Multiple Factor Analysis for Contingency Tables in the FactoMineR Package

by Belchin Kostov, Mónica Bécue-Bertaut and François Husson

Abstract We present multiple factor analysis for contingency tables (MFACT) and its implementation in the **FactoMineR** package. This method, through an option of the `MFA` function, allows us to deal with multiple contingency or frequency tables, in addition to the categorical and quantitative multiple tables already considered in previous versions of the package. Thanks to this revised function, either a multiple contingency table or a mixed multiple table integrating quantitative, categorical and frequency data can be tackled.

The **FactoMineR** package (Lê et al., 2008; Husson et al., 2011) offers the most commonly used principal component methods: principal component analysis (PCA), correspondence analysis (CA; Benzécri, 1973), multiple correspondence analysis (MCA; Lebart et al., 2006) and multiple factor analysis (MFA; Escofier and Pagès, 2008). Detailed presentations of these methods enriched by numerous examples can be consulted at the website `http://factominer.free.fr/`.

An extension of the `MFA` function that considers contingency or frequency tables as proposed by Bécue-Bertaut and Pagès (2004, 2008) is detailed in this article.

First, an example is presented in order to motivate the approach. Next, the mortality data used to illustrate the method are introduced. Then we briefly describe multiple factor analysis (MFA) and present the principles of its extension to contingency tables. A real example on mortality data illustrates the handling of the `MFA` function to analyse these multiple tables and, finally, conclusions are presented.

Motivation

MFACT was initially conceived to deal with international surveys including open-ended questions answered in different languages such as the survey designed by Akuto (1992) for comparing the dietary habits in Tokyo, Paris and New York. Three samples of individuals, each of about thousand respondents, were taken in the three cities and answered a classical questionnaire and also the open-ended question: “Which dishes do you like and eat often?”. So three groups of free-text answers in three different languages have to be analyzed.

If a unique sample (for example, New York) were analysed, the table crossing gender × age and words could be considered to see how the dietary habits change with age and gender. The age variable is divided into clusters such as under 30, $30-50$, and over 50.

For this kind of table, correspondence analysis is a reference method (Lebart et al., 1998). Through a superimposed map of the categories and word-dishes, CA allows us to visualize

1. The similarities between the gender × age categories: two categories are closer if they often eat the same dishes
2. The similarities between dishes: two dishes are closer if they are often chosen by the same categories;
3. The attractions (respectively, the repulsions) between categories and dishes: a category is at the pseudo-centroid of the word-dishes used in the answers belonging to it; a word-dish is at the pseudo-centroid of the categories that mention it.

In order to analyse and to compare several samples (New York, Tokyo and Paris), we can consider the frequency table that juxtaposes row-wise the three contingency tables. This large frequency table can then be analysed through MFACT (Pagès and Bécue-Bertaut, 2006). In this case, we can answer the following questions:

1. Which are the gender × age categories that are globally similar (through all the samples)?
2. Which are the similarities between dishes?
3. Which dishes are associated with which gender × age categories?

These three questions are the same as when only one sample was considered, but considering now all the samples. In order to simultaneously consider the three samples, MFACT centres each table on its own centroid and balances the influence of each sample in the global analysis to prevent one table playing a dominating role. Moreover, MFACT allows us to compare the three samples in a common framework:

- The gender × age category structures can be represented as issued from every sample thanks to the partial representations: it indicates how similar or different are the category structures from one city to another, from the point of view of the dish preferences.
- The methodology also allows us to study the similarities between cities: two cities are closer if the attractions (respectively, the repulsions) between categories and dishes induced by each table are similar.

The type of result provided by MFACT makes this method useful to globally analyze and compare several contingency tables. It is able to deal with any number of tables and could be used to compare the dietary habits in a large number of cities.

The dataset

In this paper, we illustrate the method through data on "the causes of mortality" provided by the "Centre d'épidémiologie sur les causes médicales de décès – Cépidc" (http://www.cepidc.vesinet.inserm.fr). These data concern the mortality data in France from 1979 to 2006. The registration of the causes of mortality is mainly motivated by prevention: identifying and quantifying the causes of mortality in order to take action to reduce avoidable mortality. Therefore, these data are used to produce one of the health indicators most commonly used.

Our aim is to study the evolution of the causes of mortality from 1979 to 2006. For each year, a crossed table contains 62 mortality causes (in rows) and age intervals (in columns). At the intersection of a row-mortality cause and a column-age interval, the counts of deaths corresponding to this cause and this age interval during this year is given. Note that the cause AIDS has been removed, since this cause did not exist in the mortality registers in 1979. Likewise, sudden infant death syndrome (SIDS) and perinatal infection, specific causes of mortality for children, have been removed.

The data are read from the **FactoMineR** package:

```
> library(FactoMineR)
> data(mortality)
```

Multiple factor analysis for contingency tables (MFACT)

Recall on multiple factor analysis

Multiple factor analysis (Escofier and Pagès, 2008) deals with a multiple table, composed of groups of either quantitative or categorical variables. MFA balances the influence of the groups on the first principal dimension by dividing the weights of the variables/columns of a group by the first eigenvalue of the separate analysis of this group (PCA or MCA depending on the type of the variables). The highest axial inertia of each group is standardized to 1.

MFA provides the classical results of principal component methods. PCA characteristics and interpretation rules are kept for the quantitative groups and those of MCA for the categorical groups. MFA offers tools for comparing the different groups such as the partial representation of the rows. This representation allows us to compare the typologies provided by each group in a common space. A graphic of the groups allows us to globally compare the groups and to evaluate if the relative positions of the rows are globally similar from one group to another. It also permits the comparison of the partial groups and assess whether they provide the same information. Another graph gives the correlations between the dimensions of the global analysis (the MFA dimensions) and each separate analysis (the PCA dimensions for quantitative groups, the MCA dimensions for categorical groups and the CA dimensions for frequency groups).

Multiple tables and notation

The multiple contingency table X_G (the causes of mortality in 1979 and 2006), of dimensions $I \times J$ (62 mortality causes × 18 age intervals in total), juxtaposes row-wise several contingency tables $X_1, ..., X_t, ..., X_T$ (one table for each year, in the example $T = 2$) of dimension $I \times J_t$ (62 mortality causes × 9 age intervals) for table t (see Figure 1). Columns can differ from one table to another (which is not the case in this example).

X_G is transformed into a proportion table (proportion of the causes of mortality for each age interval), globally computed on all the tables. Thus p_{ijt} is the proportion of row-mortality cause i $(i = 1, ..., I)$ for column-age interval j $(j = 1, ..., J_t)$ of table t $(t = 1, ..., T)$; $\sum_{ijt} p_{ijt} = 1$. The row and column margins of table X_G are respectively $p_{i\bullet\bullet} = \sum_{jt} p_{ijt}$ and $p_{\bullet jt} = \sum_{i} p_{ijt}$. The intra-table

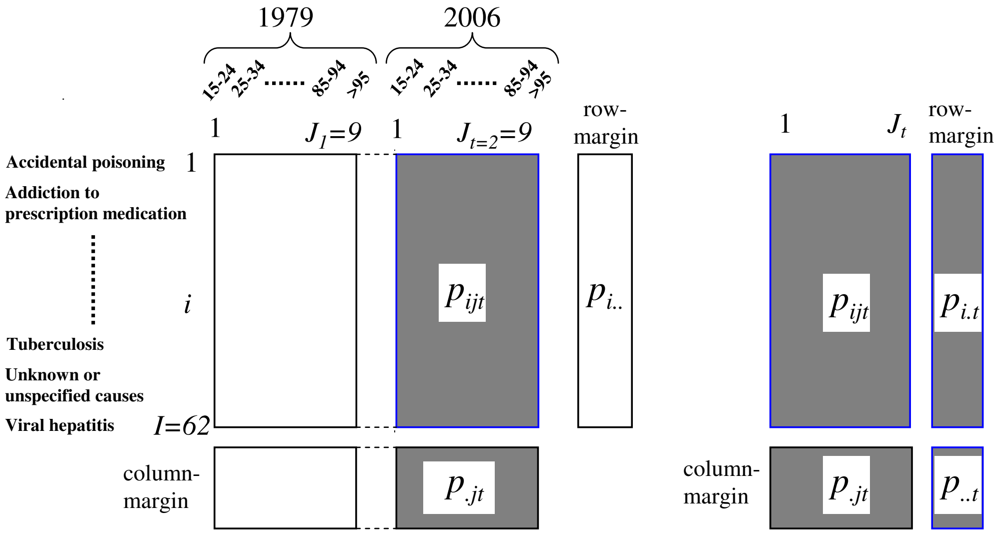

Figure 1: Multiple contingency table. T contingency tables with the same rows are juxtaposed row-wise. On the left, the global table and margins; on the right the table t and its margins. In the example, $T = 2$, $I = 62$, $J_1 = 9$, $J_2 = 9$.

independence model is considered (relations between causes of mortality and age intervals in each year). It is filled in a matrix with general term

$$\left(\frac{p_{i\bullet t}}{p_{\bullet\bullet t}}\right)\times\left(\frac{p_{\bullet jt}}{p_{\bullet\bullet t}}\right)$$

where $p_{i\bullet t} = \sum_j p_{ijt}$ is the row margin of table t (the percentage of the causes of mortality i in year t with respect to the sum of causes of mortality for all years) and $p_{\bullet\bullet t} = \sum_{ij} p_{ijt}$ is the sum of the terms of table t inside table X_G (the percentage of the sum of causes of mortality in year t with respect to the sum of causes of mortality for all years). Then table Z whose general term is the weighted residual with respect to the intra-table independence model is built. Note that Z is divided into T subtables Z_t for $t = 1, ..., T$.

$$Z = \frac{p_{ijt} - \left(\frac{p_{i\bullet t}}{p_{\bullet\bullet t}}\right) p_{\bullet jt}}{p_{i\bullet\bullet} \times p_{\bullet jt}} = \frac{1}{p_{i\bullet\bullet}}\left(\frac{p_{ijt}}{p_{\bullet jt}} - \frac{p_{i\bullet t}}{p_{\bullet\bullet t}}\right)$$

The algorithm

MFACT consists of a classical MFA applied to the multiple table Z assigning the weight $p_{i\bullet\bullet}$ to the row i ($i = 1, ..., I$) and the weight $p_{\bullet jt}$ to the column jt ($j = 1, ..., J_t$, $t = 1, ..., T$). Thus the observed proportions are compared to those corresponding to the intra-table independence model. This model neutralizes the differences between the separate average column profiles.

It is easy to verify that the weighted rows are centred as well as the weighted columns of each table t. The influence of each subtable in the global analysis is balanced in a MFA-like way. MFA classical results are obtained and CA characteristics and interpretation rules are kept.

The **FactoMineR** package (from version 1.16) includes MFACT in the `MFA` function.

Results

The analysis of one contingency table by means of CA was discussed in Husson et al. (2011). If we want to deal with two contingency tables, one of the tables could be favoured and chosen as active while the other would be considered as supplementary. However, that would lead to principal dimensions only based on the active contingency table. When both tables, that is, the mortality data of both years, have to play a balanced role to enhance the evolution of the mortality causes, MFACT turns out to be necessary.

We have used MFACT to evaluate the evolution of the mortality causes considering changes at the age profiles (between 1979 and 2006). MFACT preserves the internal structure of the relationships between causes of mortality and age intervals in each table and balances their influences on the global representation (Bécue-Bertaut and Pagès, 2004).

Implementation in FactoMineR

MFA is performed on the multiple table "causes of mortality" as follows:

```
> mfa <- MFA(mortality, group=c(9, 9), type=c("f", "f"), name.group=c("1979", "2006"))
```

In this call, only few arguments were used. The `group` argument specifies that each table has 9 columns, `type` specifies the type of the tables, here frequency tables, `name.group` gives the name of each table. By default, all the individuals are considered as active (the `ind.sup` argument is equal to `NULL`), all the groups are active (`num.group.sup` argument is equal to `NULL`) and the plots will be drawn for the first two dimensions.

Outputs

The outputs of MFA are both a series of numerical indicators (results of the separate analysis, eigenvalues, coordinates, contributions, correlations) and a series of graphics (individuals/rows, superimposed representation, variables/columns, partial axes, groups). Most of them are presented hereinafter along with an interpretation of the results.

Eigenvalues

The sequence of eigenvalues identifies two dominating dimensions that, together, account for 81.69% of the total inertia (see below). That leads to focus on the first principal plane.

```
> round(mfa$eig,3) [1:4,]
        eigenvalue   percentage cumulative \%
                    of variance  of variance
comp 1       1.790       52.420       52.420
comp 2       0.999       29.269       81.689
comp 3       0.262        7.659       89.348
comp 4       0.149        4.367       93.715
```

We recall that in MFA the first eigenvalue varies between 1 and the number of groups (two in this case). A first eigenvalue close to the maximum means that the first dimension of the separate analyses (here the CAs on each contingency table) are very similar. Thus, the value of the first eigenvalue (equal to 1.79) indicates that the first principal component is an important dispersion dimension in both tables, and is similar to the first axes of the two separate CAs. Then, we can say that the similarity between both groups justifies their simultaneous analysis and that there are enough differences to resort to a method able to highlight common and specific features.

Representation of rows and columns

Figure 2 visualizes the age intervals columns on the first principal plane. The trajectories of the age intervals of both years are drawn in different colours to ease the comparison.

```
> plot(mfa, choix="freq", invisible="ind", habillage="group")
> lines(mfa$freq$coord[1:9, 1], mfa$freq$coord[1:9, 2], col="red")
> lines(mfa$freq$coord[10:18, 1], mfa$freq$coord[10:18, 2], col="green")
```

The first dimension perfectly ranks the age intervals, opposing the youngest to the oldest. The second dimension opposes the extreme to the medium-age intervals. The homologous age intervals corresponding to 1979 and 2006 are very close. Both age trajectories are almost identical. This means that, the relationship between mortality causes and age intervals are very similar in 1979 and 2006. However, we will see hereinafter some interesting differences provided by the comparison of the groups.

In both years, the youngest intervals differ more from one another that the oldest: the distances between two consecutive age intervals are longer in the case of the youngest ages (Figure 2). That indicates that the predominant causes of mortality in young people change rapidly with age.

The visualization of the mortality causes on the principal plane shows that some causes are clearly separated from the others (Figure 3).

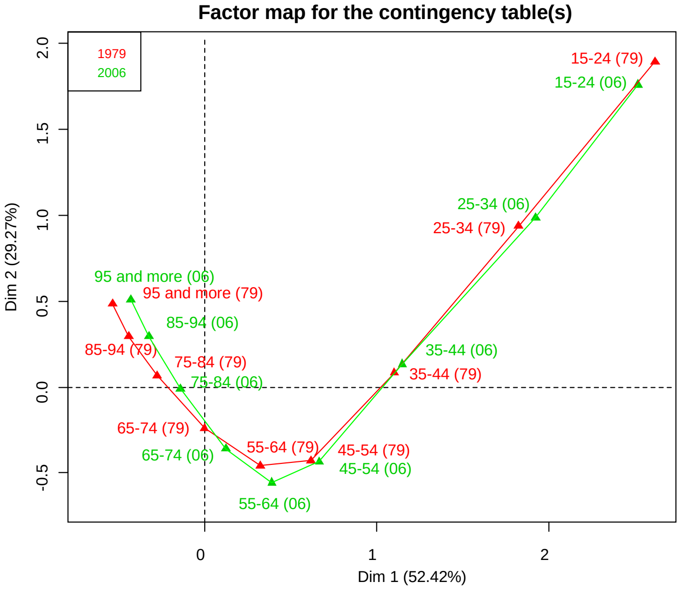

Figure 2: Trajectories of the age intervals on the first principal plane.

```
> sel <- c(2, 8:10, 15, 38, 58)
> plot(mfa, lab.ind=FALSE, habillage="group")
> text(mfa$ind$coord[sel, 1], mfa$ind$coord[sel, 2], rownames(mortality)[sel],
     pos=c(2, 2, 2, 2, 4, 2, 4))
```

This finding makes sense when related to the age intervals through the usual transition relationships. Indeed, an age interval is attracted (or repelled) by the causes that are more (or less) associated with it than if there were independence between causes and age intervals of either 1979 or 2006.

An *arch effect*, also called *Guttman effect*, is identified on the first principal plane both on rows and columns (Figures 2 and 3), pointing out that the phenomenon under study is mainly unidimensional and that both rows and columns share ranking. From the transition relationships, we can conclude that the first component ranks the causes from the “young causes” (on the right) to the “old causes” (on the left). The second component opposes the causes that affect the extreme age intervals to those concerning, mainly, the medium age intervals.

The causes highly associated with young age are very specific (Figure 3): complications in pregnancy and childbirth, addiction to prescription medication, road accidents, meningococcal disease, homicides, congenital defects of the nervous system and congenital defects of the circulatory system. They are responsible of a high proportion of the young deaths: 30% of the deaths under 35 are due to these causes. At the same time, nearly half of the deaths due to these causes are among the young: from a total count of 17211 due to these causes, 7891 are under 35.

These causes, except road accidents, are relatively unimportant within all mortality counts and thus their contributions to the construction of the dimensions are low:

```
> round(mfa$ind$contr[c(2, 8:10, 15, 38, 58), 1:2], 3)
Addiction to prescription medication  0.998  0.448
Complications in pregnancy & childb.  0.685  0.527
Congenital defects circulatory system 0.692  0.176
Congenital defects nervous system     0.179  0.070
Homicides                             1.802  0.657
Meningococcal disease                 0.105  0.084
Road accidents                       34.295 23.364
```

In the case of the road accidents, a general trend is observed: as age increases, death counts caused by road accidents decrease. This phenomenon is observed in both 1979 and 2006.

```
> mortality[58, 1:9]  # road accidents in 1979
 15-24    25-34    35-44    45-54    55-64     65-74    75-84    85-94    95 and more
  3439     1666     1195     1328     966       1117     757      135              4
> mortality[58, 10:18] # road accidents in 2006
```

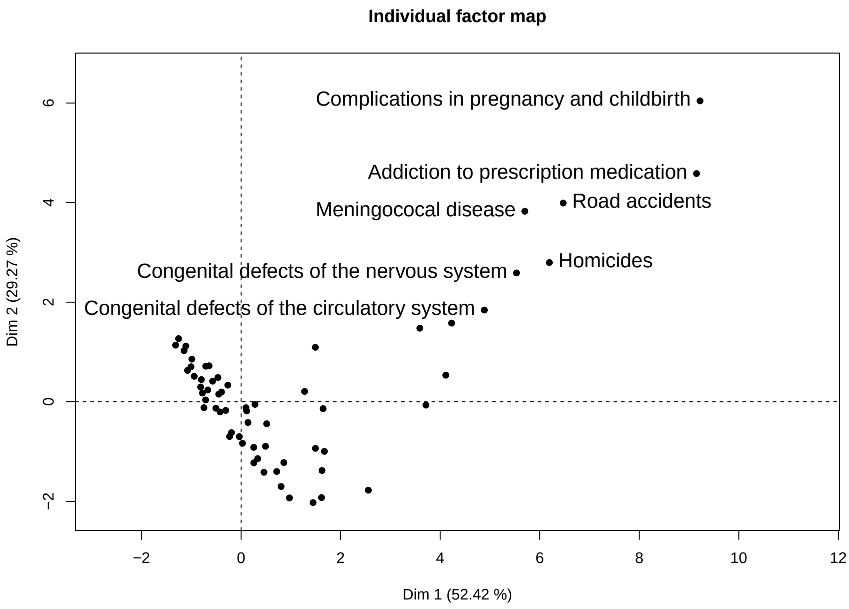

Figure 3: Mortality causes representation on the first principal plane. Only the "young" causes are labelled.

```
 15-24   25-34   35-44   45-54   55-64    65-74   75-84   85-94   95 and more
  1214     785     646     599     443      362     454     137             8
```

Dimensions of the separate analyses

The dimensions of the separate analyses can be represented through their correlations with the MFA dimensions. The first and second dimensions of the global analysis are very correlated with the first and second dimensions of the separate analyses (Figure 4).

Synthetic representation of the groups

As there are only two groups, their representation provides little information. However, in the case of a high number of groups, this kind of representation would be very useful for a global comparison.

Figure 5 shows that both groups have coordinates on dimension 1 that are very close to 1 showing that they are sensitive to the age ranking as reflected by this dimension. There are some differences between the two groups on dimension 2: the oppositions between causes, highlighted on the second dimension, are slightly more pronounced in 2006.

Superimposed representation

The superimposed representation of the partial rows (Figure 6) shows the more important changes between 1979 and 2006.

```
> sel <- c(2, 10, 41, 58)
> plot(mfa, lab.ind=FALSE, habillage="group", partial=rownames(mortality)[sel])
> text(mfa$ind$coord[sel,1], mfa$ind$coord[sel,2], rownames(mortality)[sel],pos=4)
```

Some important changes concerning the distribution of mortality causes among the age intervals between 1979 and 2006 are identified. Addiction to prescription medication is the cause of death showing the highest difference. It moves from close to the centroid (1979) to a position with a high coordinate (2006) on the first dimension: deaths related to this cause concern younger people more in 2006 than in 1979.

Table 1 ratifies this change. In 1979, about 60% of all deaths due to addiction to prescription medication are over 44. However, in 2006, 80% of deaths associated to this cause are between 25 and 44.

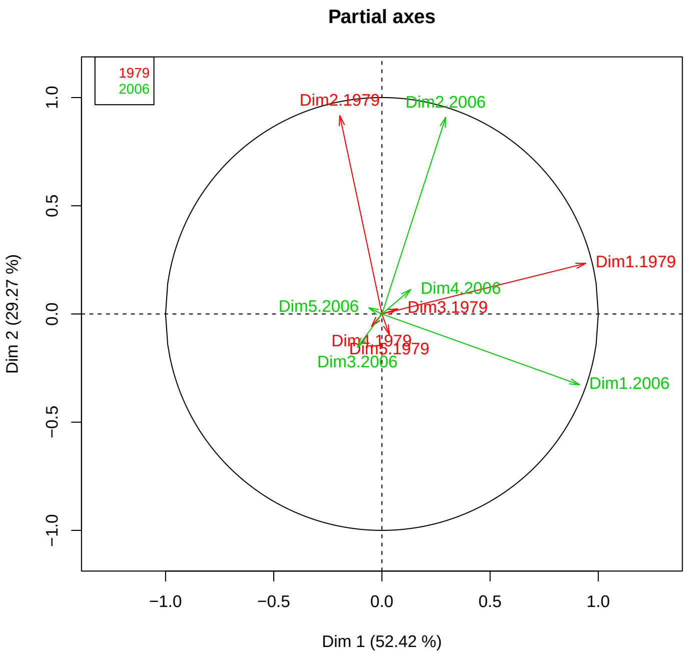

Figure 4: Dimensions of the separate analyses on the first principal plane issued from MFA.

Addiction to prescription medication	[15-24]	[25-34]	[35-44]	[45-54]	[55-64]	[65-74]	[75-84]	[85-94]	[95 or more]	Total
1979	7	4	2	6	6	2	6	0	0	33
%	21.2	12.1	6.1	18.2	18.2	6.1	18.2	0	0	100
2006	18	77	72	15	4	2	1	0	0	189
%	9.5	40.7	38.1	7.9	2.1	1.1	0.5	0	0	100

Table 1: Counts and percentages of deaths related to addiction to prescription medication in 1979 and 2006.

Conclusions

The study of mortality causes in France in 1979 and 2006 has shown the kind of results offered by MFA extended to deal with a multiple contingency table.

Concerning these data, we can conclude that the mortality structure by age changes very little from 1979 to 2006, apart from some interesting nuances. In both years, younger age is associated with very specific mortality causes. The most important change is due to addiction to prescription medication. This cause turns to be, at the same time, more important (with respect to the total count) and associated more to young ages in 2006 compared to 1979.

MFACT can be used in many different fields, such as text mining, public health, sensometrics, etc. In the example used to illustrate the motivation, MFACT has been applied to compare the favourite menus in different countries (Pagès and Bécue-Bertaut, 2006). To give a few examples, MFACT has also been used to cluster the Spanish regions depending on their mortality structure (Bécue-Bertaut et al., 2011), and to characterize food products from free-text descriptions (Kostov et al., 2011).

The MFA function offers options, tools and graphical outputs that ease the interpretation of the results. Furthermore, this function allows the analysis of multiple tables with a mixture of quantitative, categorical and frequency groups as detailed in Bécue-Bertaut and Pagès (2008). This latter possibility is largely used in studies such as:

- The relationship between the bacterial community structures and hydrogeochemical conditions in a groundwater study (Imfeld et al., 2011).
- The influence of package shape and colour on consumer expectations of milk desserts using word association and conjoint analysis (Ares and Deliza, 2010).
- Customer portfolio composition (Abascal et al., 2010).
- The answers to an open-ended question to identify drivers of liking of milk desserts (Ares et al.,

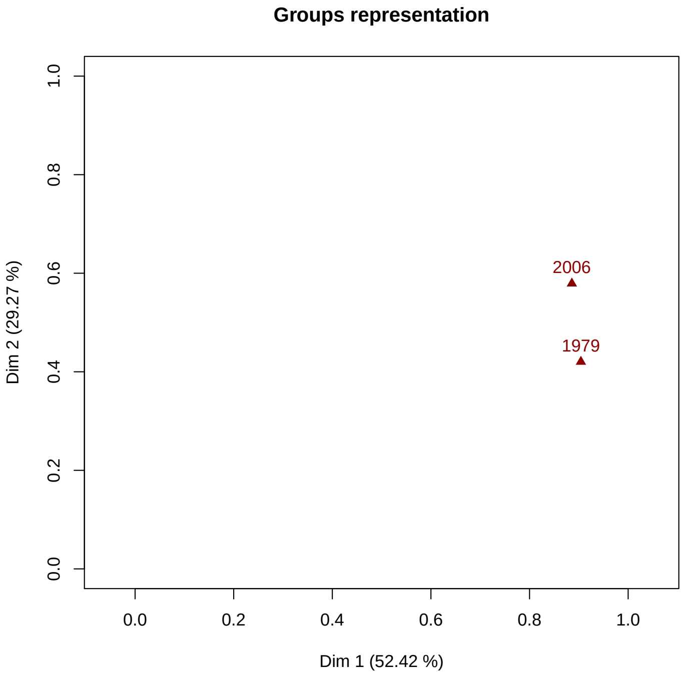

Figure 5: Synthetic representation of the groups on the first and second dimensions.

2010).

It is possible to enrich the analysis using supplementary variables, including quantitative, qualitative and frequency variables.

Summary

MFACT is a useful method to analyze multiple contingency tables. Here, we present an application of this method to French mortality data to compare the relationships between mortality causes and age intervals in two different years. We show how this method helps to compare the same information on two different occasions. It is possible to also apply MFACT to multiple tables integrating frequency, categorical and quantitative groups of variables observed on the same individuals.

Bibliography

E. Abascal, I. G. Lautre, and F. Mallor. Tracking customer portfolio composition: A factor analysis approach. *Applied Stochastic Models in Business and Industry*, 26:535–550, 2010. [p35]

H. Akuto. *International Comparison of Dietary Cultures*. Nihon Keizai Shimbun, 1992. [p29]

G. Ares and R. Deliza. Studying the influence of package shape and colour on consumer expectations of milk desserts using word association and conjoint analysis. *Food Quality and Preference*, 21:930–937, 2010. [p35]

G. Ares, R. Deliza, C. Barreiro, A. Giménez, and A. Gámbaro. Use of an open-ended question to identify drivers of liking of milk desserts. Comparison with preference mapping techniques. *Food Quality and Preference*, 21:286–294, 2010. [p35]

M. Bécue-Bertaut and J. Pagès. A principal axes method for comparing multiple contingency tables: MFACT. *Computational Statistics and Data Analysis*, 45:481–503, 2004. [p29, 32]

M. Bécue-Bertaut and J. Pagès. Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Computational Statistics and Data Analysis*, 52:3255–3268, 2008. [p29, 35]

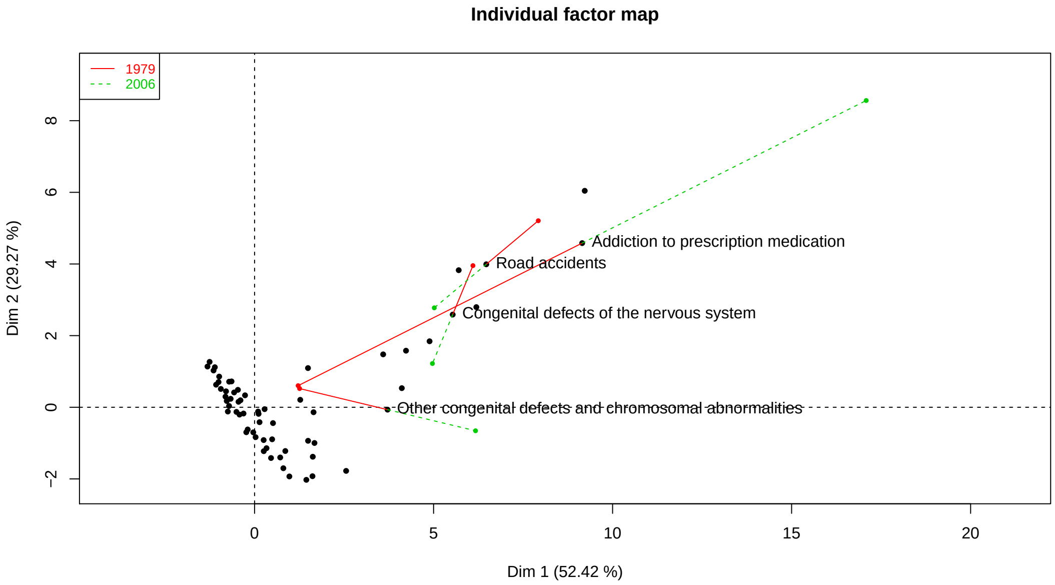

Figure 6: Excerpt of the superimposed representation of the partial and global mortality causes.

M. Bécue-Bertaut, M. Guillén, and J. Pagès. Clasificación de las regiones españolas según sus patrones de mortalidad. In *Métodos Cuantitativos en Economía del Seguro del Automóvil*, pages 97–111. Universidad de Barcelona, 2011. [p35]

J. Benzécri. *Analyse des Données*. Dunod, 1973. [p29]

B. Escofier and J. Pagès. *Analyses Factorielles Simples et Multiples*. Dunod, 4th edition, 2008. [p29, 30]

F. Husson, S. Lê, and J. Pagès. *Exploratory Multivariate Analysis by Example Using R*. Chapman & Hall / CRC Press, 2011. [p29, 31]

G. Imfeld, H. Pieper, N. Shani, P. Rossi, M. Nikolausz, I. Nijenhuis, H. Paschke, H. Weiss, and H. H. Richnow. Characterization of groundwater microbial communities, dechlorinating bacteria, and in situ biodegradation of chloroethenes along a vertical gradient. *Water, Air, & Soil Pollution*, 221(1–4): 107–122, 2011. [p35]

B. Kostov, M. Bécue-Bertaut, J. Pagès, M. Cadoret, J. Torrens, and P. Urpi. Verbalisation tasks in Hall test sessions. Presented at the "International Classification Conference (ICC)", St. Andrews, UK, 2011. [p35]

S. Lê, J. Josse, and F. Husson. FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18, 3 2008. ISSN 1548-7660. URL http://www.jstatsoft.org/v25/i01. [p29]

L. Lebart, A. Salem, and L. Berry. *Exploring Textual Data*. Kluwer Academic Publishers, 1998. [p29]

L. Lebart, M. Piron, and A. Morineau. *Statistique Exploratoire Multidimensionnelle*. Dunod, 2006. [p29]

J. Pagès and M. Bécue-Bertaut. Multiple factor analysis for contingency tables. In M. Greenacre and J. Blasius, editors, *Multiple Correspondence Analysis and Related Methods*, pages 299–326. Chapman & Hall / CRC Press, 2006. [p29, 35]

Belchin Kostov
Transverse group for research in primary care, IDIBAPS Primary Health Care Center Les Corts, CAPSE
Mejia Lequerica, s / n.
08028 Barcelona
Spain
badriyan@clinic.ub.es

Mónica Bécue-Bertaut
Department of Statistics and Operational Research

Universitat Politècnica de Catalunya
North Campus - C5
Jordi Girona 1-3
08034 Barcelona
Spain
monica.becue@upc.edu

François Husson
Agrocampus Rennes
65 rue de Saint-Brieuc
35042 Rennes France
husson@agrocampus-ouest.fr