Statistical analysis of QS World University Rankings 2021 university rankings using Scopus/SciVal databases

Imre Dobos
Budapest University of Technology and Economics
Hungary
Email: dobos.imre@gtk.bme.hu

Péter Sasvári
(corresponding author)
University of Public Service and University of Miskolc
Hungary
Email: sasvari.peter@uni-nke.hu

The rankings of universities around the world were created with the aim of measuring the performance of higher education institutions as well as the quality of institutions. Such lists provide a basis for better informed decisions by applicants to the higher education market but can also be an important source of information for decision-makers in individual states on how each country's institutions are performing in the field of international scientific competitiveness. There are numerous examples of such rankings. The authors aimed to select a leading one, namely, the QS World University Rankings 2021, and to examine the ranked universities based on statistical variables obtained from the Scopus and SciVal databases created by the Dutch academic publishing company Elsevier. Thus, the aim was not to rank institutions but instead to closely examine the statistical variables and criteria of the universities ranked by QS with the help of multivariate statistics. The results show that Scopus/ SciVal data can be used to examine not only researchers but also universities. The results also show a high degree of similarity.

Keywords:
QS university ranking, international scientific competitiveness, multivariate statistics

Introduction

The issue of international scientific competitiveness is one of the most important science policy topics today. The issue is not only addressed by professionals, but the widely known and increasingly prominent international university rankings now reach a wide range of stakeholders. Science policy itself addresses not only the regulation of higher education institutions but also that of research centers and workshops and scientific academies. With respect to all three dimensions, competitiveness has become the defining concept in recent years; the heads of institutions and

governments now routinely submit all measures to competitiveness assessment. International competitiveness can be well measured by rankings, as they show results that are quickly transparent and easy to interpret, providing empirical support for stakeholders' arguments and a basis for their decision-making. Rankings in turn trigger several international processes, of which only one involves students making better-informed decisions concerning their further studies. It is also the case that top-ranked institutions can engage the best-performing faculty and researchers more effectively, as well as build a brand for the institution both nationally and internationally. Thus, achieving a rise in the rankings for their institutions has been articulated as a policy objective by the governments of many countries. Thus, we can state at the outset that the importance of rankings is undeniable: they play a very significant role.

There are countless university rankings available around the world. Four of these are listed in Table 1. The first three rankings, the Academic Ranking of World Universities (ARWU), the QS World University Rankings (QS) and the Times Higher Education World University Rankings (THE), are called the "big three", while in recent years, US News has started to catch up with the other three. All four rating agencies use a myriad of ranking indicators. The indicators are then weighted. The indicators include both soft and hard indices. Research performance is classified as a hard indicator. Research performance is then determined by each of the four classifiers based on data held in large databases. The QS and THE lists take their performance from the Scopus database, while the ARWU and US News lists use the Web of Science dataset to determine the values of their indicators. Table 1 shows the contribution of the sum of the research indicators to the calculation of the ranking.

Table 1

Weight of research indicators

Research indicators	The weight of research output, %	Database of research output
QS	20.0	Scopus
ARWU	40.0	Web of Science
THE	60.0	Scopus
US News	62.5	Web of Science

Source: Mammadli (2021).

As the table shows, the QS list relies the least heavily on the evaluation of research results. However, the other factors with an indicator weight of 80% are not always made freely available by the QS. The other three university rating agencies give more weight to research but do not make other indicators available.

This gap raises the question of whether other freely available databases can be used to estimate the ranking of universities. One of these freely available databases is the SciVal dataset, which takes its indicators, i.e., statistical variables, from Scopus.

Note that the SciVal database is available to universities on a purchase-by-purchase basis. In this study, the question of how closely the QS list can be approximated by the statistical variables taken from SciVal is an interesting hypothesis because, in principle, only 20% of research performance can be estimated using Scopus-based indicators. However, it also raises the question of whether the other 80%, and hence the ranking, can be attributed to a non-SciVal-derived indicator.

In the present study, we try to model the QS international ranking with statistical tools and predict the expected rankings of institutions using publication data. After the introduction, we provide a theoretical overview of the relationship between international university rankings and competitiveness, scientific productivity, and critique of rankings. This theoretical overview contributes greatly to the foundation of the problem statement and to the analysis as well. Next, we present the compilation of the dataset from the SciVal database, and then we perform statistical analysis on it. The analysis starts with a correlation calculation to determine how strong the linear relationship is between the indicators taken from SciVal. Then, a principal component analysis is performed, which seeks to answer the question of whether the variance can be reduced by returning the variance with latent variables. Multicollinearity is examined using the VIF (variance inflation factor) index. This analysis aims to see if it is possible to reduce the number of variables obtained from SciVal by dropping some of them. Then, collinear variables and QS rankings are estimated using linear regression. If we omit statistical variables, the question arises as to how we can then reproduce the omitted variables using the ones we have left. With our analysis, we can answer this question. Causal relationships are examined by partial correlation calculation, which is used to investigate the cause-and-effect relationship between statistical variables. In this case, we cannot determine the direction of the causal relationship, only the relationship itself. Finally, universities are clustered to determine which groups the universities on the QS list fall into. This analysis resembles earlier analyses performed by Dobos et al. (2021), with the important difference being that it was applied to data obtained from Central and Eastern European economic researchers. Finally, we summarize our results and make suggestions as to how institutions can advance in their rankings.

International university rankings and competitiveness

As stated in the introduction, international university rankings are good tools for measuring scientific competitiveness. Rankings are thus, in this respect, the "gold standard" measure of universities, where all institutions aim to achieve a placing that advertises their own research and educational potential to the international community. Higher education is becoming increasingly prevalent as an element that helps the economy and prosperity, as well as the competitiveness of the nation state (OECD 2015). Increasing competition in today's knowledge-driven economies also

exacerbates the so-called "brain race" phenomenon, entailing a brain drain of the most talented professionals with the greatest knowledge or reputation from less advanced to more advanced countries. This is especially true in the fields of natural sciences, engineering, and mathematics. With the intensification of this competition, international university rankings are also becoming increasingly prevalent tools, as they can position different institutions along predefined indicators in a clear way (Hezelkorn 2015). Universities play a primary role in driving economic growth and are responsible for increasing the innovation potential of states through their research. For them, advancing in international university rankings is also the key to success as well as acquiring resources (Safon 2013). However, the rankings currently in use place heavy emphasis on research. This means that they do not pay as much attention to the educational pillar, so institutions that excel at education start at a relative disadvantage in the rankings (Liu–Cheng 2005).

However, the quality of higher education is a complex concept, and university rankings represent an effective proxy for it only to a limited extent. The quality of higher education – whether we choose to consider individual universities or instead nation states' systems as a whole – is also difficult to determine due to its sheer complexity and the likely limitation of measurement to only a few indicators. The situation is further complicated by the fact that the stakeholders in the higher education sector all act according to their own differing interests and goals; each of them views the quality of each institution differently. Thus, it is worth involving all stakeholders in the development of unified measurement systems and the definition of larger strategies (Bobby 2014).

Green (1994) highlights another problem in defining quality. In her view, quality can only be imagined in a complex and multidimensional system that cannot be compressed into a single concept. In this context, a third problem arises, according to which quality is not a constant but instead a dynamically changing process that is only worth analyzing in a larger social, economic, and political environment (Bobby 2014). Two strategies suggest themselves for defining quality:

- One of the strategies formulates a single unified goal that institutions can interpret as missions in their operations (Bogue 1998).
- The other strategy allows for the development of more specific indicators so that institutions can measure the invested resources and outcomes in different ways (Barker 2002).

From the two strategies, international university rankings, which enjoy great attention and popularity today, use the latter, measuring and ranking institutions according to a complex evaluation system. It is worth emphasizing that there is currently a strong international convergence in the measurement of scientific quality, which has been developed along the indicator system of rankings (Buela-Casal et al. 2007). In addition to striving for a uniform definition of quality, rankings have significant promotional value (Hazelkorn et al. 2014). Higher education institutions around the world are competing with each other based on these rankings and thus

also receive feedback on their own performance. With the appreciation of rankings, the goal of becoming a "world-class" university has emerged (Salmi 2009, Lee et al. 2020). Salmi (2009) makes the critical observation that in their pursuit of improved rankings, institutions tend to focus only on the aspects named by the rankings, adapting their activities to demands of these indicator systems. Aithal et al. (2016) found that rankings also play a significant role in shaping a performance-based culture. The formulation of the concept of a world-class university creates a competitive situation among the flagship institutions of developed economies and the emerging institutions of developing economies. The concept no longer solely encompasses improving the quality of education and increasing research performance but also focuses on sustainability and the continued competitiveness of the institution (Liu et al. 2019). Altbach's (2012) study showed that rankings in a globalizing and competition-oriented market have become permanent and indispensable actors in the higher education and academic spheres. It is worth noting here that both globalization and the internationalization of higher education have helped advance the rankings, as they provide good feedback on the position of individual institutions in the international area. Essentially, they allow the measurement of institutions in two respects, based on their performance and their international reputation. As a result, an increasing number of institutions are making internationalization their main goal (De Wit 2015). In their study, Marginson–van der Wende (2009) concluded that the internationalization of universities no longer solely appears as an objective at the institutional level; instead, it can be found at higher state and policy levels as well, primarily because of its enormous news value. Altbach (2012) highlighted that international university rankings, despite their short history, have now become iconic, representing a kind of status symbol on which politics also relies heavily. In essence, the rankings thus provide governments with feedback worldwide on the performance of their own higher education system (Hazelkorn et al. 2014). Pietrucha (2018) explains in his study that the size and performance of a given state's economy is a key factor in the ranking of each institution because it determines how much grant money a particular university receives to support its research activities. Meanwhile, in their analysis, Benito et al. (2020) did not focus solely on universities as institutions but looked instead at their broader social, economic, and political contexts. Their results showed that higher education institutions are also much more competitive in economically and politically stable states – democratic values and scientific freedom are considered highlighted values. These were confirmed in light of economic indicators by Feranecová–Krigovská (2016) in their study.

In their study, Sheeja et al. (2018) collected what goals international university rankings serve. On this basis, the following objectives have been set. The rankings should:

- measure the efficiency of higher education institutions (Shin et al. 2011),
- assist decision-makers in resource allocation and prioritize key research and education goals among institutions, thereby promoting the focused use of resources (Ioannidis et al. 2007),

– enhance the quality and performance of institutions,
– provide free and highly monitored feedback worldwide, thereby providing publicity to institutions ranked at prominent places (Yerbury 2006),
– help identify and differentiate each type of institution, and discipline through their discipline lists, thereby also taking into account the differences among disciplines in ranking (Harvey 2008), and
– also support the branding activities of the listed institutions (Yeravdekar–Tiwari 2014).

Scientific productivity as one of the pillars of international university rankings

International university rankings show a complex picture in their measurement methodology, and even the best-known and internationally the most recognized rankings differ greatly according to their measured indicators (Abramo–D'Angelo 2014). Examples of such indicators are the number of publications and citations, academic reputation, quality of education, educational environment and student satisfaction, student-lecturer ratio, ratio of foreign students and lecturers, industrial relations, and number of Nobel laureates (Halaweh 2020). Basically, however, they agree that the pillar of scientific productivity appears in all of them. According to Altbach's (2013) results, international university rankings focus primarily on the research pillar because this pillar is the easiest to quantify and measure. In this context, Salmi (2011) demonstrated that due to the dominance of the research pillar, universities with outstanding research potential perform higher in the rankings. Institutions therefore clearly focus on strengthening their research potential in relation to their three main missions – research, education, and industrial knowledge sharing (Laredo 2007).

The measurement of scientific performance is based on the number of publications, the foundations of which Lotka (1926) described in his work. Based on this, the measure of productivity is the number of publications, and the measure of scientific effect is the number of citations to publications. The only problem with the publication-based approach to productivity is that it accounts for all publications with equal value, although this is by far not the case in practice. The tradition of measuring scientific productivity comes from the natural sciences, where this means the number of publications published in scientific journals. Today, there are two generally recognized databases for measurement, which provide a clear, transparent catalog of internationally listed publications. These are the Elsevier-owned Scopus and Web of Science run by Clarivate Analytics. Most bibliometric analyses rely on the latter, as there is a long tradition of Web of Science and impact factor-based measurement, although it is worth emphasizing that Scopus offers a much greater immersion from subscribed journal publications, books, and conference publications, especially in the

field of social sciences and humanities (Halaweh 2020). Regarding the measure of publication-based productivity, Abramo et al. (2008), in agreement with the preliminary measurements, called attention to the fact that as each discipline has different productivity intensities, their representatives can only be compared with other representatives of their own discipline. It is also worth mentioning here that there are also differences in citation habits, which ultimately measure the phenomenon of knowledge dissemination and knowledge spillover (Glänzel 2008).

In their study, Lowry et al. (2007) write that rankings were originally intended to help students make further learning choices, which politics and the media have mistakenly identified as measuring academic excellence. In their view, scientific quality instead constitutes a complex system that cannot be accurately measured by mere selected indicators. The results of Kaba (2020) highlight that citation numbers are also used as measures of scientific impact on productivity defined by rankings. In her publication, Halaweh (2020) attempts to create a unified and comprehensive indicator that can accurately measure scientific quality and productivity at both national and international levels. The author recommends the indicator as a complement to the THE and QS ranking measurement method, fitting into the research pillar. The prepared indicator consists of the ratio of the weighted number of publications to the number of researchers employed in the institution.

Criticism and problem statement of international university rankings

International university rankings, currently considered a measure of scientific excellence, have been subject to several criticisms as well, chiefly because of their measurement methodology. The primary problem in this regard is that rankings are not edited by scientometrics professionals and do not apply a prudent methodology; however, despite its importance, measurement methodology is not considered by either politics or the media when communicating rankings (Loughran 2016). In parallel, King (2009) found that the institutions that rank best in the rankings tend to be named the best institutions by these actors, and thus, the rankings also serve the branding and policy goals of each university. However, it is important to emphasize that the institutions involved in the measurement only stand out along certain predefined indicator systems. The study of Doğan–Al (2019) shows that the leaders of the institutions, knowing the indicators of the rankings, drive the university to comply with them, which is most often realized by achieving the highest possible publication numbers. Here, it is again worth emphasizing that most rankings are research-focused, so educational and other activities are given less weight in the measurement. In their publication, they compared five international rankings, and the study sought to identify, by statistical means, the indicators that decision-makers should focus on when allocating resources, thereby effectively improving institutional

competitiveness. In their analysis, they identify two indicators: the number of researchers with a high number of citations and the number of publications in Nature or Science. Based on their additional indicator correlation calculations, this result is redundant; however, the authors highlighted that measuring the overall quality of universities cannot be achieved with purely statistical tools. Their result is that there are only slight differences in the data of the institutions in the top 200 rankings. Kivinen (2017) also reached the same conclusion, with the addition that in terms of disciplines, much smaller differences can be observed in the natural sciences, while in terms of the social sciences and humanities, these differences are already more significant. Kivinen's other finding is that the research pillar data have the lowest (20%) ranking in the QS ranking, while the same is 40% for THE and 60% for ARWU.

Previous papers have already investigated the variation in QS ranking using indicators extracted from the SciVal database (Dobos et al. 2022). The results show that QS ranking can be very well estimated using these indicators. However, this analysis did not address the statistical and probability relationship between the indices. One attempt has been made to explore the relationships between statistical variables (Dobos–Sasvári 2021), but these analyses did not consider the spatiality of the results.

In the current study, we examine the QS ranking based on publication data from Scopus and SciVal. Essentially, these bibliometric data are also based organically on the research pillar, so in principle, with respect to the above information, the QS ranking is based on only 20% of the data. In our analysis, we try to predict institutional rankings in the QS ranking using these publication data alone, testing the predictive value of hard factors versus soft factors.

Compilation of the dataset

When compiling the dataset, the QS ranking was considered given because it can be freely downloaded from the institute's website (QS World University Rankings 2021). However, the data available there still need to be made user-friendly because they contain multiple ties. The resolution of ties was solved in the usual way in statistics. Where there was a tie, by adding the serial numbers of each tied university in the rankings, we replaced the position in the rankings with their average. At that time, the universities in the tie were given the same ranking value.

As we are examining how to approximate the QS ranking with the indicators available on the SciVal/Scopus pages, we included freely available data for estimation. The variables also included indicators of publication, citation, and author numbers for each university that were available. These variables are as follows, with abbreviations in parentheses:

- number of total publications (PUB),
- total number of authors (AUT),

– field-weighted citation impact (FWCI),
– all citations (CIT),
– the five-year Hirsch index (H5-I),
– the university's academic faculty staff (AFS), and
– the ranked position (QS-R).

Data and variables were recorded on September 21, 2020. From the variables, the FWCI certainly needs further explanation, while the others, including the Hirsch index, are well known. The FWCI basically shows how often the author's publications are cited. If the FWCI is greater than one, more citations are expected from the publication compared to other publications in similar subject areas. The calculation algorithm for the FWCI index can be found on the Elsevier (2019) site and presented in the article of Purkayastha et al. (2019).

Statistical analysis of the dataset

The QS 2021 list includes 1,003 universities that are examined along the above seven variables. The distribution of QS-ranked universities by country can be found in the appendix. First, we generate the correlation matrix and measure the linear relationships between the variables. In the following analysis, we examine variable reduction by principal component analysis. We then consider the multicollinearity between the variables using the VIF indicator. In the fourth subsection, we estimate multicollinear variables with independent variables and examine the extent to which QS ranking can be estimated by using other variables. Our regression equations are generated by the stepwise method. After that, we present a causal study using partial correlation.

Correlation analysis

As summarized in Table 2, we measured a very high correlation between the selected variables, except for the *FWCI* index. *FWCI* shows a very weak linear correlation with four variables and a weak moderate linear correlation with the *H5-I* and *CIT* variables. However, this is not surprising because both *H5-I* and *CIT* are citation characteristic variables. A medium to strong linear relationship can be detected between the other six variables.

Table 2

Correlation between variables

Variables		AUT	FWCI	CIT	H5-I	AFS	QS-R
PUB		0.724	0.428	0.961	0.893	0.565	–0.622
	Significant (2-sided)	0.000	0.000	0.000	0.000	0.000	0.000
	N	1,002	1,002	1,002	1,001	979	1,002
AUT			0.192	0.645	0.585	0.599	–0.419
	Significant (2-sided)		0.000	0.000	0.000	0.000	0.000
	N		1,002	1,002	1,001	979	1,002
FWCI				0.525	0.683	0.109	–0.477
	Significant (2-tailed)			0.000	0.000	0.001	0.000
	N			1,002	1,001	979	1,002
CIT					0.911	0.502	–0.609
	Significant (2-sided)				0.000	0.000	0.000
	N				1,001	979	1,002
H5-I						0.482	–0.685
	Significant (2-sided)					0.000	0.000
	N					978	1,001
AFS							–0.326
	Significant (2-sided)						0.000
	N						980

Source: own compilation based on the Scopus database.

Another interesting feature of the correlations is that *H5-I* shows a relatively strong correlation with all variables. The correlation matrix suggests that the variables can be divided into several groups. All correlation coefficients are significant. Correlation coefficients were determined from relationships with different numbers of items, as a total of 25 universities had a missing value. These were not replaced because the number of items was still large enough to determine the correlation.

Principal component analysis

In the principal component analysis of the seven variables, we obtained three components that accounted for 87.620% of the variance. The fit of the model according to the Kaiser–Meyer–Olkin test is 0.804, which means a strong moderate model according to the accepted categorization.

Table 3

Components of variables, and a rotated component matrix of variables

Variables	Component		
	1	2	3
AFS	**0.855**	0.008	0.074
AUT	**0.842**	0.127	0.208
PUB	**0.729**	0.481	0.385
FWCI	–0.017	**0.930**	0.176
H5-I	**0.529**	**0.707**	0.408
CIT	**0.645**	**0.598**	0.354
QS-R	–0.224	–0.291	**–0.918**

Methods used: principal component analysis and Varimax rotation with Kaiser normalization.
Source: own compilation based on the Scopus database.

As might be expected from the correlation analysis, due to the high correlation coefficient of the seven variables, the number of citations and the variables calculated from it, i.e., the *CIT*, *H5-I* and *FWCI* variables, were included in all three main components. The variables of the first component, which essentially include the number of academic faculty members and the number of publications (*PUB*, *AUT*, and *AFS*), explain 38.828% of the variance. The second component shows a strong correlation with the *CIT*, *H5-I* and *FWCI* variables, explaining 29.326% of the variance. Finally, the third component essentially contains the QS ranking and accounts for 19.466% of the variance. An interesting feature of the component model is that the *CIT* and *H5-I* variables were included in essentially all three components.

Since the correlation coefficient between the seven variables is relatively high, we may expect high collinearity between them, so we need to test it.

Examination of multicollinearity with the VIF index

There is no uniform rule in the literature above in which value variables can be considered collinear, although there are certain empirically tested VIF thresholds that deviate from 2.5 to 10. In the case of filtering out redundancy, there is no set of theoretical/logical rules by which these can be reliably determined. We therefore made a decision in this regard by choosing to accept the recommendations of several articles (Lafi–Kaneene 1992, Liao–Valliant 2012, O'Brien 2007): we chose five as the threshold.

Table 4 shows the sequential filtering of the variables. It is worth noting here that there is no deterministic algorithm for filtering collinear variables. As a first step, it is recommended to filter out the variable with the highest VIF value, but any variable above the threshold is appropriate to take the first step. In the next step, there are again two options: either we select the element with the highest VIF value again or

the variable with the largest decrease in the value of VIF. In our case, we treated multicollinearity sequentially by choosing the highest value. This is because in the first step, the VIF value of the PUB variable has the largest value of 23.025, so we eliminated it. In the second step, the variable H5-I was removed because of the two VIF values greater than five, specifically the one with a higher value of 10.612. With these two steps, our algorithm also ended because the VIF value of each remaining variable remained below three.

When examining the initial VIF values, it is immediately revealed that the values of the variables AUT, FWCI, AFS, and QS-R are lower than the initial threshold of 5, so these variables could not be included in the collinear variables to be eliminated due to the stepwise decrease in the VIF value.

Table 4

Evolution of VIF values during the algorithm

Variables	Step		
	0	1	2
PUB	**23.025**	–	–
AUT	2.502	2.133	2.129
FWCI	2.944	2.354	1.574
CIT	**17.942**	**7.064**	2.804
H5-I	**12.260**	**10.612**	–
AFS	1.722	1.712	1.649
QS-R	1.922	1.894	1.695

Source: own compilation based on the Scopus database.

This means that the number of publications (*PUB*) and the Hirsch index (*H5-I*) variables depend linearly on the other variables.

Linear regression estimation of collinear variables and QS ranking

The filtered two variables are estimated with the remaining five variables. For the regression estimation, we do not use the usual "enter" method but the "stepwise" regression, in which the variable with a nonsignificant parameter is filtered by the algorithm. We do the same in the case of linear estimation of QS ranking.

When estimating the number of publications in Scopus with the remaining five variables, the R^2 value of the estimate became 0.950, which can be considered very high. The linear equation of the estimate is as follows:

$$PUB = 8{,}677.222 + 0.059 \cdot AUT + 0.173 \cdot FWCI - 2{,}997.675 \cdot CIT - 3.720 \cdot AFS + 0.368 \cdot QS\text{-}R \quad (1)$$

This shows that the number of authors, the *FWCI* index, and the place in the QS ranking increase, while the number of citations and the total number of academic faculty staff decrease the total number of dissertations. Because we cannot establish

a logical relationship, we cannot determine the causal relationship in this step. Each of our parameters is significant at the 0.000 level, which supports the interpretability of the model.

The estimation of the Hirsch index with four variables resulted in a model where R^2 was 0.906. The estimate of the variable is as follows:

$$H5\text{-}I = 23.389 + 31.400 \cdot FWCI + 0.0002 \cdot CIT + 0.003 \cdot AFS - 0.025 \cdot QS\text{-}R \quad (2)$$

The same can be stated for this index as for all publications. However, our variables cannot be estimated by the number of authors. The *FWCI*, the number of citations, and all professional staff increase the Hirsch index, but their place in the QS ranking decreases it. The parameters are significant at the 0.000 level.

Finally, the position in the QS ranking is estimated. The value of R^2 was 0.469, which can be considered moderate. The linear equation of our estimate is as follows:

$$QS\text{-}R = 830.634 - 3.865 \cdot H5\text{-}I \quad (3)$$

This shows that only one variable, i.e., the Hirsch index, is enough to estimate the place in the QS ranking. The coefficients of the other variables are so insignificant that omitting the variables does not significantly reduce the value of R^2 either. If the nonsignificant variables were included in the QS ranking estimate, the value of R^2 would be 0.480, but the other variables would not be significant, while the *H5-I* parameter is significant at the 0.000 level. It is worth noting that the value of the multiple correlation coefficient is 0.693, suggesting a strong correlation between the QS ranking and the variables collected from Scopus and SciVal.

Partial correlation analysis: Cause and effect

Partial correlation is suitable for filtering out the effect of other variables when determining the correlation between two variables in a linear model. This can also be interpreted by mapping the causal relationship between the two variables. Table 5 shows the partial correlations that help describe the causal relationships.

When exploring causal relationships, partial correlation values above 0.25 in absolute value are considered. There are three values between 0.44 and 0.78, while there are four additional values between 0.25 and 0.4. In Table 5, we coloured the examined partial correlations.

Table 5

Partial correlations

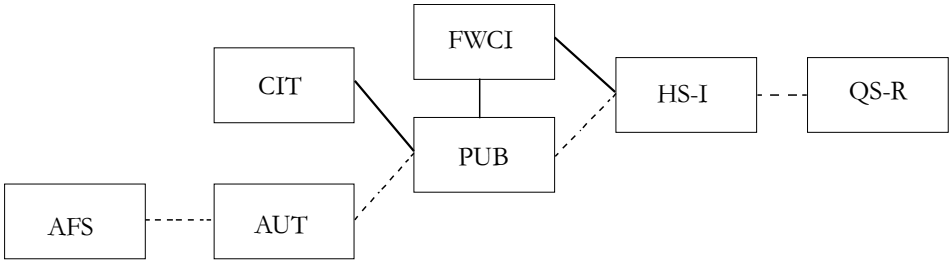
Variables		AUT	FWCI	CIT	H5-I	AFS	QS-R
PUB		**0.384**	**–0.448**	**0.779**	**0.367**	**0.076**	**–0.122**
	Significant (2-sided)	0.000	0.000	0.000	0.000	0.017	0.004
	N	971	971	971	971	971	971
AUT			**0.035**	**–0.158**	**–0.100**	**0.307**	**–0.003**
	Significant (2-sided)		0.277	0.000	0.002	0.000	0.937
	N		971	971	971	971	971
FWCI				**0.219**	**0.644**	**–0.163**	**–0.059**
	Significant (2-tailed)			0.000	0.000	0.000	0.065
	N			971	971	971	971
CIT					**0.168**	**–0.075**	**0.134**
	Significant (2-sided)				0.000	0.019	0.000
	N				971	971	971
H5-I						**0.151**	**–0.255**
	Significant (2-sided)					0.000	0.000
	N					971	971
AFS							**0.028**
	Significant (2-sided)						0.381
	N						971

Source: own compilation based on the Scopus database.

Figure 1 shows the causal relationships between the variables. The relationships between 0.44 and 0.78 are indicated by solid lines and correlations between 0.25 and 0.4 by dotted lines. The figure immediately shows that the citation block, i.e., all citations, Hirsch index, and *FWCI* index, depends on the total number of publications. This highlights that the number of publications shows a strong correlation with the evolution of citations. At the same time, the number of authors is positively related to publication indices, i.e., all publications.

Figure 1

Causal relations between the variables



Source: own compilation based on the Scopus database.

In summary, that causal relationship can be described thus: an increase in the number of coauthors increases the number of publications. Nevertheless, the number of publications can increase the number of citations and then the Hirsch index. As a result, positions in the QS ranking through the Hirsch index may also improve.

Grouping of universities by cluster analysis

We also attempted to group universities. This study aimed to determine whether university groups are recognizable in the dataset. Grouping was performed by using the Quick Cluster technique. The advantage of this technique is that it determines the centers of each cluster, which allows each group to be typified.

Table 6 shows that the number of items of the 13 selected clusters is very different. Eight of the clusters had fewer than eleven items. These eight clusters thus included a total of 34 universities. In the other five clusters, however, the number of items was at least 43. The number of clusters was set relatively high, but even in this way, our dataset was not divided into interpretable clusters, i.e., groups. However, approximately 81% of the dataset was not further decomposed by the algorithm.

Table 6

Number of universities in the 13 clusters

Cluster	Case number	Cluster	Case number
1	4	9	109
2	2	10	5
3	1	11	1
4	11	12	7
5	3	13	71
6	43	total:	978
7	471	missing value:	25
8	250		

Source: own compilation based on the *QS* database.

The 8 clusters of 34 universities are presented in Table 7. These eight clusters may include universities with "outstanding" data having been identified by using cluster centers with the exception of Mexico's Universidad Nacional Autónoma de México (UNAM), which, based on the values of indicators, composes a cluster of its own. Of the 33 remaining universities, 4 are Chinese, 2 are French and 1 is Danish. The remaining 26 universities are all in the Anglo-Saxon language area. These are the United States (17), the United Kingdom (4), Australia and Canada (both 2) and Singapore (1).

Table 7

The eight clusters with the fewest items and the universities in them

Cluster number	Quantity	Name of university	Country
3	1	Harvard University	United States
11	1	Universidad Nacional Autónoma de México (UNAM)	Mexico
2	2	University of Toronto	Canada
		Stanford University	United States
5	3	University of Oxford	United Kingdom
		UCL (University College London)	United Kingdom
		Johns Hopkins University	United States
1	4	University of Paris-Saclay	France
		Massachusetts Institute of Technology (MIT)	United States
		University of Michigan-Ann Arbor	United States
		University of Washington	United States
10	5	University of Melbourne	Australia
		Peking University	China
		University of Copenhagen	Denmark
		Sorbonne Université	France
		University of California, Berkeley (UCB)	United States
12	7	Tsinghua University	China
		University of Cambridge	United Kingdom
		Imperial College London	United Kingdom
		University of Pennsylvania	United States
		Columbia University	United States
		University of California at Los Angeles	United States
		University of California at San Diego	United States
4	11	University of Sydney	Australia
		University of British Columbia	Canada
		Shanghai Jiao Tong University	China
		Zhejiang University	China
		National University of Singapore (NUS)	Singapore
		University of Chicago	United States
		Yale University	United States
		Cornell University	United States
		Northwestern University	United States
		Duke University	United States
		University of Pittsburgh	United States
Total:	34		

Source: own compilation based on the *QS* database.

We characterized the 34 universities using the cluster means. The cluster means were not identified by the mean values but by the order of the individual variables, i.e., the values measured on at least the interval scale were transformed to an ordinal scale. On the ordinal scale, it was best for *PUB*, *FWCI*, *CIT*, and *H5-I* to have values as large as possible. However, we considered the lowest value for our *AUT*, *AFS*, and *QS-R* variables first. The results are shown in Table 8.

Table 8

Transformed values of cluster means

Clusters	PUB	AUT	FWCI	CIT	H5-I	AFS	QS-R	Quantity
3	1	12	2	1	1	8	1	1
11	9	13	13	11	11	13	8	1
2	2	11	1	2	2	12	3	2
5	3	9	3	3	3	11	2	3
1	4	10	4	4	4	6	9	4
10	6	6	7	6	6	7	6	5
12	5	8	5	5	5	10	4	7
4	7	7	6	7	7	9	5	11
6	8	5	8	8	8	5	7	43
13	10	4	9	9	9	4	10	71
9	11	3	10	10	10	3	11	109
8	12	2	11	12	12	2	12	250
7	13	1	12	13	13	1	13	471
Total								978

Source: own compilation based on the Scopus database.

It is immediately apparent that the third cluster is excellent along almost every variable, but this has been achieved by the high number of lecturer–researcher staff and the many authors. The same can be said for the second and fifth clusters as well. However, the eleventh cluster seems to be the worst along almost all variables. In the first, fourth, tenth, and twelfth clusters, a very narrow circle follows the top four universities. The other five clusters include smaller and medium-sized universities, which is also quite clear from the rankings.

Figure 2

The geographic distribution of universities belonging to the eight remaining clusters

Source: own compilation based on the *QS* database.

As noted above, the 34 universities in the eight sharply diverging clusters are all excellent along the indicators examined, the one exception being a Mexican university

in the eleventh cluster. The distribution of the other 33 universities across countries is illustrated in Figure 2. Half of the universities are from the United States of America, followed by China and the United Kingdom with 4 universities each. If we look at the universities in terms of which ones are from native English-speaking countries, there are 26.

These countries are the United States (17), the United Kingdom (4), Australia (2), Canada (2) and Singapore (1).

In Europe, the top-ranked universities are France (2) and Denmark (1), in addition to the UK, making a total of seven universities from Europe.

Conclusion

In this paper, a positive answer is given to the question of whether the ranking published by QS can be estimated using indicators taken from the Scopus/SciVal database. However, the examination of the linear relationship between the variables included also showed that not all variables are needed because of the high multicollinearity between the variables. The clustering, or cluster analysis, indicates that most of the universities considered to be outstanding are from Anglo-Saxon countries.

University rankings are one of the tools often used by education decision-makers and science politicians to prepare decisions. This paper examined one of these prominent sources of information, the QS World University Rankings 2021. Rankings use a variety of information with different weights, such as research performance through publications and citations, educational performance, or the university's ability to attract industrial R&D funds. From these "legs", the analysis considered only the research, and within that, only those data that could be extracted from the freely available Scopus/SciVal databases. Thus, the purpose of the study was twofold. On the one hand, we analyzed the linear relationships between the variables extracted from the datasets, and on the other hand, we grouped universities into groups using cluster analysis. Mapping of linear relationships between variables was performed using five techniques. The correlation analysis showed that there is a relatively strong linear relationship between the selected variables. All this points in the direction that the variables can be grouped using principal component analysis. The correlation matrix of the seven variables was returned using three components. This reproduced nearly 88% of the variance. The first component shows a strong relationship with headcount data and the number of publications. The second component contained the citations and the indicators that could be derived from them until eventually the QS sequence alone was included in one component. Knowing this, we were able to reduce the seven variables with the variance inflation factor. The H5 index and number of publications show strong collinearity with the remaining five variables. Interestingly, we estimated the QS ranking using a regression model, which gave a

high R^2 value of 0.469. Another interesting feature of the estimate is that the order in the case of stepwise regression depends only on the five-year Hirsch index. Finally, a causal relationship was revealed by partial correlation analysis. This study essentially confirmed the results of the principal component analysis and our variables attributable to the factors. Accordingly, in addition to the relationship between the headcount data and the number of publications, the citation indicators were combined, and at the end of the chain, the QS university ranking was linked to the five-year Hirsch index. This result was also supported by the regression.

In the cluster analysis, the result was that large and well-known universities form clusters in smaller groups and numbers. As shown, 34 out of the 1003 universities were included in eight clusters, indicating low density. The other five clusters then included universities with a high number of items, making it more difficult to distinguish between them. Groups were represented by cluster means, and mean values were transformed to an ordinal scale. This showed that the clusters yielded almost the same order along five variables, while in the case of headcount data, even if it was rendered in reverse order, it resulted in a similar order. Smaller universities included in the QS list were divided into different groups, which are among the smaller groups in the international comparison. This fact also shows that smaller universities, given their current size, can only rise significantly in such rankings if they merge with larger organizations.

Subsequent research could address whether the latter statement is also met for the other two major international rankings, i.e., the ARWU and THE, and US News lists, if the rankings are predicted using the Scopus/SciVal databases.

Appendix

Table A1

Distribution of QS-ranked universities by country

Country	Frequency	%
United States	151	15.1
United Kingdom	84	8.4
China (Mainland)	51	5.1
Germany	45	4.5
Japan	41	4.1
Australia	36	3.6
Italy	36	3.6
South Korea	29	2.9
France	28	2.8
Russia	28	2.8
Canada	26	2.6
Spain	26	2.6
India	21	2.1
Malaysia	20	2.0
Taiwan	16	1.6
Poland	15	1.5
Brazil	14	1.4
Argentina	13	1.3
Netherlands	13	1.3
Mexico	12	1.2
Colombia	11	1.1
Chile	10	1.0
Czech Republic	10	1.0
Kazakhstan	10	1.0
Saudi Arabia	10	1.0
Switzerland	10	1.0
Belgium	9	0.9
Finland	9	0.9
Turkey	9	0.9
Austria	8	0.8
Hungary	8	0.8
Indonesia	8	0.8
Ireland	8	0.8
Lebanon	8	0.8
New Zealand	8	0.8
Sweden	8	0.8
Thailand	8	0.8
United Arab Emirates	8	0,8
Hong Kong SAR	7	0.7
Pakistan	7	0.7

(Table continues on the next page.)

(Continued.)

Country	Frequency	%
Portugal	7	0.7
South Africa	7	0.7
Greece	6	0.6
Israel	6	0.6
Ukraine	6	0.6
Denmark	5	0.5
Iran, Islamic Republic of	5	0.5
Egypt	4	0.4
Jordan	4	0.4
Lithuania	4	0.4
Norway	4	0.4
Philippines	4	0.4
Slovakia	4	0.4
Uruguay	4	0.4
Venezuela	4	0.4
Costa Rica	3	0.3
Ecuador	3	0.3
Estonia	3	0.3
Kuwait	3	0.3
Latvia	3	0.3
Peru	3	0.3
Singapore	3	0.3
Bahrain	2	0.2
Bangladesh	2	0.2
Belarus	2	0.2
Brunei	2	0.2
Croatia	2	0.2
Cuba	2	0.2
Iraq	2	0.2
Macau SAR	2	0.2
Romania	2	0.2
Slovenia	2	0.2
Vietnam	2	0.2
Bulgaria	1	0.1
Cyprus	1	0.1
Georgia	1	0.1
Malta	1	0.1
Oman	1	0.1
Panama	1	0.1
Qatar	1	0.1
Total	1003	100.0

Source: own compilation based on the *QS* database.

REFERENCES

ABRAMO, G.–D'ANGELO, C. A. (2014): How do you define and measure research productivity? *Scientometrics* 101: 1129–1144. https://doi.org/10.1007/s11192-014-1269-8

ABRAMO, G.–D'ANGELO, C. A.–DI COSTA, F. (2008): Assessment of sectoral aggregation distortion in research productivity measurements *Research Evaluation* 17 (2): 111–121. https://doi.org/10.3152/095820208X280916

AITHAL, P. S.–SHAILASHREE, V. T.–KUMAR, P. M. (2016): The study of new national institutional ranking system using ABCD framework *International Journal of Current Research and Modern Education (IJCRME)* 1 (1): 389–402. http://dx.doi.org/10.5281/zenodo.161077

ALTBACH, P. G. (2012): The globalization of college and university rankings *Change: The Magazine of Higher Learning* 44 (1): 26–31. https://doi.org/10.1080/00091383.2012.636001

ALTBACH, P. G. (2013): *The international imperative in higher education* Sense Publishers, Rotterdam. http://doi.org/10.1007/978-94-6209-338-6

BARKER, K. C. (2002): *Canadian recommended e-learning guidelines* Futur Ed for Canadian Association for Community Education and office of Learning Technologies, HRDC, Vancouver, BC.

BENITO, M.–GIL, P.–ROMERA, R. (2020): Evaluating the influence of country characteristics on the higher education system rankings' progress *Journal of Informetrics* 14 (3): 101051, https://doi.org/10.1016/j.joi.2020.101051

BOBBY, C. L. (2014): *The abcs of building quality cultures for education in a global world* Paper presented at the International Conference on Quality Assurance, Bangkok, Thailand.

BOGUE, G. (1998): Quality assurance in higher education: The evolution of systems and design ideals *New Directions for Institutional Research* 99: 7–18. https://doi.org/10.1002/ir.9901

BUELA-CASAL, G.–GUTIÉRREZ-MARTINEZ, O.–BERMÚDEZ-SÁNCHEZ, M. P.–VADILLO-MUNOZ, O. (2007): Comparative study of international academic rankings of universities *Scientometrics* 71 (3): 349–365. https://doi.org/10.1007/s11192-007-1653-8

DE WIT, H. (2015): Is the international university the future for higher education? *International Higher Education* 80: 7. https://doi.org/10.6017/ihe.2015.80.6133

DOBOS, I.–MICHALKÓ, G.–SASVÁRI, P. (2021): The publication performance of Hungarian economics and management researchers: a comparison with the Visegrád 4 countries and Romania *Regional Statistics* 11 (2): 165–182. https://doi.org/10.15196/RS110207

DOBOS, I.–SASVÁRI, P.–URBANOVICS, A. (2022): The predictability of QS ranking based on Scopus and SciVal data. The predictability of QS ranking based on Scopus and SciVal data. *KOME – An International Journal of Pure Communication Inquiry* Forthcoming. https://doi.org/10.17646/KOME.75672.85

DOBOS, I.–SASVÁRI, P. (2021): A QS World University Rankings 2021 vizsgálata a Scopus-/SciVal-adatbázisok segítségével *Statisztikai Szemle* 99 (9): 874–900. https://doi.org/10.20311/stat2021.9.hu0874

DOĞAN, G.–AL, U. (2019): Is it possible to rank universities using fewer indicators? A study on five international university rankings *Aslib Journal of Information Management* 71 (1): 18–37. https://doi.org/10.1108/AJIM-05-2018-0118

FERANCOVÁ, A.–KRIGOVSKÁ, A. (2016): Measuring the performance of universities through cluster analysis and the use of financial ratio indexes *Economics and Sociology* 9 (4): 259–271. https://doi.org/10.14254/2071-789X.2016/9-4/16

GLÄNZEL, W. (2008): Seven myths in bibliometrics. About facts and fiction in quantitative science studies *COLLNET Journal of Scientometrics and Information Management* 2 (1): 9–17. https://doi.org/10.1080/09737766.2008.10700836

GREEN, D. (1994): *What is quality in higher education?* Society for Research into Higher Education, London.

HALAWEH, M. (2020): Research Productivity Index (RPI): a new metric for measuring universities, research productivity *Information Discovery and Delivery* 49 (1): 29–35. https://doi.org/10.1108/IDD-01-2020-0003

HARVEY, L. (2008): Rankings of higher education institutions: a critical review *Quality in Higher Education* 14 (3): 187–207. https://doi.org/10.1080/13538320802507711

HAZELKORN, E.–LOUKKALA, T.–ZHANG, T. (2014): *Rankings in institutional strategies and processes: Impact or illusion?* European University Association, Brussels.

IOANNIDIS, J. P.–PATSOPOULOS, N. A.–KAVVOURA, F. K.–TATSIONI, A.–EVANGELOU, E.–KOURI, I.–CONTOPOULOS-IOANNIDIS, D. G.–LIBEROPOULOS, G. (2007): International ranking systems for universities and institutions: a critical appraisal *BMC Medicine* 5 (1): 30. https://doi.org/10.1186/1741-7015-5-30

KABA, A. (2020): Global research productivity in knowledge management: an analysis of Scopus database *Library Philosophy and Practice* 3920. https://digitalcommons.unl.edu/libphilprac/3920

KING, R. (2009): *Governing universities globally: organizations, regulation and rankings* Edward Elgar Publishing Cheltenham, UK.

LAFI, S. Q.–KANEENE, J. B. (1992): An explanation of the use of principal-components analysis to detect and correct for multicollinearity *Preventive Veterinary Medicine* 13 (4): 261–275. https://doi.org/10.1016/0167-5877(92)90041-D

LAREDO, P. (2007): Revisiting the third mission of universities: toward a renewed categorization of university activities? *Higher Education Policy* 20 (4): 441–456. https://doi.org/10.1057/palgrave.hep.8300169

LEE, J.–LIU, K.–WU, Y. (2020): Does the Asian catch-up model of world-class universities work? Revisiting the zero-sum game of global university rankings and government policies *Educational Research for Policy and Practice* 19: 319–343. https://doi.org/10.1007/s10671-020-09261-x

LIU, N. C.–CHENG, Y. (2005): The academic ranking of world universities *Higher Education in Europe* 30 (2): 127–136. https://doi.org/10.1080/03797720500260116

LIU, Z.–MOSHI, G. J.–AWUOR, C. M. (2019): Sustainability and indicators of newly formed world-class universities (NFWCUs) between 2010 and 2018: empirical analysis from the rankings of ARWU, QSWUR and THEWUR *Sustainability* 11 (10): 2745. https://doi.org/10.3390/su11102745

LOTKA, A. J. (1926): The frequency distribution of scientific productivity *Journal of the Washington Academy of Sciences* 16 (12): 317–324. https://www.jstor.org/stable/24529203

LOWRY, P. B.–KARUGA, G. G.–RICHARDSON, V. J. (2007): Assessing leading institutions, faculty, and articles in premier information systems research journals *Communications of the Association for Information Systems* 20: 142–203. https://doi.org/10.17705/1CAIS.02016

MAMMADLI, A. (2021): Global university rating indicators and suggestion for establishment of entrepreneur universities in Azerbaijan *InterConf* 42: 192–210. https://doi.org/10.51582/interconf.19-20.02.2021.016

MARGINSON, S.–VAN DER WENDE, M. (2009): The new global landscape of nations and institutions. In: OECD (ed.): *Higher education to 2030 Volume 2: Globalisation* pp. 17–62., OECD, Paris. https://doi.org/10.1787/9789264075375-en

O'BRIEN, R. M. (2007): A caution regarding rules of thumb for variance inflation factors *Quality & Quantity* 41 (5): 673–690. https://doi.org/10.1007/s11135-006-9018-6

OECD (2015): *Education at a Glance 2015: OECD Indicators* OECD Publishing, Paris. https://doi.org/10.1787/eag-2015-en

PIETRUCHA, J. (2018): Country-specific determinants of world university rankings *Scientometrics* 114: 1129–1139. https://doi.org/10.1007/s11192-017-2634-1

PURKAYASTHA, A.–PALMARO, E.–FALK-KRZESINSKI, H. J.–BAAS, J. (2019): Comparison of two article-level, field-independent citation metrics: field-weighted citation impact (FWCI) and relative citation ratio (RCR). *Journal of Informetrics* 13 (2): 635–642. https://doi.org/10.1016/j.joi.2019.03.012

SAFON, V. (2013): What do global university rankings really measure? The search for the X factor and the X entity *Scientometrics* 97: 223–244. https://doi.org/10.1007/s11192-013-0986-8

SALMI, J. (2009): *The challenge of establishing world-class universities* World Bank, Washington, DC. https://doi.org/10.1596/978-0-8213-7865-6

SALMI, J. (2011): Nine common errors when building a new world class university *Dyna* 78 (168): 5–7. https://doi.org/10.6017/ihe.2011.62.8529

SHEEJA, N. K.–MATHEW K., S.–CHERUKODAN, S. (2018): Impact of scholarly output on university ranking *Global Knowledge, Memory and Communication* 67 (3): 154–165. https://doi.org/10.1108/GKMC-11-2017-0087

SHIN, J. C.–TOUTKOUSHIAN, R. K.–TEICHLER, U. (2011): *University rankings: theoretical basis' methodology and impacts on global higher education* (Vol. 3), Springer, Dordrecht.

YERAVDEKAR, V. R.–TIWARI, G. (2014): Internationalization of higher education in India: how primed is the country to take on education Hubs? *Procedia – Social and Behavioral Sciences* 157 (27): 165–182. https://doi.org/10.1016/j.sbspro.2014.11.020

INTERNET SOURCES

ELSEVIER (2019): *Research metrics guidebook.* https://www.elsevier.com/research-intelligence/resource-library/research-metrics-guidebook (downloaded: February 2023)

LIAO, D.–VALLIANT, R. (2012): Variance inflation factors in the analysis of complex survey data *Survey Methodology* 38 (1): 53–62. https://www.rti.org/publication/variance-inflation-factors-analysis-complex-survey-data/fulltext.pdf (downloaded: 21/04/2021)

LOUGHRAN, G. (2016): Why university rankings may be harming higher education *The Irish Times*. www.irishtimes.com/news/education/why-university-rankings-maybe-harminghigher-education-1.2793532 (downloaded: February 2023)

QS WORLD UNIVERSITY RANKINGS (2021): https://www.topuniversities.com/university-rankings/world-university-rankings/2021 (downloaded: February 2023)

YERBURY, D. (2006): Spreading universities' foreign risks *The Age*. https://www.theage.com.au/national/spreading-universities-foreign-risks-20060112-ge1k4w.html (downloaded: February 2023)

DATABASE/WEBSITE

Scopus database: https://www.scopus.com