Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

EnQuireR: analyzing questionnaires with R

M. Cadoret
Agrocampus Ouest

G. Fournier
Agrocampus Ouest

F. Le Poder
Agrocampus Ouest

J. Bouche
Agrocampus Ouest

O. Fournier
Agrocampus Ouest

S. Lê
Agrocampus Ouest

Abstract

The **EnQuireR** package focuses on categorical variables and provides many tools to automate the survey process. It includes both univariate and multivariate data analyses comprising Multiple Correspondence Analysis (MCA), clustering analysis and semantic marking. The package also offers an easier view of the results by the automatic generation of a *.pdf* report and of a *Beamer* type presentation *via* the use of *Sweave*. An example is used throughout this article to illustrate the package functionalities on a real dataset.

Keywords: categorical variables, univariate data analysis, multivariate data analysis, clustering, semantic marking.

1. Introduction

In many fields (psychology, consumer market studies, politics, science education, food science and so on), the use of categorical variables is commonplace when making surveys. The objective of the **EnQuireR** package is twofold: first, to automate the analysis of questionnaires by mean of a predefined sequence of univariate and multivariate analyses dedicated to categorical data; second, to automate the writing of *.pdf* reports and of *Beamer* type presentations *via* the use of *Sweave* (Leisch (2002)).

The **EnQuireR** package targets a wide range of users from students to scientists, and is designed to be accessible to anyone with a basic knowledge of statistics.

We will first present the dataset used throughout the paper then the methods implemented in the package and finally we'll make some comments on the structure of the documents automatically generated.

2. An illustrative example

We will develop an example throughout this paper using the "*tea*" dataset included in the package. The data used here refer to a survey carried out on a sample of 300 tea consumers. They were asked about how they consume tea (usage and attitude), the image they have of the product (perception) and some descriptive information:

- Which type of tea do you consume the most often? (black, green or perfumed tea)
- How do you consume tea the most often? (pure, with lemon, with milk, other)
- On which form do you consume tea? (tea bag, bulk, tea bag plus bulk)
- Do you sweeten your tea? (yes, no)
- Where do you buy tea? (large-scale retail stores, specialized shop, both)
- How much do you spend for tea? (downmarket, supermarket, famous brands, upscale, variable)
- How often do you drink tea? (more than twice a day, once a day, three to six times a week, once to twice a week)
- Six questions concerning the place where they consume tea: at home? at work? in a teahouse? at friends? at the restaurant? in a pub? (yes or no for each question)
- Six questions concerning the moment when they taste tea: at breakfast? in the afternoon? in the evening? after lunch? after dinner? anytime? (yes or no for each question)
- Concerning the image they have of the product, twelve questions were asked: do you associate tea with escape or exoticism? Do you associate tea with spirituality? Is tea healthy? Is tea diuretic? Do you associate tea with conviviality? Does tea prevent from iron absorption? Do you think tea is feminine? Do you think tea is refined? Do you think tea is slimming? Is tea stimulating? Is tea relaxing? Does tea have an effect on health? (yes or no for each question)
- Concerning descriptive variables, consumers were asked about their age, their gender, their socio-economic group (worker, employee, middle-class executive, top executive, other active, non active, student) and if they were used to practice a sport regularly (yes or no).

3. Univariate and bivariate analyses

3.1. Graphical representations

When dealing with categorical variables, one of the first questions of interest to get a global view of the dataset is how individuals are spread among categories. To do so, the commonly used graphical tool is the bar plot. From this perspective, the **EnQuireR** package provides

two "bar plot" functions: a first one, *ENbarplot*, designed for an analysis of the variables one by one; a second one, *XvsYbarplot*, designed for an analysis of the variables one by one but conditionally to a second one.

The *ENbarplot* function creates an horizontal bar plot with colour shading from the smallest to the highest frequency. The bar plot can be sorted by alphabetical order (cf. Figure 1) when a lot of categories are to be displayed and when the user wants to have a look at a particular one easily and quickly.

```
> data(tea)
> ENbarplot(tea, 20, numr=1, numc=1)
```

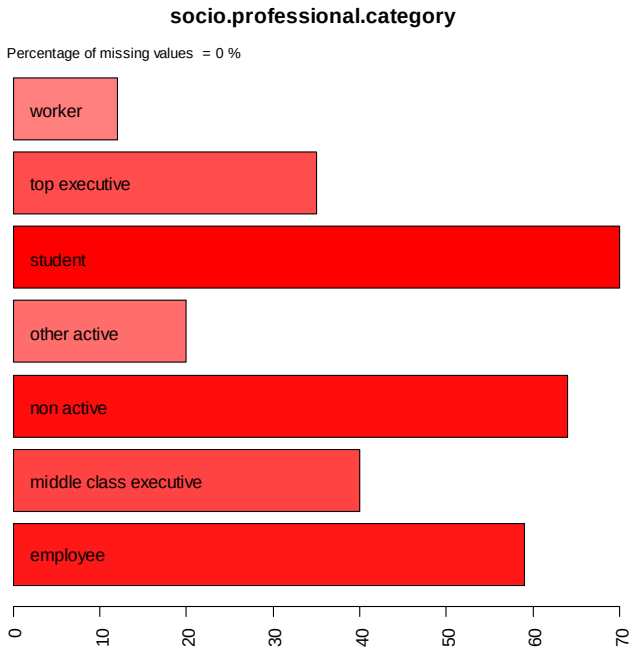

Figure 1: bar plot sorted by alphabetical order

It can also be sorted by frequency (cf. Figure 2) when the user wants to have a quick overview of the categories that are the least or the most chosen.

```
> data(tea)
> ENbarplot(tea, 20, spl=TRUE, numr=1, numc=1)
```

For each variable, the number of missing values is counted, then the percentage of missing values is printed at the top of the graphical device.

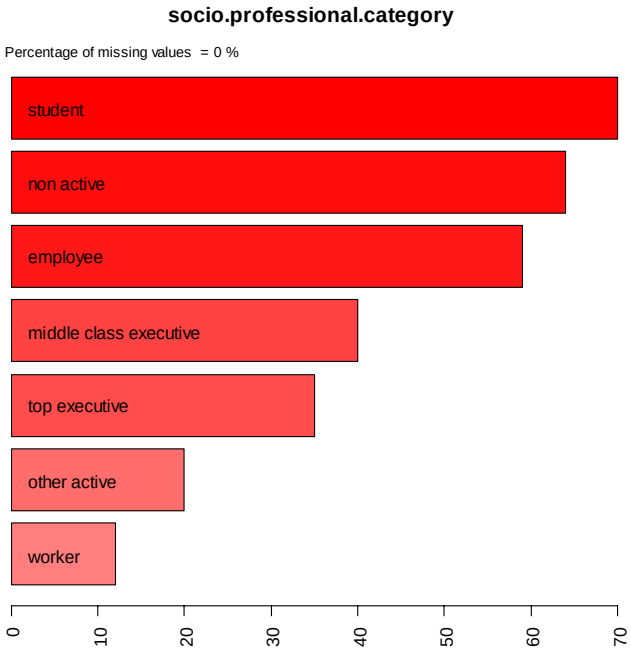

Figure 2: bar plot sorted by frequency

As for the *XvsYbarplot* function, it creates a bar plot for a given variable X conditionally to another one Y. In other words, it provides a bar plot of X for each subpopulation induced by the categories of Y. Each category of Y has its own colour shading from the smallest to the highest frequency of X (cf. Figure 3): this functionality has its importance in a multivariate analysis context when describing the typology issued from clustering techniques.

```
> data(tea)
> XvsYbarplot("socio.professional.category","sex",tea, legend.text=TRUE)
```

In this example, we're interested in the representativity of our sample by visualizing the socio-economic groups (first variable) according to gender (second variable).

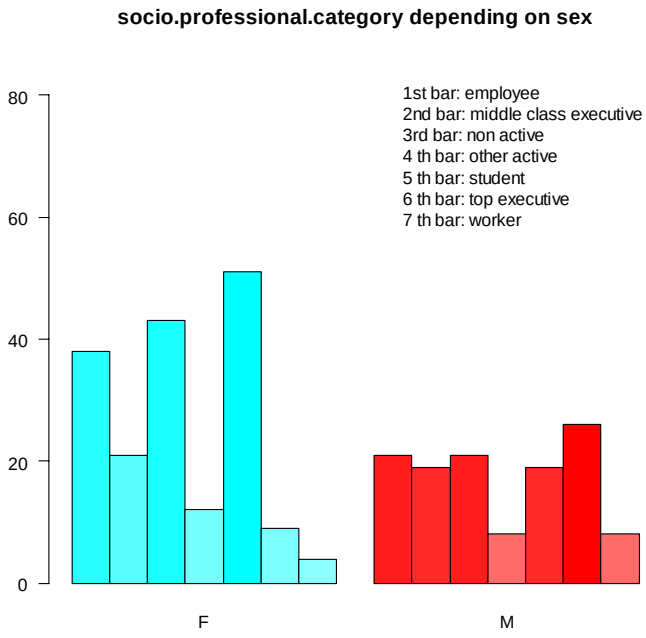

Figure 3: the social and economic category according to the gender

But beyond this graphical study, it is quite frequent to confront blocks of questions (*e.g.* in a marketing context, we often want to confront usage and attitude towards a product on the one hand, perception on the other hand) and to look for the variables of one block that are *significantly* linked to those of the other block. For that purpose, we developed the *chisq.desc* function described in the following section.

3.2. Inferential aspects

The *chisq.desc* function takes as input two groups of quantitative variables and returns a

description in terms of dependence of each variable of the first group with respect to each variable of the second one.

First, from a global point of view, the *chisq.desc* function perfoms a χ^2 test for each couple of variables from one and the other group and returns a table of distances from the situation of independence. In Figure 4, the rows represent the variables of the first group and the columns those of the second one, cells are coloured in light red when the *p-value* associated with the test is significant. The hypotheses associated with the test are: H_0: the two variables are independent; H_1: the two variables are linked. Under H_0, the test statistic

$$\chi^2_{obs} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

follows a χ^2 distribution with $(I-1)(J-1)$ degrees of freedom, where I (*resp.* J) denotes the number of possible answers for the question of the first group (*resp.* second group), n_{ij} denotes the number of people that have chosen answers i and j, $n_{i.}$ (*resp.* $n_{.j}$) denotes the number of people that have chosen answer i (*resp.* j) and n the number of people that have been surveyed.

Second, from a more local point of view, in other words for each χ^2 test, the *chisq.desc* function returns:

- a table of contributions, where the rows represent the categories of a given variable of the first group, the columns those of another given variable of the second group, each cell the contribution of the couple of categories to the χ^2 distance between the two variables;
- a table of *p-values*, where the rows represent the categories of a given variable of the first group, the columns those of another given variable of the second group, each cell the significance of the couple of categories in the χ^2 distance between the two variables.

For the contributions table, the sum of the contributions over the cells is equal to the χ^2 distance between the two variables. As for the *p-values* table, it is obtained from a test based on the hypergeometric distribution criterion (Lebart *et al.* (2006) and Lê *et al.* (2008)). This test compares the percentages of individuals possessing the category i among those possessing the category j ($n_{ij}/n_{.j}$) to the percentage of individuals possessing the category i within the whole population ($n_{i.}/n$).

Let's say we want to compare some "*usage and attitude*" type of variables with some "*perception*" type of variables.

```
> data(tea)
> chisq.desc(tea,13:17,31:35)
```

According to Figure 4, we can say that the way people drink tea (variable *shape*) is not independent from the way they perceive tea (variables *refined* and *slimming*).

But looking at variables two by two has its own limits, when for instance the number of questions is relatively large (which happens quite often in consumer surveys) or when the objective is to get a typology of surveyed people described by their answers to the questions they've been asked, hence the need for using a multivariate strategy.

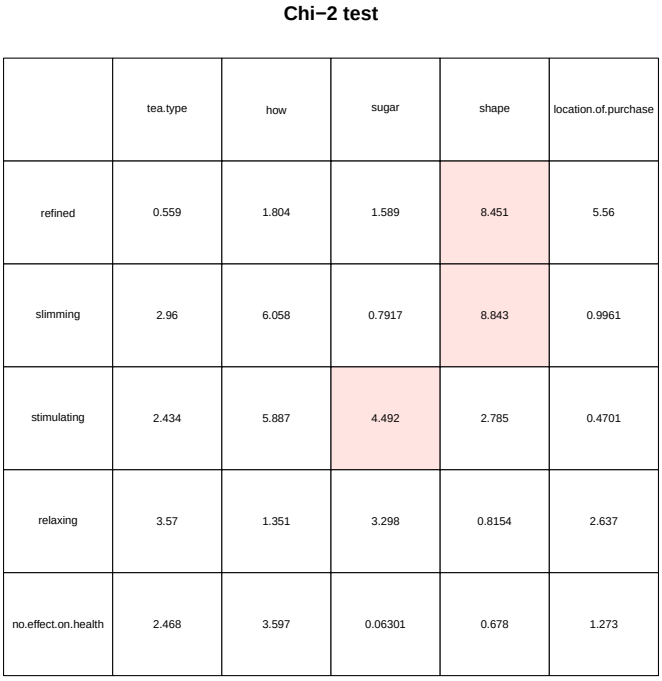

Figure 4: usage and attitude variables *versus* image

4. Multivariate analysis

As the main issue of questionnaires is to obtain a typology of surveyed people based on the answers they have provided, the core of our methodology is:

- to use Multiple Correspondence Analysis (MCA) to obtain a representation of the individuals based on the components issued from MCA;
- to use Hierarchical Ascending Classification (HAC) on the components.

Indeed, MCA acts as a change of basis if all components are kept, but instead of working with the categorical variables directly it's now possible to work with the components that are quantitative and therefore to apply a usual HAC on those components.

4.1. Principal dimensions of variability; Multiple Correspondence Analysis

In our context, an individual is characterized by the answers he has given, in other words by the categories he possesses (a question is considered as a categorical variable). Two individuals are all the more close as they have answered the same way, in other words as they have in common a great number of categories. In MCA, the distance between two individuals i and l is given by the following formula:

$$d^2(i,l) = \sum_k \frac{IJ}{I_k}(\frac{x_{ik}}{J} - \frac{x_{lk}}{J})^2 = \frac{1}{J}\sum_k \frac{I}{I_k}(x_{ik} - x_{lk})^2,$$

where x_{ik} is equal to 1 if the individual i has taken the category k and 0 otherwise, I is the total number of individuals, I_k is the number of individuals who have taken the category k and J is the number of variables.

The expression $(x_{ik} - x_{lk})^2$ is either equal to 0 or 1. The distance $d^2(i,l)$ grows alongside with the number of different categories for both individuals. A category k takes part in this distance formula with a weight equal to $\frac{I}{I_k}$ which corresponds to the inverse of the category's frequency. This means that individuals having a rare category are separated from all other individuals.

The routine used to perform MCA in the **EnQuireR** package is the one implemented in the **FactoMineR** package which provides a representation of the individuals and of the answers to the questions. This routine has been enhanced in a questionnaire context where missing values are frequently encountered and where a large number of people may be surveyed. Indeed, it commonly happens that people forget or simply refuse to answer some questions for privacy reasons for instance. For that purpose, we implemented in the *missmca* function the algorithm proposed by Brigitte Escofier in "Traitement des variables incomplètes en analyse des correspondances multiples" (Escofier (1990)) that takes into account missing values in MCA.

As for the graphical representation of both individuals and categories provided by the *MCA* function of the **FactoMineR** package it has been improved when figures are too crowded. When a great number of people are surveyed, too many points are to be represented (either surveyed people or categories associated with questions) which prevents the user from easy interpretation and decision-making (cf. Figure 5).

```
> res <- MCA(tea[,1:18])
```

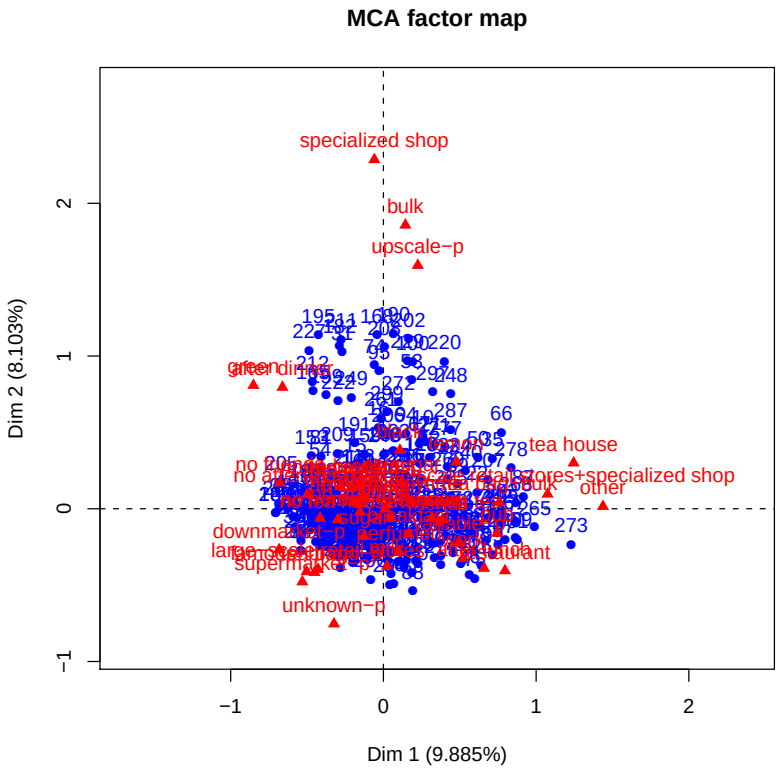

Figure 5: usual graphical representation of the individuals and the modalities provided by *MCA*

To make easier the interpretation of such graphs, we propose two points of view via the functions *ENlisib* and *ENdensity* that are complementary. The function *ENlisib()* proceeds in two steps. The first step consists in selecting the individuals with the highest coordinates (depending on a given threshold) and the categories that are significantly linked to the components provided by MCA; the second step consists in selecting randomly a given percentage of what remains. The individuals (cf. Figure 6) and the categories (cf. Figure 7) that are selected during the two steps are then plotted.

```
> res <- MCA(tea[,1:18])
> ENlisib(res,0.05,50,c(1,2))
```

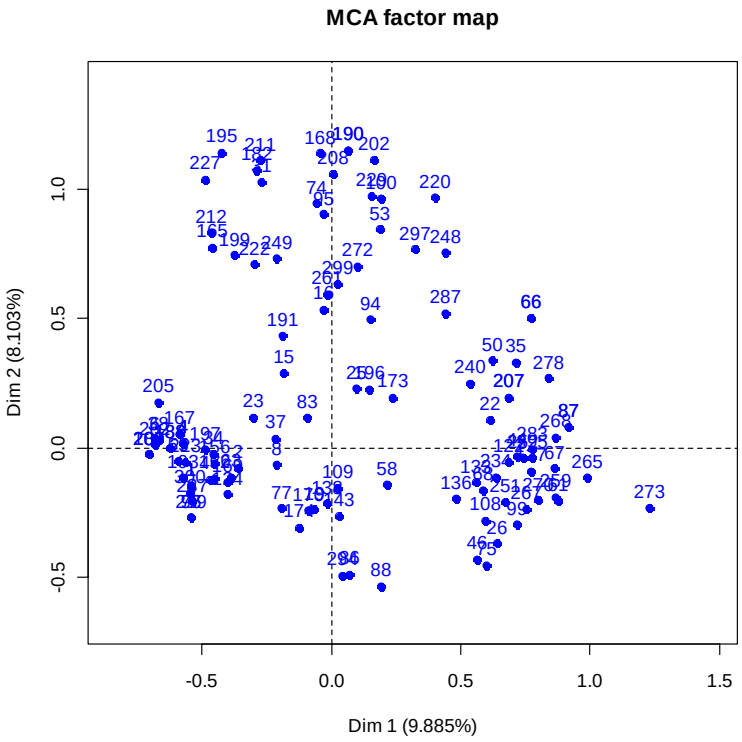

Figure 6: *ENlisib* applied on the individuals

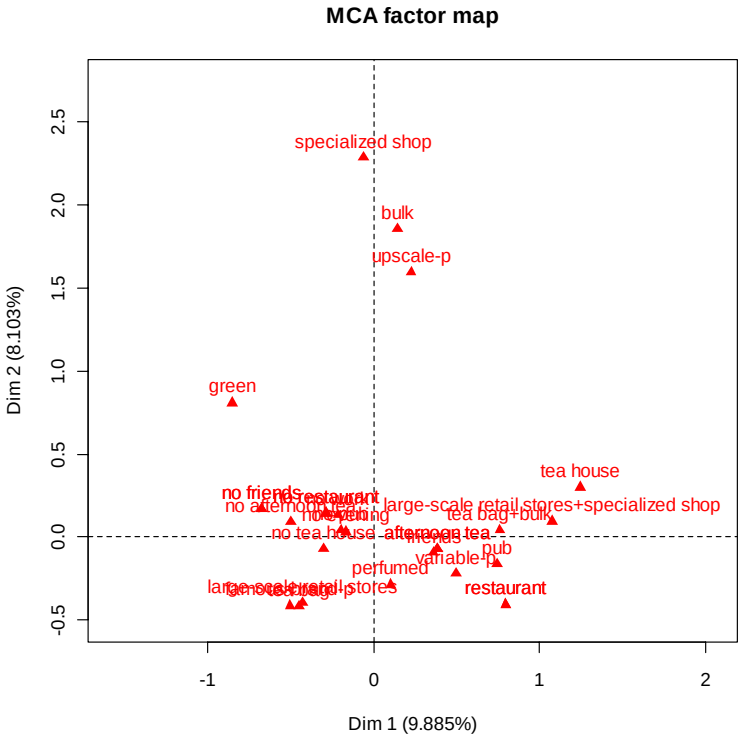

Figure 7: *ENlisib* applied on the categories

Another way to circumvent the problem of the superposition of the individuals due to their number is to use density curbs via the function *ENdensity* (cf. Figure 8). This function provides a visualization of the shape of the scatter plot function of the distribution of the coordinates of the individuals.

```
> res <- MCA(tea[,1:18])
> ENdensity(res)
```

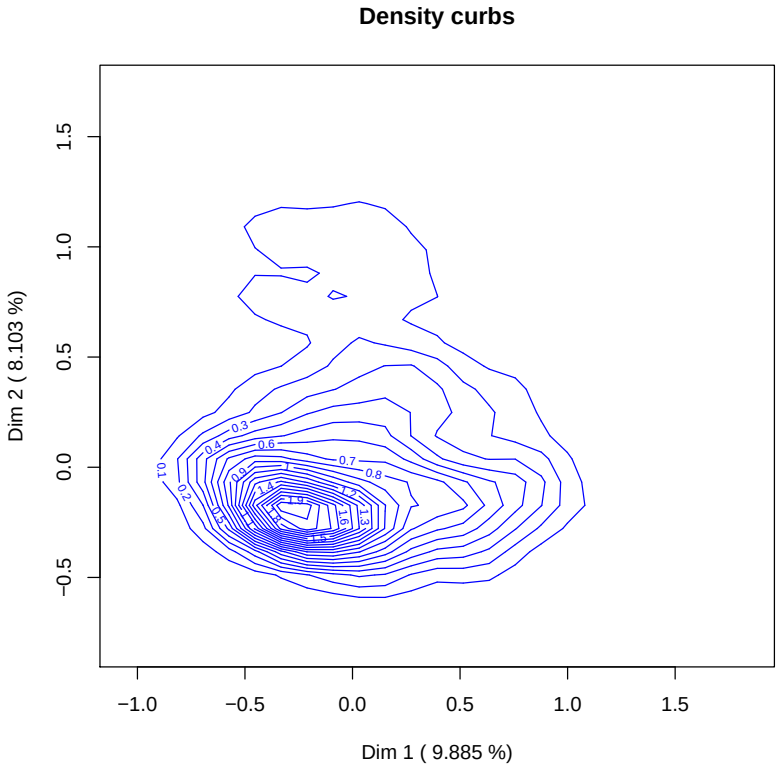

Figure 8: density curbs representing the scatter plot of the individuals

Remark It is frequent with MCA that percentages of inertia associated with components are quite low. As the inertia can be interpreted as the information associated with a component it is important to check whether the percentages really reveal a meaningful structure of the dataset. To do so, the function *p_inertia* compares thoses percentages to the ones that would be obtained by using datasets generated at random. In that view, we set the null hypothesis H_0 where the persons surveyed have answered at random and in an independent way to each question. To get such datasets, we generate as many multinomial variables as there are questions in the dataset. Each multinomial variable possesses as many categories as there are possible answers to the question it is associated with; the proportions being obtained on the basis of the frequencies observed. We then perform MCA on each of the datasets and keep the percentages of inertia in order to get their distribution under the null hypothesis and to

test whether the original structure is meaningful or not.

```
> data(tea)
> p_inertia(tea[,1:18])
```

```
         % of variance p-value
Dim.1         9.884961       0
Dim.2         8.103115       0
Plan.1-2     17.988076       0
```

Table 1: *p-values* associated with the test of the significance of the dimensions

In our example, *p-values* associated with the test of the significance of the dimensions are all null (cf. Table 1). We can conclude that even if the percentages of inertia of our dataset "tea" are quite low, they are significantly different from what we would obtain with datasets being the result of chance.

4.2. Clustering on the individuals

The ENMCA function

One of the finality of a survey is of course to draw up a typology of the surveyed people. Following a multivariate analysis such as MCA, the logical second step consists in performing unsupervised classification.

Lots of algorithm perform cluster analysis on numeric variables but it is quite rare to find algorithms which perform cluster analysis on categorical variables directly. The principle of the *ENMCA* function consists in performing MCA on the categorical variables, then performing unsupervised classification on the principal components obtained by MCA which correspond to the coordinates of the individuals after the change of basis due to MCA. Once the appropriate number of clusters is chosen by the user (cf. Figure 9), the *ENMCA* function provides outputs directly related to the clusters, i.e. that allow an easy understanding of each group of surveyed people. This function corresponds to the following sequence:

1. A MCA is performed. If there are no missing values in the dataset, MCA is performed, otherwise *missmca* is used.

2. A Hierarchical Ascendant Classification using Ward's criterion to aggregate clusters is performed on the factorial axes of the MCA.

3. The user chooses the proper number of clusters (cf. Figure 9) by simply clicking on the dendrogram.

4. The main outputs include a variable created from the clustering process that indicates the cluster individuals belong to, a plot where the individuals are coloured depending on their value for this variable, a description of each group (cf. Figure 10).

```
> data(tea)
> ENMCA(tea[,1:18])
```

Remark The choice of the Ward's criterion to aggregate clusters is to be put in relation with MCA itself which principle is to maximize the inertia of the cloud of the individuals: indeed, Ward's criterion consists in aggregating clusters by minimizing the inertia within the cluster thus obtained and fits perfectly MCA's objective.

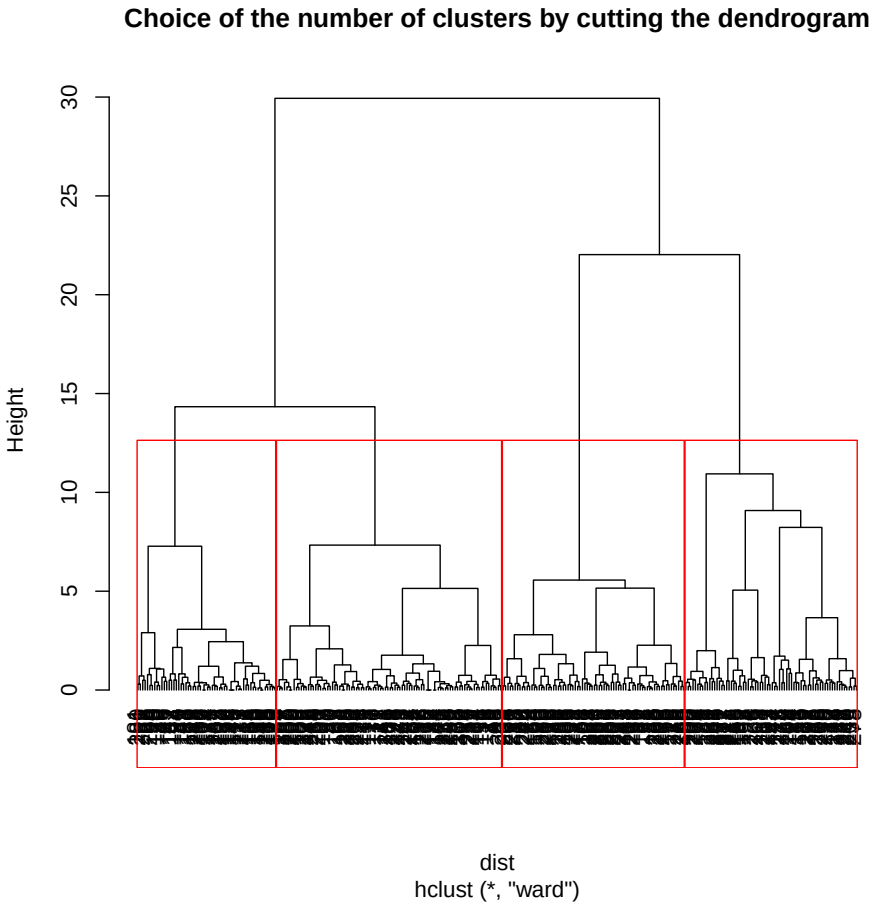

Figure 9: the user has to define a number of clusters

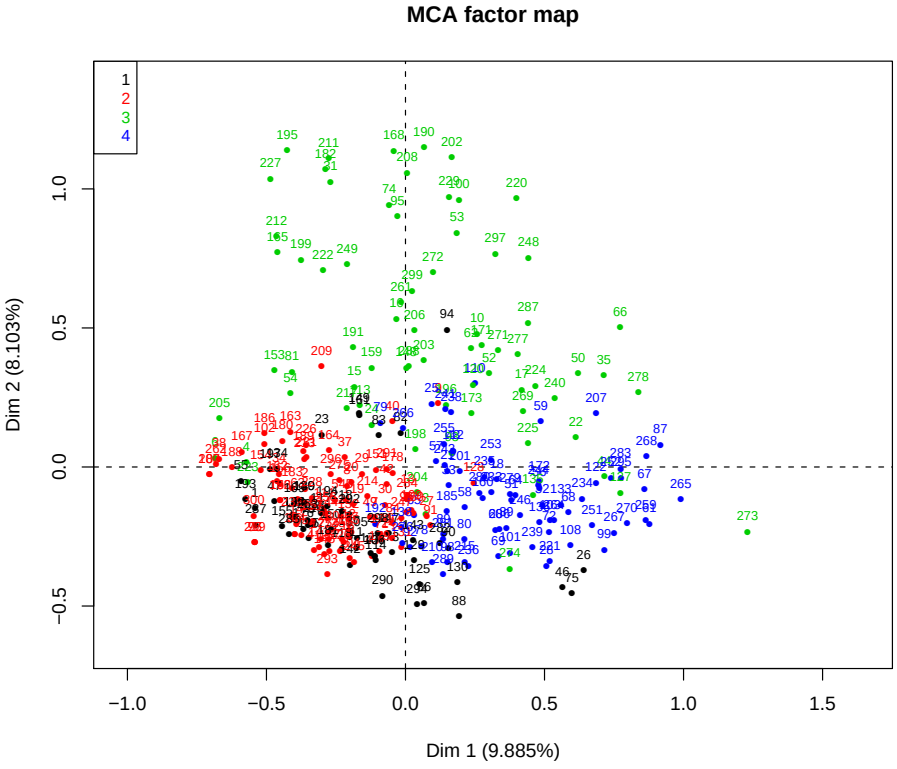

Figure 10: MCA factor map, for each cluster a specific color is displayed

Remark Once the clusters obtained it can be interesting to study their stability by visualizing confidence ellipses around their barycenters. Those ellipses answer partially to the following question: where would the barycenters of the clusters be if we were working on another population (*i.e.* obtained by resampling the original population)? To get those ellipses, the *ENellipse* function proceeds the following way:

- perform MCA on the dataset and get the coordinates of the individuals on the components;
- get the coordinates of the clusters' barycenters;
- use resampling techniques such as bootstrap to pick individuals at random;
- recalculate the coordinates of the barycenters;
- repeat the two previous steps 500 times (for instance);
- draw ellipses around the resampled barycenters (cf. Figure 11).

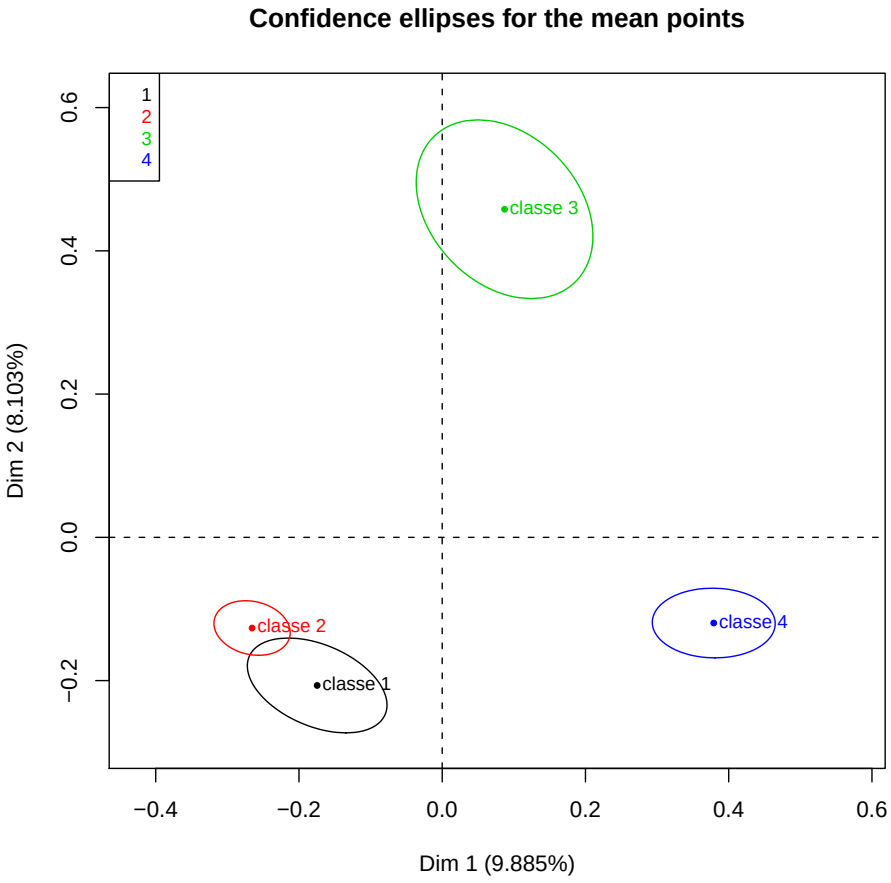

Figure 11: confidence ellipses around barycenters of clusters

Semantic marking

The automatic description of a subpopulation induced by one category of a given categorical variable can be easily obtained with the *catdes* function of the **FactoMineR** package (Lebart *et al.* (2006) and Lê *et al.* (2008)). For instance, the following code line provides a description of the *Gender* variable first (global point of view), then on the subpopulations induced by the categories *Female* on the one hand, *Male* on the other hand (local point of view):

```
> res.catdes <- catdes(tea, num.var=19)

$test.chi2
                            p.value df
feminine                    0.0000  1
afternoon.tea               0.0000  1
conviviality                0.0001  1
socio.professional.category 0.0001  6
sugar                       0.0004  1
age                         0.0008  4
tea.house                   0.0043  1
frequency                   0.0114  3
after.dinner                0.0119  1
sport                       0.0274  1
pub                         0.0279  1
location.of.purchase        0.0411  2

$category
$category$F
                                      Cla/Mod Mod/Cla Global p.value v.test
feminine=feminine                       80.62   58.43  43.00    0.00   6.55
afternoon.tea=afternoon tea             71.60   67.98  56.33    0.00   4.81
conviviality=conviviality               64.88   88.20  80.67    0.00   3.82
sugar=no sugar                          69.03   60.11  51.67    0.00   3.43
frequency=more than 2/day               70.08   50.00  42.33    0.00   3.15
tea.house=tea house                     75.86   24.72  19.33    0.01   2.76
age=15-24                               70.65   36.52  30.67    0.01   2.55
socio.professional.category=student     72.86   28.65  23.33    0.01   2.53
after.dinner=no after dinner            61.29   96.07  93.00    0.02   2.27
sport=no sportive                       66.94   45.51  40.33    0.04   2.09
pub=pub                                 71.43   25.28  21.00    0.04   2.08
location.of.purchase=specialized shop   40.00    6.74  10.00    0.04  -2.06
pub=no pub                              56.12   74.72  79.00    0.04  -2.08
sport=sportive                          54.19   54.49  59.67    0.04  -2.09
after.dinner=after dinner               33.33    3.93   7.00    0.02  -2.27
tea.house=no tea house                  55.37   75.28  80.67    0.01  -2.76
sugar=sugar                             48.97   39.89  48.33    0.00  -3.43
conviviality=no conviviality            36.21   11.80  19.33    0.00  -3.82
age=25-34                               37.68   14.61  23.00    0.00  -4.01
socio.professional.category=top executive 25.71  5.06  11.67    0.00  -4.11
afternoon.tea=no afternoon tea          43.51   32.02  43.67    0.00  -4.81
feminine=no feminine                    43.27   41.57  57.00    0.00  -6.55

$category$M
```

	Cla/Mod	Mod/Cla	Global	p.value	v.test
feminine=no feminine	56.73	79.51	57.00	0.00	6.55
afternoon.tea=no afternoon tea	56.49	60.66	43.67	0.00	4.81
socio.professional.category=top executive	74.29	21.31	11.67	0.00	4.11
age=25-34	62.32	35.25	23.00	0.00	4.01
conviviality=no conviviality	63.79	30.33	19.33	0.00	3.82
sugar=sugar	51.03	60.66	48.33	0.00	3.43
tea.house=no tea house	44.63	88.52	80.67	0.01	2.76
after.dinner=after dinner	66.67	11.48	7.00	0.02	2.27
sport=sportive	45.81	67.21	59.67	0.04	2.09
pub=no pub	43.88	85.25	79.00	0.04	2.08
location.of.purchase=specialized shop	60.00	14.75	10.00	0.04	2.06
pub=pub	28.57	14.75	21.00	0.04	-2.08
sport=no sportive	33.06	32.79	40.33	0.04	-2.09
after.dinner=no after dinner	38.71	88.52	93.00	0.02	-2.27
socio.professional.category=student	27.14	15.57	23.33	0.01	-2.53
age=15-24	29.35	22.13	30.67	0.01	-2.55
tea.house=tea house	24.14	11.48	19.33	0.01	-2.76
frequency=more than 2/day	29.92	31.15	42.33	0.00	-3.15
sugar=no sugar	30.97	39.34	51.67	0.00	-3.43
conviviality=conviviality	35.12	69.67	80.67	0.00	-3.82
afternoon.tea=afternoon tea	28.40	39.34	56.33	0.00	-4.81
feminine=feminine	19.38	20.49	43.00	0.00	-6.55

From a global point of view, we can say that gender is significantly linked to variables related to the image (*feminine*, *conviviality*) as well as to the usage (*afternoon.tea*, *sugar*, *tea.house*, *frequency*, *after.dinner*, *pub*, *location.of.purchase*) people have of tea.

From a local point of view, we can say for instance that:

- 58.43 % of the women polled think that tea is feminine (43.00 % in the whole population and the difference between the two proportions is significant)
- 67.98 % of the women polled drink tea in the afternoon (56.33 % in the whole population and the difference between the two proportions is significant)
- 88.20 % of the women polled associate tea with conviviality (80.67 % in the whole population and the difference between the two proportions is significant)
- 60.11 % of the women polled drink tea with no sugar (51.67 % in the whole population and the difference between the two proportions is significant)
- 50.00 % of the women polled drink tea more than twice per day (42.33 % in the whole population and the difference between the two proportions is significant)

The idea of the semantic marking (HoTu *et al.* (1988) and Gettler-Summa (2000)) is to generalize this kind of description to associations of categories (in the **EnQuireR** package to couples and triplets of categories). To do so, we will apply the *catdes* function to the original variables first, then to couples and triplets of variables, as explained below.

- from the original dataset, apply the *catdes* function and select the ten variables that are the most linked to the variable of interest (S_1)

- generate *new* categorical variables by crossing variables from S_1
- from the *new* categorical variables, apply the *catdes* function and select the ten variables that are the most linked to the variable of interest (S_2)
- generate *new* categorical variables by crossing variables from S_1 and S_2
- from the *new* categorical variables, apply the *catdes* function and select the ten variables that are the most linked to the variable of interest (S_3)

For instance, the following code line provides a description of the *Gender* variable first (global point of view), then on the subpopulations induced by the categories *Female* on the one hand, *Male* on the other hand (local point of view) at three different levels:

```
> res.semantic <- ENmarking(tea,19)
```

```
$F
$F$lev_1
$F$lev_1$marking
      Variable(s)
 [1,] "feminine"
 [2,] "afternoon.tea"
 [3,] "socio.professional.category"
 [4,] "age"
 [5,] "conviviality"
 [6,] "sugar"
 [7,] "frequency"
 [8,] "tea.house"
 [9,] "after.dinner"
[10,] "sport"

$F$lev_1$catdes
                                        Cla/Mod Mod/Cla Global p.value V-test
feminine=feminine                          0.81    0.58   0.43    0.00   6.55
afternoon.tea=afternoon tea                0.72    0.68   0.56    0.00   4.81
conviviality=conviviality                  0.65    0.88   0.81    0.00   3.82
sugar=no sugar                             0.69    0.60   0.52    0.00   3.43
frequency=more than 2/day                  0.70    0.50   0.42    0.00   3.15
tea.house=tea house                        0.76    0.25   0.19    0.00   2.76
age=15-24                                  0.71    0.37   0.31    0.01   2.55
socio.professional.category=student        0.73    0.29   0.23    0.01   2.53
after.dinner=no after dinner               0.61    0.96   0.93    0.01   2.27
sport=no sportive                          0.67    0.46   0.40    0.02   2.09
frequency=1/day                            0.52    0.28   0.32    0.04  -1.73
sport=sportive                             0.54    0.54   0.60    0.02  -2.09
after.dinner=after dinner                  0.33    0.04   0.07    0.01  -2.27
tea.house=no tea house                     0.55    0.75   0.81    0.00  -2.76
sugar=sugar                                0.49    0.40   0.48    0.00  -3.43
conviviality=no conviviality               0.36    0.12   0.19    0.00  -3.82
age=25-34                                  0.38    0.15   0.23    0.00  -4.01
socio.professional.category=top executive  0.26    0.05   0.12    0.00  -4.11
afternoon.tea=no afternoon tea             0.44    0.32   0.44    0.00  -4.81
```

```
feminine=no feminine                    0.43    0.42   0.57    0.00  -6.55


$F$lev_2
$F$lev_2$marking
      Pair(s)
 [1,] "feminine_tea.house"
 [2,] "feminine_afternoon.tea"
 [3,] "feminine_conviviality"
 [4,] "feminine_after.dinner"
 [5,] "feminine_sugar"
 [6,] "feminine_frequency"
 [7,] "afternoon.tea_conviviality"
 [8,] "feminine_sport"
 [9,] "afternoon.tea_tea.house"
[10,] "afternoon.tea_sugar"

$F$lev_2$catdes
                                                Cla/Mod Mod/Cla Global p.value V-test
feminine_conviviality=feminine_conviviality        0.83    0.53   0.38    0.00   6.71
feminine_after.dinner=feminine_no after dinner     0.81    0.57   0.41    0.00   6.59
feminine_afternoon.tea=feminine_afternoon tea      0.87    0.38   0.26    0.00   5.99
feminine_frequency=feminine_more than 2/day        0.89    0.33   0.22    0.00   5.75
afternoon.tea_conviviality=afternoon tea_conviviality 0.74 0.62   0.50    0.00   5.24
feminine_tea.house=feminine_no tea house           0.80    0.44   0.32    0.00   5.15
feminine_sugar=feminine_no sugar                   0.84    0.33   0.23    0.00   4.92
feminine_sport=feminine_no sportive                0.83    0.30   0.22    0.00   4.44
afternoon.tea_sugar=afternoon tea_no sugar         0.77    0.42   0.32    0.00   4.35
feminine_sport=feminine_sportive                   0.78    0.28   0.21    0.00   3.40
afternoon.tea_tea.house=afternoon tea_tea house    0.79    0.21   0.16    0.00   2.98
feminine_sugar=feminine_sugar                      0.76    0.25   0.20    0.00   2.87
afternoon.tea_tea.house=afternoon tea_no tea house 0.69    0.47   0.40    0.00   2.58
feminine_tea.house=feminine_tea house              0.81    0.15   0.11    0.01   2.57
feminine_frequency=feminine_3 to 6/week            0.92    0.06   0.04    0.02   2.15
feminine_frequency=feminine_1/day                  0.75    0.13   0.11    0.04   1.75
feminine_afternoon.tea=feminine_no afternoon tea   0.71    0.20   0.17    0.05   1.65
```

```
feminine_sport=no feminine_no sportive                          0.48    0.15    0.19    0.04  -1.72
feminine_frequency=no feminine_3 to 6/week                      0.36    0.04    0.07    0.02  -2.04
afternoon.tea_conviviality=no afternoon tea_conviviality        0.49    0.26    0.31    0.01  -2.20
feminine_after.dinner=no feminine_after dinner                  0.25    0.02    0.05    0.00  -2.60
feminine_conviviality=no feminine_conviviality                  0.48    0.35    0.43    0.00  -3.20
feminine_frequency=no feminine_1/day                            0.40    0.14    0.21    0.00  -3.40
afternoon.tea_conviviality=no afternoon tea_no conviviality     0.29    0.06    0.13    0.00  -3.88
feminine_conviviality=no feminine_no conviviality               0.28    0.07    0.14    0.00  -4.34
afternoon.tea_sugar=no afternoon tea_sugar                      0.34    0.14    0.24    0.00  -4.86
afternoon.tea_tea.house=no afternoon tea_no tea house           0.42    0.29    0.40    0.00  -4.87
feminine_sport=no feminine_sportive                             0.41    0.26    0.38    0.00  -5.01
feminine_after.dinner=no feminine_no after dinner               0.45    0.39    0.52    0.00  -5.10
feminine_sugar=no feminine_sugar                                0.30    0.15    0.29    0.00  -6.37
feminine_afternoon.tea=no feminine_no afternoon tea             0.26    0.12    0.27    0.00  -6.92
feminine_tea.house=no feminine_no tea house                     0.39    0.31    0.48    0.00  -7.05


$F$lev_3
$F$lev_3$marking
      Triplet(s)
 [1,] "feminine_sugar_tea.house"
 [2,] "feminine_afternoon.tea_tea.house"
 [3,] "feminine_conviviality_after.dinner"
 [4,] "feminine_tea.house_after.dinner"
 [5,] "feminine_afternoon.tea_after.dinner"
 [6,] "feminine_afternoon.tea_sugar"
 [7,] "feminine_frequency_after.dinner"
 [8,] "feminine_tea.house_sport"
 [9,] "feminine_afternoon.tea_sport"
[10,] "feminine_conviviality_tea.house"

$F$lev_3$catdes
                                                                Cla/Mod Mod/Cla Global p.value V-test
feminine_conviviality_after.dinner=feminine_conviviality_no after dinner          0.83    0.52   0.37    0.00   6.44
feminine_afternoon.tea_after.dinner=feminine_afternoon tea_no after dinner        0.87    0.37   0.25    0.00   5.80
feminine_frequency_after.dinner=feminine_more than 2/day_no after dinner          0.89    0.33   0.22    0.00   5.75
```

feminine_conviviality_tea.house=feminine_conviviality_no tea house	0.85	0.40	0.28	0.00	5.64
feminine_tea.house_after.dinner=feminine_no tea house_no after dinner	0.82	0.43	0.31	0.00	5.33
feminine_afternoon.tea_tea.house=feminine_afternoon tea_no tea house	0.91	0.27	0.18	0.00	5.32
feminine_sugar_tea.house=feminine_no sugar_no tea house	0.86	0.25	0.17	0.00	4.37
feminine_afternoon.tea_sport=feminine_afternoon tea_sportive	0.89	0.19	0.12	0.00	4.03
feminine_afternoon.tea_sugar=feminine_afternoon tea_no sugar	0.86	0.21	0.15	0.00	3.99
feminine_tea.house_sport=feminine_no tea house_no sportive	0.85	0.22	0.15	0.00	3.83
feminine_afternoon.tea_sugar=feminine_afternoon tea_sugar	0.88	0.17	0.11	0.00	3.68
feminine_afternoon.tea_sport=feminine_afternoon tea_no sportive	0.85	0.20	0.14	0.00	3.66
feminine_tea.house_sport=feminine_no tea house_sportive	0.76	0.22	0.17	0.00	2.64
feminine_tea.house_after.dinner=feminine_tea house_no after dinner	0.81	0.14	0.10	0.01	2.43
feminine_conviviality_tea.house=feminine_conviviality_tea house	0.80	0.13	0.10	0.01	2.30
feminine_afternoon.tea_sugar=feminine_no afternoon tea_no sugar	0.81	0.12	0.09	0.01	2.19
feminine_frequency_after.dinner=feminine_3 to 6/week_no after dinner	0.92	0.06	0.04	0.02	2.15
feminine_sugar_tea.house=feminine_sugar_no tea house	0.74	0.19	0.15	0.02	2.06
feminine_afternoon.tea_tea.house=feminine_afternoon tea_tea house	0.80	0.11	0.08	0.02	2.04
feminine_afternoon.tea_after.dinner=feminine_no afternoon tea_no after dinner	0.73	0.20	0.16	0.03	1.96
feminine_afternoon.tea_sport=feminine_no afternoon tea_no sportive	0.79	0.11	0.08	0.03	1.89
feminine_afternoon.tea_tea.house=no feminine_afternoon tea_tea house	0.78	0.10	0.08	0.04	1.74
feminine_tea.house_sport=feminine_tea house_sportive	0.85	0.06	0.04	0.05	1.66
feminine_sugar_tea.house=feminine_sugar_tea house	0.85	0.06	0.04	0.05	1.66
feminine_conviviality_after.dinner=no feminine_no conviviality_after dinner	0.17	0.01	0.02	0.04	-1.72
feminine_tea.house_sport=no feminine_no tea house_no sportive	0.45	0.12	0.16	0.02	-2.08
feminine_afternoon.tea_sugar=no feminine_afternoon tea_sugar	0.42	0.09	0.13	0.02	-2.12
feminine_frequency_after.dinner=no feminine_1/day_no after dinner	0.45	0.14	0.19	0.01	-2.32
feminine_tea.house_after.dinner=no feminine_no tea house_after dinner	0.27	0.02	0.05	0.01	-2.36
feminine_conviviality_after.dinner=no feminine_conviviality_no after dinner	0.50	0.33	0.39	0.01	-2.53
feminine_afternoon.tea_sugar=no feminine_no afternoon tea_no sugar	0.34	0.06	0.11	0.00	-2.83
feminine_afternoon.tea_sport=no feminine_no afternoon tea_no sportive	0.29	0.04	0.08	0.00	-2.90
feminine_frequency_after.dinner=no feminine_1/day_after dinner	0.00	0.00	0.02	0.00	-2.94
feminine_afternoon.tea_after.dinner=no feminine_no afternoon tea_after dinner	0.15	0.01	0.04	0.00	-3.02
feminine_conviviality_after.dinner=no feminine_no conviviality_no after dinner	0.30	0.06	0.12	0.00	-3.71
feminine_conviviality_tea.house=no feminine_conviviality_no tea house	0.43	0.25	0.35	0.00	-3.99
feminine_conviviality_tea.house=no feminine_no conviviality_no tea house	0.27	0.06	0.14	0.00	-4.37
feminine_afternoon.tea_sport=no feminine_no afternoon tea_sportive	0.25	0.08	0.19	0.00	-5.64
feminine_tea.house_sport=no feminine_no tea house_sportive	0.35	0.19	0.32	0.00	-5.65

```
feminine_afternoon.tea_after.dinner=no feminine_no afternoon tea_no after dinner  0.28  0.11  0.22  0.00  -5.70
feminine_afternoon.tea_sugar=no feminine_no afternoon tea_sugar                    0.21  0.06  0.16  0.00  -5.78
feminine_tea.house_after.dinner=no feminine_no tea house_no after dinner           0.40  0.29  0.43  0.00  -5.88
feminine_afternoon.tea_tea.house=no feminine_no afternoon tea_no tea house         0.27  0.12  0.26  0.00  -6.51
feminine_sugar_tea.house=no feminine_sugar_no tea house                            0.25  0.11  0.25  0.00  -6.93
```

5. Reporting

The **EnQuireR** package provides two kinds of reports automatically generated using *Sweave* that are put in the *EnQuireR* folder created in the working directory.

A first "detailed" and exhaustive report gathers all the different results (numerical and graphical outputs) provided by the functions *ENbarplot*, *chisq.desc* and *ENMCA* in a *.pdf* document.

From a univariate point of view, the following lines will create a *.pdf* document in which each categorical variable is displayed in a separate page.

```
> data(tea)
> res.enbarplot <- ENbarplot(tea,c(18,20,23:27),report=TRUE)
```

From a bivariate point of view, the following lines will create a *.pdf* document which indicates whether a given set of categorical variables depend on another given set of categorical variables. For each variable of the first set, a range of statistical indicators (Pearson's Chi square coefficient, contribution) is computed in order to evaluate the relationship with the variables of the other set.

```
> data(tea)
> res.chisq.desc <- chisq.desc(tea,c(1,2),c(3,4),report=TRUE)
```

Finally, from a multivariate point of view, the following lines will create a *.pdf* document which focuses on how the individuals and categories are spread in the correspondence map and which provides a description of the two first axes that consists of a list of meaningful categories sorted according to their p-values. The report concerns also the partition in clusters. For each cluster, the report provides a list of categories that best describe the cluster; it provides also several graphical representations of the clusters issued from the *ENlisib* and the *ENellipses* functions in particular: level curbs and confidence ellipses around the barycenters of the groups, for instance.

```
> data(tea)
> res.enmca <- ENMCA(tea[,1:18],report=TRUE)
```

The report is structured the following way:

- Quick overview of the questionnaire
- Multivariate exploration of the questionnaire (graphical representation of the questionnaire, highlights on the two principal axes of variability)
- Typology of the individuals (choice of the number of clusters, simultaneous comparison of the clusters, description of each cluster)

In addition, a more succinct report is created, made of slides only: the *Beamer* type report. This report concerns the same information presented previously from a multivariate point of view but more synthetically and is therefore available with the *ENMCA* function only. It is

divided into two main parts. First, the multivariate exploration of the questionnaire; second, the typology of the individuals.

The first part aims to answer to the following questions:

- How is my dataset "structured"?
- How does my dataset look like?
- How can the main axes of variability be interpreted?

The second part aims to answer the following questions:

- How many groups are there in my dataset?
- How can the groups be displayed?
- How different are the groups?
- How can the groups be described?

Therefore, the user disposes of two complementary documents which are both very useful at the crucial moment of the dataset interpretation. Of course, results from the automatically generated reports cannot replace the user's expertise.

Technically, the use of this functionality requires to install MikTeX and TeXnicCenter. To facilitate the installation for the user, the *.sty* and *.cls* files needed for the generation of the reports have been integrated to the package; hence its size.

6. Concluding remarks

This paper presented the **EnQuireR** package designed for the studies of categorical variables. Our contribution to the study of surveys does not consist in a collection of tools but more in the way those tools are articulated and integrated in a logical sequence of statistical analyses. This logical sequence naturally leads to the idea of automatic reports provided by our package. Some further works related to this package could include a methodology allowing the comparison of different partitions on the same individuals. Moreover, in order to provide a friendly interface in the **Rcmdr** environment (Fox *et al.* (2008)), we are working on a **Rcmdr** plug-in.

References

Escofier B (1990). "Traitement des variables incomplètes en analyse des correspondances multiples." *Modulad*, **5**, 1–12.

Fox J, with contributions from Michael Ash, Boye T, Calza S, Chang A, Grosjean P, Heiberger R, Kerns GJ, Lancelot R, Lesnoff M, Messad S, Maechler M, Murdoch D, Neuwirth E, Putler D, Ripley B, Ristic M, , Wolf P (2008). *Rcmdr: R Commander*. R package version 1.3-15, URL http://www.r-project.org,http://socserv.socsci.mcmaster.ca/jfox/Misc/Rcmdr/.

Gettler-Summa M (2000). "Marking and Generalization by Symbolic Objects in the Symbolic Official Data Analysis." *Ed. Kiers, H.A.L., Rasson, J.P., Groenen, P.J.F. et al. : Proc. of IFCS'00, Namur, Belgium.*

HoTu B, Diday E, Gettler-Summa M (1988). *Generating rules for expert system from observations.*

Lê S, Josse J, Husson F (2008). "FactoMineR: an R package for multivariate analysis." *Journal of Statistical Software*, **25 (1)**, 1–18.

Lebart L, Piron M, Morineau A (2006). *Statistique exploratoire multidimensionnelle*. Dunod.

Leisch F (2002). "Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis." In W Härdle, B Rönz (eds.), "Compstat 2002 — Proceedings in Computational Statistics," pp. 575–580. Physica Verlag, Heidelberg. ISBN 3-7908-1517-9, URL http://www.stat.uni-muenchen.de/~leisch/Sweave.

Affiliation:

Sébastien Lê
Agrocampus Rennes
UMR CNRS 6625
65 rue de Saint-Brieuc
35042 Rennes Cedex, France
E-mail: sebastien.le@agrocampus-ouest.fr
URL: http://www.agrocampus-ouest.fr/math/le/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd