sciendo

# Are Teaching and Research activities mutually exclusive? A Data Mining study on European Universities

**Simona GRĂDINARU**
*The Bucharest University of Economic Studies, Bucharest, Romania*
simonaegradinaru@gmail.com

**Anamaria ALDEA**
*The Bucharest University of Economic Studies, Bucharest, Romania*
anamaria.aldea@csie.ase.ro

**Levida BEŞIR**
*The Bucharest University of Economic Studies, Bucharest, Romania*
levida.besir@gmail.com

**Crişan ALBU**
*The Bucharest University of Economic Studies, Bucharest, Romania*
crisan.albu@csie.ase.ro

**Abstract.** *Universities all around the world operate by following several institutional missions, with a central purpose on teaching and research activities. The importance of each aspect alongside the connection between them provide a disputed topic in the literature, many authors confirming or rejecting the intuitive inverse relationship by using various means, more or less quantitative. This paper aims to examine the teaching and research dimensions of the research-active European universities from a data mining perspective. For each dimension previously considered we employ the K-means Clustering in order to identify the groups of similar higher education institutions and we analyze the insights produced by the results. In addition, we build some target variables considering the teaching and research outputs and we investigate their drivers by employing the Logistic Regression. Furthermore, we explore the controverted relationship between the two institutional missions we considered through the use of Correspondence Analysis. Preliminary results illustrate that the dataset contains two types of universities: a category of very large and prestigious institutions and a second group of small and medium sized institutions, quite different from each other. Interest insights are given by the third part of the study, in which the Correspondence Analysis confirms an inverse relationship between teaching and research activities. Unfortunately, this is very likely a consequence of the time constraint – both activities require the same limited resources and therefore increasing the teaching burden for academics may diminish the time and energy dedicated to research.*

**Keywords**: data mining, cluster analysis, logistic regression, correspondence analysis, education, European universities.

## Introduction

Higher Education Institutions (HEIs) operate by ensuring several institutional missions, with a strong focus on teaching and research activities. They also play a significant role in the advancement of economies and societies not only by fostering economic growth and strengthening the technological progress, but also by educating proactive citizens in societies

and contributing to the personal development of individuals. While the first part takes shape of universities research, the second one is more likely related to the teaching activities.

Both teaching and research prowess are unanimously accepted to play an essential part in determining institutional performance. Back in the medieval period when the first universities were born, their purpose might have been dedicated to teaching only. But as time has passed, the expectations for HEIs grew in terms of delivering knowledge, innovation and public goods for society. Governments provide financial support for universities to produce more graduates, whereas the private sector is funding the research and innovation that provide advancements to various industries or looking for a qualitative workforce into specific fields. Very often the teaching funds overcome the amounts for conducting research, particularly due to tuition fees. Yet academics are promoted and obtain prestige for their success in research, while the ability to teach undergrads remains scarcely valued, especially at research dedicated universities.

With such an amount of pressure and responsibility, HEIs are struggling with finding a balance between teaching and research. When being performed by the same people, these activities are rather rival, requiring limited time, a concept that many authors have tried to measure it through various means. Teaching activities cover more time than expected, ranging from preparing the courses and keeping them up to date to reviewing projects and guiding theses. On the other hand, research needs large continuous periods of time as new ideas does not appear overnight and nor does finding evidence in this direction so it requires continuous time spent on this topic, without multitasking with other duties. There may be some cases in which research and teaching improve each other, but those ones are very particular circumstances and it is very unlikely to be a result of being conducted in the same department or by the same academics.

This paper aims to examine the teaching and research dimensions of the research-active European universities from a data mining perspective. We employ the K-means Partitioning Clustering in order to identify the groups of similar HEIs and we investigate teaching and research drivers by employing the Logistic Regression. Furthermore, we aim to explore the controverted relationship between the two institutional missions through the use of Correspondence Analysis. Interest insights confirm an inverse relationship between teaching and research activities.

## Literature review

HEIs have been the object of study for many authors using a variety of approaches, more or less quantitative. Some authors applied robust nonparametric techniques on European universities in exploring the trade-off effect between teaching and research and found that a proper educational efficiency does not weaken research efficiency and beyond a specific threshold increasing the quality of publication also improves the educational efficiency (Bonaccorsi et. al, 2006). Another study identified within European universities clear direction for improvement between education and amount or quality of research, but it left unanswered the relationship between teaching and research, since it was providing only unidimensional rankings (Daraio et. al, 2015). It has also been confirmed a trade-off between teaching and research for Romanian universities with increased efforts and investing in research with the purpose of improving the university ranking (Stoica and Aldea, 2016). Taking a step further and examining the rivalry between industry and academic research, Calderini and Franzoni (2004) discovered that scientific performances are likely to amplify

before or after a patent event, confirming that the development of industrial application also have an impact on the scientific community.

As a proof of the continuous struggle of HEIs, a survey on US public and private research universities confirmed that academic administrators endeavor for balancing undergraduate teaching and research (Gray et. al, 1992). With respect to the same presumable relationship, the work of Hattie and Marsh (1996) indicated that increased research may drive some improvements to teaching, but never the opposite. With this mindset, their study used an analogy referring the relationship between these dimensions to a marriage: if the universities manage to marry the two and consume the marriage, the connection between the attributes is likely to increase. Their recommendation for the HEIs is to start rewarding commitment, creativity and critical analysis and to value more these attributes especially when they occur in both directions.

Logistic regression is a popular Machine Learning classification algorithm especially in medical fields, social sciences, marketing research or even banking. A widespread example is given by a credit institution assessing whether a company will become bankrupt or not, in order to understand if it is profitable or not to grand a loan to that company (Westgaard and Wijst, 2001). A recent study has examined the impact of participating in dance or music lessons in the educational aims of high school students through the use of logistic regression (Cabrera et. al, 2019).

Correspondence analysis has become well-known in ecology, medical work, marketing research and social sciences, gaining recent popularity in psychology too. One of the first studies in this direction, provided by Hoffman and Franke (1986), in which the authors explored the beverage purchase and consumption of 34 respondents and they discovered some specific segments among them. Doey and Korta (2011) proved the utility of correspondence analysis in psychological research by confirming some associations between types of risk (substance abuse, drop out, violence or mental health) and age categories.

## Methodology

*Cluster analysis*, a form of unsupervised learning in machine learning, refers to a broad set of tools for building groups (also referred as *clusters*) in a data set (James, 2013). The aim is to find distinct groups with observations as homogenous as possible within each group, while the groups should be as heterogeneous as possible from each other. Clustering is a popular technique in many fields and various clustering methods exist in the literature. We employed the *K-means* clustering algorithm, a wide used technique in unsupervised learning, which partitions a data set with $n$ observations and $p$ features into $k$ clusters, previously specified.

*Logistic regression*, often used as a classifier (Hastie et. al, 2017), applies a logistic function to estimate a binary dependent variable based on a set of exploratory variables. The response variable has two possible values, typically in the form of win/lose, success/failure or below average/above average, one of them being the target event, whereas the independent variables could be either binary or continuous variables. The provided outcome of a logistic regression is the likelihood for the target event to occur, a value between 0 and 1, for each instance of the data set. By setting a threshold to these probabilities, class membership can be predicted for each observation. Logistic regressions can be binary (categorical response with only two possible outcomes), multinomial (more than two categories without ordering) or ordinal (more than two categories ordered). Equation (1) explains the binary logistic regression we employed in this study.

$$log\left(\frac{Pr(G=1 \mid X=x)}{Pr(G=K \mid X=x)}\right) = \beta_{10} + {\beta_1}^T x \quad (1)$$

*Correspondence analysis* is a tool for analyzing the possible associations between rows and columns of contingency tables (Matei Maer, 2018), which contain the joint frequencies of two categorical variables. It is also related to dimension reduction, in the same manner as principal components analysis, but using qualitative variables. The main idea of this tool is to build simple indices or new dimensions that will illustrate properly the relations between the categories represented on rows and columns. Extracting them in decreasing order of importance is essential for summarizing the information in smaller dimension spaces. These indices provide the coordinates of each row and column within the table, which are simultaneously displayed in the same graph. When applying it for a contingency table with $n$ rows and $p$ columns or vice versa, correspondence analysis selects a number of $min(n-1, p-1)$ new dimensions. Although it has been discussed that this method applies only on qualitative variables, continuous variables could also be treated by defining some categories regarding some intervals or classes.

## Data description

We explored the data set provided by RISIS – ETER facility, a database with various indicators on Higher Education Institutions, including students and graduates, personnel, finances and research activities. The information on 2 764 universities was initially collected, covering 12 European countries, taking into account only 2014 data. Constraints on the missing data reduced the data set to less than 300 institutions, for which we applied a filter on the institution type and we kept into analysis only 264 universities. Table 1 displays a preview of the variables we selected to describe the universities in terms of teaching and research.

**Table 1. Variables selected to describe universities**

| Category | Variable | Description |
|---|---|---|
| Teaching | ACSTAF | Total number of academic staff (headcount) |
| | GOVAL | Basic government allocation (million euro) |
| | GRAD | Total number of graduates at ISCED 5-7 |
| | TEACHLOAD | Students enrolled at ISCED 5-7 divided by total number of academic staff |
| Research | PROF | Number of full professors (headcount) |
| | PHDSTUD | Total students enrolled at ISCED 8 |
| | CIT | Average number of citations of publications |
| | EUFP | Number of participations to European Framework Programs (EU-FP) from the reference year |
| | PUB | Count of publications (article and review) from the Thomson Reuters' Web of Science database, with at least one author affiliated to the institution |

Source: RISIS-ETER facility and authors' own research.

Table 2 provides the summary statistics for the entire data set, according to the dimensions selected above. Interested insights can be easily drawn by detecting the universities with extreme values. The Katholieke Universiteit Leuven (KU Leuven) reaches the maximum academic workforce, an institution that boasts with a long tradition of high-quality education and pioneering research. Federal Institute of Technology Zürich (ETHZ)

owns the largest government allocation, which is not a surprise since this institution was founded and continues to receive a lot of support from the Swiss Government.

**Table 2. Summary statistics**

| Category | Variable | Min | Mean | St. dev. | Max |
|---|---|---|---|---|---|
| Teaching | ACSTAF | 76 | 1 551 | 1 438.208 | 9 069 |
| | GOVAL | 0.140 | 102.575 | 130.855 | 1 010.720 |
| | GRAD | 117 | 4 633 | 3 341.207 | 21 973 |
| | TEACHLOAD | 0.942 | 14.482 | 7.277 | 36.554 |
| Research | PROF | 10 | 132.500 | 182.820 | 1 160 |
| | PHDSTUD | 10 | 509.500 | 1 010.480 | 5 870 |
| | CIT | 0.622 | 4.781 | 1.833 | 13.835 |
| | EUFP | 1 | 43.070 | 65.270 | 389 |
| | PUB | 2 | 477.703 | 615.626 | 3 380 |

Source: RISIS-ETER facility data and authors' own research.

Regarding the ISCED 5-7 graduates, Sapienza University of Rome registered the highest value, a well-known institution with a long history and strong background. Surprisingly or not, the largest number of full professors is offered by University of Oxford, together with the highest number of participations to Framework Programs, which supports various research and development activities with coverage across almost all scientific disciplines. University College London gathers the highest number of PhD students from our data set, which again it is not a coincidence since this institution is committed to research as a fundamental part of its mission and therefore attracting students passionate about research and willing to build a career in this direction. Federal Institute of Technology Lausanne holds the highest citation score, mainly due to their focus in research and innovation, proved by numerous patents.

At a glance, the summary statistics prove some heterogeneity, confirmed by the large data ranges. This remark led us to conduct further a cluster analysis with the aim of identifying groups of similar universities and discover the patterns among them.

## Results and discussions

First part of the analysis consists in identifying the groups of similar universities, by each dimension. Let's study into more depth the teaching aspect, for which we chose as describers the following variables: ACSTAF, GOVAL and GRAD. Our judgement was that Higher Education Institutions need to produce graduates as an outcome, with a focus on the undergraduate level (ISCED 5-7), considering their disposable inputs: academic staff and the government allocation funds. Of course, each institution has a different manner of producing the output using the inputs (also named as *production function*), but we will explore that into a following study of nonparametric efficiency analysis.

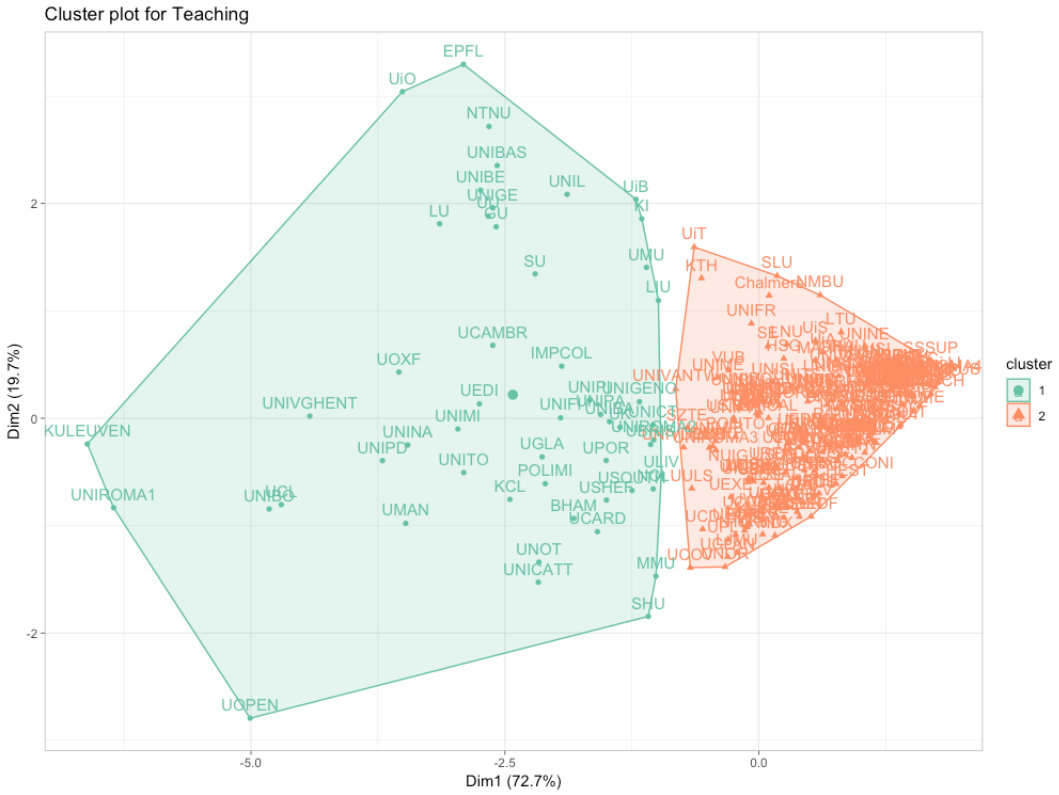


**Figure 1. Graphical representation of the two-cluster solution obtained for Teaching**
Source: Author's research and representation in RStudio 1.1.463 with *factoextra* package.

Before running the cluster analysis, we need to understand the data and identify the outliers, if any. We employed an initial principal component analysis in RStudio 1.1.463 and discovered that almost 94% of the original variance can be represented in only 2 dimensions. When representing the universities into the new subspace, two outliers were spotted: University of Zürich and Federal Institute of Technology Zürich. We decided to exclude them, based on the fact that they are too right biased in terms of inputs.

The remaining of 262 universities were introduced into the cluster analysis. A preliminary dendrogram suggested the two clusters and we used this number into the k-means algorithm. After varying this number and studying the quality of the solutions produced, it turned out that the dendrogram suggestion was actually the best solution, obtaining two groups with 53 and 209 universities. The larger cluster with 209 institutions contains small and medium-sized universities whereas the first one, which is more dispersed as represented below in Figure 1, is composed of large and prestigious universities, including University of Oxford, University of Cambridge, together with the highest values previously identified. First cluster has the average for the government allocation of 146.5 million euros, while the average of the small to medium sized universities allocation reaches only 91.4 million euros. The difference is also perceptible in terms of academic workforce, with an average of 2 104 for the large universities and 1 410 for the medium institutions. A similar distinction appears for ISCED 5-7 graduates, the large universities cluster having an average of 5 328 academics, while the second cluster including only 4 432 academics.

With respect to research dimension, the describers we decided to use were: ACSTAF, PROF, PHDSTUD, EUFP and CIT. Unfortunately, PUB could not be used in its raw shape due

to extremely high correlation with the other research variables, but we will be using it later with a different purpose. Both academic workforce and professors need to continue their research in order to get promoted, while PhD students learn the research process during their study time and are required to publish papers in order to graduate. On the other hand, EU-FP participations provide the proof of contributing to the industry research and also plays a role in increasing the institutional citation score. In this manner, we thought of the last variable as a target outcome, whereas the other variables could be perceived as inputs.

Almost 90% of the original information can be represented in 2 dimensions obtained by principal components analysis. The outlier detection ran on these dimensions identified two universities as outliers: Federal Institute of Technology Lausanne and University of St. Cyril and Methodius in Trnava, first one as being extremely biased on all aspects and the second one due to a very large citation score in comparison with the according inputs.

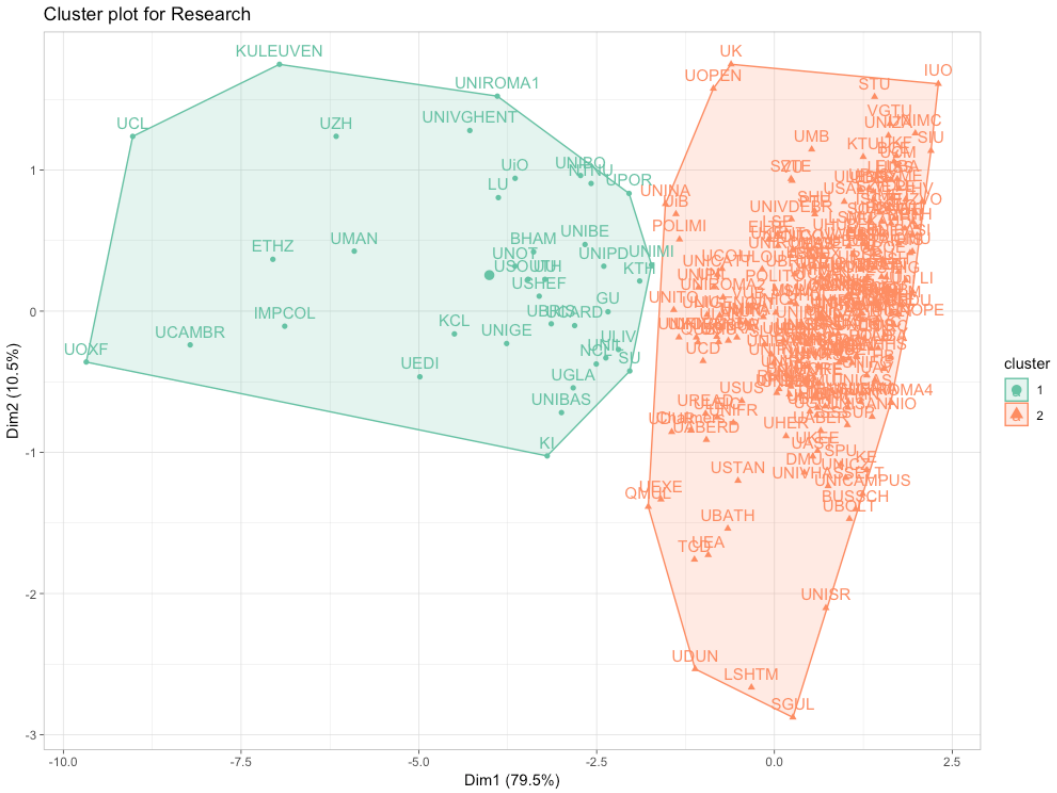


**Figure 2. Graphical representation of the two-cluster solution obtained for Research**
Source: Author's research and representation in RStudio 1.1.463 with *factoextra* package.

The remaining 262 universities were grouped by k-means into two groups, very similar to the previous analysis: a cluster with 225 institutions gathering small and medium sized European universities and a second group with 37 large and well-known universities, typically with variables double than in the previous category. Any solutions except to the presented one produce weaker results, both in terms of cluster allocation and interpretability. Looking at the indicators we used for building the clusters, the large universities have an average of 2 664 academic staff, whereas the small and medium institutions only reach and average of 1 367 academics, almost half of the previous value. The distinction is maintained for the number of professors, first cluster having an average of 287

full professors, compared to 185 for the second group. The large universities produce an average of 1 572 graduates, compared to only 774 degrees offered from the medium institutions. The research indicators lead to the same distinction: the prestigious universities gather an average of EU-FP participations around 77, more than double than the medium sized universities, which only reach a mean of 37. Not surprisingly, the large and prestigious institutions also reach a higher citation score, averaged for this cluster at around 5.57, in comparison with a mean value of 4.64 for the second cluster.

Subsequently, we have decided to build two target variables, one for each aspect previously studied. Teaching is highly influenced by the workload of the academics, hereinafter referred as teaching burden or teaching load, whilst research is mainly reflected in the number of publications. Our interest was to explore which variables have an impact on the previously defined response variables, for each dimension. Therefore, we transformed teaching load and publications into binomial variables and used further the logistic regression for calculating odds ratio for each observation to belong to a class defined by the target event.

**Table 3. Logistic regression output for Teaching**

| Terms | Estimate | Std. Error. | z value | Pr( > \|z\| ) | Significance | Log odds |
|---|---|---|---|---|---|---|
| (Intercept) | 0.1411 | 0.3242 | 0.4351 | 0.6635 | | 1.1515 |
| ACSTAF | -0.0074 | 0.0010 | -7.2203 | 0.000 | *** | 0.9926 |
| GOVAL | 0.0290 | 0.0048 | 6.0427 | 0.000 | *** | 1.0294 |
| GRAD | 0.0015 | 0.0002 | 6.9988 | 0.000 | *** | 1.0015 |

Note on Significance codes: 0'***' ; 0.001'**' ; 0.01'*' ; 0.05'.' ; 0.1' '.
Source: Author's research results obtained in RStudio 1.1.463.

Table 3 gives the output of the logistic regression we built for Teaching, illustrating that all coefficients are statistically significant at the level of 0.05 (5%), except for intercept. Last column in Table 3 illustrates the log odds for the target event, which were obtained from the estimated coefficients. In order to facilitate the interpretation, the change in log odds could be easily translated into change of odds. Consequently, increasing the academic workforce with a single employee would reduce the teaching load with 0.74% chances. Attracting another million euros of government allocation drives the teaching load to increase with an odd of 2.94%, whereas increasing the number of ISCED 5-7 graduates also drives a growth of the teaching burden 0.15% odds. This actually proves the pressure and responsibility the universities have to face against increasing budget funds.

The graph in Figure 3 illustrates the curve of the Receiver Operating Characteristics (ROC), the area under this curve (AUC) and the classifier performance for each probability interval painted by the right axis. An ideal classifier needs to be as close as possible to the upper left part of the graph, with AUC near 1, meaning a perfect discrimination between classes. The current AUC for the Teaching Logistic Regression has a value of 0.855, considered adequate in comparison to the value of 1 for a perfect classifier.

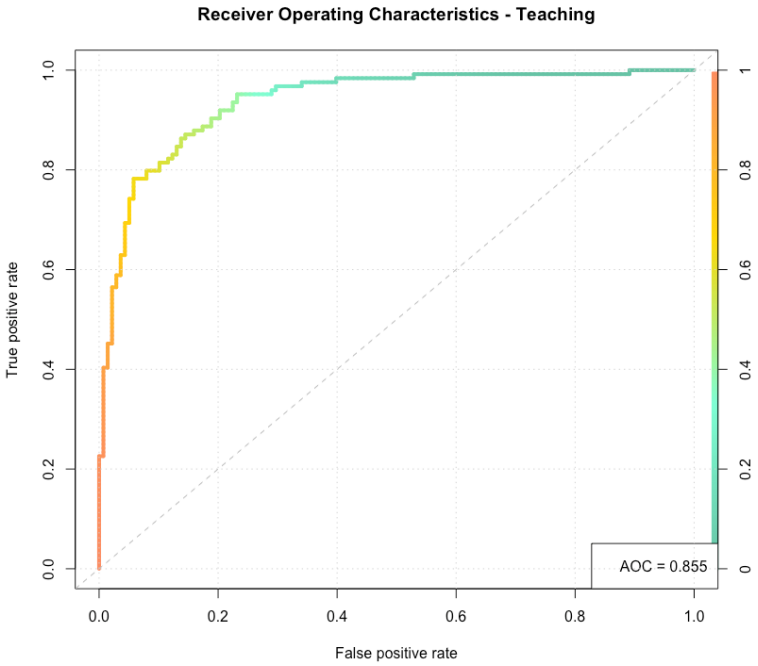


**Figure 3. Receiver Operating Characteristics for Logistic Regression – Teaching**
Source: Author's research and representation in RStudio 1.1.463.

With respect to the research dimension of universities, Table 4 shows the output of its according logistic regression. We used academic staff, professors, PhD students, EU-FP participations and mean citation score in modeling the number of publications above or below the mean as a binomial variable. This time, only three explanatory variables are statistically significant at the significance threshold of 0.05.

**Table 4. Logistic regression output for Research**

| Terms | Estimate | Std. Error. | z value | Pr( > \|z\| ) | Significance | Log odds |
|---|---|---|---|---|---|---|
| (Intercept) | -9.3992 | 1.7784 | -5.2853 | 0.0000 | *** | 0.0001 |
| ACSTAF | 0.0002 | 0.0004 | 0.4424 | 0.6582 | | 1.0002 |
| PROF | 0.0179 | 0.0043 | 4.1479 | 0.0000 | *** | 1.0180 |
| PHDSTUD | 0.0015 | 0.0007 | 2.2931 | 0.0218 | * | 1.0015 |
| EUFP | 0.0650 | 0.0196 | 3.3162 | 0.0009 | *** | 1.0672 |
| CIT | 0.2787 | 0.2541 | 1.0969 | 0.2727 | | 1.3214 |

Note on Significance codes: 0'***' ; 0.001'**' ; 0.01'*' ; 0.05'.' ; 0.1' '.
Source: Author's research results obtained in RStudio 1.1.463.

An increase of one employee in the academic staff leads to an increase in publications above the mean with an almost neglectable insignificant odd of only 0.02%. If number of professors expand with one more person, the publications are likely to increase with a chance of 1.8%, but if the PhD students increase with one, the chance of expanding the publications is only 0.15%. It is not surprisingly that a professor weights 12 times more than a PhD student in improving the number of publications. Nevertheless, the highest impact is provided by EU-FP participations: accumulating a single more participation may lead to an increase in

publications with an odd of 6.72%, almost four times larger than the contribution of an additional single professor.

Analogous to the previous analysis, Figure 4 displays the ROC curve and proves an AUC value of 0.943, very close to 1, underlying an excellent classifier for the research dimension.

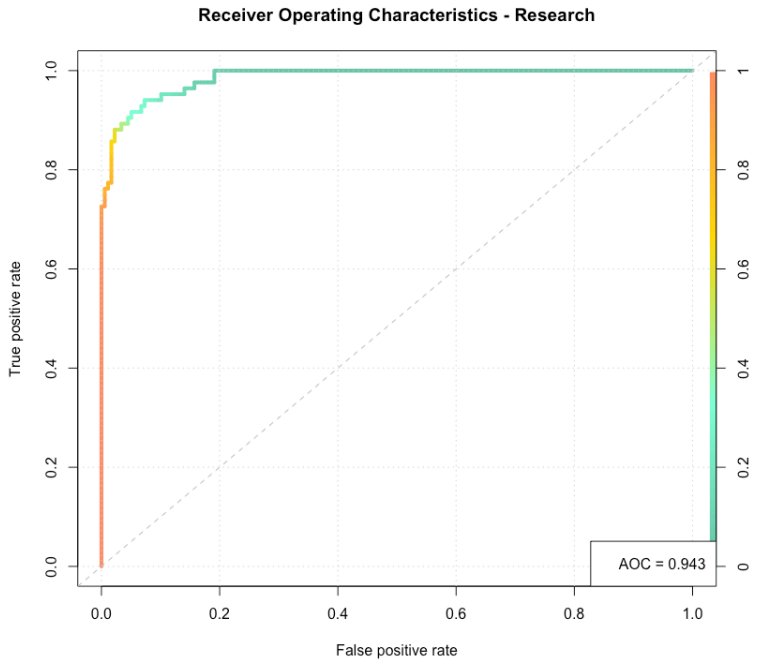


**Figure 4. Receiver Operating Characteristics for Logistic Regression - Research**
Source: Author's research and representation in RStudio 1.1.463.

Last but not the least, the third aim of this paper was to explore any connections between the teaching and research dimensions. Considering that all variables we used were continuous, we needed to treat some of them by defining some categories. After taking a closer look to all variables of interest for our purpose, we defined intervals based on quartiles for teaching load and publications, providing four categories for the rows and columns of the contingency table, as shown in Table 5. A closer look at this table may easily spot a possible correlation between high publications and low teaching load or between low publications and very elevated teaching load.

**Table 5. Contingency table between Teaching Load and Publications**

| | Low publications | Modest publications | Medium publications | High publications | Total |
|---|---|---|---|---|---|
| Low teaching load | 7 | 11 | 6 | 41 | 65 |
| Moderate teaching load | 16 | 14 | 18 | 17 | 65 |
| High teaching load | 19 | 22 | 12 | 12 | 65 |
| Very elevated teaching load | 23 | 18 | 11 | 13 | 65 |
| **Total** | 65 | 65 | 47 | 83 | 260 |

Source: Author's research processing and results.

After applying correspondence analysis on the contingency table previously built, the results only confirm the intuitive thinking from above. Figure 5 proves the existing relationships between teaching load and publication. First two dimensions selected by correspondence analysis gather together 97.8% of the initial variance, a proof of the fact that insights are not missed by representing the categories in the new subspace.

All categories are far away from the origin, indicating how discriminant they are from each other. In the left part of the graph, two categories are very close: low teaching load and high publications, underlying that they are associated with each other. All the other categories are represented in the opposite part of the graph, highlighting that they are most likely negatively correlated with the left-side categories.

Focusing on the upper right part of the chart, it appears this quadrant to be defined by high values for teaching load, while the publications remain low or modest in this condition. The bottom right quadrant presents the associations from the middle values: a moderate teaching load is associated with a medium number of publications. Therefore, Figure 5 is a clear representation of the intuitive inverse relationship between the teaching load and publications, caused by the time constraint of academics.

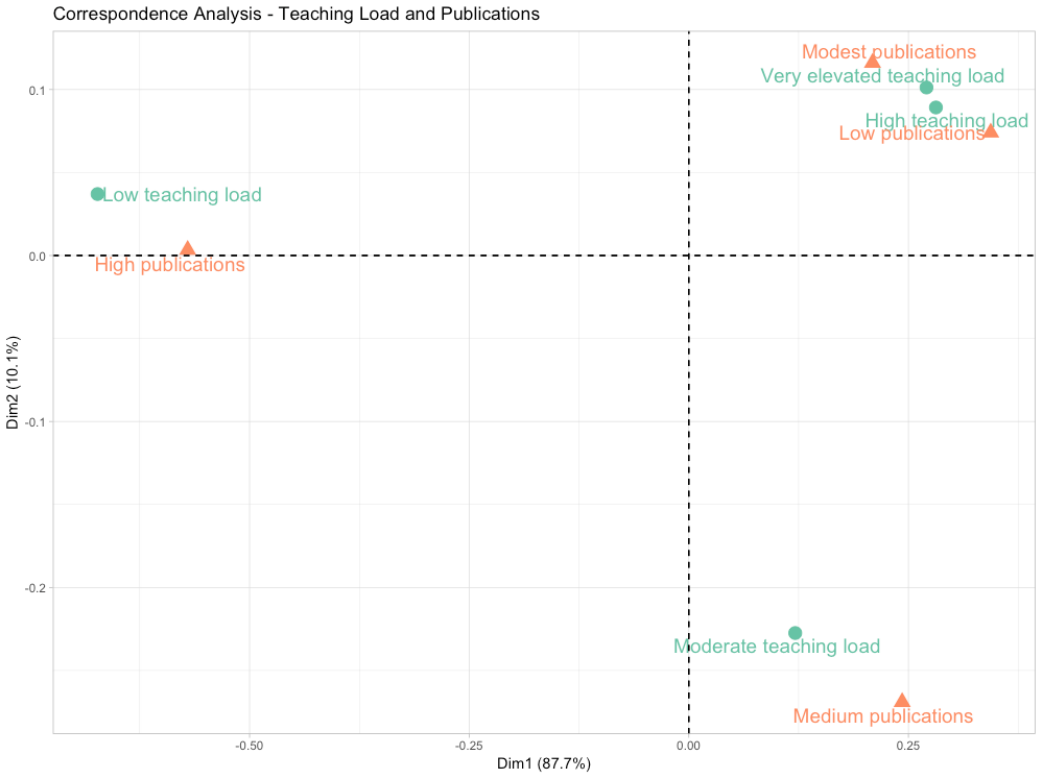


**Figure 5. Correspondence Analysis graphical representation between Teaching Load and Publications**

Source: Author's research and representation in RStudio 1.1.463 with *factoextra* package.

## Conclusion

Universities all around the world have a central purpose on teaching and research, alongside some other institutional missions they need to deliver through their activity. With such an amount of pressure and responsibility, HEIs are struggling with finding a balance between

teaching and research, a popular topic in the literature that was approached in various ways, more or less quantitative.

The main purpose of this study was to examine the drivers of teaching and research activities, together with exploring the controversial relationship between them. Before studying these, the data set went through an outlier detection method (principal components analysis) and was analyzed with k-means partitioning clustering, revealing two groups of universities within the heterogenous initial data: a cluster of small and medium sized HEIs and group of large and prestigious universities. Two logistic regression models were built for each dimension, revealing three major drivers for each aspect considered: academic workforce, government allocation and undergraduates for teaching burden on one teaching side and professors, PhD students and EU-FP participations on publications as a research outcome.

A Correspondence Analysis was applied to the intervals generated for the previously selected target outcomes: teaching load and publications. Results illustrated that low teaching loads and high amounts of publications are negatively correlated all other categories considered. Interest insights were given by categories grouped nearby each other, including here high and very elevated teaching burdens with low and modest amounts of publications, which together with the previously mentioned similar categories proved and intuitive inverse relationship between the teaching load and publications, very likely caused by the time constraint of academics and their lack in simultaneously focusing on both directions.

Future directions of research may involve attaching some descriptors for the teaching quality, although data sources are pretty scarce for these measures. It would be also interesting to explore the production function for each HEI and to assess the performance and efficiency of universities for both dimensions in order to go into more depth and understand if there is indeed a trade-off between teaching and research.

## References

Bonaccorsi, A., Daraio, C., Simar, L. (2006). Advanced indicators of productivity of universities. An application of robust nonparametric methods to Italian data. *Scientometrics*, 66(2), 389-410.

Cabrera, J.C., Karl, S.R., Rodriguez, M.C. (2019). *Predicting College Enrollment for Students Who Partake in Music or Dance Lessons Using Propensity Score Matching and Logistic Regression*. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada.

Calderini, M., Franzoni, C. (2004). Is academic patenting detrimental to high quality research? An empirical analysis of the relationship between scientific careers and patent applications. Paper presented to the 4th workshop on *Economic Transformation in Europe,* Sophia Antipolis, January 29-30, 2004.

Daraio, C., Bonaccorsi, A., Simar, L. (2015). Rankings and university performance: A conditional multidimensional approach. *European Journal of Operational Research*, 244(3), 918-930.

Doey, L., Kurta, J. (2011). Correspondence Analysis applied to psychological research. Tutorials in Quantitative Methods for Psychology, 7(1), 5-14.

Gray, P., Froh, R., Diamond, R. (1992). *A National Study of Research Universities: On the Balance between Research and Undergraduate Teaching*. Center for Instructional Development, Syracuse University.

Hastie, T., Tibshirani, R., Friedman, J. (2017). *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Second edition, Springer.

Hattie, J., Marsh H.W. (1996). The relationship between teaching and research: A meta-

analysis. *Review of Educational Research*, 66(4), 507-542.
Hoffman, D., Franke, G. (1986). *Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research.* Journal of Marketing Research, 23(3), 213-227.
James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. New York, Springer.
Maer Matei, M.M. (2018). *Analiza datelor cu R.* Editura Universitara, Bucuresti.
RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.
Stoica, M., Aldea, A. (2016). Efficiency of teaching and research activities in Romanian universities: An order-alpha partial frontiers approach. *Economic Computation and Economic Cybernetics Studies and Research*, 50(4), 169-186.
Westgaard, S., Wijst, N. (2001). Default probabilities in a corporate bank portfolio: A logistic model approach. *European Journal of Operational Research*, 135(2), 338-349.