

Key Courses of Academic Curriculum Uncovered by Data Mining of Students' Grades

Ł.G. GAJEWSKI^a, J. CHOLONIEWSKI^{a*} AND J.A. HOLYST^{a,b}

^aFaculty of Physics, Center of Excellence for Complex Systems Research, Warsaw University of Technology, Koszykowa 75, PL-00662 Warsaw, Poland

^bITMO University, 19, Kronverkskiy av., 197101 Saint Petersburg, Russia

Learning is a complex cognitive process that depends not only on an individual capability of knowledge absorption but it can be also influenced by various group interactions and by the structure of an academic curriculum. We have applied methods of statistical analyses and data mining (principal component analysis and maximal spanning tree) for anonymized students' scores at Faculty of Physics, Warsaw University of Technology. A slight negative linear correlation exists between mean and variance of course grades, i.e. courses with higher mean scores tend to possess a lower scores variance. There are courses playing a central role, e.g. their scores are highly correlated to other scores and they are in the centre of corresponding maximal spanning trees. Other courses contribute significantly to students' score variance as well to the first principal component and they are responsible for differentiation of students' scores. Correlations of the first principal component to courses' mean scores and scores variance suggest that this component can be used for assigning ECTS points to a given course. The analysis is independent of declared curricula of considered courses. The proposed methodology is universal and can be applied for analysis of students' scores and academic curriculum at any faculty.

DOI: [10.12693/APhysPolA.129.1071](https://doi.org/10.12693/APhysPolA.129.1071)

PACS/topics: 89.65.-s, 89.65.Ef

1. Introduction

Teaching and learning play a pivotal role for development of an individual as well as of the human civilization [1]. In fact, we spend about 18 years for our education and complexity of this process can be studied at different levels, e.g. scores of an individual pupil/student, scores of learning groups, scores related to different courses, changes of scores during the study periods but also early drop-out prediction [2] etc. [3]. Advancements in various platforms for e-learning (such as massive open on-line courses [4]) make some analyses obvious, but a progress in digital tools related to educational processes (such as USOS [5] — virtual deanery) and their growing popularity make possible new kinds of analyses where data mining methods can be applied. Results of such analyses can lead to improvement of teaching curricula and study plans at schools/universities.

In the present paper, we analyse grades at Faculty of Physics, Warsaw University of Technology during the first five semesters of the first level studies leading to the engineering degree in applied physics. The core of curriculum consists mostly of courses in physics, mathematics, computer science and electrical engineering. After the 5th semester a student selects one of specializations: optoelectronics, materials and nanostructures (formerly solid state physics), computer physics, or medical physics. We are interested to learn dependences between students' scores at different courses. To uncover these

dependences we shall use specific tools of data mining methods which are widely known and applied in complex systems analyses, namely correlation matrices [6, 7], correlation-based networks [8], minimal (maximal) spanning trees [9–11], and principal component analysis [12–15].

As far as we know, this is the first study of teaching results at Polish universities using data mining tools. A similar methodology was applied for learning data at Singaporean primary schools [16] but it was focused on dynamics of single lessons rather than on a whole teaching curriculum. We hope that the framework proposed in this paper can be applied to data from other faculties and can be helpful to better understand and organize teaching/learning process.

2. Data acquisition and filtering

The data set has been received from the deanery of Faculty of Physics at Warsaw University of Technology and it has contained information about students' final grades of particular courses as well as students' group and specialization. The data contains records of engineering degree candidates since 2010 until 2015.

We have excluded from the analysis all students who did not pass the first five semesters and we have chosen three consecutive students' classes: J, K and L. The class J means all students that started their academic studies in 2010, K in 2011 and L in 2012. Those classes have been chosen because only for them there was a large number of records of core courses that every student must complete in order to get a degree. After removing general University courses like physical education or foreign languages we have got 27 core courses (see Table I). Each student (and course) in our analysis is represented as a vector of

*corresponding author; e-mail: choloniewski@if.pw.edu.pl

numbers — grades. The total number of students considered in the analyses is 217 (class J — 76 students, K — 66, L — 75).

Since there have been still some missing records for individual students, we have filled those gaps with an average score at that course in all three classes. Amounts

of missing grades at each class are: class J — 5.46%, K — 14.03%, L — 1.68%, and for all classes combined — 6.72%.

Abbreviations of courses names as well as their mean grades and variances of grades (and other calculated parameters) are listed in Table I.

TABLE I

Core courses chosen for analysis — names, abbreviations used in figures, semester at which they take place, various parameters calculated in the paper. Two highest values in columns 4–8 are **bold**, and the two lowest are underlined. In the last column, only the highest value has been highlighted. 1st PCC stands for “1st Principal Component contribution”.

Course	Abbreviation	Semester	Mean grade	Variance	Sum of correlations	Explained variance	1st PCC	Node degree (MST)
Algebra and Geometry	AG	1	3.72	0.45	10.93	0.033	0.21	2
Analysis of Experimental Data	AED	1	4.12	0.25	8.93	0.018	0.12	1
Fundamentals of Information Technology	FIT	1	4.58	<u>0.18</u>	7.45	<u>0.013</u>	<u>0.090</u>	1
Fundamentals of Physics 1	FP1	1	3.51	0.48	12.29	0.035	0.24	9
Mathematical Analysis 1	MA1	1	3.49	0.64	10.18	0.046	0.25	1
Basics of Programming	BP	2	4.18	0.53	8.06	0.038	0.18	1
Fundamentals of Physics 2	FP2	2	3.62	0.51	11.52	0.037	0.24	3
Mathematical Analysis 2	MA2	2	<u>3.45</u>	0.73	11.2	0.053	0.29	3
Physics Laboratory 1	PL1	2	4.07	0.3	7.92	0.021	0.12	1
Fundamentals of Electronics	FE	3	3.56	0.67	10.1	0.048	0.24	1
Mathematical Analysis 3	MA3	3	<u>3.43</u>	0.79	10.03	0.057	0.27	1
Mechanics	M	3	3.85	0.49	10.18	0.035	0.21	3
Programming Languages	PL	3	4.03	0.89	8.31	0.064	0.24	3
Electrodynamics	E	4	4.37	0.35	8.03	0.025	0.14	1
Electronics in Physical Experiment	EPE	4	4.55	<u>0.19</u>	7.09	<u>0.014</u>	<u>0.077</u>	2
Mathematical Methods of Physics	MMP	4	4.16	0.54	8.68	0.039	0.18	1
Object-Oriented Programming	OOP	4	4.21	0.47	8.06	0.034	0.15	1
Probability	P	4	3.56	0.63	9.25	0.046	0.23	1
Quantum Physics	QP	4	3.70	0.97	9.64	0.070	0.29	2
Chemistry	C	5	4.02	0.38	9.55	0.027	0.16	2
Engineering Graphics	EG	5	3.72	0.51	<u>5.97</u>	0.036	0.11	1
Fundamentals of Optics	FO	5	3.62	0.5	<u>5.01</u>	0.036	0.091	1
Fundamentals of Virtual Devices Design	VDD	5	3.48	0.6	8.58	0.044	0.20	2
Introduction to Solid State Physics	ISSP	5	3.97	0.6	7.18	0.044	0.18	1
Introduction to Nuclear Physics	INP	5	3.92	0.28	8.03	0.020	0.11	2
Physics Laboratory 2	PL2	5	4.24	0.27	7.54	0.020	0.11	2
Statistical Physics and Thermodynamics	SPT	5	4.09	0.67	8.21	0.048	0.19	3

3. Methodology

As a number of core courses is the same for each student, the dataset can be presented as $N_S \times N_C$ matrix of grades $G_{s,c}$ (where N_S — number of students, N_C — number of courses, s — student index, C — course index). Positions of columns and rows are arbitrary but set.

Such a representation allows us to easily define a measure of a coincidental deviations from the means between

courses as a Pearson product-moment correlation coefficient [17]:

$$C_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}, \tag{1}$$

where in our case $X = G_{*,X}$ and $Y = G_{*,Y}$ are vectors of grades for given courses X and Y obtained by students which can be perceived as series of observations of random variables, $\text{cov}(X,Y)$ is a covariance between these grades X and Y , and σ_α is a standard deviation from the

mean of the variable α . A positive (negative) correlation $C_{X,Y}$ between two courses is when variations from corresponding means tend to have the same (opposite) sign for a given student.

A courses' network can be built using calculated values $C_{X,Y}$ as links' weights. Such a network is unfortunately a hardly readable weighted complete graph. To overcome this problem we have applied the inverted Kruskal algorithm [18] and we have reduced the complete graph to the related maximal spanning tree (MST). Let us remind that the MST possesses the same number of nodes as the original network and the inverted Kruskal algorithm sorts a list of edges from the original network by correlations $C_{X,Y}$ descending and iterates through the list adding an edge if it does not cause a cycle to occur.

The principal component analysis [19] is a method of variables transformation equivalent to a special rotation of coordination axes. Let us write scores of a given student as a vector variable \mathbf{x} where the dimension of this vector equals to the number of all courses the student attended. One should start with centring the analysed variable \mathbf{x} :

$$\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{E}(\mathbf{x}). \quad (2)$$

Here $\mathbf{E}(\dots)$ stands for the vector of expected values, that is calculated as a vector of means over all students for each course's score. The first principal component is a direction $\gamma_{(1)}$ in the new framework where the variance of all observed data points projected onto this direction is maximal

$$\text{var}(\gamma_{(1)}^T \tilde{\mathbf{x}}) = \max_{\mathbf{a} \in \mathbb{R}^p, \|\mathbf{a}\|=1} \{\text{var}(\mathbf{a}^T \tilde{\mathbf{x}})\}, \quad (3)$$

where $\gamma_{(1)}^T$ — the transverse vector defining the first axis of transformed data, $\text{var}(\dots)$ — variance of a given variable. The axis of the second principal component $\gamma_{(2)}$ is orthogonal to the first one

$$\gamma_{(1)}^T \gamma_{(2)} = 0 \quad (4)$$

and it maximises the remaining part of variance

$$\text{var}(\gamma_{(2)}^T \tilde{\mathbf{x}}) = \max_{\mathbf{a} \in \mathbb{R}^p, \|\mathbf{a}\|=1} \{\text{var}(\mathbf{a}^T \tilde{\mathbf{x}})\} \quad (5)$$

and so on. In fact the transformation is equivalent to finding eigenvectors of covariance matrix corresponding to maximal eigenvalues (they are real because the matrix is real and symmetric). In such a way one can reduce an original highly dimensional space to a few directions corresponding to largest variances [20].

For the sake of comparison, two types of shuffling have been performed. The first type (S — “by students' grades”) is to replace each row of a matrix of grades G_{s*} with its random permutation. Similarly the second type (C — “by courses' grades”) is to replace each column of the matrix G_{*c} with its random permutation. While the type S shuffling keeps the mean grade of a student the same, one can expect it destroys possible relations between courses. Exact opposite is for the type C shuffling (keeping courses' means and destroying possible relations between students).

4. Results and discussion

Correlation coefficients given by Eq. (1) have been calculated for each pair of courses and are presented as a heat map in Fig. 1a. The courses' grades are quite well correlated, the mean correlation (excluding diagonal elements) and its standard error is 0.326 ± 0.006 . There is a triple of courses FP1, FP2 and AG that are strongly correlated one to another, i.e. $C_{\text{FP1,FP2}} = 0.73$, $C_{\text{FP1,AG}} = 0.7$ and $C_{\text{FP2,AG}} = 0.62$. Furthermore, courses from this group are the most correlated with other courses (sum of correlation coefficients $C_X = \sum_{i \neq X} C_{X,i}$): $C_{\text{FP1}} = 12.29$, $C_{\text{FP2}} = 11.52$ and $C_{\text{AG}} = 10.93$. On the other hand, courses FO and EG are least correlated with others; $C_{\text{FO}} = 5.01$ and $C_{\text{EG}} = 5.97$ (for all C_X — see Table I). FO is the only course that tends to negatively correlate with some other courses: PL (-0.12), VDD(-0.10) and SSP (-0.06), (for all C_X — see Table I).

In Fig. 1b and c correlation matrices are shown for shuffled data. The effect of C-shuffling is a complete destruction of correlation matrix. New correlations values are close to zero and the mean correlation is 0.0001 ± 0.007 .

The S-shuffling diminishes the original correlations by 20%, and their mean value is equal to 0.258 ± 0.005 . It can suggest that a given student gets similar grades for several courses and interchanging his own grades does not change dramatically correlations between grades of different courses. The S-shuffling destroys, however, high positive correlations of the triple FP1, FP2 and AG as well negative correlations for the course FO.

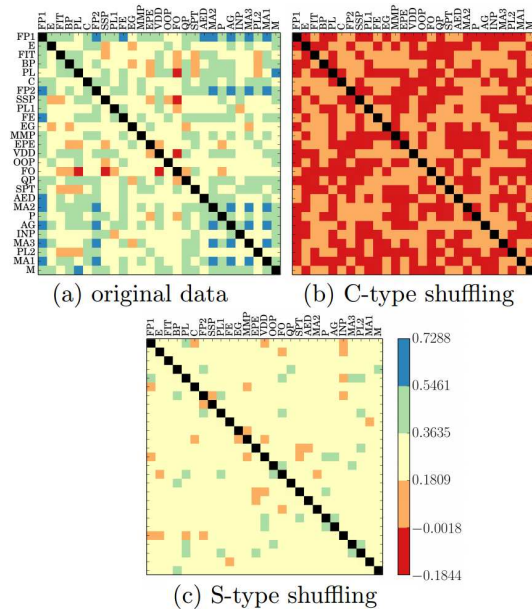


Fig. 1. The course-course Pearson correlation coefficients in a form of heatmap. In the C-type shuffling all grades for a given course are randomly exchanged between students and in the S-type shuffling all grades of a given student are randomly exchanged between courses.

In order to further investigate the correlations we have constructed a complete graph of courses where weights of links between nodes are defined by Eq. (1). Corresponding maximal spanning trees (MST) for this graph are presented in Fig. 2 (merged data from classes J, K, L) and Fig. 3 (only students of L-class).

For better readability labels above links are $C_{X,Y}$ and thickness of a given link is proportional to that value. One can clearly see that the course FP1 is a central hub of the network. Its degree in Fig. 2 is $k_{JKL}^{MST} = 9$ and in Fig. 3 $k_L^{MST} = 6$. One can see also that courses FP2 and AG from the highly correlated triple FP1, FP2, AG are in the centres of MST in Figs. 2 and 3. At the second figure these two courses are local hubs of degrees 4 and 5, respectively. We need to stress that observed MSTs of different classes are different and certain courses can appear in different neighbourhood depending on what class we consider. In all cases, however, the course FP1 remains a central node of MST and the courses FP2 and AG are close to it. It suggests that this triple of courses lays a foundation of knowledge for the students and perhaps in some way it verifies students' knowledge acquired in secondary schools since it is on the first two semesters. We shall call these three courses *Central (C)*.

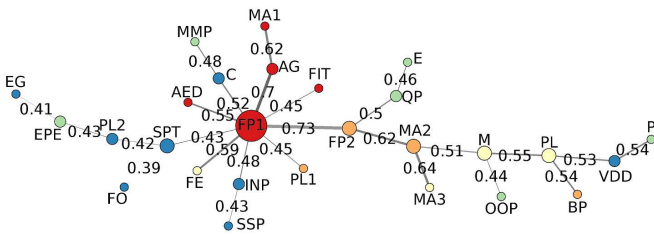


Fig. 2. Maximal spanning tree for the network of courses where links are defined via correlations between the nodes. Symbols correspond to different courses (see Table I), numbers indicate correlations coefficients between courses and colours are indicators of a corresponding semester for that course (i.e. red - semester 1, blue - semester 5 etc., see Table I). The network has been constructed for all three classes (J,K,L) combined. Size of a node is proportional to its degree, i.e. to a number of its nearest neighbours in the MST.

Up to now we have focused on correlations between scores of various courses. Important information is contained also in scores' variances and in spectral properties of entire covariance matrix. Variances of each course have been collected in Table I and principal component analysis (PCA) has been applied to get insight about a possible dimensionality reduction, i.e. what variables are needed to describe student's scores accurately enough. We have taken grades of students as observations and courses as features (observables). A comparison of variances explained by variables (sorted by its variance descending) for the original data, PCA-transformed original and shuffled data is shown in Fig. 4. Matrix in Fig. 5 presents contributions of each course to principal components (PC). The first PC explains almost 40% of total

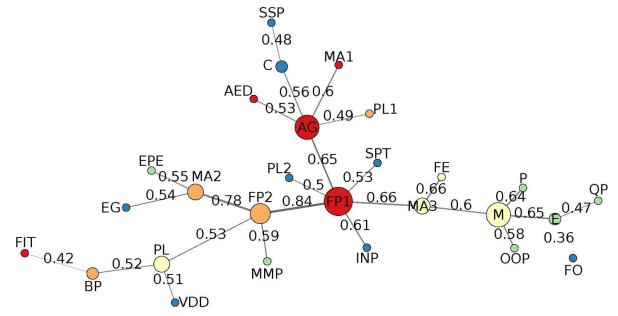


Fig. 3. Maximal spanning tree for one class (L) only. Meaning of symbols, labels and colours is the same as for Fig. 2.

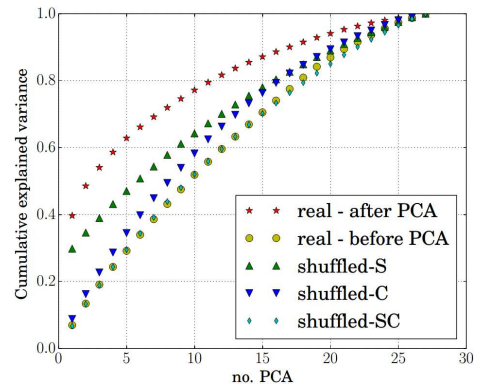


Fig. 4. Plot of explained cumulative variance. We accumulate variances of each course (labelled *real-before* PCA) in descending order, normalized in such a way that the total sum is 1.0. After PCA decomposition we do similarly with principal components (points labelled *real-after* PCA). Points labelled *shuffled-S/C/SC* are results of PCA decomposition after a particular shuffling. S is for S-type shuffling, Central for C-type shuffling and SC for both shuffles together.

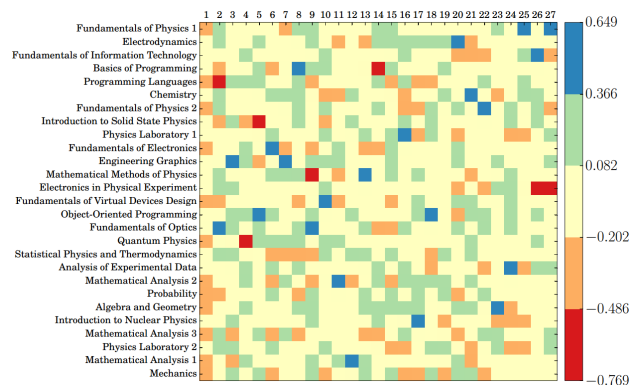


Fig. 5. Transformation matrix from original variables to principal components. It shows what contribution each original variable (a course) has in a particular principal component.

variance (Fig. 4) and all courses contributions to it have the same sign (Fig. 5). It suggests that PC is connected to a student's mean grade and can be used as a simple discrimination between students. Further principle components explain slightly higher amount of variance than most variate courses (namely QP — 7%). The second principal component has unlike contributions of PL and FO which was visible in the correlations analysis as well. The highest contribution to the third PC has been made by EG, to the fourth — QP and so on. The S-shuffling of students grades diminishes the role of the first PC that explains 30% of the total variance (Fig. 4) and the first PC after C-shuffling explains only 10% of the total variance. These facts are in agreement with observations of shuffling influence on Pearson's correlation coefficients (Fig. 1a-c) where the S-shuffling was less destructive than the C-shuffling. Let us remind also that after combining S- and C-shuffling the plot of cumulative variance is indistinguishable from the corresponding plot using variances of original study courses (Fig. 4).

The aim of Fig. 6 is to present relations between various measures describing considered courses scores.

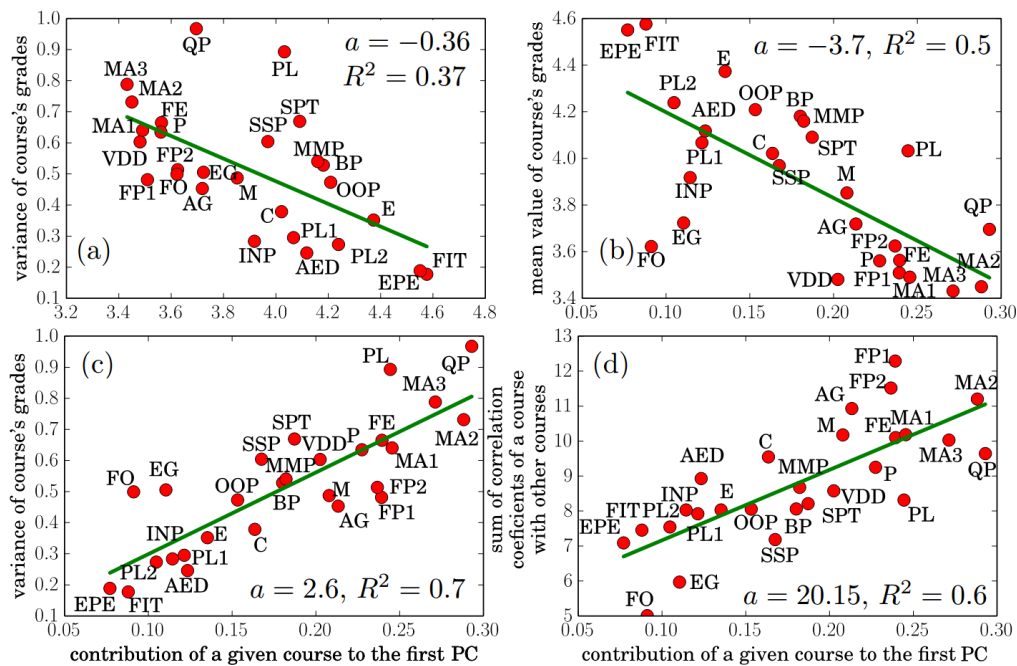


Fig. 6. (a) variances and means of courses' grades, (b): courses' mean grades and a courses' contribution to the first PC, (c): courses' variances and a courses' contribution to first PC, (d): sum of Pearson's r correlations and courses' contribution to the first PC. Green lines are linear fits ($y = ax + b$), R^2 — coefficient of determination.

5. Concluding remarks

Statistical analysis combined with data mining methods applied for students' grades make possible to disclose several dependences between different courses forming academic curriculum at Faculty of Physics, Warsaw University of Technology. A slight negative correlation exists between mean and variance of course grades (Fig. 6a) and between course's contribution to 1st PC and mean scores

(Fig. 6b). Stronger positive correlations are found between the course's contribution to 1st PC and variance as well as sum of correlation coefficients (Fig. 6c,d). In Fig. 6c the Central group of courses lies below and the Diff group lies mostly above the trend line. The opposite situation is in Fig. 6d. Courses from Diff group possess the largest contributions to 1st PC and contributions of Central courses are higher than average.

A slight negative linear correlation ($R^2 = 0.37$) has been found between mean and variance of course grades (Fig. 6a), i.e. courses with higher mean scores tend to display lower scores variances. Three courses from the Central group lie below the trend line at this plot. On the other hand, there is group of four courses: QP, MA2, MA3 and PL that possess highest variances and are far above the trend line. We shall call this group *differentiating* (Diff) since the high variance of scores differences students group.

A negative linear correlation ($R^2 = 0.5$) exists also between the course's contribution to 1st PC and the course mean (Fig. 6a). The largest outlier from the trend is the course FO that was negatively correlated to other courses in Fig. 1. Stronger positive correlations are between contributions to 1st PC and variance and sum of correlation coefficients (Fig. 6c,d). In Fig. 6c the Central group of courses lies below and the Diff group lies mostly above the trend line. The opposite situation is in Fig. 6d. Courses from Diff group possess the largest contributions to 1st PC and contributions of Central courses are higher than average.

Stronger positive correlations are found between the course's contribution to 1st PC and variance as well as sum of correlation coefficients (Fig. 6c,d).

The above correlations together with observations of maximal spanning trees (Figs. 2, 3) suggest the existence of at least two specific groups of courses in the considered curriculum.

(Fig. 6b). Stronger positive correlations are found between the course's contribution to 1st PC and variance as well as sum of correlation coefficients (Fig. 6c,d).

The above correlations together with observations of maximal spanning trees (Figs. 2, 3) suggest the existence of at least two specific groups of courses in the considered curriculum.

The first group can be called Central and it consists of two courses *Fundamental of Physics I* (FP1), *Fundamental of Physics II* (FP2) from semesters 1 and 2, and the course *Algebra with Geometry* (AG) from the semester 1. The Central group is represented by nodes in centres of maximal spanning trees (Figs. 2 and 3) since these three courses possess large correlations coefficients with other ones (Table I and Fig. 1a). The Central group possesses higher than average contributions to 1st PC (Fig. 6) and in Fig. 6d it is far above the trend line. It means sum of correlations coefficients for these courses is much larger than it can be expected from contributions to 1st PC. However the Central group lies below the trend line in Fig. 6c,d, i.e. the courses possess smaller variances of students' scores than it can be expected from their mean grades or their contributions to 1st PC. In other words the courses from the Central group reduce differences in initial students' knowledge. One can understand this result as a proof of the proper organization of the considered academic curriculum since lecture topics presented at above courses should make possible a smooth launch of studies.

The second group that can be called Diff consists of courses *Quantum Physics* (QP; semester 5), *Mathematical Analysis II* (MA2; semester 2), *Mathematical Analysis III* (MA3; semester 3) and *Programming Languages* (PL; semester 3). This group contributes significantly to students' score variance and to the first principal component (Table I and Fig. 6). It follows the above courses highly differentiate the teaching process and their scores separate less skilled students from better ones. This interpretation is confirmed by the fact that above courses possess rather low mean scores (Table I, Fig. 6a,b) i.e. corresponding exams are difficult to pass for a large part of students. The Diff group lies above the trend line in Fig. 6a and below the trend line in Fig. 6d. The last results mean that this group is less correlated to other courses as one could expect from its contribution to 1st PC.

Correlations of the first principal component to other courses' parameters seen in Fig. 6b–d suggest that the contribution to the first principal component measures a course significance and can be used for assigning ECTS points to a given course. Let us stress that our analysis has been based entirely on student grades and is independent of declared curricula of considered courses.

In conclusion, we have shown that data mining methods when applied for students scores can be useful tools for uncovering key courses in academic curricula. The framework consists of comparing results from the following methods: (a) observations of positions of courses in the maximal spanning tree corresponding to a correlations matrix, (b) calculations of courses' mean scores and corresponding variances, (c) calculations of courses' contributions to the 1st principal component. We suggest that when outcomes from above methods are similar then such a combined framework can be used at Universities and at other schools as an additional tool to optimize

teaching strategies since it does not need a priori knowledge of assumed curriculum aims.

Acknowledgments

This research has received funding as *RENOIR* Project from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 691152. J.A.H. has been also partially supported by Russian Scientific Foundation, proposal #14-21-0013.

References

- [1] J. Mulhern, *A History of Education: A Social Interpretation*, 2nd ed., Ronald Press Co., New York 1959.
- [2] C. Marquez-Vera, A. Cano, C. Romero, A.Y.M. Noaman, H. Mousa Fardoun, S. Ventura, *Expert Syst.* **33**, 107 (2016).
- [3] C. Romero, S. Ventura, *IEEE Trans. Syst. Man Cybern. Central Appl. Rev.* **40**, 601 (2010).
- [4] J. Kay, P. Reimann, E. Diebold, B. Kummerfeld, *IEEE Intell. Syst.* **28**, 70 (2013).
- [5] *University Study-Oriented System USOS* (on-line; accessed 2016.03.07).
- [6] G. Wang, C. Xie, S. Chen, J. Yang, M. Yang, *Phys. A* **392**, 3715 (2013).
- [7] J. Speth, S. Drożdż, F. Grümmer, *Nucl. Phys. A* **844**, 30c (2010).
- [8] T.K. Dal'Maso Peron, F.A. Rodrigues, *Euro Phys. Lett.* **96**, 48004 (2011).
- [9] A.Z. Górski, S. Drożdż, J. Kwapien, *Euro Phys. Lett. B* **66**, 91 (2008).
- [10] M. Tumminello, T. Aste, T. Di Matteo, R.N. Mantegna, *Proc. Natl. Acad. Sci.* **102**, 10421 (2005).
- [11] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, *Euro Phys. Lett. B* **30**, 285 (2002).
- [12] A. Giuliani, J.P. Zbilut, F. Conti, C. Manetti, A. Micheli, *Physica A* **337**, 157 (2004).
- [13] M.J. Disney, J.D. Romano, D.A. Garcia-Appadoo, A.A. West, J.J. Dalcanton, L. Cortese, *Nature* **455**, 1082 (2008).
- [14] B.A. Suslick, L. Feng, K.S. Suslick, *Anal. Chem.* **82**, 2067 (2010).
- [15] J.D. Power, D.A. Fair, B.L. Schlaggar, S.E. Petersen, *Neuron* **67**, 735 (2010).
- [16] W.P. Goh, D. Kwek, D. Hogan, S.A. Cheong, *EPJ Data Sci.* **3**, 36 (2014).
- [17] R.V. Hogg, E.A. Tanis, D.L. Zimmerman, *Probability and Statistical Inference*, 9th ed., Pearson Education Ltd., Upper Saddle River 2015.
- [18] J.B. Kruskal, *Proc. Am. Math. Soc.* **7**, 48 (1956).
- [19] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York 2013.
- [20] I.H. Witten, E. Frank, M.A. Hall, *Data Mining. Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publ., Burlington 2011.