A REGULATORY IMPACT ANALYSIS (RIA) APPROACH BASED ON EVOLUTIONARY ASSOCIATION PATTERNS

Alfonso Iodice D'Enza

Dipartimento di Scienze Economiche, Università di Cassino
Via S. Angelo Folcara, Cassino (ITALY)
iodicede@gmail.com

Francesco Palumbo

Dipartimento di Istituzioni Economiche e Finanziarie,
Università di Macerata, Via Crescimbeni 20, Macerata (ITALY)
palumbo@unimc.it

Abstract

The present paper focuses on ex post analysis to asses the impact of an adopted policy by measuring system performance. Since accurate impact assessment requires in-depth knowledge of the structure underlying the system, this contribution proposes a suitable use of multidimensional data analysis (MDA) to investigate the associations characterizing the indicators/attributes of the system. The general aim is to identify homogeneous subsets of objects that are described by subsets of attributes. This approach was planned to study students performance in Italian universities: the focus is on student careers. The example data set is a data mart selected from the University of Macerata data base and refers to the students at the Economics Faculty from 2001 to 2007.

Keywords: *Regulatory Impact Analysis, Association Study, Correspondence Analysis, Supplementary projection*

1. INTRODUCTION

Regulatory policies aim to improve the overall performance of a system by implementing specific actions. Any policy takes up human and financial resources and presupposes a positive impact on the system where it has been put into practice. If the effects of regulation are important to the policy maker, it should be possible to identify the effects of regulatory policies by measuring the system reaction. In this direction, policy makers need detailed information about the system. Regulatory Impact Analysis (RIA) refers to all activities to evaluate policy effects. It is a very hard task to find an accepted definition of RIA. Scholars can only agree on a broad definition: RIA aims to evaluate whether and how far the regulatory policy advances the objectives of the overall regulatory process. In other words, what is clear is that RIA output is not necessarily restricted to one or more quantitative indexes. A multivariate analysis of the whole system is suitable to assess changes due to the regulatory policy. Statistical analysis and quantitative mathematical models can help the analyst to implement RIA, which can be performed *ex ante* or *ex post*: the former refers to the analysis and the studies done before the policies are put into practice; the latter refers to analyses performed after the regulatory action.

The present paper focuses on *ex post* analysis to assess the impact of an adopted policy by analysing the changes in the system structure. In statistical analysis, complex economic and social systems are described by a set of continuous or qualitative variables (binary and categorical) and by the association among such variables. Studying the changes, which are assumed to be due to the regulatory policy, in the average values, in the variability and the association among these variables permits the regulatory impact and hence the policy effectiveness to be assessed. This paper proposes a multidimensional data analysis (MDA) approach to investigate the association structure in a given set of attributes. MDA techniques represent a mostly graphical exploratory tool to detect correlation/association structures in multivariate data sets. In dealing with continuous variables Principal Component Analysis (PCA, (Jolliffe, 2002)) represents the best known and most widely used method to synthesise and graphically display a set of variables. The graphical representations allow us to analyse and interpret the correlation structure underlying a multivariate quantitative data set. Correspondence Analysis (CA (Greenacre, 2007)), with the help of the graphical representations, as for PCA, aims to study multiple associations characterizing categorical attributes. The key element in the success of MDA techniques is the strong contribution of visualization: it exploits the human ability to perceive 3-D

space. Cartesian spaces permit one to visualize positions of a set of dimensionless points. The concepts of far and close are innate concepts. It is not necessary to be a mathematician to understand them. Distance represents the measure of closeness in space (Palumbo *et al.* 2008).

Given a complex social or economic system described by a set of p attributes $X_1, X_2, \ldots, X_p$, and assuming that attributes values can be recorded before and after the regulatory intervention, the method described in the present proposal aims to detect patterns of change in the attribute association. This approach was implemented to study Italian university student performance: in particular, the focus is on student careers. The example data set is a data mart selected from the University of Macerata (UniMc) data base from 2001 to 2007, and refers to students at the Economics Faculty.

The present paper is structured as follows: section 2 introduces the basic notation and definitions; in section 3 the approach to the association study of partitioned data is explained: in particular, basics of CA are given in section 3.1, whereas in section 4 a comparative association-based quantification is described. Section 5 shows an application on real data to analyse the policy effects.

2. PRESENTATION OF THE PROBLEM

In the Italian higher education (HE) system, students are fairly free to choose in what order to take examinations (this is mostly true in the case of social sciences faculties). However, they are only allowed to take the examination after the course has been taught. In the case of the Economic Faculty of the UniMc, students can take examinations in their current year even if they have not yet completed the previous year's examinations. In other words, second year students can start taking second year examinations without completing those of first year. These regulations lead most of the students to defer those examinations they feel more difficult to pass. Although this behaviour results in a larger number of examinations being passed at first, it causes an undesired effect in their whole career. By so doing, in most cases students defer the study of economic foundations to the end. The Economics Faculty of UniMC, in the academic year 2002/2003 incorporated the Ministerial Decree n. 509 (November 1999), that changed the entire organization of university courses where the Bachelor-Master[1] structure was adopted. At the same time, the Italian system introduced *modules*, the base unit for the accumu-

[1] *In Italian Master stands for a postgraduate diploma, whereas Laurea magistrale (or specialistica) corresponds to the European Master diploma.*

lation of credits, which in turn are the common unit of measurement across the national HE system. It should be noted that the expected *output* of the teaching process was radically modified by the reforms: no bachelor degrees nor credit system existed. The Bachelor-Master structure encourages students to be on schedule with their career. Moreover, the Economics Faculty introduced a bonus system for students completing the examination on schedule, in each year. Our expectation was to find changes in the students' careers in terms of their behaviour in taking examinations.

In the Italian HE system comparisons between the students' behaviour before and after the regulatory intervention proves very difficult due to the adoption of *modules* and the credit system since there is no equivalence between courses before and after the intervention. However, the Economics Faculty of UniMc is a very special case in the Italian HE system: it introduced *modules* in the academic year 1999/2000, thanks to the Italian HE autonomy principle. Thus it is possible and effective to make comparisons of students' behaviours before and after the reform. The comparison is implemented focusing only on the students' careers. A student career is coded as a binary sequence where the value 1 indicates whether the student passed the examination. Available data refer to academic years from 2001/02 to 2006/07, both for the second and third year students: data are coded according to the same scheme for each occurrence. Here the quantities considered throughout the rest of the paper follow:

– I number of binary sequences;
– J number of attributes/examinations;
– $\mathbf{N}$ $(I \times J)$ is a binary data matrix; n_{ij} indicates the presence of attribute j in sequence i. In other words, it represents a cross-tabulation of two qualitative attributes with I and J levels, respectively.

3. STUDY OF ASSOCIATION FOR PARTITIONED DATA

Identification of patterns of association in binary data is based on a profitable use of CA. In a way, the application of CA can be interpreted as a quantification of qualitative data: this approach was introduced by (Saporta, 1975). The quantification consists of a projection of starting variables on an orthogonal subspace: the number of chosen dimensions determines the degree of synthesis provided by quantification. The advantages in using CA to study associations of binary data are then to remove noise and redundancies in data as well as obtain a synthetic representation of multiple associations characterizing attribute levels.

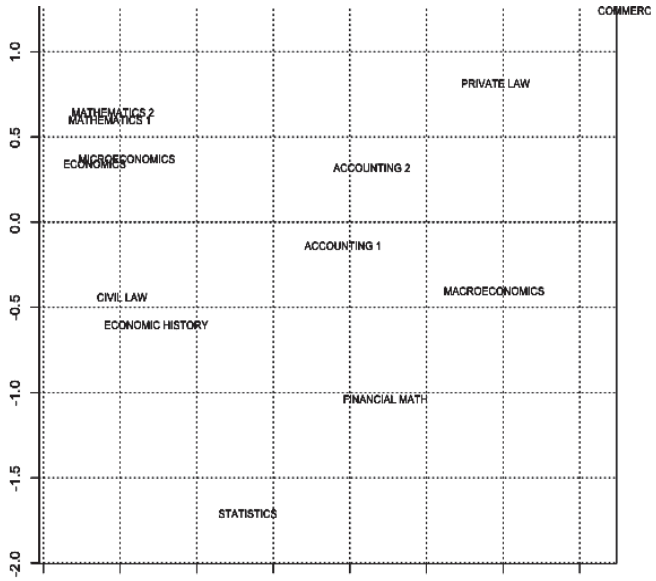

Fig. 1: Second year Students: reference map 2006/07.

3.1 Basics of CA

Implementation of CA consists of a few transformations of the frequency matrix **N** and of a singular value decomposition (SVD). The correspondence matrix is $\mathbf{P} = \frac{\mathbf{N}}{grand\ total(\mathbf{N})}$, with row and column margins denoted by **r** and **c**, respectively. A reduced rank approximation of **P** is given by the SVD of its centered version **Q**, with general element

$$q_{ij} = \frac{(p_{ij} - r_i c_j)}{\sqrt{r_i c_j}},\ i = 1,\ldots,I; j = 1,\ldots,J. \tag{1}$$

The solution is obtained through a SVD $\mathbf{Q} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\mathsf{T}$, with **U** and **V** being the *left* and *right* singular vector matrices; $\mathbf{D}_\alpha$ is the diagonal matrix of singular values. The *principal co-ordinate* of the i^{th} row point on the s^{th} dimension is obtained through $f_{is} = a_{is}\alpha_s$, with a_{is} being the corresponding *standard co-ordinate*, that is $a_{is} = \frac{u_{is}}{\sqrt{r_i}}$, α_s being the s^{th} singular value and u_{is} being the i^{th} element of the corresponding singular vector. Using an algebraic formalization, the principal coordinates of rows and columns are $\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\alpha$ and $\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}_\alpha$, respectively.

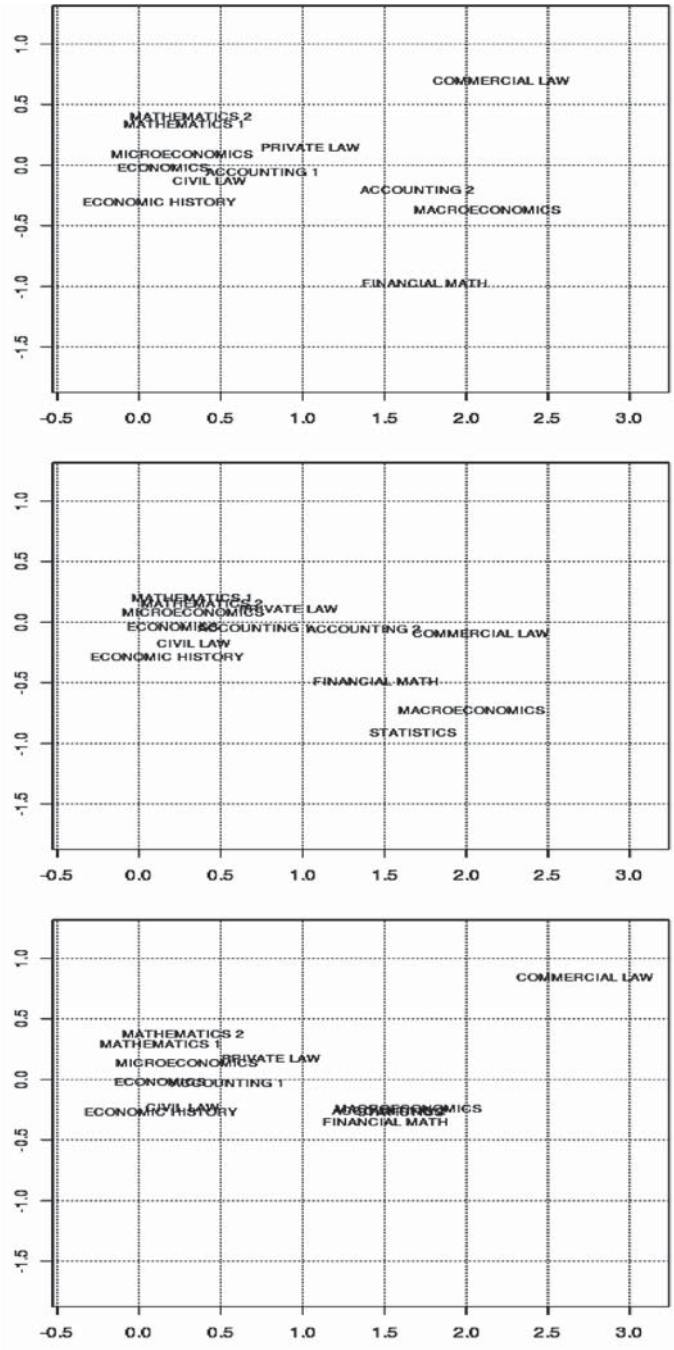

Fig. 2: Second year Students (before the intervention): from 2001/02 to 2004/05 vs. 2006/07

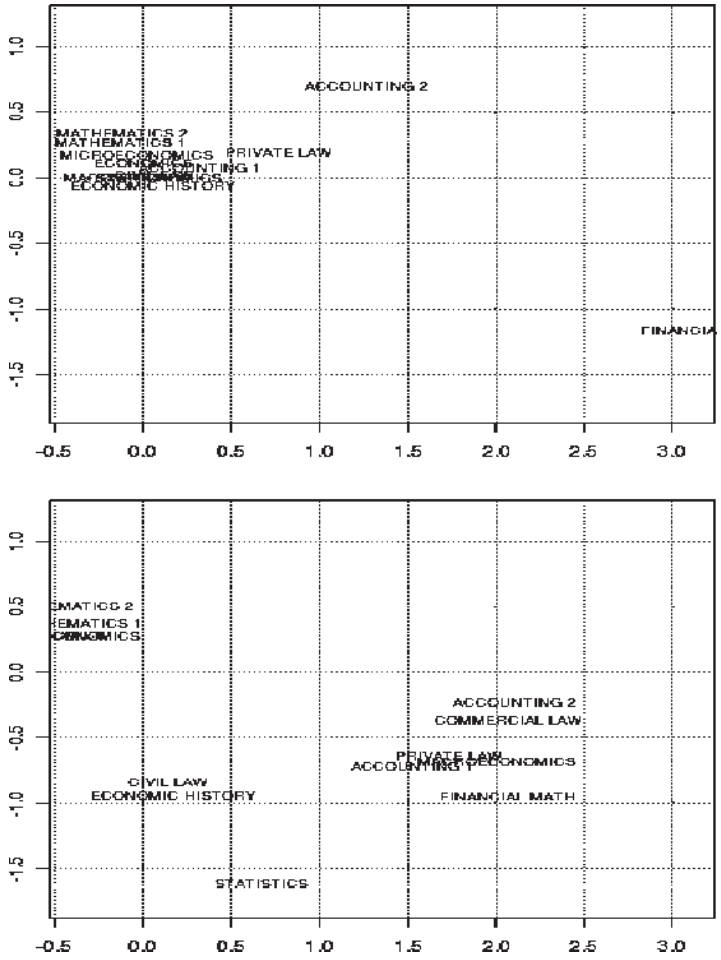

Fig. 3: Second year Students (after the intervention): from 2001/02 to 2004/05 vs. 2006/07.

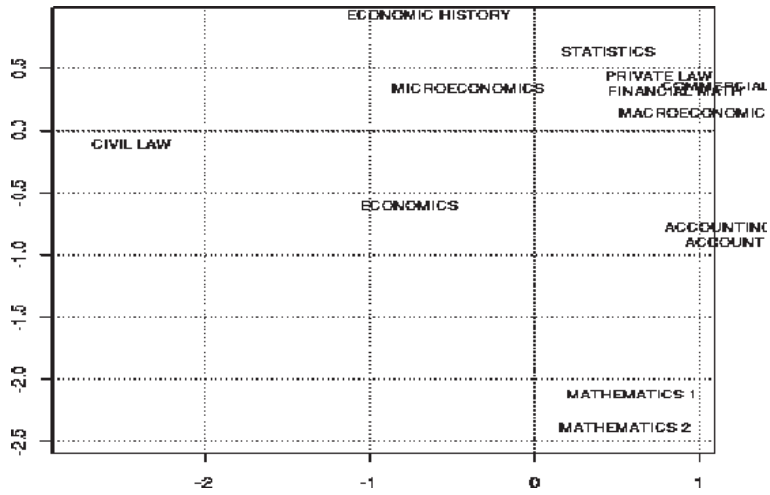

Fig. 4: Third year Students: reference map 2006/07.

4. COMPARISON OF ASSOCIATION STRUCTURES

The data in question are stratified with respect to the academic years considered. The aim is to compare the association structures along years. The CA output is the graphical representation of the multiple association structures characterizing the variables. In this context the choice falls on an MDA-based approach because students are assumed to plan their career by considering simultaneously all of the examinations. Disjoint analysis of the years in question would lead to independent solutions that cannot be compared directly.

In this paper, the considered years are analysed with respect to a suitably chosen reference year. The comparative analysis is obtained by exploiting two properties of the CA theory: the transition formulas and the supplementary projection of additional information. Note that both of these aspects depend on the linear relation characterizing the left and right singular vectors. In particular, transition formulas permit the row principal coordinates to be computed as a function of column principal coordinates and *vice versa*, formally:

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{G}\mathbf{D}_\alpha^{-1} \text{ and } \mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{P}^\mathsf{T}\mathbf{F}\mathbf{D}_\alpha^{-1}. \tag{2}$$

Furthermore, transition formulas are used to project additional rows (or columns) as supplementary information on the CA factorial display. Consider an additional J-dimensional row vector $\mathbf{k}$, and le t the sum of its elements be $(\mathbf{k1})$, where $\mathbf{1}$ is

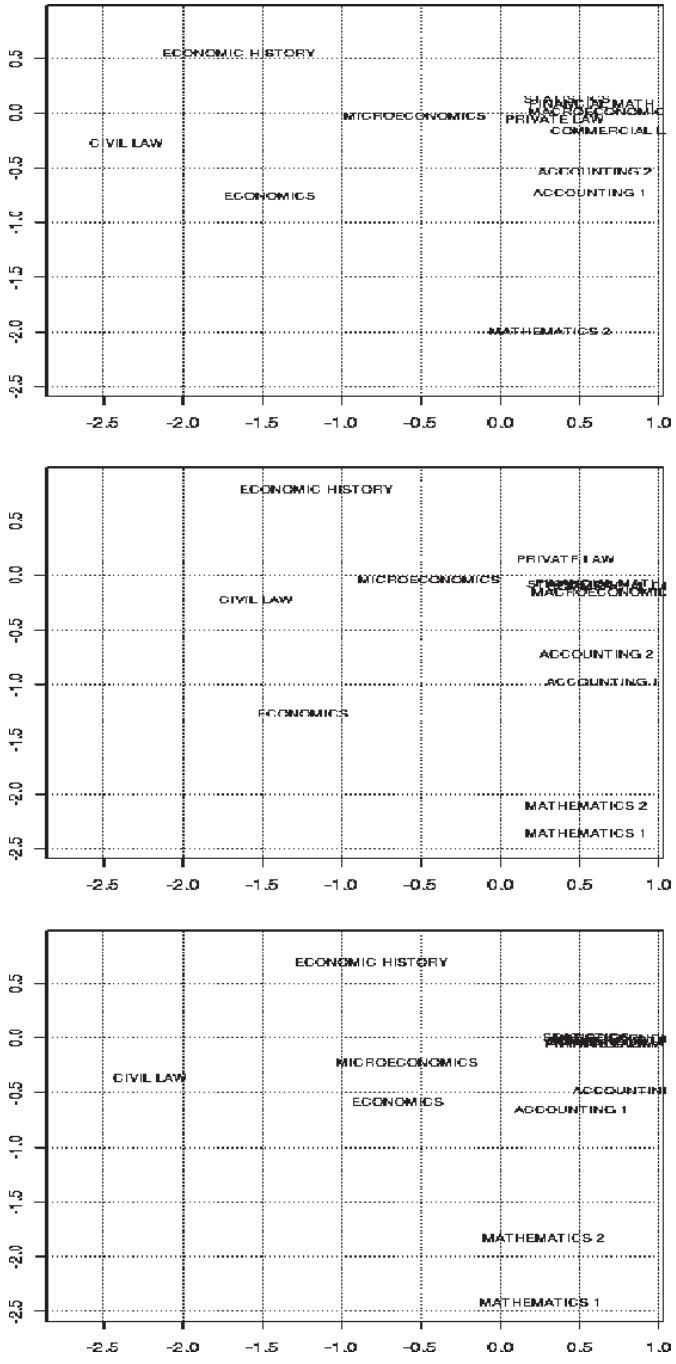

Fig. 5: Third year Students (before the intervention): from 2001/02 to 2004/05 vs. 2006/07.

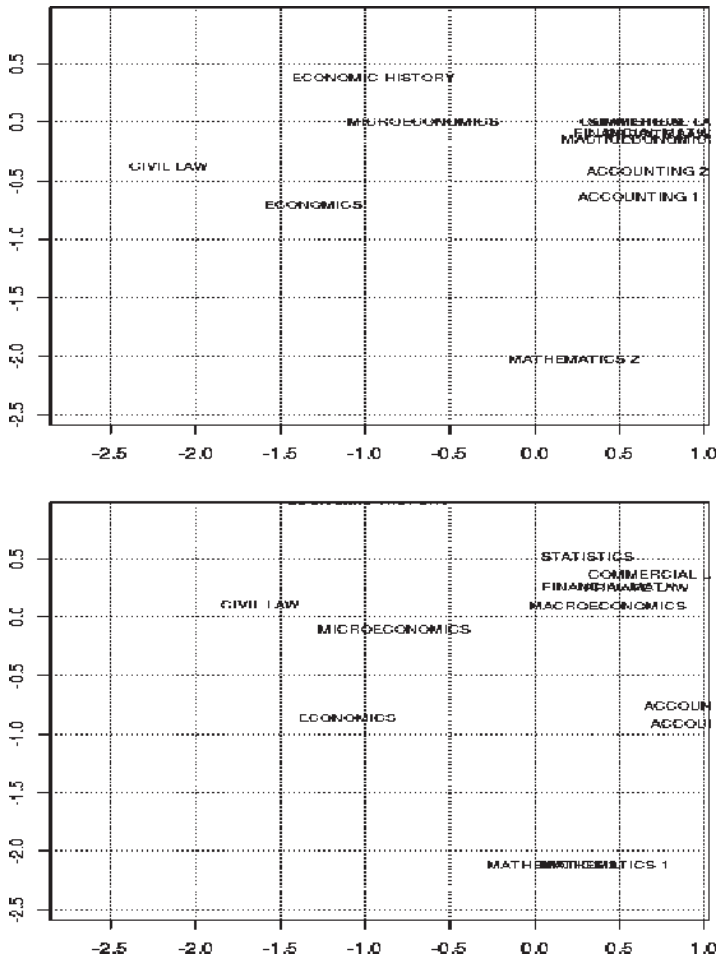

Fig. 6: Third year Students (after the intervention): from 2001/02 to 2004/05 vs. 2006/07.

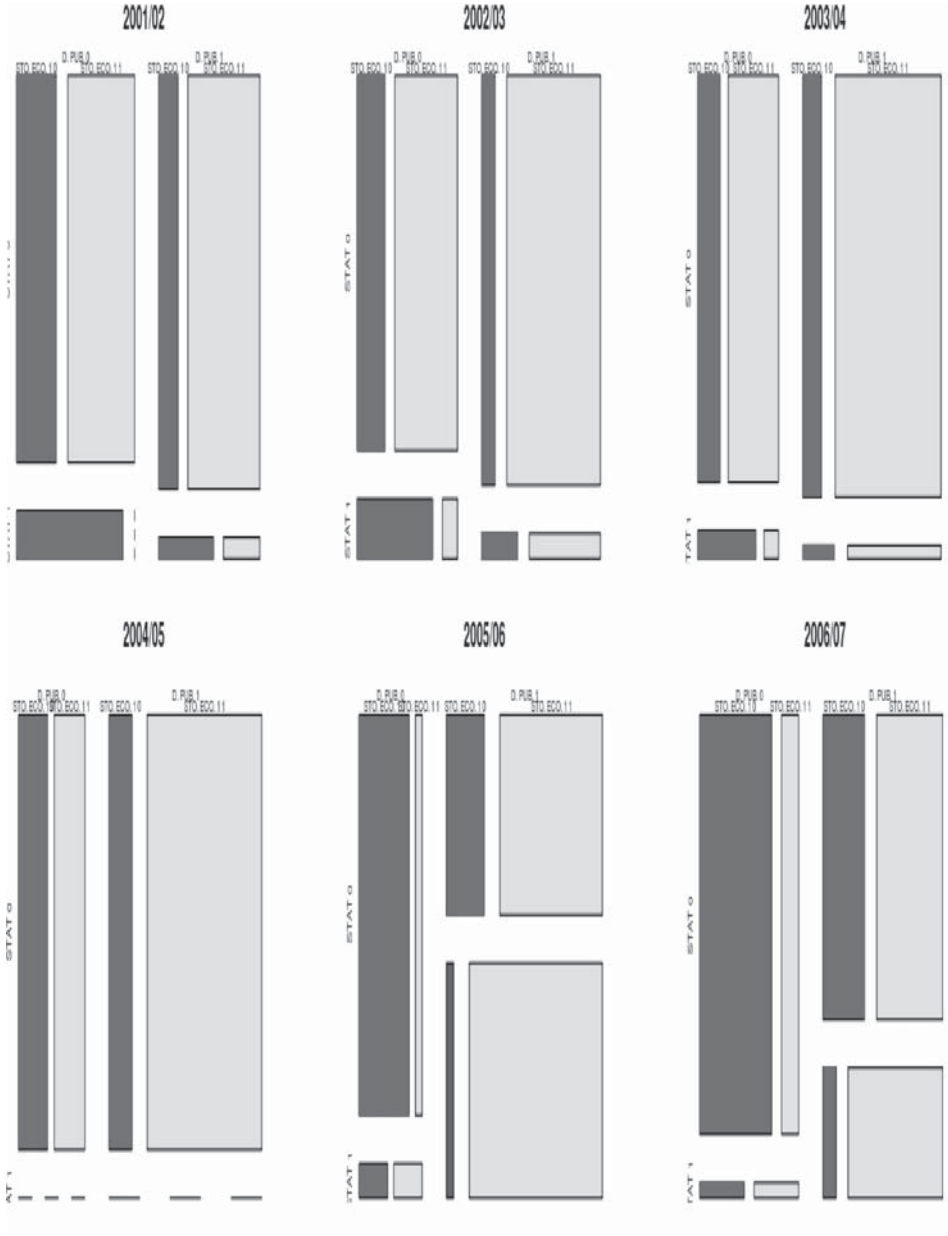

Fig. 7: Mosaic plot: Civil Law vs economic History vs Statistics.

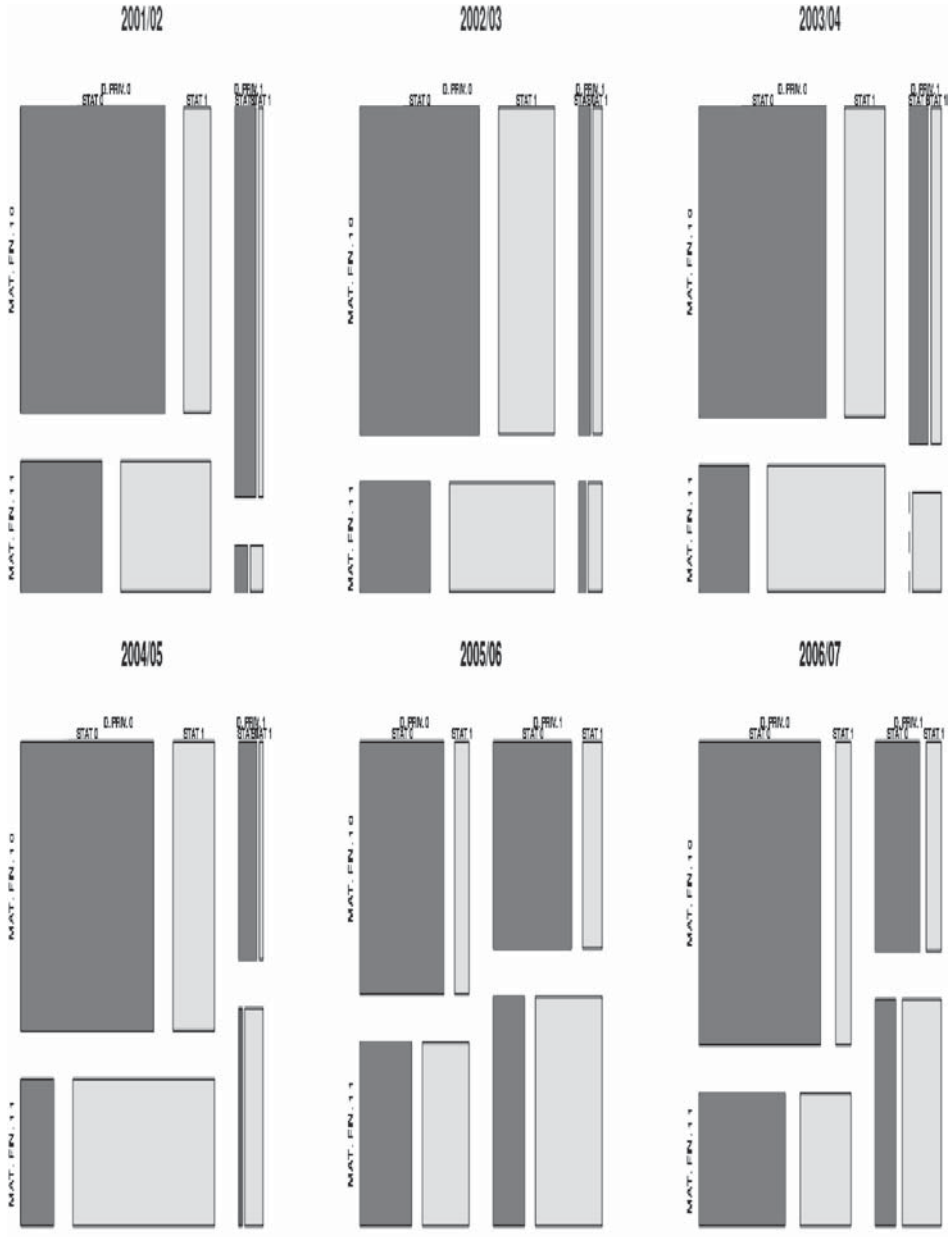

Fig. 8: Mosaic plot: Private Law vs Finantial Maths vs Statistics.

a $(J \times 1)$ vector with all elements equal to one. The corresponding row profile is $\tilde{\mathbf{k}} = \left(\frac{1}{\mathbf{k}\mathbf{1}}\right)\mathbf{k}$. Then the position of the supplementary row point on the CA map is given by

$$\mathbf{f}^{+} = \tilde{\mathbf{k}}\mathbf{D}_c^{-1/2}\mathbf{V}. \tag{3}$$

Both of the concepts above are used to obtain a visualization of the association structure in each academic year with respect to a chosen reference year. Let $\mathbf{N}$ be the $(I \times J)$ matrix containing student careers at the reference year and let $\mathbf{N}^{+}$ be the $(I^{+} \times J)$ matrix of the academic year to be compared. The additional rows in $\mathbf{N}^{+}$ are quantified according to equation 3: thus the $\mathbf{F}^{+}$ matrix of supplementary row coordinates is obtained. Since rows correspond to students, it is more interesting to quantify and visualize the columns of $\mathbf{N}^{+}$, corresponding to the examinations. In order to obtain the column coordinates $\mathbf{G}^{+}$, the transition formulas (see equation 2) are used; in particular

$$\mathbf{G}^{+} = (\mathbf{D}_c^{+})^{-1/2}(\mathbf{P}^{+})^{\mathsf{T}}\mathbf{F}^{+}(\mathbf{D}_\alpha^{+})^{-1}. \tag{4}$$

Notice that the '+' symbol indicates that the corresponding quantities, all defined in section 2, are computed with respect to the additional information.

5. EXAMPLE OF APPLICATION

The considered data are a data mart selected from the UniMc data base. In particular, data refer to the students of the Economics Faculty at UniMc. Second and third year students are analyzed separately; the considered time range goes from academic years 2001/2002 to 2006/2007 . As stated above, the Economics Faculty introduced the Bachelor-Master system in 2002/03, hence the impact of the regulatory policy can be already observed in 2003/04 for second year students and 2004/05 for third year students. As the credits system had already been introduced, 14 foundation course (eight for the first year and six for the second year) remained unchanged after the policy intervention. The aim is to assess student reaction to the regulatory policy. The generic considered binary record refers to a single student career: if the student passed an examination, then the corresponding field is '1', otherwise it is '0'.

The focus is on how the students' behaviour evolved: the aim is to ascertain whether the regulatory policy had any effect on student attitude to and capacity to pass the foundation course examinations. In order to do this, consider the reference year to be 2006/07, a CA solution is obtained on second and third year

students, respectively. The reference low-dimensional space is obtained according to the association structure characterizing 2006/2007 students: previous years' associations are then projected as supplementary information (see details in section 4).

From the top to the bottom of figure 2, the factorial maps represent the association structure of the attributes/examinations for each of the considered academic years (2001/02,...,2003/04) with respect to the last (2006/07, figure 1); second year student behaviours before the intervention are considered. Figure 3 shows the second year student behaviors after the intervention (years 2004/05 and 2005/06). What clearly emerges is the change in the association structure after the introduction of the regulatory policy (2003). Indeed, figure 2 maps are characterized by a global stability in the association structure, whereas the top map in figure 3 shows an evident change in the association pattern: most of the examination points form a circular manifold located at the centre of the map. This reflects a certain disorientation in the student behaviours. The bottom map (2005/06) shows a reaction of the students and a re-arrangement of their behaviours, due to the intervention.

Figures 5 and 6 are the same as figures 2 and 3 in all but the reference students, since third year students are considered. In this case the overall difference between the first three maps and the last two is not that evident: this is due to older students being considered. The fundamental examinations in question are all scheduled in the first two years of studies: thus, before *and* after the intervention, students are supposed to have already passed all the considered examinations.

The representation of association structure on the factorial displays is now compared to a further tool for the visualization of associations, the mosaic plot (Hartigan and Kleiner, 1984). Mosaic plots are a straightforward way to display multi-way contingency tables. In particular, joint frequencies are represented by patches: the larger the patch, the higher the co-occurrence of the two considered modalities. When the number of considered attributes increases, the readability of the mosaic plot deteriorates quite rapidly. Thus, we consider a mosaic plot representation of attributes for second and third year student data. The choice of the attributes to display depends on the factorial map representations: with respect to the last year factorial map, attributes that are both far from the centre and close to each other are selected. According to this criterion, the considered attributes for second year data are Civil Law, Economic History and Statistics; the considered attributes for third year data are Private Law, Financial Maths and Statistics. In both figures 7 and 8 two aspects emerge. The overall association structure varies

if considered before and after the policy intervention, as the shapes of the patches show. Furthermore, consider the bottom-right corner patch, which corresponds to the proportion of students that passed all three examinations in question. This patch is considerably larger in 2005/06 and 2006/07 than in previous years. It may be concluded that the associations pointed out by the factorial maps are consistent with the mosaic plot representations; in addition, factorial maps permit all 14 attributes to be considered simultaneously, whereas a mosaic with five or more attributes would be unhelpful as it would be too hard to interpret.

REFERENCES

GREENACRE,M.J., 2007. Correspondence Analysis in Practice, second edition. *Chapman and Hall/CR.*

HARTIGAN J.A. and KLEINER B.,1984. A mosaic of television ratings. *The American Statistician*, **38**, 32-35.

JOLLIFFE, I.T., 2002. Principal Component Analysis. *Springer*.

IODICE D'ENZA A. and PALUMBO F., 2007. Binary data flow visualization on factorial maps. *revue Modulad*, n. 36.

MACQUEEN J., 1967. Some methods for classification and analysis of multivariate observations. *5th Berkeley Sym. on Mathematical Statistics and Probability procs.*, L.M. Le CAM and J. NEYMAN, eds, Univ. of California Press.

PALUMBO, F., VISTOCCO D., MORINEAU A., 2008.Hugemultidimensional data visualization: back to the virtue of principal coordinates and dendrograms in the new computer age. *Handbook of Data Visualization*, Chen C., Hrdle W., Unwin A. eds. Springer-Verlag, pp. 3-44.

SAPORTA G., 1975. Liason entre plusieurs ensembles de variables et codages de données qualitatives. *Thése de III cycle, Univ. de Paris VI*, Paris.

STUDIO DELL'EVOLUZIONE DI STRUTTURE ASSOCIATIVE PER L'ANALISI DI IMPATTO DELLA REGOLAMENTAZIONE

Riassunto

L'obiettivo del presente lavoro è valutare ex-post l'impatto di una politica di intervento misurando opportunamente la variazione di performance di un sistema oggetto di studio. In particolare, si propone di utilizzare opportune tecniche di analisi multidimensionale dei dati (AMD) per studiare le strutture di associazione che caratterizzano l'insieme di attributi che descrive il sistema. L'impatto della politica di intervento viene pertanto valutato in base alle variazioni nella struttura associativa alla base del sistema.

Il contesto applicativo in cui tale approccio è stato sviluppato è l'analisi delle performance degli studenti universitari attraverso il monitoraggio delle loro carriere. I dati analizzati fanno riferimento alle carriere degli studenti della Facoltà di Economia dell'Università di Macerata dal 2001 al 2007.