Identification of Behavior Patterns Within Graduated Students and Undergraduate Modules at the Technical University of Cartagena, Spain

A. Molina-García[1], M. Kessler[2] and A. Botía[3]
[1]Department of Electrical Eng., Universidad Politécnica de Cartagena, 30202 Cartagena, Spain
[2]Department of Applied Mathematics and Statistics, Universidad Politécnica de Cartagena, 30202 Cartagena, Spain
[3]Research Results Transfer Office, Universidad Politécnica de Cartagena, 30202 Cartagena, Spain

Keywords: Undergraduate Curricula, Learning Behaviour Patterns.

Abstract: As a consequence of the governmental decision to adapt the Spanish graduate and post-graduate studies to converge to the 'European Higher Education Area', the goal of the so-called Bologna Process, committees of experts were set up at the Technical University of Cartagena, located south of Spain, to design the new curricula that would build up the restructured offer of courses. It was decided to provide as supporting material to these committees statistical information about the academic behaviour and results of the students in modules of the existing courses. In this paper the main aspects of this study are presented, discussing the set of variables selected to characterize modules and students. Information about the structure of variability between students on one hand and between modules on the other hand is presented, based on a principal component analysis. Finally patterns were identified among modules and among students using a cluster procedure. The influence of relevant factors like gender, course and marks obtained at the School Leaving certificate on the resulting groups composition was explored as well.

1 INTRODUCTION

The Technical University of Cartagena, located south of Spain, offers a generalist engineering education, with emphasis in engineering fundamentals and practices attending approximately 6000 students. Nowadays, these undergraduate degrees usually involve five academic years and allow the students to continue their university education through the Doctor of Philosophy (PhD) degree. It is possible for the students to focus on a specific field of interest: mechanical, electrical, chemical, civil; within the last stages. A complete restructuration of the courses was undertaken in the last years as a consequence of the governmental decision to adapt the Spanish graduate and post-graduate studies to converge to the 'European Higher Education Area', the goal of the so-called Bologna Process, the inter-governmental process that promotes reforms in higher education with 47 countries. Consequently committees were set up on one hand on a national level, where experts from both academia and private sectors would establish a list of recommendations for the design of the new syllabuses, but also on a local level within each university, where representatives of each department involved in the teaching programs would concretely decide about the modules that would build up the curricula.

A statistical study of academic indicators computed from data collected during the last decade was then launched in our university. Two datasets were built: dataset S, consisting of more than 1000 students that successfully graduated from the School for Industrial Engineering at the Technical University of Cartagena, and dataset M, describing more than 200 modules along the last decade as well. The purpose of this study was to provide updated diagnostic regarding patterns of behaviour within the students and, simultaneously, identify groups of modules within the different courses that are offered, which would present homogeneity in terms of student behaviour. It must be emphasized that within the Spanish system, for a given module, each student has three opportunities to take the associated exam along a given academic year.

Molina-García A., Kessler M. and Botía A..
Identification of Behavior Patterns Within Graduated Students and Undergraduate Modules at the Technical University of Cartagena, Spain.
DOI: 10.5220/0004964305820587
In *Proceedings of the 6th International Conference on Computer Supported Education* (CSEDU-2014), pages 582-587
ISBN: 978-989-758-020-8

Moreover, it is not compulsory to pass successfully all the modules corresponding to a given stage in order to proceed to the next stage. As a consequence, a significant variability is observed between the student behaviours and strategies, resulting into variability between their trajectories within the university studies. Variability is also present therefore between modules. In this paper, the main results and conclusions of the statistical study are presented.

The rest of the paper is structured as follows. Section 2 gives a brief description about the most relevant variables used to provide, on one hand, information about the students' patterns of academic behaviour and, on the other hand, information that could help discriminating between modules. Section 3 presents preliminary results of an exploratory analysis of these variables. Multidimensional analysis techniques were then applied, and the results of a principal component analysis are discussed in Section 4. Finally, relevant conclusions are listed in Section 5.

2 VARIABLES AND FACTORS

As mentioned in Section 1, two multi-dimensional datasets S and M (Students data and Modules data respectively) are constructed by merging several university records databases. The essential one is the examination records database, but a student personal information database, is used to recover data like birth-date, gender or marks achieved at the School Leaving Certificate.

2.1 Data-Set *M*

For each module and each academic year from 2002/2003 to 2008/2009, the following variables have been computed:

- *TOOK_EXAM*: Within the students that registered for that module that year, proportion of students that took the exam at least in one out of the three possible opportunities.
- *PASSED_EXAM_REGISTERED*: Within the students that registered for that module that year, proportion of students that passed the exam.
- *PASSED_EXAM_TOOK_EXAM*: Within the students that took the exam for that module at least once out of the three possible opportunities, proportion of students that passed the exam.
- *OPP_TAKE_EXAM*: Within the students that took the exam for the first time that year, average number of opportunities they have had to take that same exam previously. This variable may require a little bit more explanation: since it is not compulsory for a student that has registered for a given module to take the exam, it happens that some students decide eventually not to take the exam in the first opportunity they have, neither in the second opportunity, or even end up not taking the exam at all that year. It is therefore of interest in particular to assess the perception that the students have regarding the difficulty of passing the exam associated to the module, to check, within the students that took the exam for the first time, how many times have they waited before daring to do it.
- *AVG_MARK*: Within the students that passed the exam that year, the average mark (on a 0 to 10 scale).
- *NUM_EXAM_PASS*: Within the students that passed the exam that year, the average number of times they had to take the exam to actually pass the module. It therefore amounts to the number of times they failed plus one.

2.2 Data-Set *S*

For each student that successfully graduated within the years 2001/2002 to 2008/2009 at the Industrial Engineering School, the following variables have been computed:

- *DURATION*: Relative duration of the studies, i.e. the number of years it took for the student to graduate divided by the number of stages in the course.
- *MARK*: Weighted average mark on a 0-10 scale. The weights are proportional to the ECTS assigned to each module.
- *OPP_TAKE_EXAM*: Within all compulsory modules, average number of opportunities that the student used to actually take the exam for the first time. (For more explanation, see Database M).
- *NUM_EXAM_PASS*: Within all compulsory modules, average number of times that the student had to take the exam to pass it (see as well Database M).

A series of additional variables or factors were also included to check their association with the variables of interest: gender, age at graduation, identification of the High School of origin, mark achieved at the School Leaving Certificate. (Called "*Selectividad*" in Spanish). The dataset S contains 1087 students.

3 FIRST DESCRIPTIVE INDICATORS

As a first step into the data exploration, a descriptive analysis was carried out making an intensive use of graphics and numerical indicators. A brief summary is presented in this Section to provide a sense of the orders of magnitude and the variability of the different variables.

3.1 Data-Set *M*

The dataset *M* contains the evolution of almost 200 modules over the considered academic years. Compulsory modules were only considered in this dataset, since optional modules present a high homogeneity in terms of student behaviours and performance. In Table 1, the first quartile, median, third quartile, mean, and standard deviation of the relevant variables are presented.

We may for example pinpoint a few figures out of Table 1: the average number of opportunities {*OPP_TAKE_EXAM*} that the students use to actually take an exam is close to 2, and some modules present really high values for that variable. On average almost 80% of the students that take at least once the exam of a module at the Industrial Engineering School pass {*PASSED_EXAM_TOOK_EXAM*}, while 70% of the registered students take the exam at least once. The number of times the students take an exam until they pass {*NUM_EXAM_PASS*} takes typically lower values than the number of opportunities that the students use to take the exam for the first time, a fact that already seems to anticipate that the perceived difficulty of a module before taking the exam is higher than the "real" difficulty to pass it.

From a quick check of the values of the Pearson correlation coefficients, we can emphasize the following:

- The highest correlation is found between the proportion of students that pass the exam with respect to the total number of registered students (*PASSED_EXAM_REGISTERED*) and the proportion of registered students that take the exam at least once: $cor(*PASSED_EXAM_REGISTERED*,*TOOK_EXAM*)= 0.86$.
- The second highest correlation happens to be found between the proportion of registered students that pass the exam and the proportion of students that pass the exam with respect to the number of students that take the exam at least once: $cor(*PASSED_EXAM_REGISTERED*,*PASSED_EXAM_TOOK_EXAM*)=0.75$. This is of course very natural; both variables depend directly on the number of students that pass the exam.
- The variables *PASSED_EXAM_TOOK_EXAM* and *NUM_EXAM_PASS* present a correlation of 0.66, which reveals a clear (and expected) association between the proportion of success when taking an exam and the average number of times the student has to take the exam before actually passing it.
- The lowest degrees of association are found between the variables *TOOK_EXAM* and *OPP_TAKE_EXAM* with the variables *NUM_EXAM_PASS*, *PASSED_EXAM_TOOK_EXAM* and *AV_MARK_10*, see values of correlations in Table 2, which tends to confirm that there is no strong relation between the perception of the student in terms of anticipating his possibilities of success at an exam (measured through his disposition to take the exam) and the actual difficulty to pass if he takes the exam.

3.2 Data-Set *S*

The dataset S contains 1087 students that successfully graduated from the Industrial Engineering School at the Technical University of Cartagena. 932 were male students while 155 were female students. In Table 2, the first quartile, median, third quartile, mean, and standard deviation of the relevant variables are presented. It may pointed out the high values that takes the variable *DURATION*, the centre of its distribution corresponding to an increment by 2 thirds of the expected "theoretical" duration before graduation.

This is explained partly by the fact that the students end their studies by a final project requiring full time investment while they also have to take modules until the end. They then usually begin an extra academic year to complete and present the project, which then adds one year to the absolute duration even if they actually use only a few extra months to do it.

On the other hand, half of the graduated students present in their academic trajectory an average close to 2 as for the number of opportunities they use before actually taking an exam, while the average of times they have to take the exam to pass it, is close to 1.

Table 1: Dataset M. Relevant variables.

NUM_REGISTERED	TOOK_EXAM	OPP_TAKE_EXAM	NUM_EXAM_PASS	PASSED_EXAM_REGISTERED	PASSED_EXAM_TOOK_EXAM	AV_MARK_10
Min. 1.0	Min 0.1667	Min 0.000	Min 0.000	Min 0.1239	Min 0.2609	5.000
1st Qu. 50.0	1st Qu. 0.5433	1st Qu. 1.443	1st Qu. 1.250	1st Qu. 0.3647	1st Qu. 0.6581	1st Qu. 5.867
Median 82.0	Median 0.667	Median 1.969	Median 1.571	Median 0.4921	Median 0.7778	Median. 6.176
Mean 90.65	Mean 0.6678	Mean 2.204	Mean 1.605	Mean 0.5273	Mean 0.7748	Mean 6.340
3rd Qu. 119.0	3rd Qu. 0.7987	3rd Qu. 2.770	3rd Qu. 1.885	3rd Qu. 0.6618	3rd Qu. 0.9104	3rd Qu. 10.00
Max. 296.0	Max. 1.000	Max. 11.33	Max. 3.091	Max. 1.000	Max. 1.000	Max. 10.00
Sd. 58.98	Sd. 0.17	Sd. 1.02	Sd. 0.43	Sd. 0.21	Sd. 0.16	Sd. 0.71

Table 2: Dataset M. Variable correlation results.

	NUM_EXAM_PASS	PASSED_EXAM_TOOK_EXAM	AV_MARK_10
TOOK_EXAM	-0.3015765	0.3496622	0.2827057
OPP_TAKE_EXAM	0.4034407	-0.3738091	-0.3646772

Table 3: Dataset S. Relevant variables.

AGE	MARK_SLC	DURATION	MARK_GRAD	OPP_TAKE_EXAM	NUM_EXAM_PASS
Min. 21	Min 5.020	Min 1.000	Min 5.350	Min 1.000	Min 1.000
1st Qu. 24	1st Qu. 6.274	1st Qu. 1.500	1st Qu. 6.100	1st Qu. 1.380	1st Qu. 1.200
Median 25	Median 7.066	Median 1.670	Median 6.390	Median 1.780	Median 1.380
Mean 25.27	Mean 7.055	Mean 1.792	Mean 6.478	Mean 1.918	Mean 1.446
3rd Qu. 26	3rd Qu. 7.752	3rd Qu. 2.000	3rd Qu. 6.755	3rd Qu. 2.310	3rd Qu. 1.615
Max. 49	Max. 9.680	Max. 4.500	Max. 9.490	Max. 6.220	Max. 3.000
Sd. 3.27	Sd. 0.97	Sd. 0.54	Sd. 0.55	Sd. 0.71	Sd. 0.34
	NA's 170000				

As for dataset M, a first glance at the correlation structure (see Table 4) reveals some interesting facts:

- The correlation between the mark achieved at the School Leaving Certificate MARK_SLC and the remaining variables is negative except with MARK_GRAD for which it is close to 0.5.
- There is a rather strong negative association between the variable *NUM_EXAM_PASS* and the final mark achieved at graduation, which seems to indicate that when the students have failed an exam, they usually do not achieve good marks when they finally pass.

A more complete exploratory analysis was performed by systematically splitting the datasets according to the levels of the considered factors, distinguishing between courses, students gender, high school of origin for dataset S and courses, stage, temporal location of the module for dataset M, and computing accordingly numeric indicators and displaying graphs.

4 PRINCIPAL COMPONENT ANALYSIS

Our interest was in particular in gaining understanding about the structure of variability between modules on one hand, and between students on the other hand. Among the considered variables, *which ones contribute the most to discriminating between modules or students?* A principal component analysis was then performed on the correlation matrix in both datasets.

4.1 Data-Set *M*

For dataset M the six variables: *TOOK_EXAM, OPP_TAKE_EXAM, NUM_EXAM_PASS, AV_MARK_10 PASSED_EXAM_REGISTERED,*

*PASSED_EXAM_TOOK_EXA*M, were considered for the principal component analysis based on their correlation matrix. The proportion of variability explained by the first, the first two and the first three components is respectively 58%, 74% and 84%, and the expression of the first three components is the following:

$$PC1 = 0.40\ TOOK\_EXAM - 0.37 OPP\_TAKE\_EXAM + (-0.39)\cdot NUM\_EXAM\_PASS + 0.50 \bullet PASSED\_EXAM\_REGISTERED + 0.43 \bullet PASSED\_EXAM\_TOOK\_EXAM + 0.34\ AV\_MARK\_10 \quad (1)$$

$$PC2 = (-0.59) \bullet TOOK\_EXAM + 0.31\ OPP\_TAKE\_EXAM + (-0.46) \bullet NUM\_EXAM\_PASS - 0.23\ PASSED\_EXAM\_REGISTERED + 0.33\ PASSED\_EXAM\_TOOK\_EXAM + 0.42 \bullet AV\_MARK\_10 \quad (2)$$

$$PC3 = 0.03\ TOOK\_EXAM + 0.45\ OPP\_TAKE\_EXAM + (-0.17) \bullet NUM\_EXAM\_PASS + 0.26\ PASSED\_EXAM\_REGISTERED + 0.46\ PASSED\_EXAM\_TOOK\_EXAM - (-0.70) \bullet AV\_MARK\_10 \quad (3)$$

The first component is clearly a weighted average of the different variables, assigning positive weights for the variables for which high values indicate good results and negative weights for the two variables for which low values indicate good results *OPP_TAKE_EXAM* and *NUM_EXAM_PASS*. The modules that score higher in this first component can thus be considered as the most successful modules, in terms of student behaviour and results. As for the second component the following interpretation is suggested: PC2 measures the difference between the difficulty perceived by the students before the exam and the actual difficulty to pass the module. Indeed it can be written as the sum of two terms: (-*TOOK_EXAM - PASSED_EXAM_REGISTERED + OPP_TAKE_EXAM*) and *(AV_MARK_10 + PASSED_EXAM_TOOK_EXAM - NUM_EXAM_PASS)*. The first term takes its highest values for modules where a low proportion of students take the exam and therefore the proportion of students that pass the module with respect to the registered students is low and the number of opportunities used to actually take the exam for the first time is high.

The influence of the different factors (Stage, temporal location in the academic year, course, and assigned ECTS value) on the principal components scores was explored. Differences between the different courses at the School for Industrial Engineering was found for the PC2, where two courses scored typically higher: the Industrial Management Engineer course and the Automatism and Industrial Electronics course, where a significant part of the students work and therefore have a higher tendency to use several opportunities before taking an exam. Principal components plots (PC2 versus PC1) were also provided to follow the temporal evolution of the module scores. As an example, the scores obtained for the considered years by the modules of the first stage of the Industrial Engineering course are shown in Figure 1. This kind of plot allows monitoring the evolution, for a given module, of the student behaviour and results and detecting possible difficulties.

4.2 Data-Set *S*

For dataset *S* five variables *MARK_SLC, DURATION, MARK_GRAD, OPP_TAKE_EXAM, NUM_EXAM_PASS* were considered for the principal component analysis based on their correlation matrix. The proportion of variability explained by the first, the first two and the first three components is respectively 55%, 75% and 85%, and the expression of the first three components is:

$$PC1 = 0.37 \cdot MARK\_SLC + (-0.47) \cdot DURATION + +0.46 \cdot MARK\_GRAD - 0.45 \cdot OPP\_TAKE\_EXAM - -0.46\ NUM\_EXAM\_PASS \quad (4)$$

$$PC2 = -0.56\ MARK\_SLC + -0.48\ DURATION - 0.37 \cdot MARK\_GRAD - 0.54 \cdot OPP\_TAKE\_EXAM + 0.19\ NUM\_EXAM\_PASS \quad (5)$$

$$PC3 = 0.74\ MARK\_SLC - 0.06\ DURATION - 0.42 \cdot MARK\_GRAD - 0.25\ OPP\_TAKE\_EXAM + 0.46 \cdot NUM\_EXAM\_PASS \quad (6)$$

As for dataset *M*, the first component is easy to interpret as a global score of the graduated student's success: it consists of a weighted average of all variables, with positive weights for variables that translate positively in terms of the student's results and negative weights for variables for which high values would indicate worst results. The second component is interpreted as the sum of two terms: *MARK_SLC+MARK_GRAD - NUM_EXAM_PASS* and *DURATION + OPP_TAKE_EXAM.*

The weights have been omitted, which reflect on one hand the student's performance (their marks and the number of times they need to take an exam to pass the module) and on the other hand their apprehension before taking an exam (*OPP_TAKE_EXAM*) which have of course an influence on the duration of their studies. The

students that score most negatively on *PC2* have achieved good or very marks at the School Leaving Certificate and during their graduate studies but have taken a rather long time to graduate and have used several opportunities before actually taking an exam.

Finally the third component can be seen as the influence of two differences: *MARK_SLC-MARK_GRAD* and *NUM_EXAM_PASS-OPP_TAKE_EXAM*: students that score high in *PC3* are students that have achieved lower marks during their graduate studies than was expected from their School Leaving Certificate and that, although they usually take the first opportunities to take an exam, frequently fail and need to repeat the exam several times to pass the module.

As for dataset *M* in the previous subsection, an exploration of the influence of the factors considered in the dataset on the principal components was carried out. As an example, interesting facts was the influence of the Gender on the *PC3* score. Consider for example the Technical Industrial Engineer course, mention Industrial Chemistry, where approximately half of the graduated students are female (concretely 49 out of 101 individuals in the dataset *S*), consider the lower part of the *PC3* scores that contains 10 % of the students, i.e. the portion of the dataset with *PC3* scores values under the corresponding 10% quartile, only 1 out of the 9 corresponding students is female, while if you consider the 10% upper part, 8 out of the 9 included students are female.

5 CONCLUSIONS

In this paper the main aspects of a statistical study about students' academic behaviour and results at the School for Industrial Engineering in the Technical University of Cartagena are presented. This study was initiated to provide supporting material to the local committees at the university, who are in charge of designing the new curricula and syllabuses as a consequence of the restructuration of the courses in the context of convergence to the 'European Higher Education Area'. Two datasets were considered, one focused on modules and the evolution of associated academic indicators, and the other one focused on graduated students.

This datasets allowed us to explore systematically and confirm (or refute) the existing impressions or empirically gained knowledge of the members of the committees based on their experience as teachers and managers at the university. Additionally, it also allowed identifying a few modules with atypical results and quantifying up to which extent they behaved really differently from the other modules in the same course, and opened thus the door to a possible action from the responsible of the School.

Multivariate analysis techniques like principal component analysis and clustering have been used to understand better the structure of the variability between modules and between students, and permitted to define sensible partitions of the datasets: five groups of modules characteristics were proposed while six profiles for the graduated students were suggested. It is then particularly interesting to check the association of relevant factors with the groups' composition: for example the differences in gender composition between the different students' profiles, or the differences in the groups relative size between the different courses.

ACKNOWLEDGEMENTS

The authors are very grateful to the Vice-Chancellor for European Convergence and Quality Standards Education, *Josefina García-León*, for her constant support; and to the Chief Computing Developer, *José Sánchez-Manzanares* for endless technical help and advice.

REFERENCES

http://europa.eu/legislation_summaries/education_training_youth/lifelong_learning/c11088_en.htm.

http://www.educacion.es/educacion/que-estudiar-y-donde/bachillerato/opciones-despues-bachillerato/pau.html.

Mardia, Kantilal Varichand, Kent, John T. and Bibby, John M; *Multivariate analysis: Probability and Mathematical Statistics: A Series of Monographs and Textbooks*; Academic Press Harcourt Brace Jovanovich Publishers, London 1979.

Industrial Engineering Curriculum, Spanish Government Reporter, (203), August 2000.

European Commission - Education and Training, Online. [Available], http://ec.europa.eu.