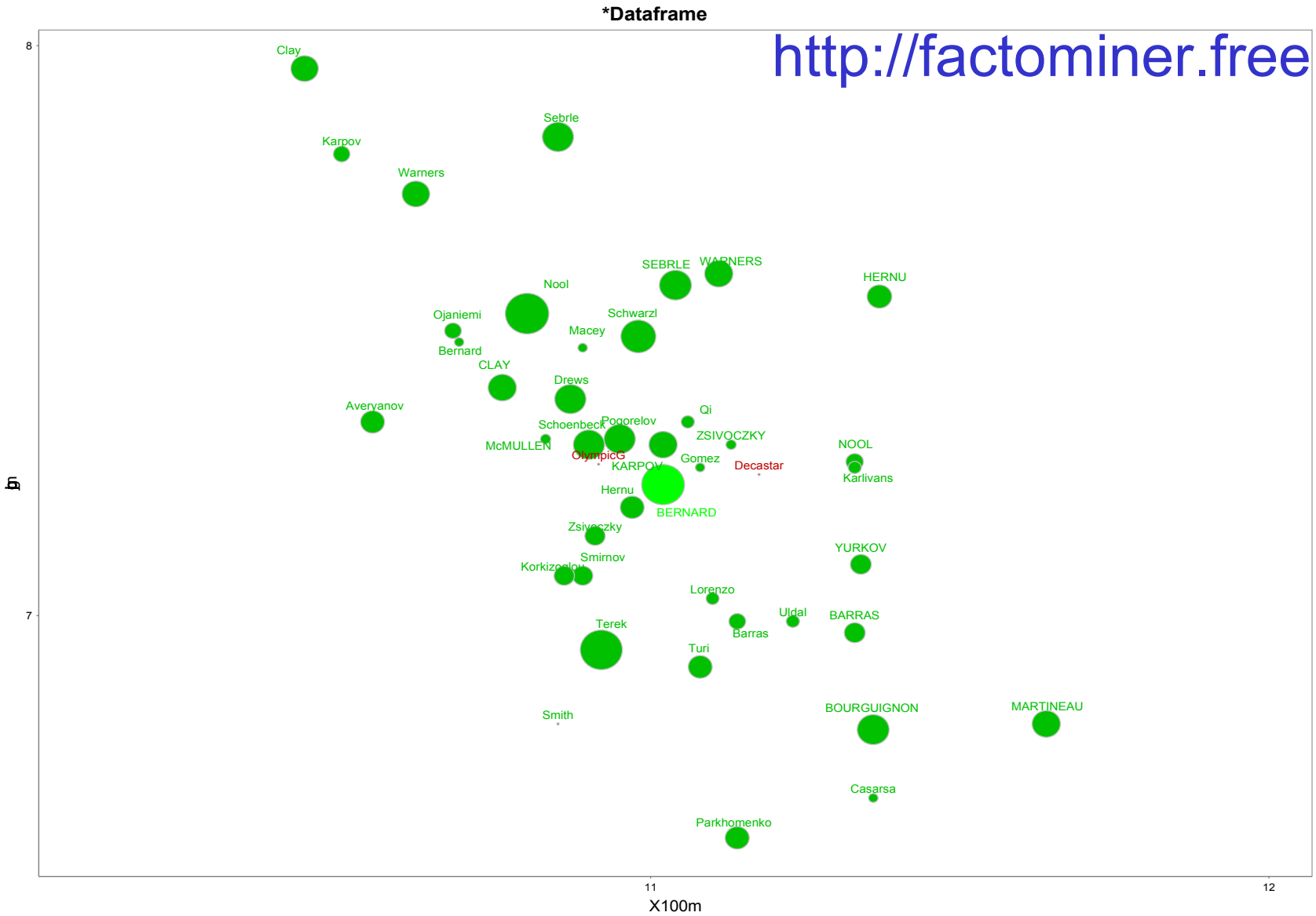
FACTOMINER



http://factominer.free.fr

AGROcampus RENNES

FACTOMINER



AGROcampus RENNES

FACTOMINER

The FactoMineR team is nearly all the time ready to improve the package

AGROcampus RENNES

FACTOMINER