XXXIII CONFERENZA ITALIANA DI SCIENZE REGIONALI

UNIVERSITY STUDENTS: WHO ARE THEY? RESULTS FROM A CLUSTER ANALYSIS BASED ON COGNITIVE SKILLS AND PERSONALITY TRAITS

Rosanna NISTICÒ[1], Michelangelo MISURACA[2]

SOMMARIO

Lo sviluppo delle competenze degli studenti è un compito complesso, che dipende non solo dal sistema educativo, ma anche da numerosi altri fattori come la qualità del contesto socio-istituzionale, il contesto economico e il background familiare. In questo lavoro sono stati analizzati i dati relativi agli studenti immatricolati presso l'Università della Calabria nell'anno accademico 2008-2009, con l'obiettivo di studiare differenze e analogie nella loro carriera scolastica e a livello personale, così come le loro aspettative. Da un punto di vista statistico, per mezzo di una strategia multivariata, i profili degli studenti universitari del primo anno sono stati analizzati e classificati in diverse tipologie. L'obiettivo principale è quello di fornire una classificazione degli studenti che possa essere utile a programmare e progettare interventi e misure specifiche, con l'obiettivo di "tagliare" la formazione del capitale umano su misura per le caratteristiche individuali di ciascun gruppo.

[1] Università della Calabria – Dip. di Economia e Statistica, via P. Bucci Cubo 0C/1C, I-87036, Arcavacata di Rende (CS), e-mail: r.nistico@unical.it
[2] Università della Calabria – Dip. di Economia e Statistica, via P. Bucci Cubo 0C/1C, I-87036, Arcavacata di Rende (CS), e-mail: michelangelo.misuraca@unical.it

1 Introduction

The importance of cognitive skills in determining individual and socio-economic performanc is broadly shared by economists. Until some decades ago, theoretical and empirical researches in the field of educational economics were concentrated on the quantity of education (years of schooling). More recently, thanks to the increased availability of survey data based on international test scores, applied studies focused mainly on students' competencies. Different international programs are designed in order to measure cognitive skills, like the *Trends in International Mathematics and Science Study* (TIMSS), the *Progress In International Reading Literacy Study* (PIRLS) and above all, for its focus on "competencies for life", the OECD *Program for International Student Assessment* (PISA). This latter carries out since 2000, with a three years interval, a survey based on standardised tests regarding the ability of 15-y.o. students in solving abstract and everyday-life problems. A number of studies (e.g., Hanushek *et al.*, 2000; Hanushek *et al.*, 2008) give evidence that cognitive skills of the population are highly related to individual earnings, to income distribution and to economic growth, providing both individual and social returns. Furthermore, territorial disparities in students' competencies suggest they are affected by the quality of the educational system, as well as by a number of other exogenous factors, such as the social environment or the students' family background (Bratti *et al.*, 2008; Checchi, 2010). Side by side to the importance relied upon cognitive skills, empirical research has recently pointed out that individual success, both in educational career and working life, are related to other intrinsic characteristics, *personality traits* – e.g., social skills, motivation and leadership – or *non-cognitive skills* (Heckman *et al.*, 2006; Duncan *et al.* 2007; Carneiro *et al.*, 2007; Brunello *et al.*, 2011).

By reason of the relevance competencies play in the economic development, governments frequently try to improve cognitive skills, for example through programs rising the quality of education or enhancing students' effort. Unfortunately these programs often slip down the policy agenda, because they pay off in the future (OECD, 2010), or they suffer of a not-in-deep investigation of the different interventions students actually need. As an example, some Italian University carry on specific programs aimed at making first year classrooms homogeneous in Mathematics, Reading, Foreign Language or other subjects, by means of preliminary courses. These courses often take place in undifferentiated classrooms or, at the best, in classrooms formed on the basis of the test scores, which do not take into account other personality traits that could facilitate or bias students' learning process and performance.

Students who attend the same class have in fact different cognitive and non-cognitive skills, including social motivation, relational abilities and political concerns. Moreover, some of them are more informed than others as regards opportunity offered by Faculties, their pro-

grams, degree employment opportunities, somebody has more higher expectations than others regarding jobs and earnings they can get once graduated, and so on.

This paper aims at contributing to the empirical literature, by investigating cognitive and non-cognitive characteristics which differentiate University students. We have analysed individual data collected on students enrolled at the University of Calabria (Unical) in the academic year 2008/2009 by a questionnaire submitted at the enrolment time.

With the aim of obtaining a classification of "in entrance" students more articulated than the usual partition in "low" and "high" ability students based only on evaluation tests, we have considered a set of 10 variables describing acquired cognitive skills as depicted by student's educational background (*High school qualification, Normalised qualification grade, Successful in passing High School classes, Evaluation of learning difficulties at High School, Matching between school votes and self-evaluation, School performance in Italian literature and literacy, School performance in Foreign Language, School performance in Mathematics, School performance in Computer Science, Extra-school experiences*), 4 variables describing family background affecting their individual development and abilities (*Father's educational level, Mother's educational level, Father's employment status, Mother's employment status*), 6 variables related to University studies' awareness and expectations (*Faculty of enrolment, Working while studying, Information on educational goals of the academic course, Information on teaching courses and skills to be acquired, Importance attached to get the university degree on time, Probability of concluding studies on time*) and 6 variables more closely related to personality traits concerning job opportunities (*Information on job opportunities with a university degree, Expectations on monthly wage with a high school qualification, Expectations on monthly wage with a first level degree, Expectations on monthly wage with a second level degree, Opinion on chances to get a job matching the first level degree skills, Opinion on chances to get a job matching the second level degree skills*). Students' cognitive skills and individual traits are then jointly considered in order to group students by means of a statistical two-step strategy of *Multiple Correspondence Analysis* and *Cluster Analysis*.

The main utility of this work relies on the fact that students' clustering, based both on cognitive skills and personal traits, could help to reduce biases in human capital development programs aimed at enhancing the quality of education and students' effort in developing skills and competencies. These programs could became in fact more efficient and effective if specifically tailored on the whole individual characteristics.

The paper is organized as follows. Section 2 describes the survey and the questionnaire submitted to enrolling University students. In Section 3 we presents the dataset used and some preliminary results, and Section 4 describe the methodology of analysis. Section 5 provides and comments the factorial representation, while Section 6 presents the cluster analysis and discusses the different typologies of students.

2 The Survey

In 2008 the Regional Government of Calabria, reacting to the results of 2006 PISA wave, designed a massive human capital intervention aimed at enhancing and aligning competencies of the first year University students in the Calabria region. In the last decade some international surveys have highlighted a remarkable gap in Italy about cognitive skill formation, in comparison with other equivalent developed countries. At a sub-national level, literacy in the areas of Reading, Mathematics and Science, as measured by PISA surveys, is heavy lower in the southern regions than in the northern ones. As concerns the Calabria region, it is still found in the last OCSE-PISA wave relating to 2009 a score in Reading competencies (448) lower than both the National (486) and the OCSE (493) average. Even more, it is also less than the score performed by other southern regions. A similar situation is depicted by results achieved in Mathematics and Science tests in 2009 and in the previous PISA waves.

As part of the program of intervention of the Regional Government of Calabria, a questionnaire to be submitted to enrolling students was prepared by the Statistical Office of the University of Calabria, in a more detailed form than in the previous academic years, since a *Survey on the enrolled students' opinions* has been carried out each academic year in the last decade. In particular, the questionnaire regarding the 2008-2009 academic year wave was extended in the sections and number of inquiries, including a number of students' personality traits such as self-esteem, information collected before choosing Faculty and University, difficulties encountered in their school career, attendance of post school courses and social activities, parents' background, work experiences. The target population has been the 5485 students enrolled in the academic programs of University of Calabria in 2008/2009.

A *Computer assisted self-interviewing* (CASI) web system has been implemented in order to collect data, with an average time for completing the questionnaire of about 15 minutes. The idea has been to obtain accurate answers from the students, due to a highly confidential method of responding, minimising the time and the costs of the survey. The proposed questionnaire has listed 85 questions divided into 10 different sections: *biographical data*; *information used in choosing Faculty*; *educational background and competencies*; *personal attitudes*; *family background*; *job experiences*; *choice of University*, *Faculty and degree course*; *first university year expectations*; *future perspectives as graduated*; *job and wage expectations*.

3 Data structure and preliminary analyses

A common problem of sample survey is how to deal with the total amount of collected data, because sometime the structure of the designed questionnaire is not specifically linked with the goals of the survey. This is typically the case of *multi-purpose surveys*, in which the complex of questions and the number of explored topics is quite large, in the idea that it is better

to have a wider quantity of information. Moreover, the use of electronic questionnaires makes easier the collection of a higher amount of data.

The questionnaire used by the University of Calabria lists at the same time different kinds of questions, from open-ended to multiple choice ones, with different types of modalities and scales. Because of this the pre-treatment procedure have been difficult. From the original 85 questions about 260 variables have been extracted, observed on 4565 units, with a coverage rate of about 83% with respect to the target population.

Data have been cross-referenced with the matriculation information listed into the registrar's office database, in order to validate personal data obtained from the questionnaire and to add other information not provided by the survey, like the *Type of High School Qualification* or the *Grade of High School Qualification*. It has been considered opportune a re-coding procedure for numerical variables by carrying out for each a distribution of classes. Moreover the qualification grade has been normalised in order to make comparable all the students.

Particular attention has been devoted to the questions related to personal *Monthly Wage Expectations*, having a High School qualification, a Bachelor degree or a Master degree. The open-ended questions related to these topics have required a special treatment for two main reasons. The first one concerns data normalisation from a technical point of view, because respondents have been the possibility of using different decimal marks and thousands separators in representing percentages and incomes. The second one concerns, as in many other surveys, the typical problem of collecting sensitive data like incomes.

The most obvious cases of ambiguousness have been manually checked and edited. After drawing data distributions it has been decided to consider a trimmed range of values. In order to preserve the higher amount of information, and use the same criterion for the different distributions, a lower threshold to 6^{th} percentile and an upper threshold to 94^{th} percentile have been set up. In this way a range of 100€ - 2000€ has been considered for income expectations with a qualification, 400€ - 3000€ with a three-years university degree and 1000€ - 6000€ with a two-years university degree. Data out of the determined boundaries has been considered missing values. The percentage of noise introduced in this way has been minor compared with the loss of information produced by a listwise deletion of units with item non-responses.

From the respondents side it has been decided to filter data from the questionnaire not completed (about 1%), using a control variable produced automatically by the CASI system, and on the base of the cross-reference with data obtained by the registrar's office. In this way it has been considered for the analysis a set of 4508 records, with a coverage rate of 82% with respect to the target population. From the variable side a set of 27 variables with 122 features has been selected, by considering some characteristic usually proposed in the domain's literature (as seen in Section 1), together with other personality traits relevant for depicting students' awareness and expectations.

A complete list of the variables involved into the analysis is showed in the following Table 1.

Table 1 - Active variables used in the analysis: labels and descriptions

Variable	Description	N. of features
SESSO	*Student gender*	2
A - EDUCATIONAL BACKGROUND (10)		
TIPO_DIPLOMA	*High school qualification*	10
MAT_NORM	*Normalised qualification grade (36-60 → 60-100)*	5
D10	*Successful in passing High School classes*	2
D11	*Evaluation of learning difficulties at High School*	5
D12	*Matching between school votes and self-evaluation*	5
D13_1	*School performance in Italian literature and literacy*	4
D13_2	*School performance in Foreign Language*	4
D13_5	*School performance in Mathematics*	4
D13_9	*School performance in Computer Science*	4
D15	*Extra-school experiences*	2
B - FAMILY BACKGROUND (4)		
D23_1	*Mother's educational level*	5
D23_2	*Father's educational level*	5
D24_1	*Mother's employment status*	4
D24_2	*Father's employment status*	4
C - UNIVERSITY EXPECTATION (6)		
FACOLTA	*Faculty of enrolment*	7
D28	*Working while studying*	2
D40_1	*Information on educational goals of the academic course*	5
D40_9	*Information on teaching courses and skills to be acquired*	5
D47	*Importance attached to get the university degree on time*	4
D49	*Probability of concluding studies on time*	3
D - JOB EXPECTATION (6)		
D40_3	*Information on job opportunities with a university degree*	5
D58	*Expectations on monthly wage with a high school qualification*	5
D56	*Expectations on monthly wage with a first level degree*	6
D57	*Expectations on monthly wage with a second level degree*	7
D60	*Opinion on chances to get a job matching the first level degree skills*	4
D61	*Opinion on chances to get a job matching the second level degree skills*	4

The 4508 units' sample is composed by a 60% of female and a 40% of male students, respectively. The dwelling place of the 98% is Calabria, with a 99% of students that declares an Italian nationality. From this viewpoint University of Calabria mainly serves, at the present as in the original idea of its founders, the regional territory. With respect to High School qualification grade (Table 2) – normalised for considering the same scale of votes for all students, both qualified before and after the 1997 legislative reform – it is really interesting to note that the percentage of 2008-2009 enrolled students with the best qualification grade (21.4%) is quite higher than the corresponding Regional percentage (10.1%; SISPICAL, 2009) and National percentage (6.1%; MIUR, 2010).

Table 2 - High school grade distribution table

MAT_NORM - *Normalised qualification grade*

Features	Abs. freq.	%
60 - 69	523	11.6%
70 - 79	879	19.5%
80 - 89	1142	25.3%
90 - 99	999	22.2%
100	965	21.4%

Analysing the evaluations made by the students with respect to learning difficulties at High School, clearly emerges that the two-third of the sample has experienced null or weak difficulties. Only 1% has declared to have experienced strong difficulties. By analysing school performances (Table 3a and 3b) it is possible to notice that the majority of the students declares a middle-high level in Italian literature and literacy, with respect to Foreign languages and Quantitative competencies that are for two-third of the sample at a middle-low level. It is also interesting to note that half of the sample declares to have never studied Computer Sciences at High School.

Table 3a - School performance distribution tables: Italian and Foreign Language

D13_1 - *School perform. in Italian*

Features	Abs. freq.	%
Italiano_MP	2330	51.7%
Italiano_SD	2003	44.4%
Italiano_NS	164	3.6%
Italiano_NA	11	0.2%

D13_2 - *School perform. in Foreign Language*

Features	Abs. freq.	%
LingueStraniere_MP	1527	33.9%
LingueStraniere_SD	2216	49.2%
LingueStraniere_NS	674	15.0%
LingueStraniere_NA	91	2.0%

Table 3b - School performance distribution tables: Mathematics and Computer Science

D13_5 - *School perform. in Mathematics*

Features	Abs. freq.	%
Matematica_MP	1302	28.9%
Matematica_SD	1895	42.0%
Matematica_NS	1247	27.7%
Matematica_NA	64	1.4%

D13_9 - *School perform. in Computer Science*

Features	Abs. freq.	%
Informatica_MP	1003	22.2%
Informatica_SD	947	21.0%
Informatica_NS	208	4.6%
Informatica_NA	2350	52.1%

As regards family background, it is possible to see a similar distribution for both parents, especially for the lower educational levels (Table 4).

Table 4 - Parents' Educationl level distribution tables

D23_1 - *Mother's Educational level*

Feature	Abs. freq.	%
Nessun Titolo	36	0.9%
Licenza Elementare	465	11.6%
Licenza Media	1266	31.6%
Diploma	2001	49.9%
Laurea	740	18.5%

D23_2 - *Father's Educational level*

Feature	Abs. freq.	%
Nessun Titolo	40	1.0%
Licenza Elementare	468	11.7%
Licenza Media	1395	34.9%
Diploma	1929	48.2%
Laurea	676	16.9%

4 Methodological Framework

The analysis of complex systems of data, in which concur both a huge number of observations and a considerable number of features, implies often substantial critical states. Investigating and understanding the association within a set of variables, where we are interested in how strongly and in which way these variables are interrelated, requires hundreds if not thousands of cross-tabulations and makes really challenging the study of multifaceted phenomena

Sample surveys, in which data are typically collected via questionnaires, are a common case. In this context it becomes necessary to carry out a statistical treatment of data which allows to extract the maximum amount of information for the observed phenomenon, and light the key aspects of its global structure. The classical statistical techniques related to the analysis of contingency tables do not lend themselves to the required *changing up*, because they often leave unused most of the inherent (and significant) collected information.

It is than necessary to consider a strategy that eliminates redundant cross-tabulations and highlights the ones that probably would not have been considered, aiming at reaching a gen-

eral assessment of the questionnaire without *a priori* choices, in an exploratory fashion, and offering new working hypothesis and viewpoints.

A suitable approach that fulfil these latter research needs is the well-known *Multiple Correspondence Analysis* (MCA), originally developed by Benzécri (1973) and Lebart *et al.* (1984) in the frame of the so called French School of *Analyse de Données*.

This method – far from being merely a generalisation of the *Correspondence Analysis* for studying relations of higher order than the simple bivariate one – is able to perform an appropriate reduction of dimensionality onto a *n x p* table, where *n* is the number of observations (i.e., the respondents) and *p* a set of categorical variables (i.e., the questions). Consider then a *n x p* table **X** and the corresponding *n x s* indicator matrix **Z**, where *s* represents the different features of the *p* variables. Each element in column can be actually seen as a dummy variable with 0 and 1 in case of absence/presence of the feature. MCA performs an eigenvalue decomposition of the square matrix **B**=**Z`Z**, known as *Burt matrix*. The original variables are replaced by linear combinations of themselves known as *latent variables* or *factors*.

The functional connections between the quantities are then transformed into geometric relationships, providing a planar representation that allows an overall view and an immediate visualisation, both of the characterizing features and units on which data have been collected.

The factors and the factorial maps can be read in terms of *percentage of explained inertia*, a measure of variability that expresses the amount of original information represented in the reduced space obtained by MCA. Because of the nature of the analysis, an orthogonal relation between features belonging to a same variable is introduced, producing in this way an artificial sphericity. This means that the explained inertia represents a pessimistic measure of the explanatory power of the obtained synthesis.

It is possible to compute an *adjusted explained inertia*, by considering Benzécri's correction:

$$\lambda_i^* = \left(\frac{p}{p-1}\right)^2 \left(\lambda_i - \frac{1}{p}\right)^2 \tag{1}$$

where λ_i is the *i-th* eigenvalue obtained by decomposing the Burt matrix ($\lambda > 1/p$). The *percentage of adjusted explained inertia* is then calculated as:

$$\frac{\lambda_i^*}{\sum_{i=1}^{s-p} \lambda_i^*} \tag{2}$$

The parameters of MCA are estimated by pooling the data across units, under the implicit assumption that all the observations come from a single, homogenous group.

However, it often seems more realistic to assume that units come from heterogeneous groups, so that they are different with respect to their attitudes and behaviours, and the other characteristics of interest. The presence of groups depends clearly on the association structure over the data, but often MCA visualisation is not really intuitive and easy to read.

To cope with these issues, French School proposed a two-step sequential approach called *tandem analysis* (Arabie *et al.*, 1996): after performing a factorial method for transforming the original variables, a clustering method onto the factors is carried out. In a general acceptation *Cluster Analysis* (CA) can be seen as a multivariate technique that try to organize information about a set of variables in order to discover homogeneous groups, but it has to be considered more properly as a variety of methods that attempt to form classes with an *internal cohesion* and an *external isolation*.

This strategy has the advantage that it works on a reduced number of variables that are orthogonal and ordered with respect to the borrowed information. The choice of the factorial method is an important and tricky phase because it will affect the results, but taking into account this critical state it achieves an improvement of the overall quality of clustering.

A satisfying way for obtaining non-overlapping clusters, is to consider a *hierarchical* CA, in which different levels of aggregation (i.e., the number of clusters) are investigated at the same time (Gordon, 1999). Agglomerative hierarchical clustering methods start from *n* classes, each formed by a single unit, and arrive through successive steps at a unique class containing all the units. Divisive hierarchical methods start instead from the global class, the class containing all the units, and arrive at defining the *n* classes. Differences arise in the way used to compute the distance between a cluster and the units, or across classes, that is the chosen agglomerative or divisive criterion.

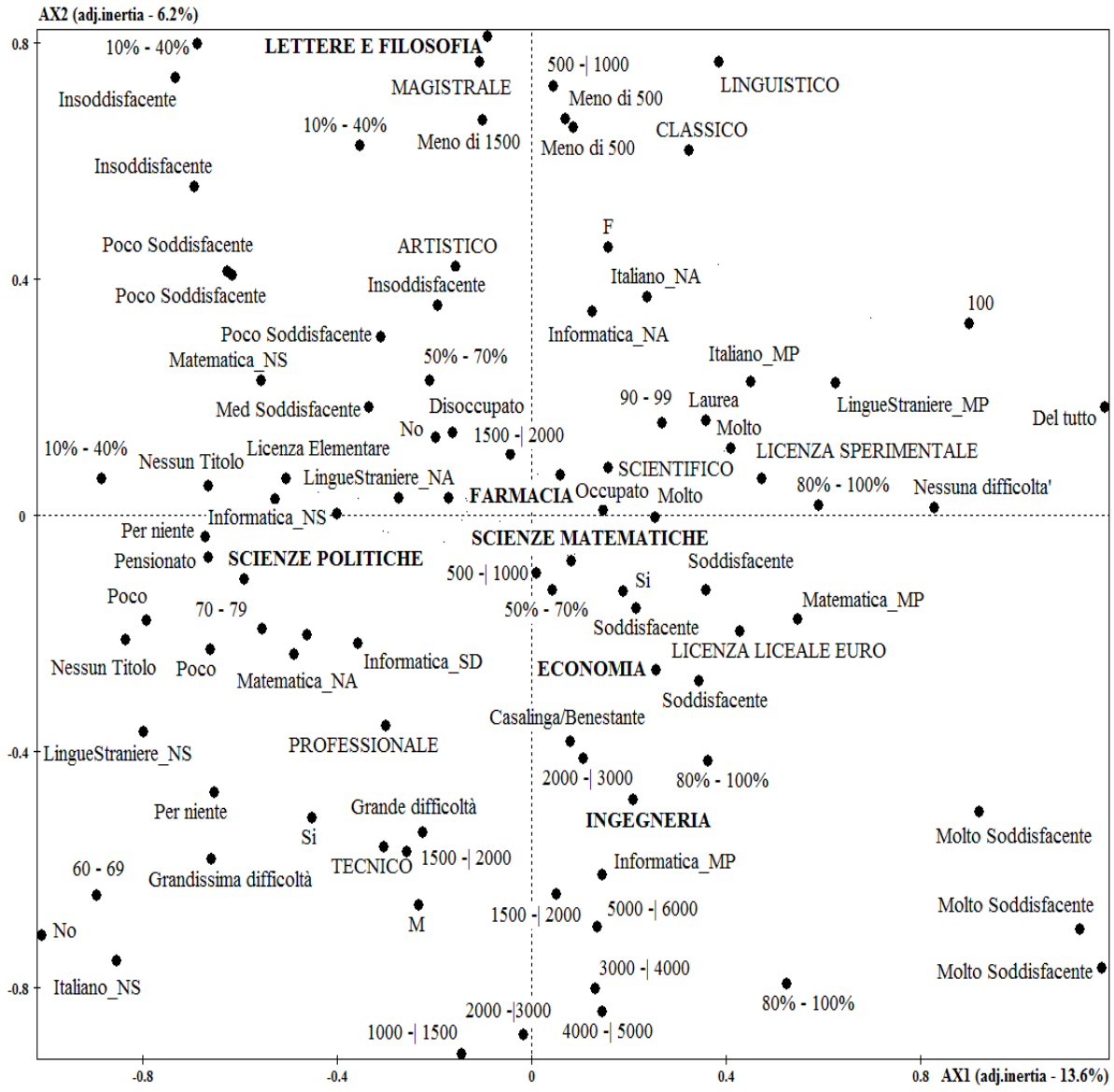
In practical applications, the agglomerative methods are the most used as they allow to construct the hierarchy of partitions with a significantly reduced computational cost. The results can be represented by a *dendrogram*, a tree structure representing the complete hierarchy of partitions. The optimal ones can be chosen by considering a suitable aggregation index. A partition with a higher aggregation index means that the distance between the two closest clusters is large, so they are well separated. Cutting the tree at a level corresponding to a significant “jump” in the index level leads to a good partition.

5 Factorial representation and interpretation

In order to analyse and visualise the latent association structure of the dataset, the matrix cross-tabulating the 4382 respondents and the 27 active variables have been considered. Once the Burt matrix has been calculated we have decided to not consider the missing value.

The first factorial plane (Figure 1) explain about the 20% of the total adjusted inertia.

Figure 1 - MCA: most significant active features on the first factorial plane



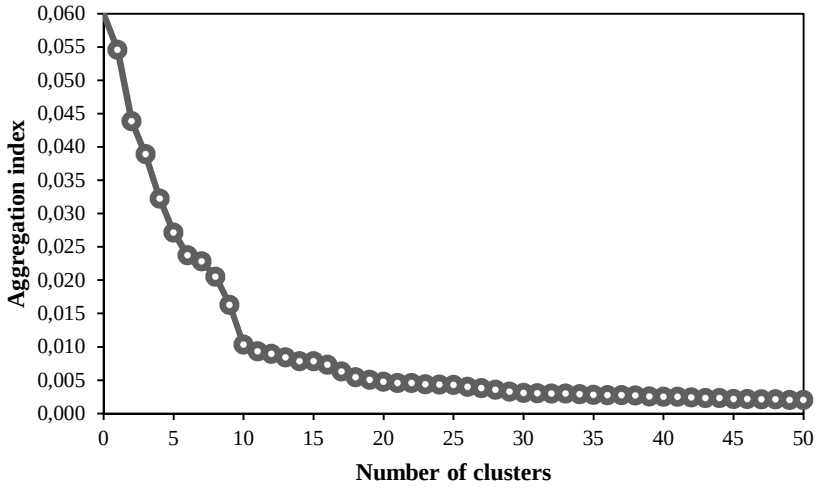
On this factorial plane is it possible to see on the first axis – from the left to the right side – an opposition between students with a poor educational and family background and students with good ones (e.g., *school performances* in different matters). On the second axis instead, it is possible to see – from the bottom to the top – an opposition between students with optimistic expectation in finding a job and earning middle-high wages, and students with more pessimistic expectations in terms of finding a job and earning good wages. Particularly interesting, even if not surprising, is in this latter case the strong connection between expectations and the domain both of the educational background and the chosen Faculty.

It would be possible to read the map in terms of quadrants, but in the frame of tandem analysis approach it is more appropriate to perform a Cluster Analysis on the MCS factors.

6 University Students: Who Are They?

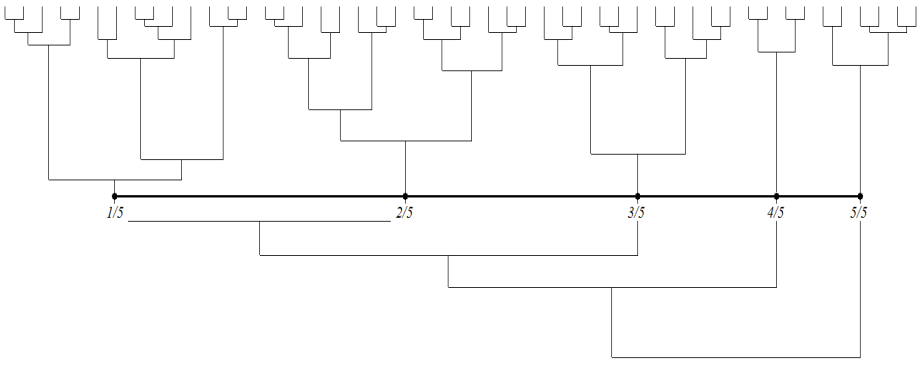
The clustering procedures performed is a hierarchical cluster analysis based on the *Ward criterion* for aggregating the units at the different steps. At each step of the criterion leads to aggregate together those groups for which there is a lower increase of deviance within the groups or, alternatively, the greatest decrease of deviance between the groups. The algorithm used in the clustering procedure is the *nearest-neighbours chain* (Benzécri, 1982).

Figure 2 - Scree plot of the number of clusters and the aggregation level



In this case a partition with 5 clusters seems to be the optimal solution, as confirmed by the *scree-plot* (Figure 2). This solution is confirmed looking at the dendrogram (Figure 3)

Figure 3 - Dendrogram representation with the partition in 5 clusters



The 5 clusters obtained from the chosen partition can be described by considering for each the most relevant features. To decide if a given feature *j* is a relevant characteristic of a cluster *k* it

is necessary to verify if it is significantly more present in *k* than in the sample. To deal with this problem a statistical significance test is performed: the null hypothesis H_0 assumes an equal proportion of *j* in the *k* and in the sample (against an unusually high proportion among the individuals in *k* with respect to the sample). By considering the N hypergeometric random variable, number of feature *j* observed in *k*, it is possible to calculate a p-value:

$$p_k(j) = \mathrm{P}\{N \geq n_{jk} \mid H_0\} \tag{3}$$

The higher is n_{jk} the lower is the hypergeometric probability, so that the null hypothesis is more doubtfully true. The results can be easier read if we consider the value, known as *test-value*, assumed by a Gaussian variable for the same probability $p_k(j)$ of the hypergeometric variable:

$$test-value = \frac{n_{jk} - n_k \frac{n_j}{n}}{\sqrt{n_k \frac{n-n_k}{n-1} \frac{n_j}{n}\left(1-\frac{n_j}{n}\right)}} \tag{4}$$

The higher is the test-value the more the feature characterise the cluster. It has empirically shown that for having relevant feature it is necessary to consider |test-value| > 2.
Here in the follow the results of the Cluster Analysis are presented in terms of test-values, ordered in descending terms of importance. For a correct interpretation it is necessary to underline that the features have to be read in terms of logical disjunctions.

Table 5 - CLUSTER 1/5: Weaks and Discouraged (abs.: 1333, rel.: 29.6%)

Variable	Feature	% of feature in the cluster	% of feature in the sample	test-value	% of the cluster in the feature
D13_5	Matematica_NS	48.16	27.66	19.42	51.48
D60	10% - 40%	52.44	31.57	19.15	49.12
D61	10% - 40%	26.26	12.53	17.13	61.95
D57	Meno di 1500	42.39	25.11	16.83	49.91
D40_1	Poco Soddisfacente	24.61	13.02	14.28	55.88
FACOLTA	Lettere e Filosofia	31.06	18.21	13.95	50.43
SESSO	F	74.34	59.29	13.54	37.07
D56	500 -| 1000	38.41	24.71	13.45	45.96
D40_3	Poco Soddisfacente	22.28	11.93	13.23	55.20
D23_1	Licenza Elementare	19.28	10.32	12.20	55.27
D12	Abbastanza	48.31	34.80	12.16	41.05
D40_9	Poco Soddisfacente	37.36	24.96	12.15	44.27
D23_2	Licenza Elementare	19.28	10.38	12.08	54.91
D49	10% - 40%	19.65	10.98	11.51	52.93
D40_3	Med Soddisfacente	42.99	32.03	10.05	39.68

To the first cluster (Table 5), namely *Weaks and Discouraged*, belongs the 29.6% of the enrolled students (1333 people). They are characterized by the weakness of their cognitive skills or the low level of their family background, in terms of education, their pessimistic expectations of being successful in searching a job, their small wage expectations, or by a very poor self-esteem. Three-quarters of units in the cluster are female students – a considerable amount if one considers that women constitute the 59% of the analysed sample. They are students mainly oriented towards literary studies – about one-third has chosen the Faculty of Humanities (18% in the sample), and nearly a half have achieved unsatisfying scores in Mathematics at High School. Most of them consider to have low chances of getting a job for which their degree is required (between 10% and 40%) or a low probability to complete their studies on time. They have also low wage expectations: about the 42% of the cluster expects to earn less than 1500 euros after obtaining the second level degree and between 500 and 1000 euro after the first-level degree. Over one-fifth of them consider incomplete the information they have as regards the educational and training goals of their degree course, or the information relating to specific courses and skills they are going to acquire. One-fifth of the cluster has a family background in which both the parents have just the primary school certificate, a weight twice than the whole sample of enrolled.

Table 6 - CLUSTER 2/5: Ambitious and Hopefuls (abs.: 1446, rel.: 32.1%)

Variable	Feature	% of feature in the cluster	% of feature in the sample	test-value	% of the cluster in the feature
SESSO	M	69.78	40.71	27.36	54.99
TIPO_DIPLOMA	Tecnico	48.96	30.92	17.70	50.79
D13_1	Italiano_SD	61.20	44.43	15.55	44.18
D56	1500 -| 2000	30.98	19.10	13.54	52.03
D58	1000 -| 1500	18.46	9.96	12.57	59.47
D61	80% - 100%	59.82	46.50	12.31	41.27
D13_9	Informatica_MP	32.43	22.25	11.03	46.76
D60	50% - 70%	63.14	51.71	10.57	39.17
D56	2000 -| 3000	14.52	8.10	10.41	57.53
D57	2000 -| 3000	34.30	24.51	10.29	44.89
D40_3	Soddisfacente	53.67	42.66	10.21	40.35
D13_2	LingueStraniere_SD	58.92	49.16	9.00	38.45
D49	50% - 70%	58.09	48.60	8.74	38.34
D60	80% - 100%	23.79	16.64	8.62	45.87
FACOLTA	Ingegneria	23.65	16.66	8.43	45.54

The second cluster (Table 6), namely *Ambitious and Hopeful*s, is the largest one: 1446 students belong to it, equal to the 32.1% of the enrolled at the University of Calabria. The 70% of included units are male students, against an incidence of 41% of total first year students. As regards the cognitive skills almost a half of this cluster has hold a technical school diploma. It is high both the percentage of students who have reported very positive scores in Computer

Science (32% here against a 22% of the enrolled) and the incidence of those who have had quite satisfactory results in Italian or in Foreign Language at High School. The variables related to personality traits describe a cluster characterized by people with medium to high wage expectations. For the 34.3% their wage would fall monthly between 2000€ and 3000€ after the second level degree; in the opinion of 31% of them, their wage could reach a value between 1500€ and 2000€ after the first level degree. About the 60% of those who thinks their wage could vary between 1000€ and 1500€ after High School qualification is in the cluster. They are students with a high level of confidence in the efficacy of the degree: the 60% believes they can find an appropriate job after obtaining a degree, both with a first or a second level graduation. Similarly about 60% estimates between 50% and 70% the probability of completing studies at the scheduled time is here. Almost one-quarter of the students is enrolled in the Faculty of Engineering.

Table 7 - CLUSTER 3/5: Skillfuls and Resolutes (abs.: 1214, rel.: 26.9%)

Variable	**Feature**	**% of feature in the cluster**	**% of feature in the sample**	**test-value**	**% of the cluster in the feature**
MAT_NORM	100	52.31	21.41	29.24	65.80
D13_1	Italiano_MP	79.98	51.69	23.70	41.67
D13_2	LingueStraniere_MP	61.78	33.87	23.59	49.12
D12	Del tutto	31.05	12.91	20.66	64.78
D49	80% - 100%	62.52	40.42	18.23	41.66
D13_9	Informatica_NA	71.09	52.13	15.64	36.72
D11	Nessuna difficoltà	34.35	18.74	15.57	49.35
D13_5	Matematica_MP	46.54	28.88	15.46	43.39
D12	Molto	48.68	30.86	15.37	42.49
SESSO	F	76.36	59.29	14.46	34.68
D40_1	Soddisfacente	55.02	37.95	14.16	39.04
D10	Si	99.67	93.12	12.79	28.82
D23_1	Laurea	28.25	16.42	12.45	46.35
D23_2	Laurea	25.86	15.00	11.84	46.45
TIPO_DIPLOMA	Scientifico	50.91	36.89	11.68	37.16

In the third cluster (Table 7), namely *Skillfuls and Resolutes*, fall 1214 students, representing slightly more than one quarter of the enrolled sample. This cluster hosts a large percentage of students with high cognitive skills: the most of the students has the best High School qualification grade (about 52%), a remarkable presence if one considers that enrolled people who have had the highest grade (100/100) are just one fifth (21%). There is a large presence of students with high skills in different subjects: Italian literature and literacy (80%), Mathematics (46%) and Foreign Language (62%). More than one-third has found no difficulties in school experiences (34%), and the 99.7% has been ever successful at High School. Almost a half has got a High School qualification in science education. They have a great self-esteem, approximated by the statement of a strong matching between school votes and self-evaluation

(64.8%). Moreover the 62% is nearly sure (with a probability between 80% and 100%) to finish university studies in time are here. They feel to have a satisfactory level of information about educational goals of the academic course (about 40% of the cases). A family background with both graduate parents is shared by the students belonging to this cluster – about 46% both has a mother and a father with a university degree – in comparison with an incidence of 16% and 15% in the whole set. It should be noted that this cluster is not characterized by a specific preference in terms of Faculty chosen.

Table 8 - CLUSTER 4/5: Students by inertia and habit (abs.: 155, rel.: 3.4%)

Variable	Feature	% of feature in the cluster	% of feature in the sample	test-value	% of the cluster in the feature
D40_1	Insoddisfacente	70.32	2.66	27.53	90.83
D40_3	Insoddisfacente	58.71	2.13	25.18	94.79
D40_9	Insoddisfacente	54.19	8.01	15.39	23.27
D61	10% - 40%	34.19	12.53	7.05	9.38
D60	10% - 40%	57.42	31.57	6.67	6.25
D49	10% - 40%	22.58	10.98	4.13	7.07
D23_2	Laurea	27.74	15.00	4.07	6.36
D23_1	Laurea	29.03	16.42	3.92	6.08
D13_5	Matematica_NS	41.94	27.66	3.81	5.21
D13_2	LingueStraniere_NS	26.45	14.95	3.70	6.08
D47	Poco	10.32	4.10	3.27	8.65
D13_1	Italiano_NS	8.39	3.64	2.66	7.93
D40_3	Poco Soddisfacente	19.35	11.93	2.62	5.58
D57	3000 -| 4000	12.26	6.50	2.57	6.48
D24_1	Pensionato	7.74	3.53	2.41	7.55

The fourth cluster (Table 8), namely *Students by inertia and habit*, is the smallest one: there are only 155 students, equal to 3.4%. The group is characterized by a high percentage of students with poor cognitive skills: almost a half has a low scores in Mathematics, a quarter has unsatisfactory competencies in Foreign Languages and the 8% is insufficient in Italian literature and literacy. Most of the students in this cluster claims to be not informed about the organization of the degree course they are going to attend: the 70.3% of them states that they have not satisfactory information about the educational goals of the degree course (they are 2.6% in the sample), the 58.7% have no adequate information on profiles and job opportunities. Over one-quarter knows very little as regards teaching courses and the skills they are going to acquire. They are poorly motivated. Many of those attach low probability to find a job which requires their university degree (34% as concerns the second level degree course, 57% as regards the first level degree course). The 22% of students think there is a low probability (between 10% and 40%) of completing university studies in the scheduled time are here represented. The cultural family background, approximated by parents' degree, is high. In particular the percentage of students whose mother has get a university degree is the highest across

clusters (29%). The poor educational background of these students, the lack of importance attached to graduation, their reduced interest in knowing the organization of their course, the low study effort, depict this group as students who enrol to University by "inertia". Probably they are pushed to continue their studies for parents' solicitation, but without intrinsic motivation. It should be also noted that in this group is absent a characteristic Faculty.

Table 9 - CLUSTER 5/5: Optimistics and Proactives (abs.: 360, rel.: 8%)

Variable	Feature	% of feature in the cluster	% of feature in the sample	test-value	% of the cluster in the feature
D40_1	Molto Soddisfacente	74.44	6.90	37.01	86.17
D40_3	Molto Soddisfacente	79.17	11.25	32.65	56.21
D40_9	Molto Soddisfacente	43.61	4.04	26.86	86.26
D60	80% - 100%	38.06	16.64	10.19	18.27
D61	80% - 100%	71.39	46.50	9.90	12.26
D49	80% - 100%	63.89	40.42	9.31	12.62
D13_1	Italiano_MP	71.39	51.69	7.87	11.03
D13_9	Informatica_MP	39.44	22.25	7.63	14.16
D13_2	LingueStraniere_MP	51.67	33.87	7.19	12.18
D11	Nessuna difficoltà	33.89	18.74	7.10	14.44
D47	Molto	78.06	61.40	6.95	10.15
D12	Del tutto	25.56	12.91	6.74	15.81
D15	Si	64.72	50.60	5.57	10.21
D13_5	Matematica_MP	41.11	28.88	5.12	11.37
D28	Si	14.72	10.18	2.76	11.55

The last cluster (Table 9), namely *Optimistics and Proactives*, includes a restricted number of students, even if more numerous than the previous one: 360 units, the 8% of the enrolled. These students are optimistic, with a very positive view of their level of information, mostly with expectations of being successful in university studies and in getting a job matching their skills. They claim to have a very positive educational background, more in Italian literature and literacy (71% of the cases) or in Foreign Language (52%), a little less in Computer Science (39%) and Mathematics (41%). Over one-third of the students has not encountered any difficulties in learning at school. More than 74% of the cluster is formed by people who think their level of information about educational goals of their degree course or about courses to be attended and skills to be acquired is satisfactory, while the 79% of them consider the knowledge about job opportunity once graduated very satisfying. They have a high level of trust in their possibility of completing university studies on time (the 64% of the cases), or in the chance to get a job matching their cognitive skills, less as regards the first level degree course (38%), more as concerns the second level degree course (71%). Most of them have had extra-school experiences (64%), while the 15% of ones with working experiences. There is no characterisation about the type of High School attended nor the choice of the Faculty.

7 References

Arabie P., Hubert L.J., De Soete G. (1996), *Clustering and Classification*, World Scientific.

Benzécri J.-P. (1973), *L'analyse des données. Vol. 2. L'analyse des correspondances*, Dunod.

Benzécri J.-P. (1982), Construction d'une classification ascendante hiérarchique par la recherche en chaîne des voisins réciproques, *Les Cahiers de l'Analyse des Données*, 7, 2: 209-218.

Bratti M., Checchi D., Filippin A. (2007), *Da dove vengono le competenze degli studenti? I divari territoriali nell'indagine OCSE PISA 2003*, Il Mulino.

Brunello G., Schlotter M. (2011), *Non Cognitive Skills and Personality Traits: Labour Market Relevance and their Development in Education & Training Systems*, IZA Discussion Paper, 5743, Forschungsinstitut zur Zukunft der Arbeit.

Carneiro P, Crawford C., Goodman A. (2007), *The Impact of Early Cognitive and Non Cognitive Skills on Later Outcomes*, University College. http://cee.lse.ac.uk/.

Checchi D. (2010), *Uguaglianza delle opportunità nella scuola secondaria italiana*, FGA working papers, 25, Fondazione Agnelli.

Duncan G.J., Dowsett C.J., Claessens A., Magnuson K., Huston A.C., Klebanov P., Pagani L.S., Feinstein L., Engel M., Brooks-Gunn J., Sexton H., Duckworth K., Japel C. (2007), School Readiness and Later Achievement, *Developmental Psycology*, 43, 6: 1428-1446.

Gordon A.D. (1999), *Classification*, Chapman & Hall/CRC.

Hanushek E., Kimbo D. (2000) Schooling, Labor-Force Quality, and the Growth of Nations, *American Economic Review*, 90, 5: 1184-1208.

Hanushek E., Woessmann L. (2008), The Role of Cognitive Skills in Economic Development, *Journal of Economic Literature*, 46, 3: 607-668.

Heckman J. (1999), Policies to Foster Human Capital, *NBER Working Paper*, 7288.

Heckman J., Sixtrud N., Urzua S. (2006), The effects of Cognitive and Non cognitive abilities on Labour Market Outcomes and Social Behaviour, *Journal of Labour Economics*, 24, 3: 411-482.

Lebart L., Morineau A., Warwick M.K. (1984), *Multivariate Descriptive Statistical Analysis*, Wiley & Sons.

MIUR (2010), *Focus in breve sull'istruzione - Aggiornamento sugli esiti degli Esami di Stato della scuola secondaria di II grado. Andamento nel quinquennio 2004/2005 - 2008/2009*, Direzione Generale per gli Studi, la Statistica e per i Sistemi Informativi.

SISPICAL (2009), *Esiti dei diplomandi nelle scuole secondarie di secondo grado della Calabria - Anno Scolastico 2008/2009*, Osservatorio Regionale sull'Istruzione e il Diritto allo Studio. http://www.regione.calabria.it/.

OECD (2010), *The High Cost of Low Educational Performance. The Long-Run Economic Impact of Improving Pisa Outcomes*, Programme for International Student Assessment.

ABSTRACT

In this paper data collected on the students enrolled at University of Calabria (Unical) have been analysed, aiming at investigating differences and analogies in their educational and personal background, as well as their expectations. In a statistical perspective, by means of a multivariate strategy, first year university students profiles have been analysed and classified in different typologies. The main goal is to provide a classification of students which can be useful to project and design specific interventions and measures (such as additional courses, peer tutoring, care and encouragement of excellent students), with the aim of upgrading human capital formation, tailored to the individual characteristics of each group.

The development of students' competencies is a complex task, which depends not only on the educational system, but also on several other factors such as the quality of the socio-institutional context, the economic environment and the family background; nevertheless, they have a crucial role in promoting economic well-being.

A number of studies give evidence that cognitive skills of the population are highly related to individual earnings, to income distribution, and to economic growth, providing both individual and social returns. As a matter of facts, governments frequently try to improving cognitive skills, for example through programs rising the quality of education. Unfortunately these programs often slip down the policy agenda because they pay off in the future (OECD, 2010), or they suffer of a not-in-deep investigation of the different interventions students actually need. Students who attend the same class, in fact, have different cognitive and non-cognitive skills, including social motivation, relational abilities, political concerns. Programs aimed at enhancing human capital formation could be more efficient and effective if specifically tailored on the whole individual characteristics, comprehensive of cognitive skills and personality traits.

The aim of this paper is to give an empirical contribution in this direction by analysing individual data on students enrolled at the University of Calabria in the 2008-2009 academic year and collected by a questionnaire submitted to students at the enrolment time.

In order to group students in different clusters on the basis of both their cognitive skills and individual traits, a Multiple Correspondence Analysis and a Cluster Analysis have been carried out. The main goal is trying to obtain a classification of students "in entrance" more articulated than the usual partition in "low" and "high" ability students based only on tests scores. The idea underlying the strategy is two-fold: on one side motivations and competencies couldn't be measured by just the test scores, but involves several other personal traits, on the other side clustering based both on cognitive skills and personal traits could help in reducing biases in human capital interventions.