

The role of textual data in a statistical approach for the evaluation of the regulatory impact

Simona Balbi, Germana Scepi and Giorgio Infante

Dipartimento di Matematica e Statistica – Università “Federico II” di Napoli

Via Cintia, Monte Sant’Angelo, 80127 Napoli, Italy; sb@unina.it; scepi@unina.it

Abstract

The evaluation of Public Intervention is currently considered a strategic element of political and administrative actions. The first step for evaluating public policies consists in applying the Regulatory Impact Analysis (RIA). Usually, an economical and/or a juridical viewpoint is adopted. The paper shows how quantitative methods can be useful for an *ex-ante* evaluation of public actions having a significant impact on people lives. A statistical approach based on Conjoint Analysis has been proposed by Scepi, *et al.* (2008). Thanks to this technique, developed in the frame of market research, it is possible to compare different potential actions, described by several indicators, and to estimate utilities for groups of judges. It is well-known that one of the main problems in performing a Conjoint Analysis consists in data collection. The most common procedure is the so called *full profile* method, which forces to deal with a small number of categorical variables (and with a small number of categories). Being based on an experimental design, it is often too rigid. Moreover, in the frame of RIA, the choice of descriptors is difficult. Here we propose a procedure based on Conjoint Analysis with textual information (Balbi, *et al.*, 2008), where judges are asked to enrich their own choice with a free description of the desired action. An empirical analysis on alternative University systems in Italy shows the effectiveness of our proposal.

Keywords: Conjoint Analysis, RIA, Textual Information

1. Introduction

The activity of evaluation of Public Intervention, or Regulation activity (study of quality and efficiency of the intervention in terms of gap between the performance and the aims), is currently considered from public administration as a strategic element of political and administrative action.

Generally, a public intervention is evaluated, with suitable indicators, looking at the following dimensions: the Utility/Social Welfare (checking the incidence of the intervention on the satisfaction of needs), the Effectiveness (comparing realization indicators with indicators related to their goals), the Efficiency (checking financial resources, structural resources and human resources necessary to the achievement of goals as well as comparing the obtained results with the employed resources), the Pertinence (checking the adequacy of the specific aims and the realization mode with respect both to the real status and to the foreseeable changes of needs), the Sustainability (analysing the capacity of preserving in time the obtained results).

The evaluation process is, however, assumed to have a composite and complex nature and it must be referred to different dimensions of the analysis. Furthermore, some dimensions cannot be objectively measured and considered as “manifest” variables. Looking at the strategy of evaluation, the impact of a regulation can be evaluated with *ex-ante*, *in itinere* and *ex-post* strategies as well as to crossly check the different phases of the process.

The contribution of this paper is in the *ex-ante* evaluation. The methodologies usually applied in this context are: the Analysis of the Conformity Costs, the Cost-Effectiveness Analysis and, particularly, the Cost-Benefit Analysis. However, these methods give a partial evaluation, because they consider the problem only by economical and juridical perspectives. It becomes necessary to supply the P.A. with statistical methods able to support the evaluation activity, taking into account its intrinsic multidimensionality.

Scepi *et al.* (2008) propose the use of a statistical strategy based on Conjoint Analysis (CA). In market research, Conjoint Analysis has been developed for estimating the importance given by each customer to each characteristic describing a product/service. In this method, the preferences of consumers (rankings or ratings) are considered the dependent variables of a multiple multivariate regression model where the explicative variables are the levels of different factors characterizing the product/service of interest.

This approach seems particularly appropriated in Regulatory Impact Analysis (RIA) because, starting from different potential regulatory options, it allows to decompose the several evaluation dimensions and to obtain the estimated utilities for a group of judges (citizens/experts). However, the data collection is a critical step in CA. It is difficult for judges comparing a huge number of potential actions, described in a complex way. The most common procedure, the so called *full profile* method, forces to deal with a small number of categorical variables (and with a small number of categories). Being based on an experimental design, it is often too rigid. In RIA, we deal with a complex and multidimensional problem, and we are not always sure about the descriptions of the different options, using a small number of variables.

Our proposal consists in integrating CA with textual information achieved by answers to an open-ended question (Balbi *et al.*, 2008) and obtaining clusters of individuals similar in their own utility structure characterized by the words they use in describing their ideal public action. In the following, after introducing the methodological context (par. 2 and par.3), an empirical analysis on alternative University systems in Italy shows the effectiveness of our proposal (4).

2. The methodological framework

The impact of regulation must be *ex-ante* designed and evaluated. At this aim, Scepi *et al.* (2008) propose to apply a statistical approach based on the Conjoint Analysis method (Green, Srinivasan, 1990). The proposed method is a Multidimensional approach to Conjoint Analysis called Factorial Conjoint Analysis (FCA, Lauro *et al.*, 1998).

Let X be the (*stimuli x levels*) experimental matrix, where some different regulatory options (*stimuli*) are described by several interest variables (organizational, financial, economic, social *levels*), or core dimensions. Let Y be the (*stimuli x judges*) matrix, where the opinions of citizens or experts (*judges*) are collected, with respect to the comparison of the different options. The judges express their preferences by ranking the proposed stimuli with respect to a criterion, such as the expected benefits, the expected public utility, indirect net benefits. The results of the C.A. are in the matrix

\mathbf{B} (*levels, judges*) with general element, b_{kj} , the utility coefficient assigned to the k -th level by the j -th judge. Therefore, the classical C.A. model can be written as a multiple multivariate regression model:

$$\mathbf{Y}=\mathbf{XB}+\mathbf{E} \quad (1)$$

where \mathbf{E} is the residual matrix which has in columns the residuals of each individual model.

The factorial approach to C.A. allows us to read directly the results on perceptive maps built on the factorial plane obtained by the diagonalisation of the \mathbf{B} matrix. At this aim, the classical interpretative rules of a factorial approach are applied and we can read and interpret the relationships among judges, normative options and levels of the interest variables.

3. External information in Conjoint Analysis

The geometrical approach to C.A. was extended by Giordano and Scepi (1999) by adding an additional experimental matrix \mathbf{Z} , as external information related to matrix \mathbf{Y} columns, i.e. known characteristics of judges. In other words, they proposed a factorial approach to regression coefficient in C.A. in which classes of individuals with similar kinds of preferences and behaviours are projected into suitable reference subspaces.

In particular, two sets of regression coefficients are considered: the set \mathbf{B} defined as partial utilities explained by the product characteristics and the set \mathbf{D} defined as the partial utilities, explained by the judges characteristics.

Therefore, together with the classical solution for \mathbf{B} coefficients in (1), they introduce another regression model:

$$\mathbf{Y}=\mathbf{DZ}'+\mathbf{F} \quad (2)$$

where \mathbf{Z} is an indicator matrix inherent to the auxiliary information about judges, and \mathbf{F} is the residual term matrix.

Starting the two regression models, the inter-relation matrix, between characteristics of judges and levels of stimuli is computed:

$$\mathbf{\Theta}=(\mathbf{ZZ}')^{-1}\mathbf{ZYX}'(\mathbf{X}'\mathbf{X})^{-1} \quad (3)$$

and a Singular Value Decomposition (SVD) of $\mathbf{\Theta}$, with respect to two different metrics (Gower, Hand, 1996) has been performed.

By applying this strategy in the RIA context, it is possible to enrich the results of the analysis and to describe the characteristics of judges with similar opinions respect to the different normative options.

Scepi et al. (2008) suggest to introduce external information on the rows of the matrix \mathbf{Y} (stimuli) for enriching the results of the Factorial Conjoint Analysis in the RIA framework. The external information are supposed as variables highly correlated with experimental factors, which affect so prevalent the choice but not previously involved in the determination of the factorial components. Sustainability, economical effort, social impact are some of the possible examples of external information.

In this paper, we propose to use as external information a peculiar type of non structured information constituted by textual data.

The effectiveness of introducing textual data, as external information in a factorial conjoint analysis strategy, has been showed by Balbi et al. (2008). Here, in the RIA context, we show how textual information can be very useful for deeply understanding the preferences of judges.

In details, we introduce an open-ended question in the questionnaire. Judges are asked to describe the ideal normative option with his/her own words. The so called full profile method often used in Conjoint Analysis sometimes seems too rigid in its experimental structure. In fact, we have registered the difficulty of judges in understanding a complex option only by a small set of levels and factors.

In the frame of our proposal the questionnaire is divided into three sections:

- 1) in the first section the judges have to answer to an open-ended question like “*What’s your ideal normative option?*”
- 2) in the second one each judge has to fill his demographic data
- 3) in the last section the judges have to describe their behaviour

3. 1 The strategy in presence of textual information

Let Y , X and Z the above defined matrices. Now, let us consider the matrix T , a lexical table cross-tabulating the G descriptions, obtained by the open-ended question on the normative option, and the V words of our vocabulary, with general element t_{id} , presence/absence of the i -th word in the d -th description.

We introduce the matrix T for linearly constraining the B matrix, as in Takane and Shibayama proposal (1991). In this way, it is possible to visualise the terms in the verbal descriptions together with the C.A. levels on a common principal plane.

Thus we estimate the Q matrix of dimension (V,L) :

$$Q = (TT')^{-1} TB' \quad (4)$$

where q_{jl} is the estimated parameter linking the attribute levels to the judges textual descriptions of the ideal product.

Now we perform the SVD of the Q matrix, with the ortho-normalizing constraints:

$$WX'XW=I_L \text{ and } VTT'V=I_V \quad (5)$$

This approach allows to visualise the terms in the verbal descriptions together with the C.A. levels on a common principal plane. The information about the judges can be projected as supplementary points (Lebart et al., 1984), so to enrich the global interpretation of C.A. results.

Textual data allows to better understand the individual preferences thanks to the use of both designed and observational information, which implies the eventual introduction of characteristics which were not considered in the design.

Here, we also propose an additional use of textual data, in order to interpret the market segmentation obtained. In the specific of RIA, the proposed strategy enable an *ex-ante* evaluation of the impact of regulation by considering both additional elements and different judgements for group of individuals described both by socio-demographic variables and by words.

In other words, individuals are clustered take into account the first coordinates obtained by the singular value decomposition of Q . In the different group the peculiar words are identified by *modal answers* and *co-occurrences* are deeply investigated, thanks to textual data analysis *software* (SPAD).

4. A case study on the Italian University System

The questionnaire has been originally proposed by Lauro et al. (2007) with the aim of knowing the opinions of the experts on several possible alternatives of the Italian university systems. These different options are combinations of the following characteristics:

- 1) Management of the university system
- 2) Teacher recruitment
- 3) Formative path
- 4) Formative target
- 5) Legal value of the degree

A fractional factorial design has been obtained in order to analyze 8 different University System Scenarios (see Tab.1).

Scenario	Management	Teacher recruitment	Formative Path	Formative Target	Legal value of the certificate
A	Public	Public examination	Standardized	Cultural	Legal Value
B	Public	Public examination	Autonomous	Cultural	NO Legal Value
C	Public/Private in a system of rules	Public examination	Standardized	Professional	NO Legal Value
D	Public/Private in a system of rules	Private Contract	Autonomous	Cultural	NO Legal Value
E	Public	Private Contract	Standardized	Professional	NO Legal Value
F	Public/Private in a system of rules	Public examination	Autonomous	Professional	Legal Value
G	Public	Private Contract	Autonomous	Professional	Legal Value
H	Public/Private in a system of rules	Private Contract	Standardized	Cultural	Legal Value

Tab.1 : The experimental design

We modify this structure and introduce at the beginning of the questionnaire the following open end question “*could you give us a brief description of the model of the Italian university in which you want to work. The point of view is the effectiveness, as the capacity of a university to implement a training process to meet the expectations of users/students (in terms of management and organization of studies, quality of teachers, interest and flexibility), and having an impact on society*”.

Conjoint Analysis section was the last one in the questionnaire, in order to avoid influences in the free description. Each judge has been asked to rate the 8 scenarios according to his/her own opinion with respect to the criteria of efficiency previously explained. The more efficient is the system, the higher is the rate.

For each judge, some characteristics, such as the role in the University (full professor, associate professor and researcher) and the number of years in this role (this information is successively aggregated in the levels ≤ 3 ; 4-9; >9) are surveyed.

4.1 Pre-treatment of textual data

The answers to the open-ended question have been normalized in order to reduce the possibility of data splitting. After carrying out a quite in-depth lexicalisation in order to avoid trivial cases of ambiguity, a stop list was considered to eliminate the instrumental terms and a special threshold was introduced for the infrequent terms. A vocabulary of 693 textual forms (Bolasco, 1994), partially marked with Part of Speech tags, has been made out.

4.2 Some Results

The Figure 1 shows the C.A. levels on the first factorial plane, which represents about the 60% of the total inertia (first axis, 29%, second axis, 31%). On the left we can see a model of University strongly centralised at national level with *standardised courses*, *standardised recruitment procedures*, in which the certificate has a *legal value*. On the right, in opposition, we can see a model characterised by *autonomy*, in recruitment, in the organisation of courses, *without legal value* of certificate. The second axis opposes a *public system* to a regulatory system where *public* and *private* universities coexist. On the side of *public*, we can find a labour-market oriented *formative target* while the other model is devoted to the *cultural* growth of students.

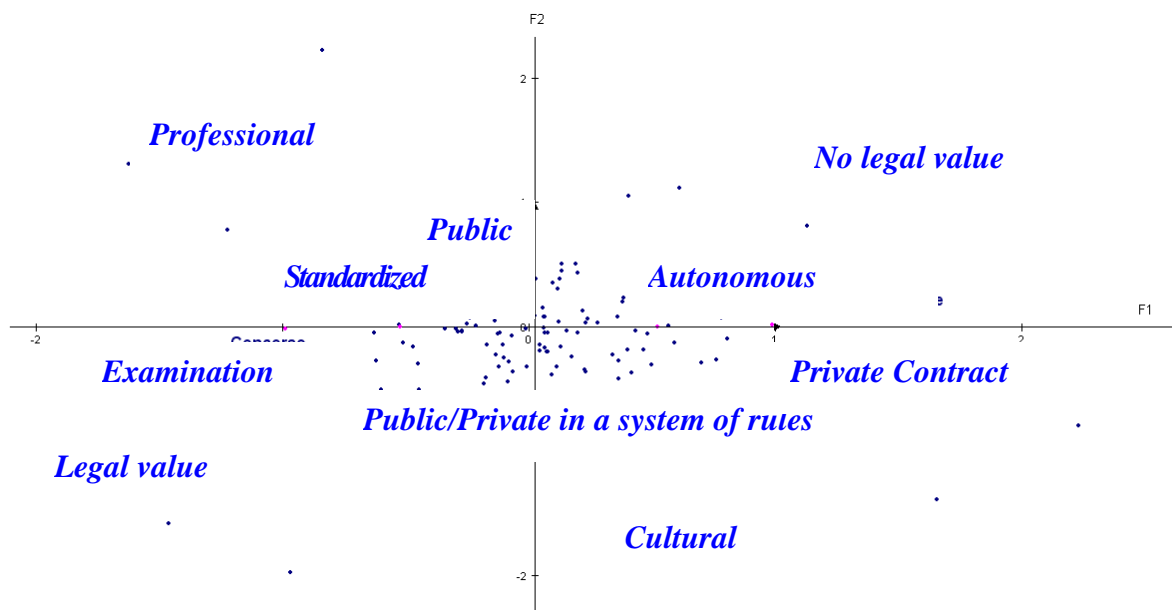


Fig.1: *The representation of levels*

In order to better understand the personal characteristics of respondents, the *role* and the *years* have been projected as supplementary points.

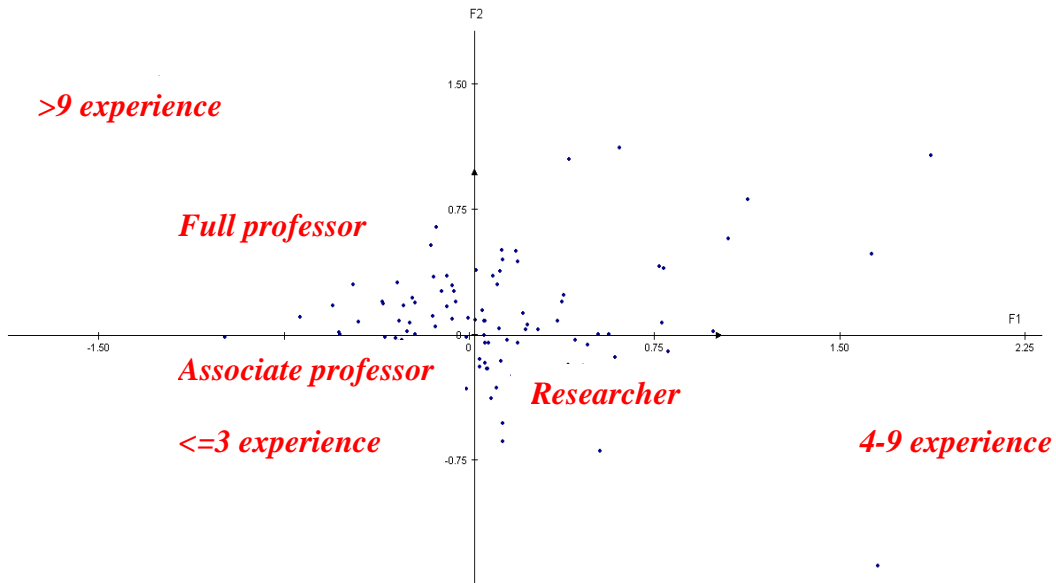


Fig.2: The representation of judges characteristics

The first factorial plane is characterised by “experience” of our experts: on the left top side we can see full and associate professors with *more than 9 years* of official activities in University, while on the bottom right side *researchers less experienced*. Taking in mind the previous interpretation of the plane, we can suggest that while younger judges show a higher utility coefficient for the level *cultural* of the variable *formative target*, professors give more importance to the *professional tools*.

In Fig. 3 the forms used to describe the natural language are projected for enriching the meaning of our results. Together with words strictly related to levels, other forms help us in considering other characteristics not included in the design.

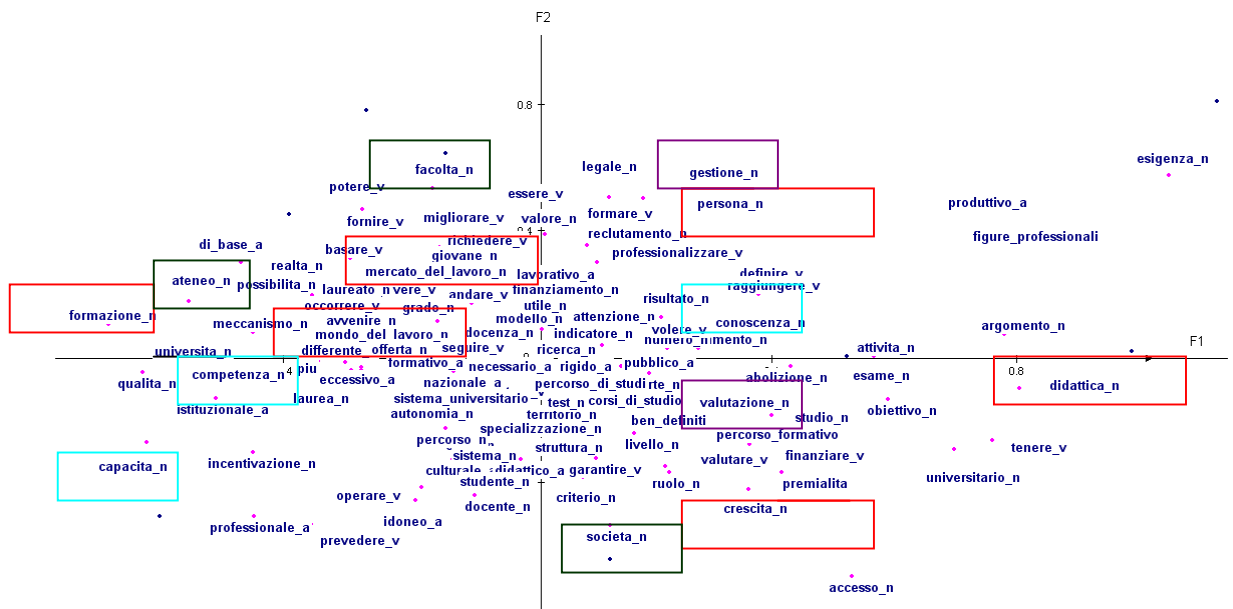


Fig.3: The representation of forms

First of all, the opposition on the first axis of *formazione, mercato del lavoro, mondo del lavoro* (on the left) and *didattica, persona, crescita* (on the right). It is very

important for discriminate the different models the opposition between *capacità* and *competenza* on the left and *conoscenza* on the right side. However, other insights are given by a model self-referential (*facoltà, ateneo*) on the top opposed to a model referred to outside (*società*), the first oriented to internal management (*gestione*), the other one to external evaluation (*valutazione*).

According to us, in RIA it is important to understand the social groups who agree, or not agree, with an eventual normative intervention. We can face this problem thanks to tools of marketing research: by segmenting our experts on the basis of their coordinates on the factorial subspace identified by FCA. Textual forms are used for characterising the identified groups.

We perform a classic cluster procedure, based on Ward algorithm, and 5 groups are detected.

Judges in the first class show to be interested in working in a university connected with the external world (*società, ruolo, sistema*), which is financed in relation to results (*valutare, finanziare*). The internal organisation (*facoltà, ateneo, gestione*) seems to be uninteresting. It is worth to be underlined that the connection is not limited to the labour market (the form *mondo del lavoro* has occurrence equal to 0)

CLASS1 LABELS OF THE FORMS	---PERCENTAGE ---		N. OF FORMS	/100 OF TOTAL	AVERAGE FOR RESPONSE	N. OF FORMS SEPARATE	/100 OF GROUP
	INTERNAL	GLOBAL					
si stema_n	4.62	1.53	130	19.94	18.6	68	52.31
val utare_v	2.31	0.46					
corso_n	3.08	1.38					
fi nanzi are_v	1.54	0.46					
ruol o_n	1.54	0.46					
soci eta_n	1.54	0.46					
gesti one_n	0.00	0.77	0	0.00	0.00	0.00	
mondo_del_lavoro_n	0.00	0.77					
val ore_n	0.00	0.77					
l aurea_n	0.00	0.92					
autonomi a_n	0.00	0.92					
facol ta_n	0.00	0.92					
ateneo_n	0.00	1.07					
esi genza_n	0.00	1.07					
uni versi ta_n	1.54	3.22					

Only the judges in the second class seems to be interested in recruitment (*reclutamento*) in a public context (*garantire, pubblico*). Note the absence of the frequent forms related to teaching (*didattica, docente*).

CLASS2 LABELS OF THE FORMS	---PERCENTAGE ---		N. OF FORMS	/100 OF TOTAL	AVERAGE FOR RESPONSE	N. OF FORMS SEPARATE	/100 OF GROUP
	INTERNAL	GLOBAL					
garanti re_v	4.44	0.92	90	13.80	18.0	56	62.22
numero_n	4.44	1.23					
recl utamento_n	3.33	0.77					
l aureato_n	2.22	0.46					
eccessi vo_a	2.22	0.46					
pubbli co_a	3.33	1.23					
si stema_n	0.00	1.53	0	0.00	0.00	0.00	
dovere_v	4.44	7.21					
di datti ca_n	0.00	1.69					
docente_n	0.00	2.91					

In the textual forms characterising the third class, we can see the peculiarity of research (*ricerca*), institutional task of universities, neglected by the other groups of judges. Not surprising the absence of the word knowledge (*conoscenza*).

CLASS3 LABELS OF THE FORMS	---PERCENTAGE ---		N. OF FORMS	/100 OF TOTAL	AVERAGE FOR RESPONSE	N. OF FORMS SEPARATE	/100 OF GROUP
	INTERNAL	GLOBAL					
grado_n	1.43	0.46	210	32.21	42.0	82	39.05
seguire_v	1.43	0.46					
prevedere_v	2.86	1.38					
risultato_n	1.90	0.77					
ricerca_n	2.38	1.23					
formativo_a	0.48	1.53	0.00	0.92	0.00	1.23	1.38
sistema_n	0.48	1.53					
garantire_v	0.00	0.92					
pubblico_a	0.00	1.23					
conoscenza_n	0.00	1.38					

The fourth class emphasised the role of the university of preparing individuals able to find a job (*di base, professionale, formativo*). The geographic reference is all the country (*nazionale*).

CLASS4 LABELS OF THE FORMS	---PERCENTAGE ---		N. OF FORMS	/100 OF TOTAL	AVERAGE FOR RESPONSE	N. OF FORMS SEPARATE	/100 OF GROUP
	INTERNAL	GLOBAL					
occorrere_v	1.71	0.46	175	26.84	21.9	73	41.71
nazionale_a	1.71	0.46					
formativo_a	3.43	1.53					
di_base_a	2.29	0.92					
realta_n	1.71	0.61					
idoneo_a	1.71	0.61					
professionale_a	1.71	0.61					
esame_n	0.00	0.92					
garantire_v	0.00	0.92	0.00	1.07	0.00	1.23	1.53
esigenza_n	0.00	1.07					
numero_n	0.00	1.23					
studio_n	0.00	1.53					

The fifth class is pragmatic: not reference to fantastic world (*potere, dovere*), the characteristic words looks at results (*produttivo, utile*) but not at students (*studente* has frequency equal to 0)!

CLASS5 LABELS OF THE FORMS	---PERCENTAGE ---		N. OF FORMS	/100 OF TOTAL	AVERAGE FOR RESPONSE	N. OF FORMS SEPARATE	/100 OF GROUP
	INTERNAL	GLOBAL					
produttivo_a	6.38	0.61	47	7.21	15.7	38	80.85
percorso_n	4.26	1.38					
utile_n	2.13	0.31					
sistema_n	0.00	1.53	0.00	2.15	4.26	7.21	2.30
formativo_a	0.00	1.53					
potere_v	0.00	2.15					
dovere_v	4.26	7.21					
studente_n	0.00	2.30					

References

- Balbi S., Infante G and Misuraca M (2008) Conjoint analysis with textual external information. In: S. Heiden et al. Eds., *JADT 2008, 9th International Conference on the Statistical Analysis of Textual Data*, 129-136, ISBN: 978-2-7297-0810-8
- Bolasco S. (1994) L'individuazione di forme testuali per lo studio statistico dei testi con tecniche di analisi multidimensionale. In: *Atti della XXXVII Riunione Scientifica della SIS*, Sanremo

- Giordano G., Scepi G. (1999) Different Informative Structures for Quality Design. *Journal of Italian Statistical Society*, 8(2-3), pages 139-149.
- Gower, J.C., Hand, D. J. (1996) *Biplots*. Chapman & Hall.
- Green P.E., Srinivasan V. (1990) Conjoint analysis in Marketing: new developments with implications for research and practise. *Journal of Marketing*, 54(4), pages 3-19.
- Lauro N.C., Giordano G. and Verde R. (1998) A multidimensional approach to conjoint analysis. *Applied Stochastic Models and Data Analysis*, pages 265-274.
- Lauro C.N., Romano E. and Giordano G., (2007) Clustering model based representation of symbolic objects, Special Topics Contributed Paper of the 56th Session of the ISI International Statistical Institute, 22 - 29 August, Lisboa
- Lebart L., Morineau A. and Warwick K.M. (1984) *Multivariate Descriptive Statistical Analysis*. Wiley & Sons.
- Scepi G., Lauro N.C. and Giordano G., (2008) The ex-ante evaluation of regulatory impact by Conjoint Analysis: some developments. In *RC33 2008 Proceedings*.
- Takane Y., Shibayama T. (1991) Principal component analysis with external information on both subjects and variables. *Psychometrika*, 56, pages 97-120.