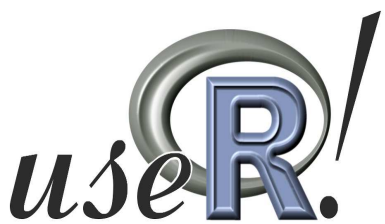
Binary attributes quantification with external information

Alfonso Iodice D'Enza*
*Università di Cassino, (Italy)
iodicede@gmail.com

The R User Conference 2009
July 8-10, Agrocampus-Ouest, Rennes, France

Outline



Binary Data

Relevance of Binary Data

During the past decade the attention to Binary Data quickly increased. There are several motivations to take into account to understand the reasons of this major interest. Among the others, binary data can be easily collected, stored and managed

Application in several fields

- Gene Expression Data
- Text Mining
- Web click-stream analysis
- Transactional Data Bases

Association Rules

A short reminder

Consider a pair of attributes (or sets of attributes) **A** and **B**: a simple association rule based on the considered attributes is:

$$\text{If} \mathbf{A} \longrightarrow \mathbf{B} = \{support = .2, confidence = .8\}$$

Sup: the 20% of sequences contain both **A** and **B** items;

Conf: the 80% of sequences containing the item **A** contain the item **B** too;

Interpretation

- the support measures the intensity of the association between **A** and **B**
- the confidence measures the strength of the logical dependence between **A** and **B**

Association rules can be easily generalised to itemsets with cardinality > 2

Association Rules

A short reminder

Consider a pair of attributes (or sets of attributes) **A** and **B**: a simple association rule based on the considered attributes is:

$$\text{If} \mathbf{A} \longrightarrow \mathbf{B} = \{support = .2, confidence = .8\}$$

Sup: the 20% of sequences contain both **A** and **B** items;

Conf: the 80% of sequences containing the item **A** contain the item **B** too;

Interpretation

- the support measures the intensity of the association between **A** and **B**
- the confidence measures the strength of the logical dependence between **A** and **B**

Association rules can be easily generalised to itemsets with cardinality > 2

Association Rules

AR mining is a NP-problem

In presence of large databases it becomes soon not feasible cause the number of rules increases exponentially:

- computational issues (not serious)
- interpretation difficulties (serious)

Association study approaches

Brute Force approach

- AR's having high/very high support are considered *trivial* rules and are discarded
- AR's with low support represent *not interesting* rules and are discarded
- defining the thresholds is a ticklish problem
  - loose thresholds determine a huge amount of output
  - tight thresholds may lead to discard interesting association patterns

Trojan horse approach

An alternative approach is to mine AR within homogeneous groups of items and/or of sequences. Homogeneous subsets can be defined through an

- exogenous criterion groups are defined according to an external categorical variable
- endogenous criterion groups are defined via a suitable cluster analysis of the sequences

Association study approaches

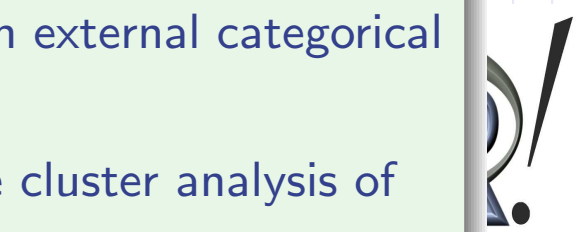
Brute Force approach

- AR's having high/very high support are considered *trivial* rules and are discarded
- AR's with low support represent *not interesting* rules and are discarded
- defining the thresholds is a ticklish problem
  - loose thresholds determine a huge amount of output
  - tight thresholds may lead to discard interesting association patterns

Trojan horse approach

An alternative approach is to mine AR within homogeneous groups of items and/or of sequences. Homogeneous subsets can be defined through an

- exogenous criterion groups are defined according to an external categorical variable
- endogenous criterion groups are defined via a suitable cluster analysis of the sequences



Data structures

A multivariate data set is given by a set of n statistical units, named *sequences* and each sequence is defined by a set of $\{I_2, I_2, \ldots, I_P\}$ binary variables, which are called *attributes* or *items*

Binary variables can assume values only in $\{0, 1\}$

To arrange these data, two possibilities exist:

presence/absence matrix S with n rows and P columns

	I_1	I_2	...	I_P
1	0	1	...	1
2	1	1	...	0
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
n	1	0	...	1

Data structures

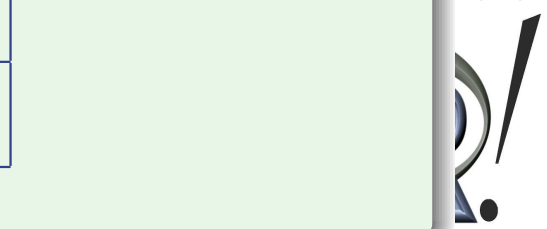
A multivariate data set is given by a set of n statistical units, named *sequences* and each sequence is defined by a set of $\{I_2, I_2, \ldots, I_P\}$ binary variables, which are called *attributes* or *items*

Binary variables can assume values only in $\{0, 1\}$

To arrange these data, two possibilities exist:

disjunctive coded matrix $\mathbf{Z}$ with n rows and $2P$ columns

	$I_1 \vdots I_1$	$I_2 \vdots I_2$	$\ldots$	$I_P \vdots I_P$
1	$0 \vdots 1$	$1 \vdots 0$	$\ldots$	$1 \vdots 0$
2	$1 \vdots 0$	$1 \vdots 0$	$\ldots$	$0 \vdots 1$
$\ldots$	$\ldots \vdots \ldots$	$\ldots \vdots \ldots$	$\ldots$	$\ldots \vdots \ldots$
n	$1 \vdots 0$	$0 \vdots 1$	$\ldots$	$0 \vdots 1$



Association measures: a different point of view

- The complete disjunctive Binary Data coding turns out extremely useful when defining the association measures
- Taking into account two general items of the matrix $\mathbf{Z}$: Z_j and $Z_{j'}$, the product $Z_j' Z_{i'}$ (with $\{j, j'\} = 1, 2, \ldots, P$) determines the following 2×2 matrix:

$$D = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

- a indicates the number co-presence
- b and c correspond to the non-matchings
- d indicates the number co-absence

Association measures: a different point of view

- The complete disjunctive Binary Data coding turns out extremely useful when defining the association measures
- Taking into account two general items of the matrix $\mathbf{Z}$: Z_j and $Z_{j'}$, the product $Z_j' Z_{i'}$ (with $\{j, j'\} = 1, 2, \ldots, P$) determines the following 2×2 matrix:

$$D = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

- a indicates the number co-presence
- b and c correspond to the non-matchings
- d indicates the number of co-absences
-
-

Association measures: a different point of view

- The complete disjunctive Binary Data coding turns out extremely useful when defining the association measures
- Taking into account two general items of the matrix $\mathbf{Z}$: Z_j and $Z_{j'}$, the product $Z_j' Z_{i'}$ (with $\{j, j'\} = 1, 2, \ldots, P$) determines the following 2×2 matrix:

$$D = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

a indicates the number co-presence

b and c correspond to the non-matchings

d indicates the number of co-absences

Association measures: a different point of view

- The complete disjunctive Binary Data coding turns out extremely useful when defining the association measures
- Taking into account two general items of the matrix $\mathbf{Z}$: Z_j and $Z_{j'}$, the product $Z_j' Z_{j'}$ (with $\{j, j'\} = 1, 2, \ldots, P$) determines the following 2×2 matrix:

$$D = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

a indicates the number co-presence

b and c correspond to the non-matchings

d indicates the number of co-absences

- using the set $\{a, b, c, d\}$ it is possible to define all the *dissimilarity/similarity* measures for binary data
-

Association measures: a different point of view

- The complete disjunctive Binary Data coding turns out extremely useful when defining the association measures
- Taking into account two general items of the matrix $\mathbf{Z}$: Z_j and $Z_{j'}$, the product $Z_j' Z_{i'}$ (with $\{j, j'\} = 1, 2, \ldots, P$) determines the following 2×2 matrix:

$$D = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

a indicates the number co-presence

b and c correspond to the non-matchings

d indicates the number of co-absences

- using the set $\{a, b, c, d\}$ it is possible to define all the *dissimilarity/similarity* measures for binary data
- the tuple $\{a, b, c, d\}$ can also be used to compute support, confidence and all of the AR interestingness measures (see [Lenca et al., 2007] for a detailed overview).

Association measures: a different point of view

- The complete disjunctive Binary Data coding turns out extremely useful when defining the association measures
- Taking into account two general items of the matrix $\mathbf{Z}$: Z_j and $Z_{j'}$, the product $Z_j' Z_{i'}$ (with $\{j, j'\} = 1, 2, \ldots, P$) determines the following 2×2 matrix:

$$D = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

a indicates the number co-presence

b and c correspond to the non-matchings

d indicates the number of co-absences

- using the set $\{a, b, c, d\}$ it is possible to define all the *dissimilarity/similarity* measures for binary data
- the tuple $\{a, b, c, d\}$ can also be used to compute support, confidence and all of the AR interestingness measures (see [1] for a detailed overview).

Association measures: a different point of view

- The complete disjunctive Binary Data coding turns out extremely useful when defining the association measures
- Taking into account two general items of the matrix $\mathbf{Z}$: Z_j and $Z_{j'}$, the product $Z_j' Z_{i'}$ (with $\{j, j'\} = 1, 2, \ldots, P$) determines the following 2×2 matrix:

$$D = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

a indicates the number co-presence

b and c correspond to the non-matchings

d indicates the number of co-absences

- using the set $\{a, b, c, d\}$ it is possible to define all the *dissimilarity/similarity* measures for binary data
- the tuple $\{a, b, c, d\}$ can also be used to compute support, confidence and all of the AR interestingness measures (see [7] for a detailed overview).

Quantification of binary attributes

Binary data marts are usually:

- high-dimensional - the considered objects are described by the presence/absence of a large number of attributes
- sparse - each object presents a sub-set of attributes which is considerably smaller than the whole set in question
- low-separability - when data are sparse, the separability of points/objects in a high dimensional space is very low

A suitable exploratory approach to study the association structure of attributes is Multiple Correspondence Analysis (MCA, [2]). The advantages of quantification are:

- reduction of dimensionality and multiple associations visualization in graphical displays
- low-dimensional description of objects for detecting homogeneous groups of objects

Quantification of binary attributes

Binary data marts are usually:

- high-dimensional - the considered objects are described by the presence/absence of a large number of attributes
- sparse - each object presents a sub-set of attributes which is considerably smaller than the whole set in question
- low-separability - when data are sparse, the separability of points/objects in a high dimensional space is very low

A suitable exploratory approach to study the association structure of attributes is Multiple Correspondence Analysis (MCA, [2]). The advantages of quantification are:

- reduction of dimensionality and multiple associations visualization on graphical displays
- low dimensional description of objects facilitates the identification of homogeneous groups of objects

Quantification of binary attributes

Binary data marts are usually:

- high-dimensional - the considered objects are described by the presence/absence of a large number of attributes
- sparse - each object presents a sub-set of attributes which is considerably smaller than the whole set in question
- low-separability - when data are sparse, the separability of points/objects in a high dimensional space is very low

A suitable exploratory approach to study the association structure of attributes is Multiple Correspondence Analysis (MCA, [2]). The advantages of quantification are:

- reduction of dimensionality and multiple associations visualization on graphical displays
- low dimensional description of objects facilitates the identification of homogeneous groups of objects

Aim of the contribution

The aim of the present contribution is to define a quantification of binary (or categorical) attributes that takes into account and emphasizes the presence of groups of homogeneous objects (binary sequences/statistical units). The proposed approach deals with both the cases of exogenous and endogenous defined groups.

exogenous information

Attributes are quantified taking into account the modalities of an external categorical attributes: it may refer to a specific feature, or it can be the result of a cluster analysis on further set of variables (e.g. socio-demographic information of customers)

endogenous information

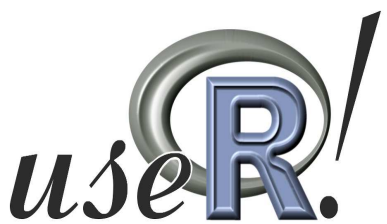
The quantification is integrated in a two-step procedure combining dimensionality reduction and clustering

NSCA-based approach: further data structures

frequencies matrix $\mathbf{F}$ of the P attributes in the $\mathbf{K}$ groups

The NSCA based quantification involves the following data structure

	1	2	...	j	...	P
1	$f_{1,1}$	$f_{1,2}$	...	$f_{1,j}$	...	$f_{1,P}$
...	...	...	...	...	...	...
k	$f_{k,1}$	$f_{k,2}$	...	$f_{k,j}$	...	$f_{k,P}$
...	...	...	...	...	...	...
K	$f_{K,1}$	$f_{K,2}$	...	$f_{K,j}$	...	$f_{K,P}$



Quantification of Binary Data: NSCA-based approach

Problem statement

Consider each of the groups to be coded via an indicator variable. Thus there will be K such indicators X_k, $k = 1, \ldots, K$, with $X_k = 1$ if the i^{th} object is in the group k ($i \Rightarrow k$), else 0. These indicators are collected together in a vector $X = (X_1, \ldots, X_K)$.

- Consider the attribute A to take values according to a generic random variable and the conditional expectation $E(X_k \mid A) = Pr((i \Rightarrow k) \mid A)$
- In case of binary attributes the reference random variable for A is Bernoulli distributed with parameter p_A [3].

Target function

Thus the target function is

$$\begin{aligned} max! \quad & E[P(X_k \mid A) - P(X_k)] \equiv \\ \equiv \quad & max!\ E[P(i \Rightarrow k \mid A)] - E[P(i \Rightarrow k)] \end{aligned} \tag{1}$$

- the problem consists in maximizing the difference between the conditional probabilities $Pr(X_k \mid A)$ and the marginal distribution.

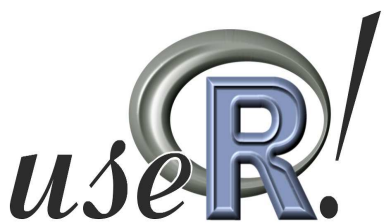
Quantification of Binary Data: NSCA-based approach

Target function re-formulation

The target function can be re-expressed as follows

$$\sum_{k=1}^{K} \left(n(k, Z_j)P(X_k \mid Z_j) - n(k)P(X_k)\right) = \tag{2}$$
$$= \sum_{k=1}^{K} \left(n(k, Z_j)\frac{P(X_k \cap Z_j)}{P(Z_j)} - n(k)P(X_k)\right).$$

the solution is obtained through a maximization of the quantity in expression 2 with respect to $\mathbf{X}$, a $(n \times K)$ matrix that assigns each sequence to one of the K groups.



Quantification of Binary Data: NSCA-based approach

Target function with p attributes

In case of p attributes the target function is

$$max! \sum_{j=1}^{p} \sum_{k=1}^{K} (n(k, Z_j)P(X_k \mid Z_j) - n(k)P(X_k)). \tag{3}$$

Let us recall the $\mathbf{F}$ matrix, then the target function in 3 is equivalent to maximize the following expression

$$max! \frac{1}{n} \sum_{k=1}^{K} \left(\sum_{j=1}^{p} \frac{f_{kj}^2}{f_{.j}} - \frac{f_{k.}^2}{n} \right). \tag{4}$$

since it results that

- $n(k, Z_j) = f_{kj}$ and $n(k, Z_j) = f_{k.}$
- $P(X_k \mid Z_j) = f_{.j}^{-1} f_{kj}$ and $P(X_k) = n^{-1} f_{k.}$

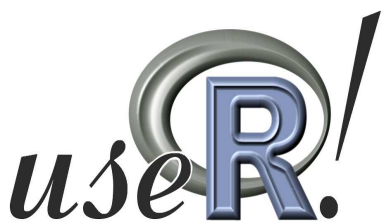
The Model and the NSCA problem

Important equality

The probability expectation can be re-expressed in terms of item frequencies as follows:

$$\frac{1}{n}\sum_{k=1}^{K}\left(\sum_{j=1}^{p}\frac{f_{kj}^2}{f_{.j}}-\frac{f_{k.}^2}{n}\right)=\frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{p}f_{.j}\left(\frac{f_{kj}}{f_{.j}}-f_{k.}\right)^2. \quad (5)$$

The right hand quantity in expression 5 corresponds to Lauro and D'Ambra's Non Symmetric Correspondence Analysis model.



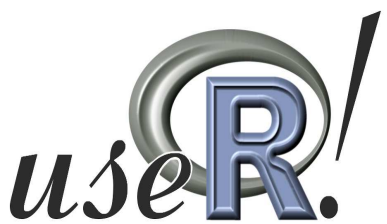
Algebraic formalization of the problem

Algebraic formalization

An algebraic formalization of the quantity in expression 5 corresponds to

$$\begin{aligned} & \mathrm{tr}\left[\mathbf{F}(\Delta)^{-1}\mathbf{F}^{\mathsf{T}} - \tfrac{p}{n}\left(\mathbf{X}^{\mathsf{T}}\mathbf{1}\mathbf{1}^{\mathsf{T}}\mathbf{X}\right)\right] \equiv \\ \equiv \quad & \mathrm{tr}\left[\mathbf{X}^{\mathsf{T}}\mathbf{Z}(\Delta)^{-1}\mathbf{Z}^{\mathsf{T}}\mathbf{X} - \tfrac{p}{n}\left(\mathbf{X}^{\mathsf{T}}\mathbf{1}\mathbf{1}^{\mathsf{T}}\mathbf{X}\right)\right] \end{aligned} \tag{6}$$

where $\Delta = \mathrm{diag}(\mathbf{Z}^{\mathsf{T}}\mathbf{Z})$ and $\mathbf{1}$ is a n-dimensional vector of ones.
The solution of the problem is in the maximization of the trace of the above matrix.



Algebraic formalization of the problem

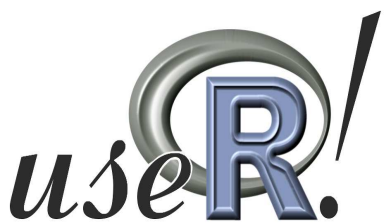
Target function

$$\frac{1}{n}\left[\mathbf{X}^{\mathsf{T}}\mathbf{Z}(\Delta)^{-1}\mathbf{Z}^{\mathsf{T}}\mathbf{X} - \frac{p}{n}\left(\mathbf{X}^{\mathsf{T}}\mathbf{1}\mathbf{1}^{\mathsf{T}}\mathbf{X}\right)\right]\mathbf{U} = \Lambda\mathbf{U} \tag{7}$$

that is to compute eigenvalues and eigenvector, in the diagonal matrix Λ and in the matrix $\mathbf{U}$, respectively.

Remark

With respect to expression 7, if no exogenous information is available matrices $\mathbf{X}$, Λ and $\mathbf{U}$ are unknown, thus a direct solution is not possible.



NSCA exogenous information: sequences and attributes quantification

Quantification of sequences

The solution of the problem in expression 7 leads to a obtain a score of the starting sequences:

$$\Psi = \left(\mathbf{Z}(\Delta)^{-1}\mathbf{Z}^{\mathsf{T}} - \frac{p}{n}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right)\mathbf{X}\mathbf{U}\Lambda^{\frac{1}{2}} \tag{8}$$

with $\mathbf{X}$ being known and defined by the exogenous criterion.

Quantification of attributes

As for sequences, the quantification of attributes is computed by

$$\Phi = \mathbf{Z}^{\mathsf{T}}\mathbf{X}\mathbf{U}\Lambda^{\frac{1}{2}}. \tag{9}$$

NSCA with endogenous information: implementation of the two-step procedure

The procedure

The algorithm runs over the following steps:

- *step 0*: pseudo-random generation of matrix $\mathbf{X}$
- *step 1*: a singular value decomposition is performed on the matrix resulting from 7, obtaining the matrix Ψ, such that

$$\Psi = \left(\mathbf{Z}(\Delta)^{-1}\mathbf{Z}^{\intercal} - \frac{p}{n}\mathbf{1}\mathbf{1}^{\intercal}\right)\mathbf{X}\mathbf{U}\Lambda^{\frac{1}{2}} \tag{10}$$

- *step 2*: matrix $\mathbf{X}$ is updated according to the results of an Euclidean squared distance based partition algorithm (K-means, [8]) on the projected sequences (Ψ matrix)

Steps 1 and 2 are iterated until the convergence: the quantity in 7 does not significant increase from one iteration to the following one.

A note on related work

Similar approaches on quantitative data

- The factorial K-means strategy is proposed by [12] to deal with the masking cluster problem in the case of multivariate continuous variables
- [13] propose constrained principal component analysis, which aims at simultaneous clustering of objects and partitioning of variables.

Similar approaches on qualitative data

- the present approach can also be defined a Non-Symmetric Factorial Discriminant Analysis (NS-FDA) proposed by [9]. The authors point out the relationship with Non-Symmetric Correspondence Analysis [6], of which NS-FDA is a special case.
- [4] propose an extension of multiple correspondence analysis that takes into account cluster-level heterogeneity in respondents' preferences/choices.

Exploited R functions

R packages

The R implementation of the procedure exploits the following packages

- base,[11]: the $svd()$ function for the singular value decomposition of the target function
- stats,[11]
  - The $hclust()$ function for the agglomerative clustering of the quantified attributes
  - The $kmeans()$ function for the K-means clustering of the quantified sequences in the iterative solution
- graphics, [11]: to obtain all of the 2D representations
- rgl, [1]: to obtain all of the 3D representations

Examples of application: the UniMC data set

The UniMC data set contains informations on the careers of bachelor students of Economics from the Università di Macerata (Italy). Each binary sequence records which of the fourteen fundamental examinations has been passed by a single student.

Data description

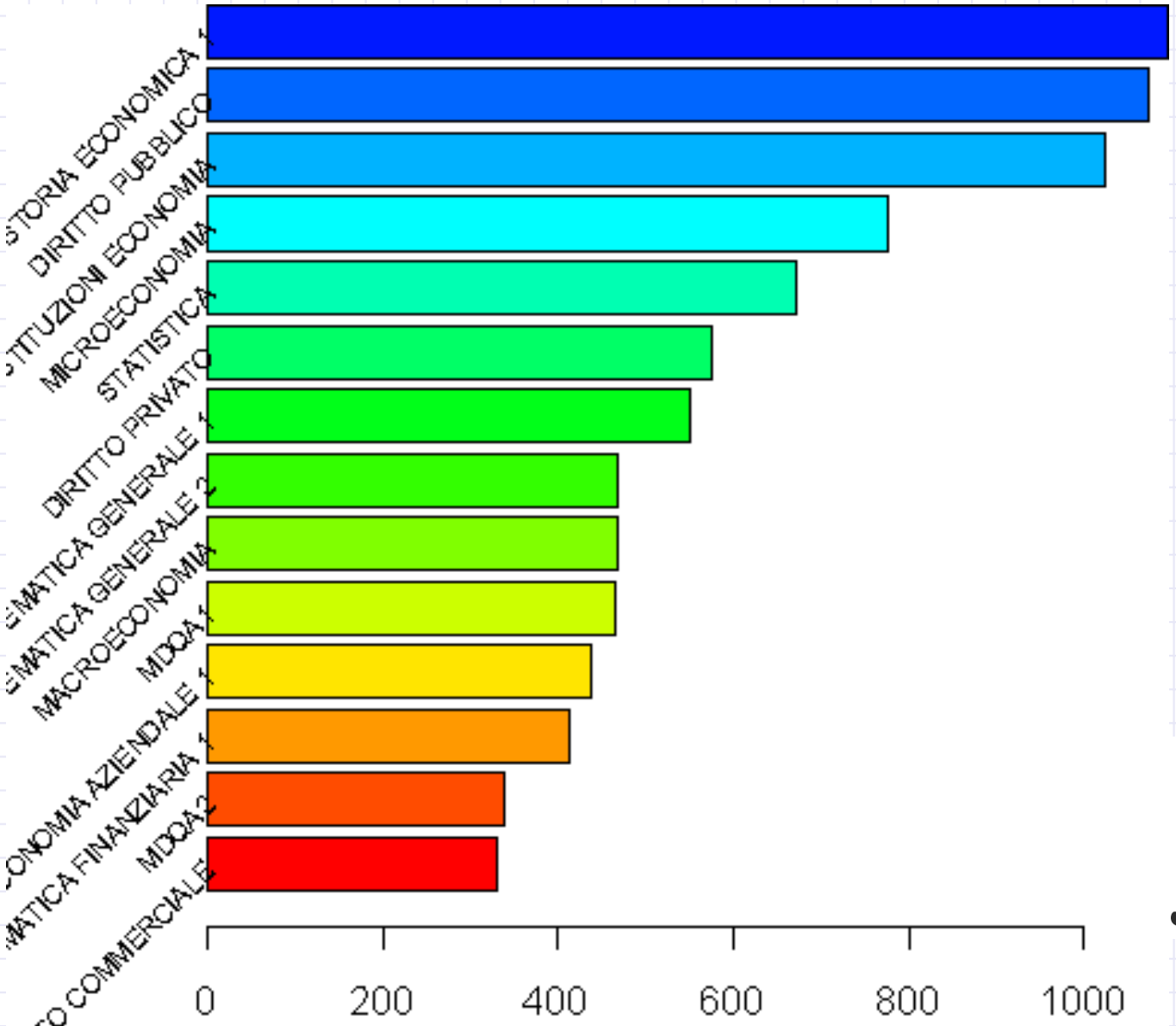
- the number of considered students (sequences) is 2421
- the number of considered examinations (attributes) is 14
- the time-range goes since 2001/2002 up to 2006/2007

Attributes

1	DIRITTO COMMERCIALE
2	DIRITTO PRIVATO
3	DIRITTO PUBBLICO
4	ECONOMIA AZIENDALE 1
5	ISTITUZIONI ECONOMIA
6	MACROECONOMIA
7	MATEMATICA FINANZIARIA 1
8	MATEMATICA GENERALE 1
9	MATEMATICA GENERALE 2
10	MDQA1
11	MDQA2
12	MICROECONOMIA
13	STATISTICA
14	STORIA ECONOMICA 1

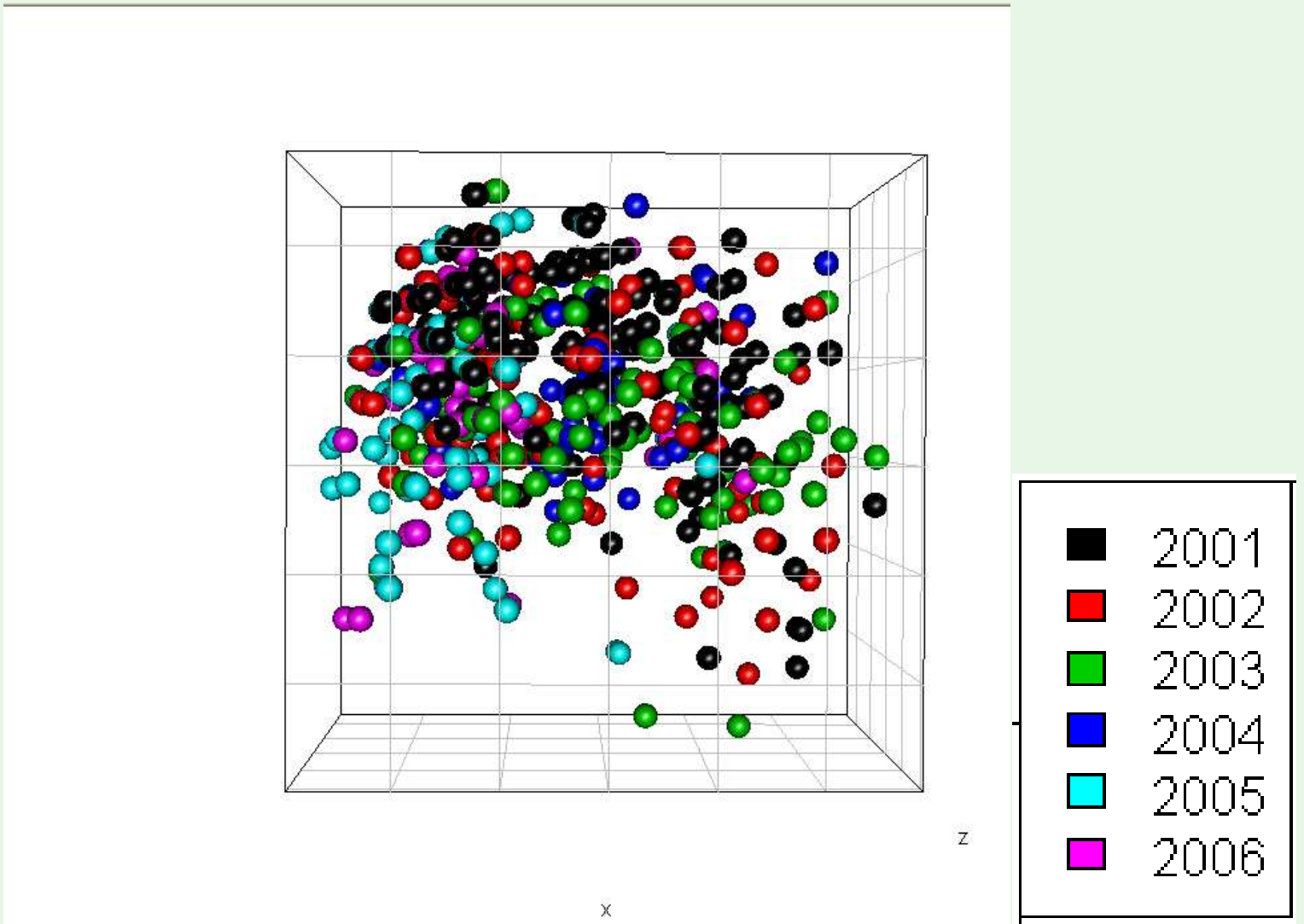
The exogenous criterion under consideration is the academic year.

Examples of application: the UniMC data set



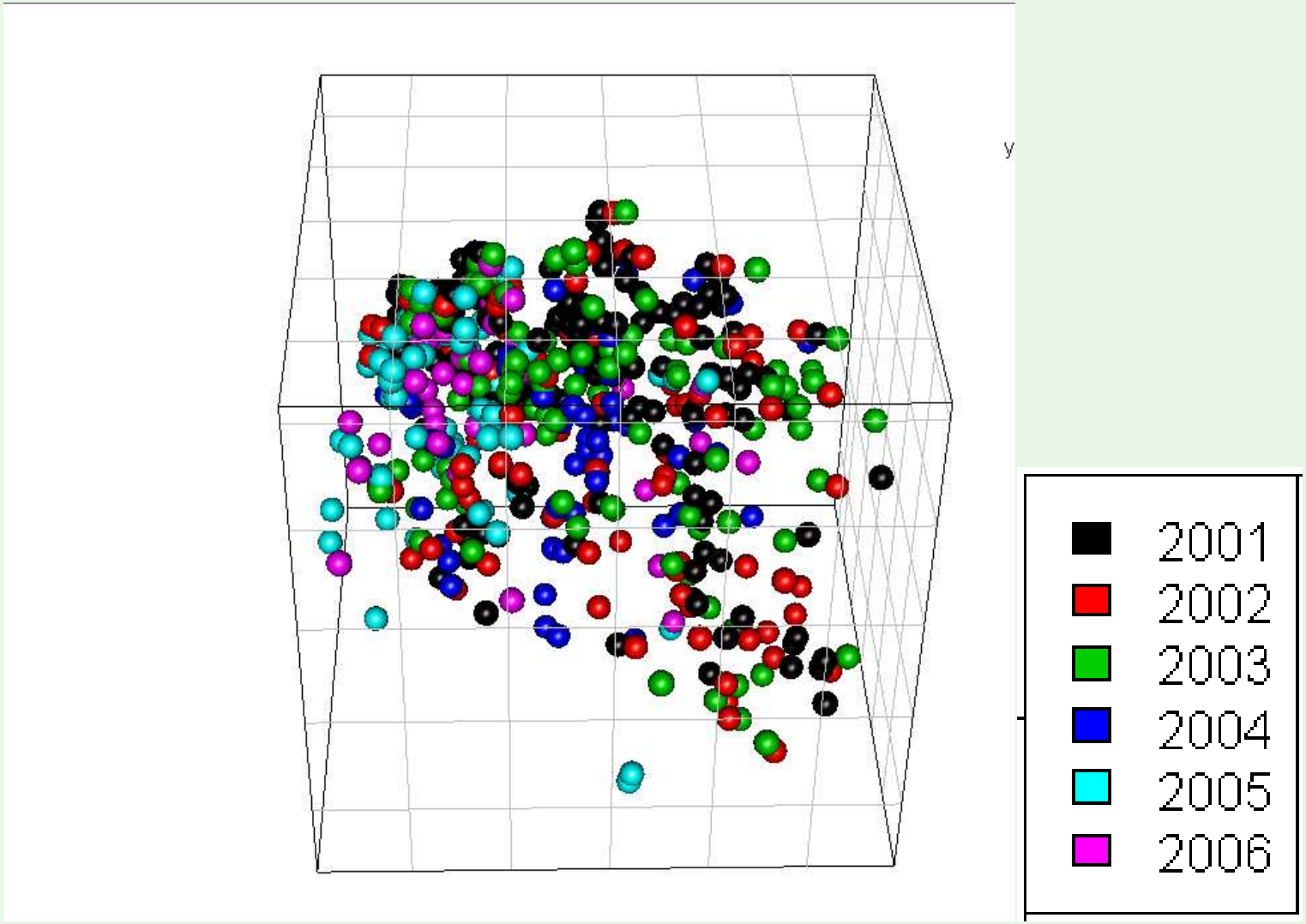
Examples of application: exogenous information approach

X vs Y



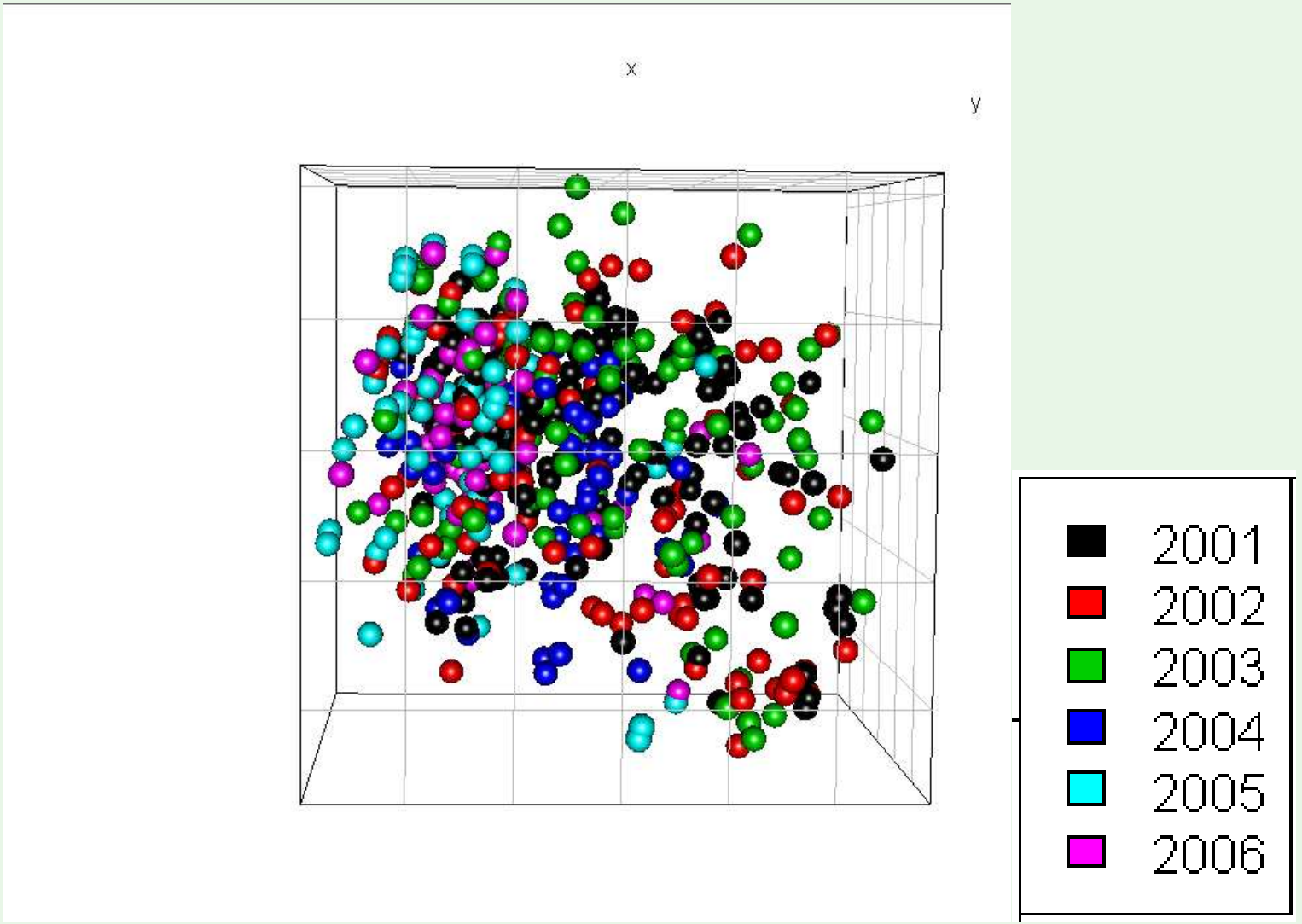
Examples of application: exogenous information approach

X vs Y $\rightarrow$ X vs Z



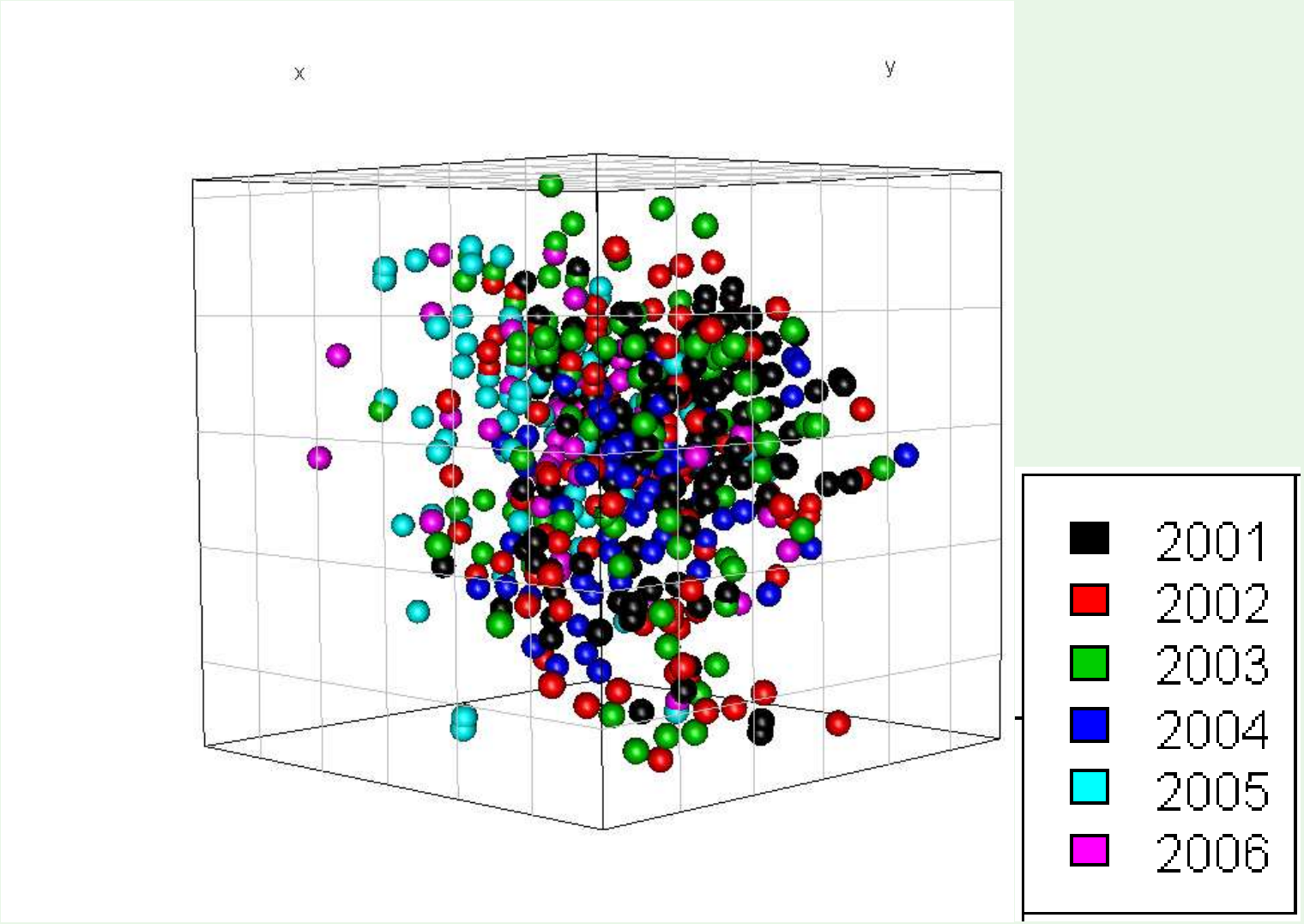
Examples of application: exogenous information approach

X vs Z



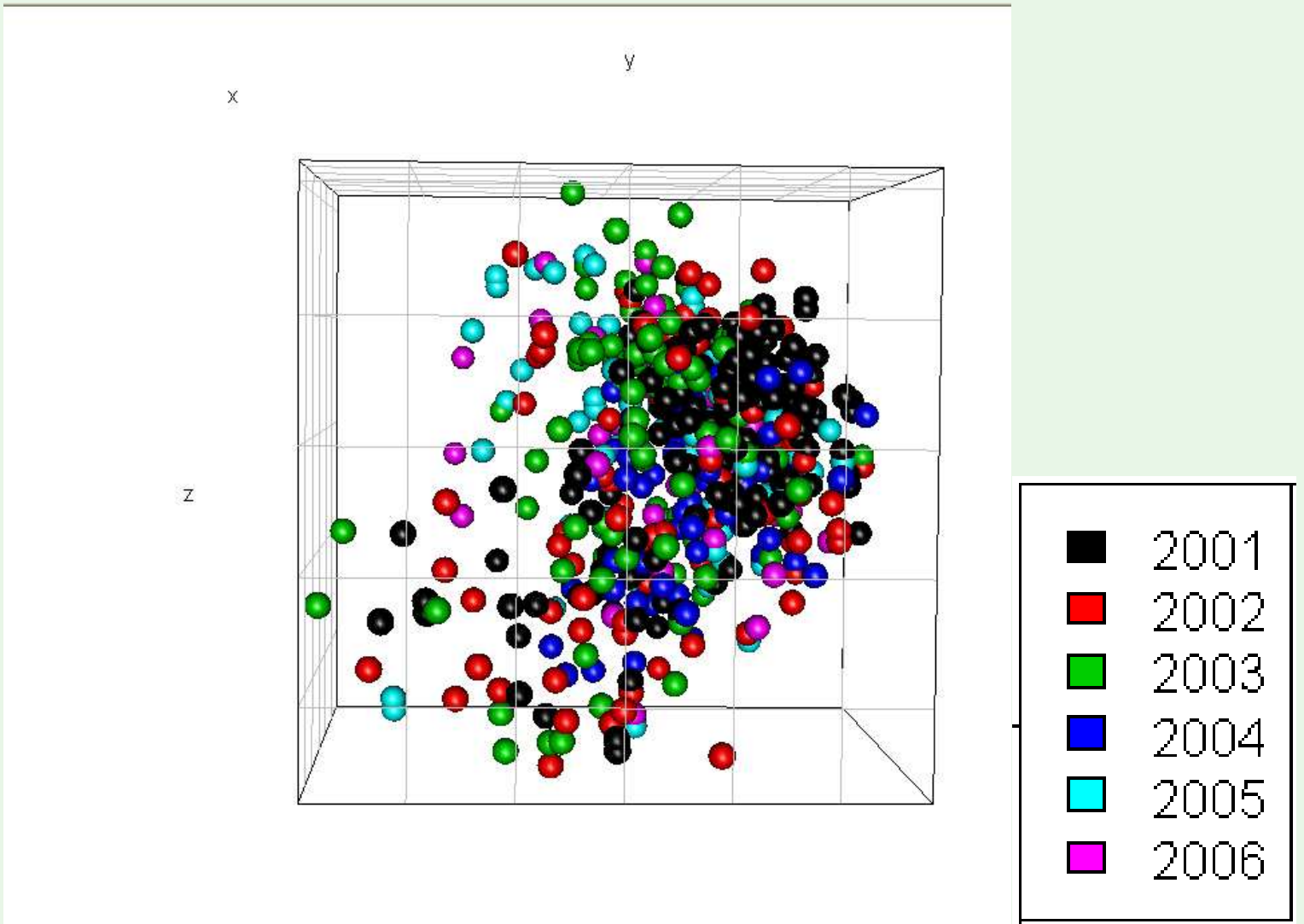
Examples of application: exogenous information approach

X vs Z $\rightarrow$ Y vs Z



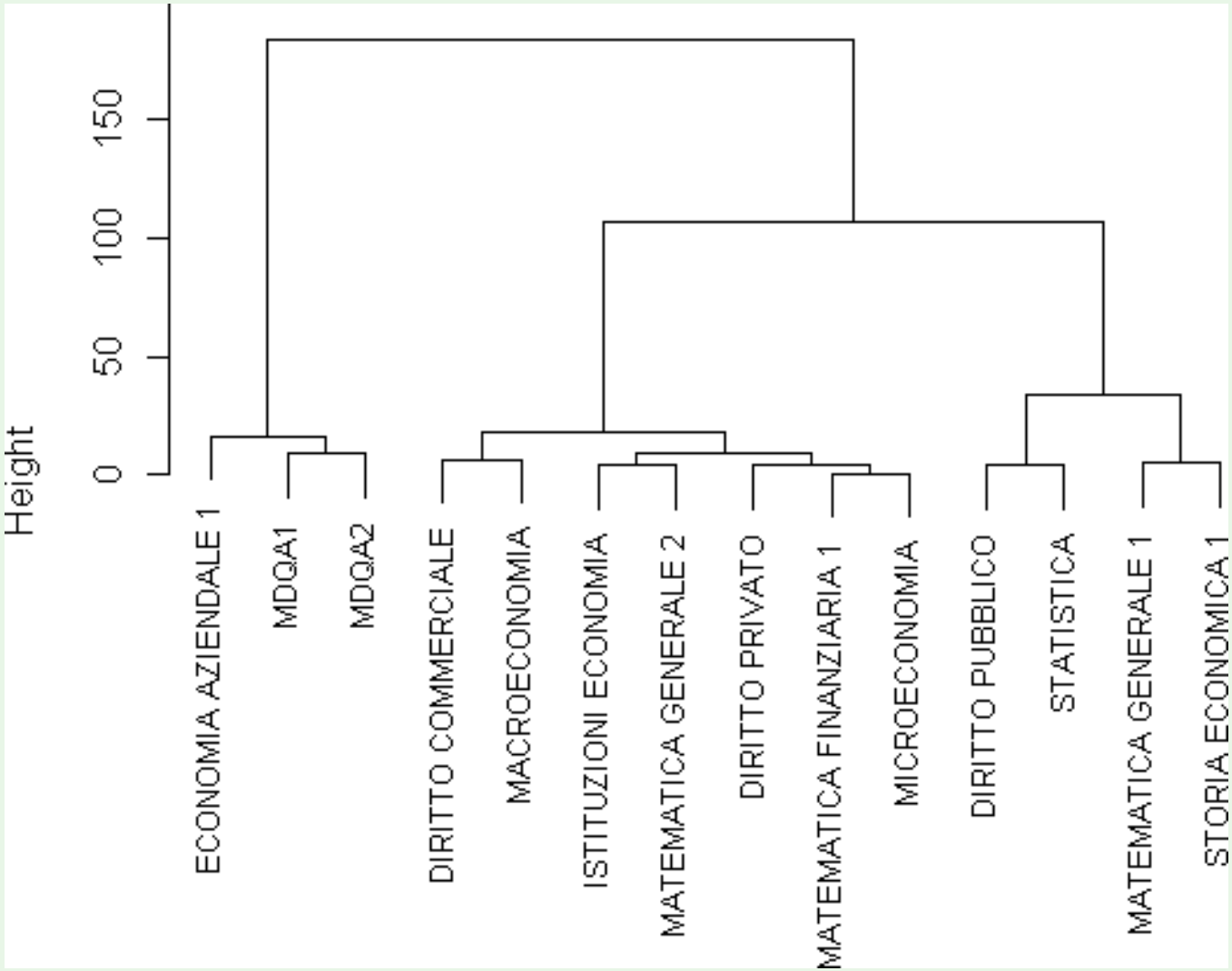
Examples of application: exogenous information approach

Y vs Z



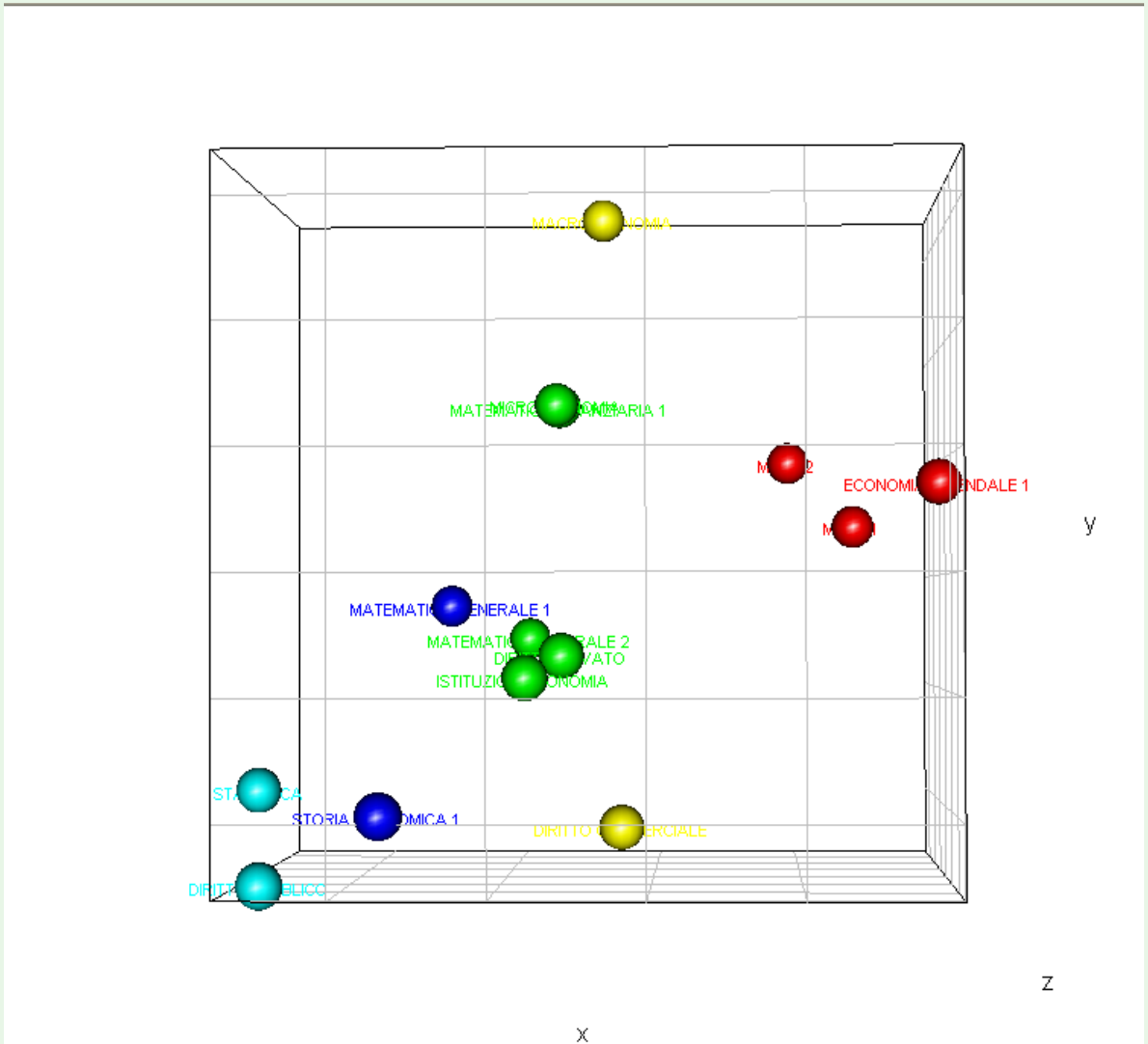
Examples of application: the UniMC data set

Dendrogram of quantified attributes



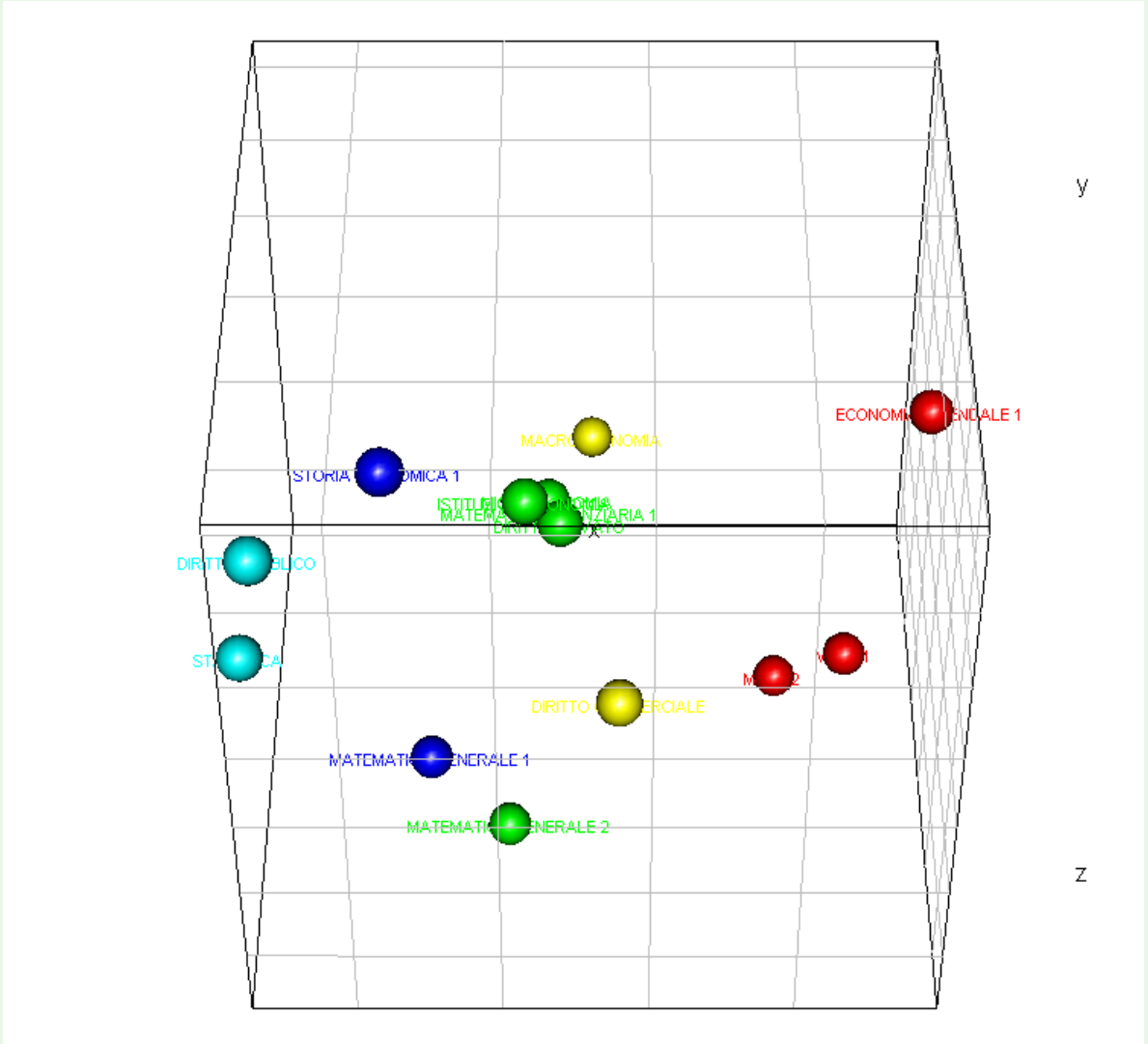
Examples of application: exogenous information approach

X vs Y



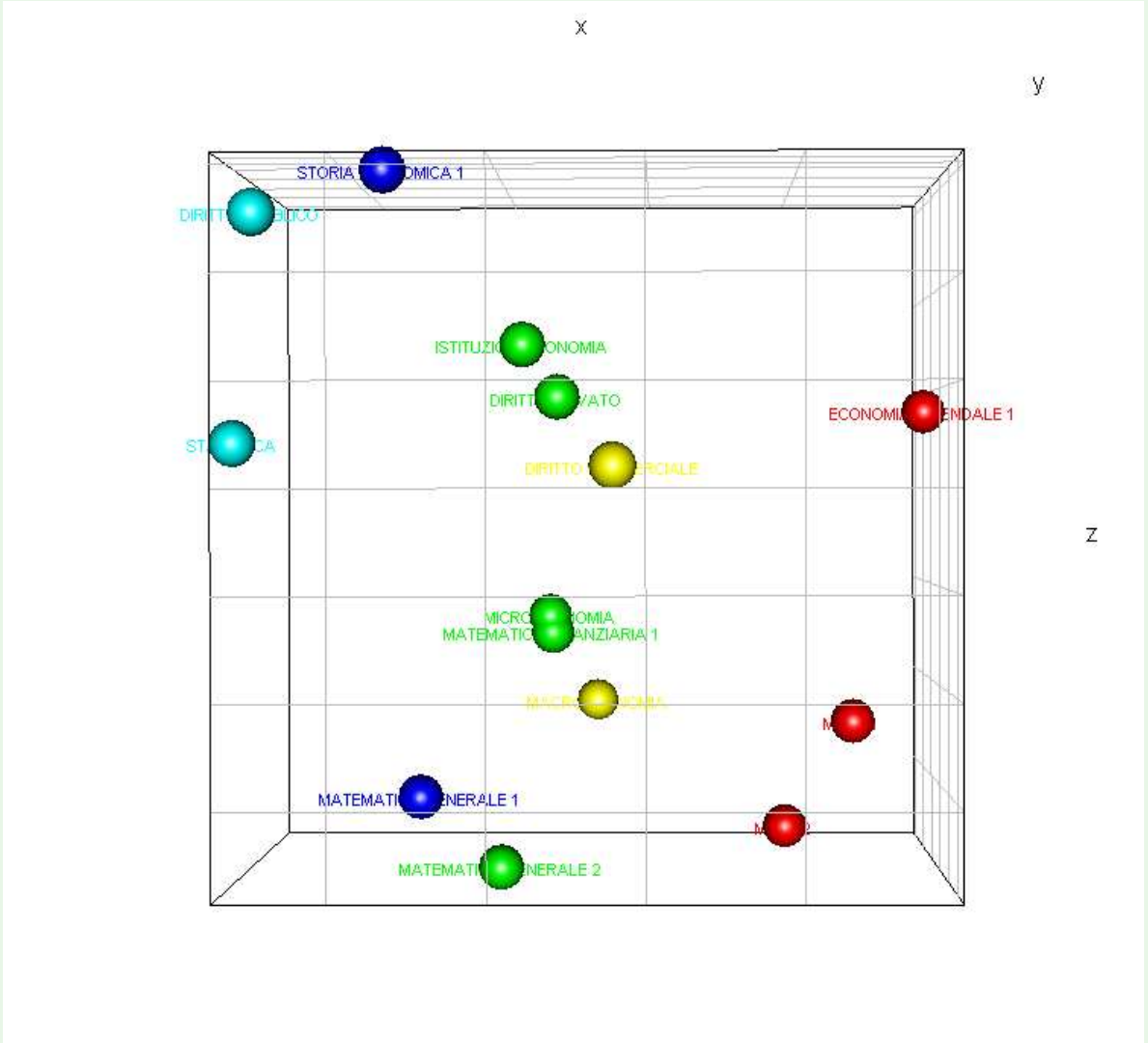
Examples of application: exogenous information approach

X vs Y $\rightarrow$ X vs Z



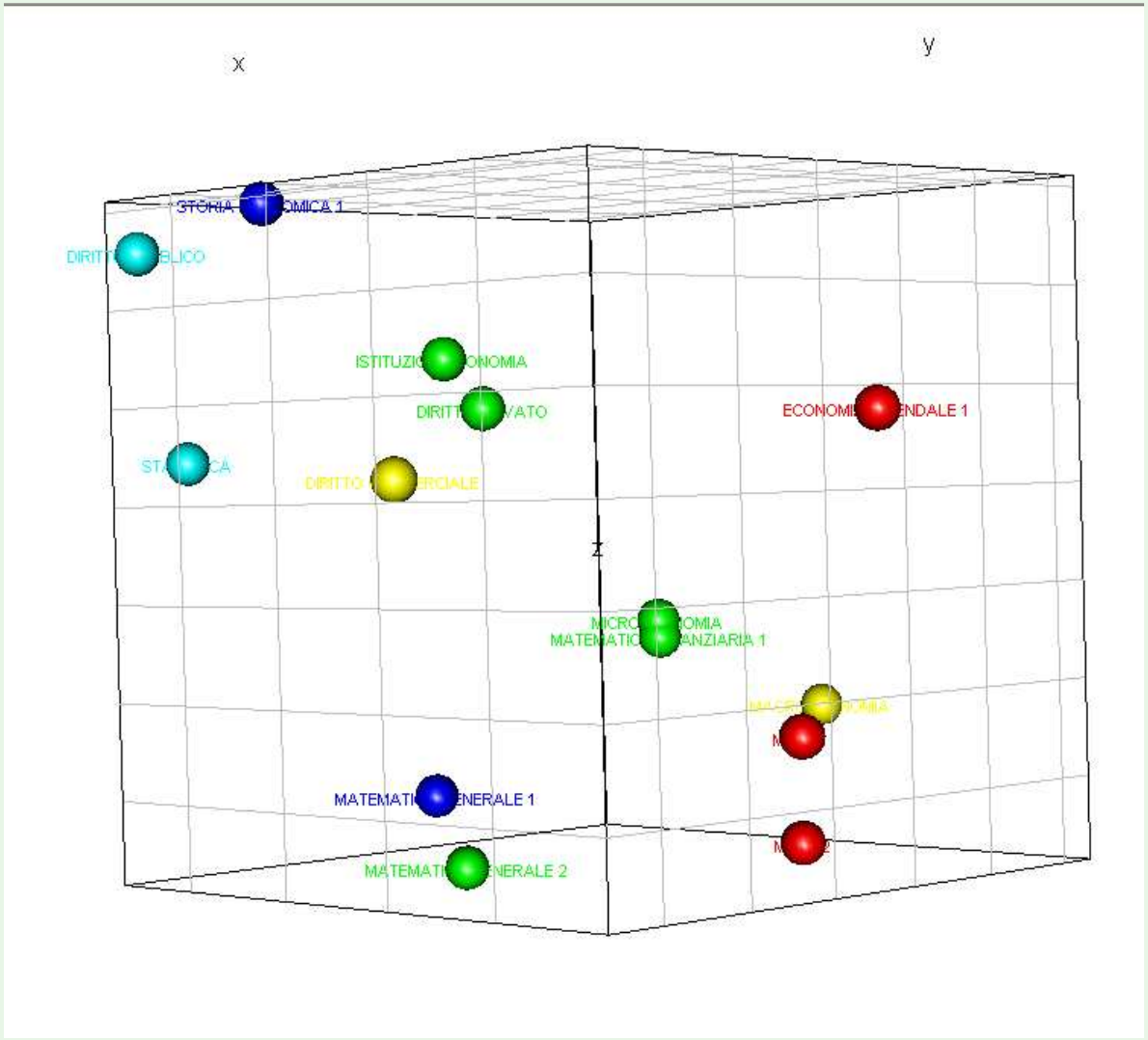
Examples of application: exogenous information approach

X vs Z



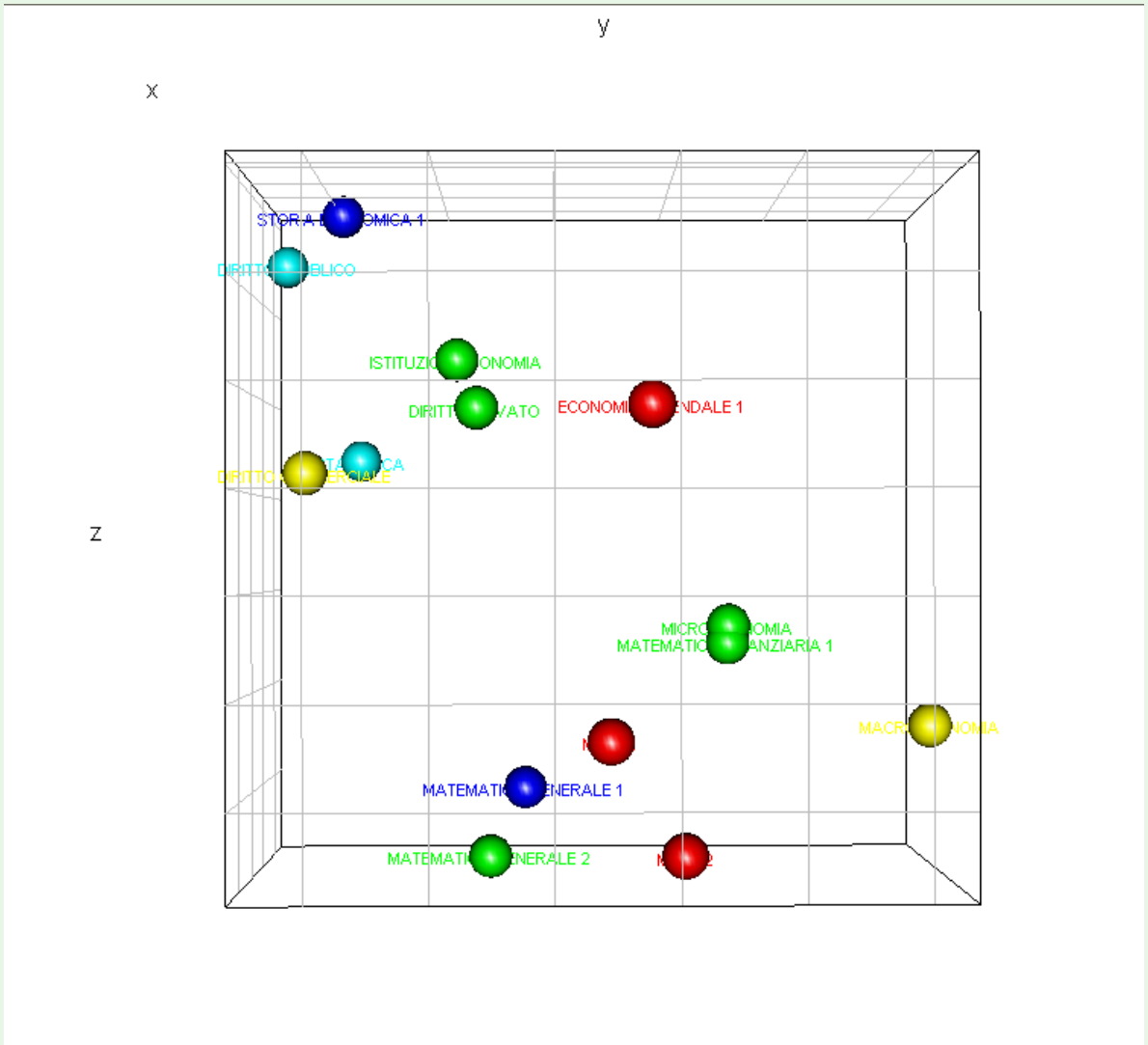
Examples of application: exogenous information approach

X vs Z $\rightarrow$ Y vs Z



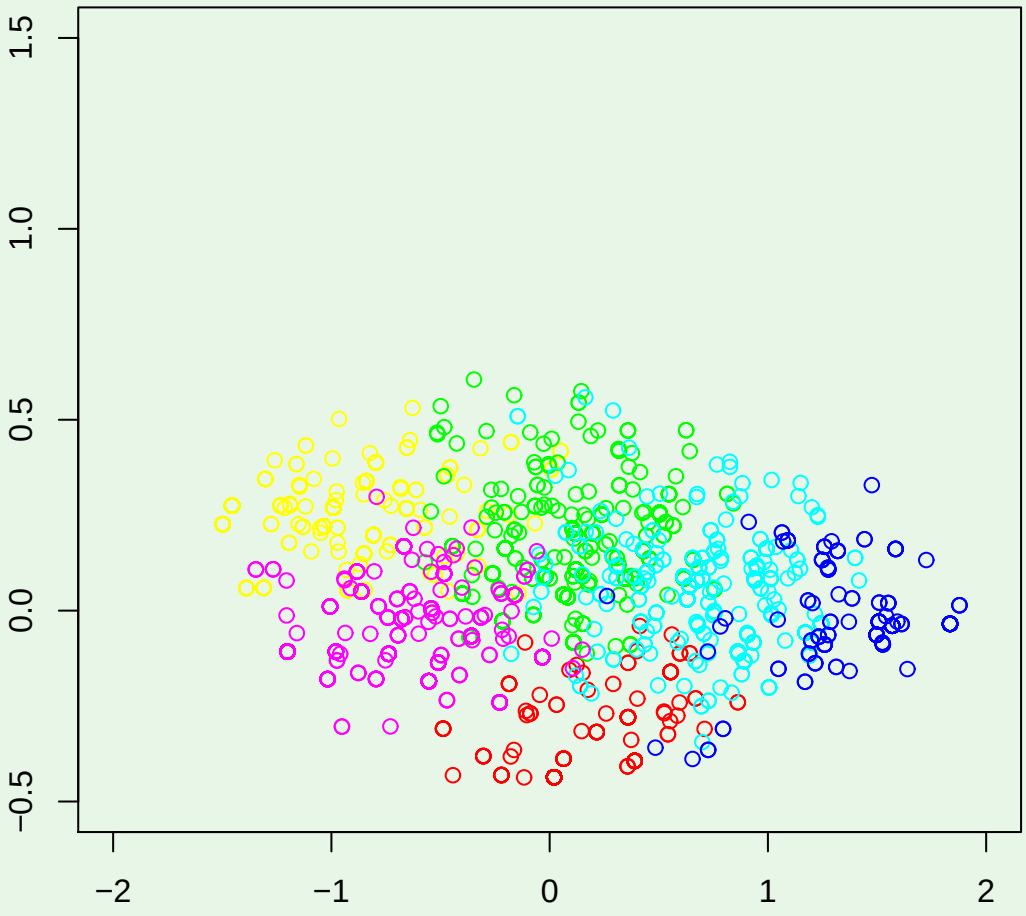
Examples of application: exogenous information approach

Y vs Z



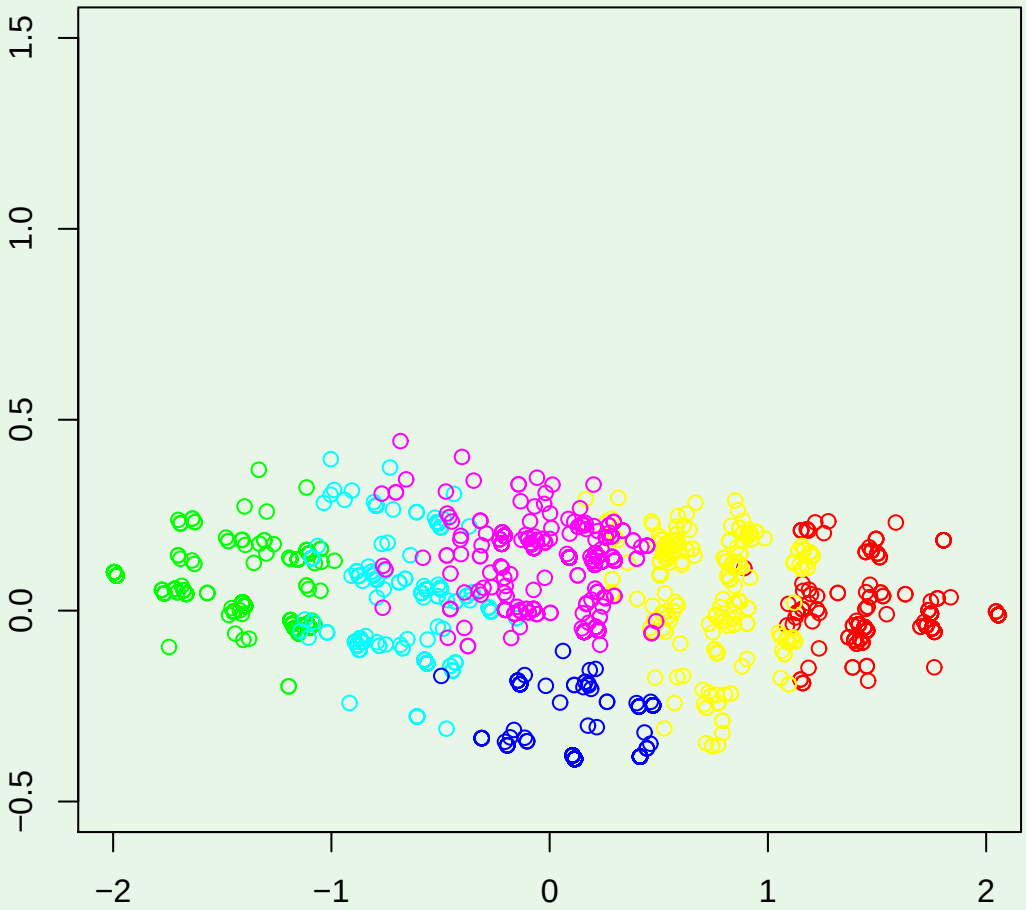
Examples of application: endogenous information approach

Sequences display: iteration 1



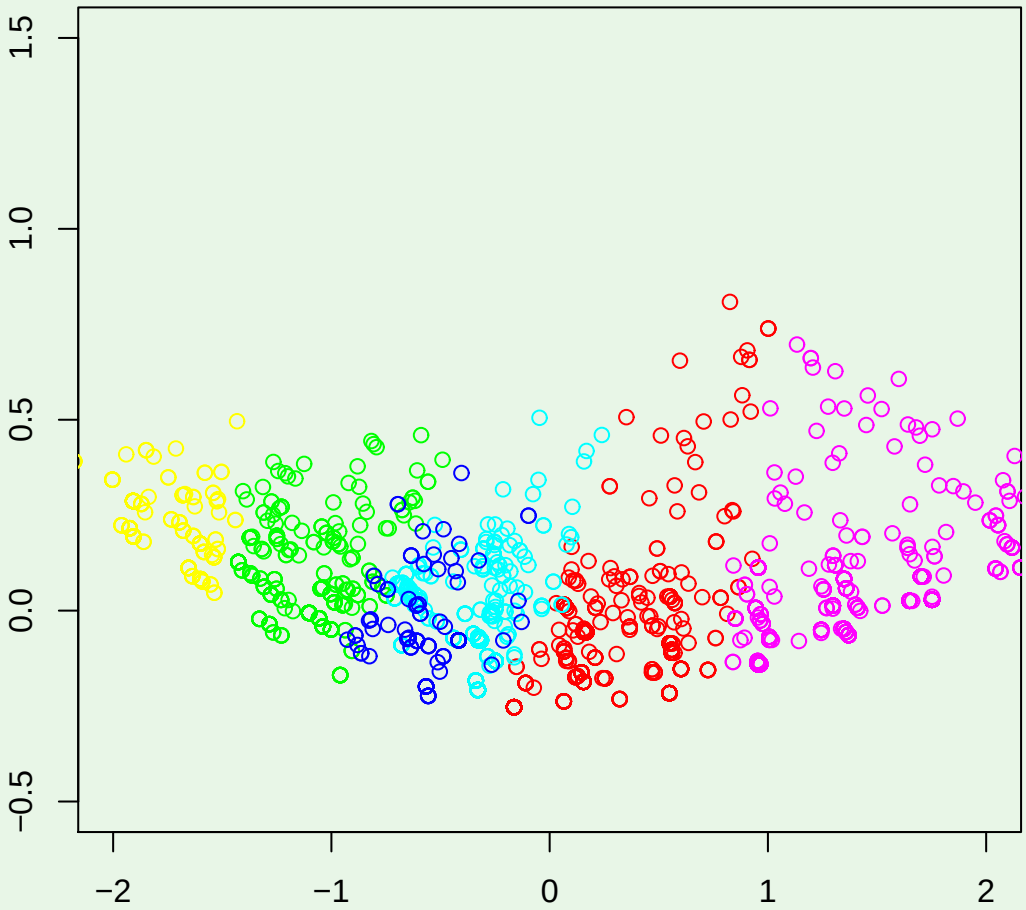
Examples of application: endogenous information approach

Sequences display: iteration 2



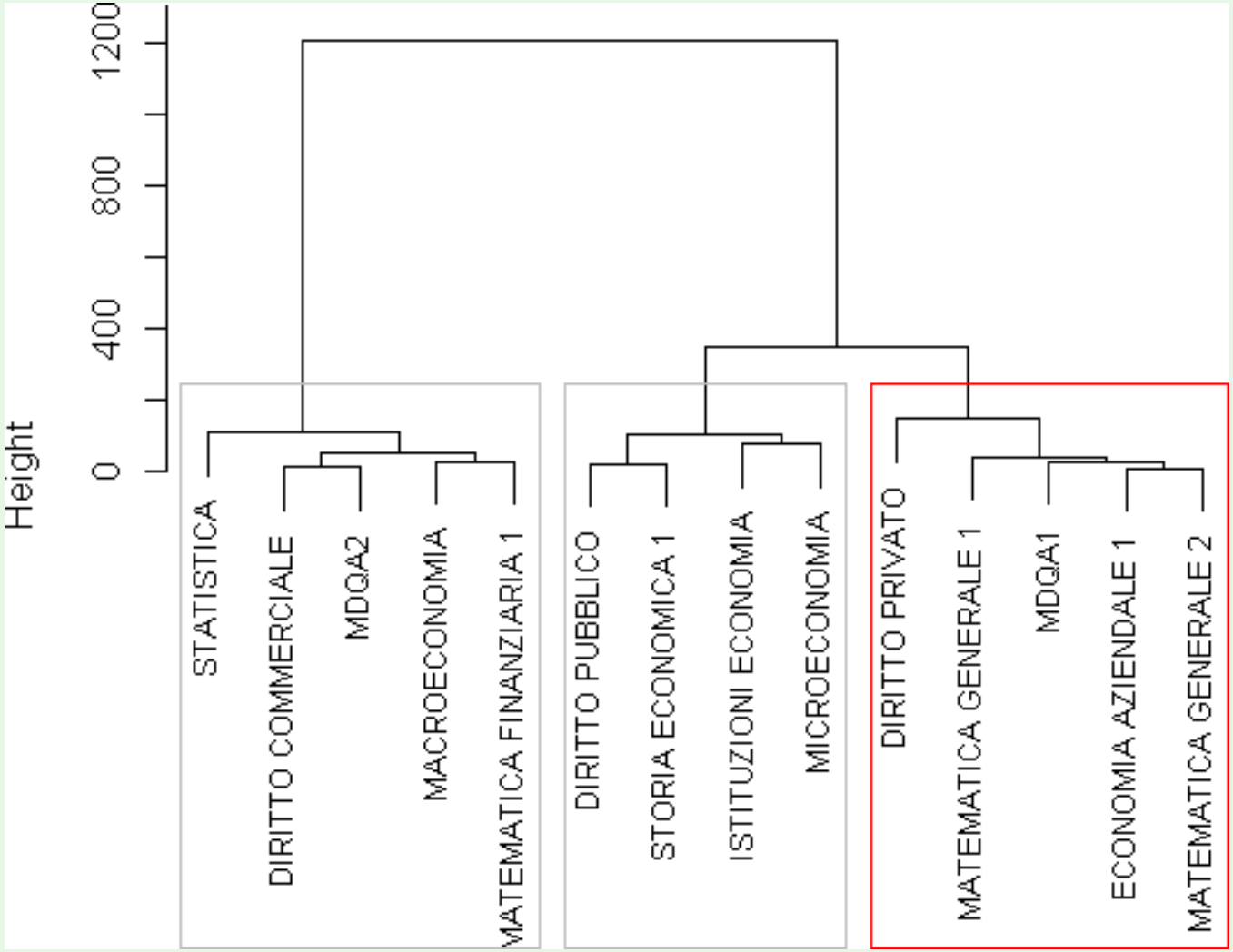
Examples of application: endogenous information approach

Sequences display: iteration 3



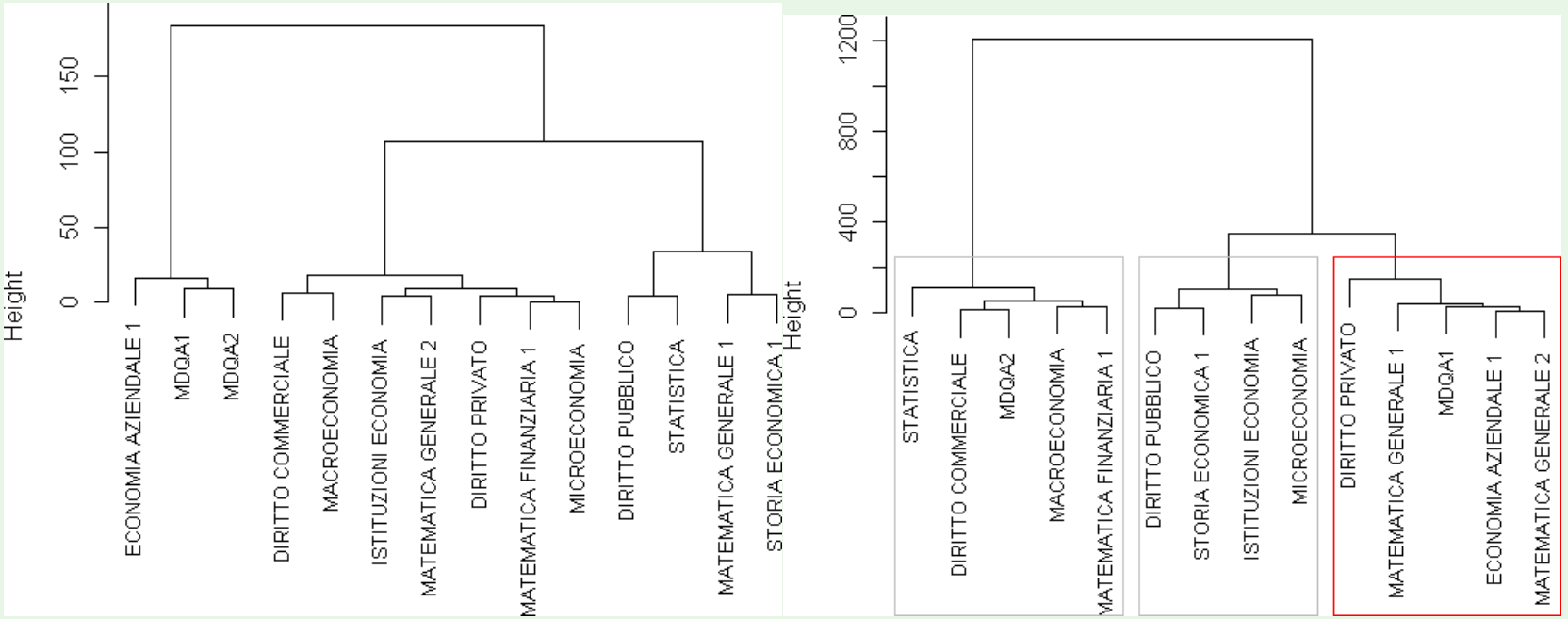
Examples of application: the UniMC data set

Dendrogram of quantified attributes



Examples of application: the UniMC data set

Dendrogram of quantified attributes



D. Adler and D. Murdoch, (2009). 'rgl: 3D visualization device system (OpenGL)'. R package version 0.84. http://CRAN.R-project.org/package=rgl.

M. J. Greenacre, (2007). 'Correspondence Analysis in Practice, second edition'. *Chapman and Hall/CR*.

T. Hastie, R. Tibshirani and J. H. Friedman, (2001). 'The Elements of Statistical Learning', *Springer*.

H. Hwang, et al., (2006) . 'An extension of multiple correspondence analysis for identifying heterogenous subgroups of respondents'. *Psychometrika* .

A. Iodice D'Enza, F. Palumbo & M. Greenacre, (2007). 'Exploratory Data Analysis Leading towards the Most Interesting Simple Association Rules'. *Computational Statistics and Data Analysis* **doi:10.1016/j.csda.2007.10.006** .

N.C. Lauro and L. D'Ambra, (1984).
L'analyse non symmétrique des correspondances.
In E. Diday et al., eds, *Data Analysis and Informatics, III*. North-Holland.

P. Lenca, et al. (2006). 'Association rule interestingness measures: e ntal and theoretical studies'. pm-pp-06-06-v01, ENST, Bretagne.

J. MacQueen, (1967). 'Some methods for classification and analysis multivariate observations'. In L. M. L. Cam & J. Neyman (eds.), *Proceedings of*

the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press.

F. Palumbo & R. Siciliano, (1999). 'Factorial Discriminant Analysis and Probabilistic Models'. In *Metron*,**MLI**,pp.185–198.

M. Plasse, et al. (2007). 'Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set'. *Comput. Statist. Data Anal.* **doi: 10.1016/j.csda.2007.02.020**.

R Development Core Team (2009). 'R: A language and environment for statistical computing'. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

M. Vichi & H. Kiers (2001). 'Factorial k-means analysis for two way data'. *Computational Statistics and Data Analysis* (37):29–64.

M. Vichi & G. Saporta, (2009). 'Clustering and disjoint principal component analysis'. *Computational Statistics and Data Analysis* (53) **doi: 10.1016/j.csda.2008.05.028** .