Scientometrics (2011) 86:77–92
DOI 10.1007/s11192-010-0280-y

On using the Shanghai ranking to assess the research performance of university systems

Domingo Docampo

Received: 6 February 2010 / Accepted: 16 August 2010 / Published online: 29 August 2010

Abstract We take a new look at the Shanghai Jiao Tong *Academic Ranking of World Universities* to evaluate the performance of whole university systems. We deal with system aggregates by means of averaging scores taken over a number of institutions from each higher education system according to the Gross Domestic Product of its country. We treat the set of indicators (measures) at the country level as a scale, and investigate its reliability and dimensionality using appropriate statistical tools. After a Principal Component Analysis is performed, a clear picture emerges: at the aggregate level ARWU seems to be a very reliable one-dimensional scale, with a first component that explains more than 72% of the variance of the sample under analysis. The percentages of variance of the indicators explained by the first component do shed light on the fact that ARWU is in fact measuring the research quality (both at the individual and collective levels) of a university system. When the second principal component is taken into account, the two principal components contribute to explain more than 90% of the variance. The rotated solution facilitates the interpretation of the components and provides clear and interesting clustering information about the 32 higher education systems under analysis.

Keywords University system · Ranking · Scale · Reliability · PCA · Clustering · Shanghai

Introduction

Ranking higher education institutions by means of subjective or objective measures goes well with the logic of these times of global competition and continuous bench-marking. Rankings are here to stay, because they attract the interest of current and future students, and the attention both of country officials and the public at large; it is no wonder that their periodic publication is subjected to close scrutiny by university authorities all over the world, and it is a fact that they stimulate personal and institutional commitment to quality.

D. Docampo (✉)
Universidad de Vigo, ETSE TELECOMUNICACION, Campus Universitario, 36310 Vigo, Spain
e-mail: ddocampo@uvigo.es

However, even the most credited global rankings are at least moderately controversial and spur academic debate on a number of issues, usually revolving around the difficulty of capturing an institutional essence with just one aggregate number.

The last decade has witnessed the emergence of the Shanghai Jiao Tong University Institute of Higher Education Academic Ranking of World Universities (ARWU), which ranks academic institutions on the basis of their research performance. As Marginson (2005) noted, "ARWU rankings are credible, based on solid, transparent numerical data of research quality and quantity—Nobel Prize winners, publications in prestigious journals, citations, etc.—and knowledge of the rankings has rapidly spread across the world". However, in spite of its having been designed to rank research universities by comparing objectively their research output, ARWU is unfortunately commencing to be used as a stick to measure institutions, and not just in relation with research. That is the major source of ARWU's troubles, one that has strongly contributed to raise the volume of the controversy around the Shanghai rankings.

In this paper we are interested in making insightful comparisons among the research quality of the different university systems around the world; we make use of ARWU because it is the only global ranking that focuses on research and does not rely on subjective data. All its indicators are open to public scrutiny, since they measure either scientific production or individual excellence recognized by very prestigious awards or by a high number of citations.[1] One thing should be clear for anyone interested in the analysis of this ranking: ARWU does not even try to assess quality at any other measurable output than research performance, and even for that single parameter, it takes a very extreme position, as one can easily realize by just taking a look at the set of indicators selected for the ranking.

Since our interest lies in assessing the quality of the research performance of university systems, we have to pose the question of how to use the information from the ranking about its institutions to summarize the performance of whole higher education systems. We will climb up to the system level by averaging over a number of institutions from each country according to its share of the World's Gross Domestic Product. This choice accommodates the information in such a way that a unified treatment can be made for countries with a very large share of the World's GDP (US, China, Japan, ...) and countries showing meager numbers (New Zealand, Hungary, Singapore, Israel, ...). In the paper we provide the aggregate measures at the university system level as entries to a scale, to then investigate its reliability and dimensionality.

A caveat must be issued concerning the data. We take the data as they are, although some pertinent concerns have been raised by Florian (2007) around their reproducibility, and some data manipulation has been performed on the original information as recognized by Liu and Cheng (2005), "the distribution of data for each indicator is examined for any significant distorting effect and standard statistical techniques are used to adjust the indicator if necessary". In particular, we have noticed that in the case of one of the indicators the authors have used the square root of the number of highly cited authors from an institution (Docampo 2008). That we can do because the number of Highly Cited authors can be downloaded from a publicly accessed database. We can say nothing about the statistical manipulations of the other indicators, since the raw data have not been made public (and it is high time they did that) by the ranking authors. Moreover, as Zitt and Filliatreau (2007) have pointed out, the indicators related to scientific production in specific journals (Science and Nature) or broad SCI publications are size-dependent, a fact that is no more than slightly corrected by means of a per-capita measure only adequately

[1] For a review of the nature of the indicators we refer to the web page of ARWU (http://www.arwu.org).

computed in universities from a few countries, and weakly weighted on the final score. This poses a problem for making sensible comparisons among individual institutions. However, the scores for the higher education systems will be computed as averages taken over a number of their universities (on a range from 2 to 71, a median of 3 and a mean value close to 8), and that procedure smoothes the effects of the diversity of sizes in such a way that no particular size-side effects should mask the main results of our analysis at least for moderate and large countries, although it is arguable that the averaging process copes with the problem in the case of very small countries.

A barrage of critics to the ranking methodologies has been recently raised by Billaut et al. (2010), based on the application of multiple criteria decision making theory. They might have missed the point: with all the inherent defects that accompany any attempt to measure anything related to higher education performance besides objective expenditures, the ARWU ranking only tries to assess the research quality of a university as measured by its scientific production and the excellence of its students and alumni. The ranking should only be judged under that basic rationale. It is our goal in this paper to find out whether the ranking provides any meaningful information about the quality of the research at the higher education system level, we think we have succeeded in that endeavor.

In a recent paper, Dehon et al. (2010) tried to uncover excellence in ARWU, and found out that, for the majority of institutions, after the effect of a number of outliers was removed by robust principal component analysis, the ranking appeared to reflect two different and apparently uncorrelated aspects of academic research: overall research output and top-notch researchers. In this paper we will see that when the information of the indicators is aggregated at the country level, the indicators of the ranking do constitute a reliable one-dimensional scale, and judging by the loads of the first principal component, it is not difficult to conclude that the scale is in fact measuring the research quality of the different university systems. When the second principal component is added, the quantity of the scientific production does clearly enter the picture, making room for a more refined interpretation of the ranking results.

The outline of the paper goes as follows. First we provide a quick and simple summary of the ARWU ranking and its methodologies. We then propose a method to aggregate measures in order to analyze the performance of whole university systems, and carry out a preliminary exploratory analysis by assessing the reliability of the so constructed scale, and the suitability of the data for the analysis. After that we perform a principal component analysis, and examine the percentage of variance explained by the first principal component to conclude that we are before a truly one-dimensional scale, with the loads of the main component pointing to the indicators related to individual and collective research quality. We therefore take the value of the first principal component as a measure of the quality of a university system. We then study the data under the light from the first and second principal components taken together, and cluster the 32 higher education systems under analysis. Finally, we present our conclusions and some future directions for our work.

ARWU ranking

ARWU ranks universities by means of the use of several indicators of academic performance, including alumni and staff winning Nobel Prizes or Fields Medals, number of highly cited researchers within each institution, articles published in Nature and Science, articles indexed in major citation indices, and the per capita academic performance of each institution.

For each indicator, the highest scoring institution is assigned the maximum value of 100, and the scores of the remaining institutions are then normalized as a percentage of the top one. All the indicators are then allocated a weight of 20% except Alumni and PCP, which receive a 10% weight. For a complete explanation of the indicators we refer to Liu and Cheng (2005) and Dehon et al. (2010). A short summary is provided next:

– Individual indicators

Alumni	Total number of graduates from an institution winning Nobel Prizes in the sciences or Fields Medals in Mathematics.
Award	Total number of the staff working at an institution at the time of winning Nobel prizes in the sciences, or Fields Medals in Mathematics.
HiCi	Total number of highly cited researchers in broad subject categories found at the web site of the Institute of Scientific Information.[2]

– Collective indicators

N&S	Total number of articles published in *Science* and *Nature* in the past five years.
PUB	Total number of articles indexed by Science Citation Index-Expanded and Social Science Citation Index in the previous year.
PCP	Total scores of the previous five indicators divided by the number of full-time equivalent academic staff.[3]

The ARWU ranking data thus rely on the history of universities in the past and current centuries (indicators Alumni and Award), in the last ten to twenty years, reflected in the number of Highly Cited Authors, and in the previous five years, as measured by the indicator N&S. It also measures the current performance in quantity of publications by means of the indicator PUB.

Because of the narrow selection of the quality journals, *Nature* and *Science*, and the overwhelming majority of Nobel prizes in the natural sciences and Medicine (The Nobel prize in Economics is only four decades old), ARWU is clearly biased towards the research performance in the natural and life sciences. However, insofar as the quality of the research in the natural and life sciences of its universities could be taken as a reasonable proxy of the research quality of a higher education system, the results of this paper will be meaningful. Given the academic diversity of universities around the world, it is arguable that just research in natural and life sciences could always be taken as a good proxy for the overall research quality of a single institution, but once the scores from a number of universities are aggregated we can safely assume that those scientific outputs will by and large be highly correlated with academic performance of a country's university system in all disciplines.

Country scores

In order to make the comparisons needed for the paper, it is thus necessary to find a procedure for aggregating scores at a country's university system level. We think that a

[2] http://www.isihighlycited.com.

[3] When the number of academic staff for the institution is not known, ARWU uses the weighted total scores of the other five indicators.

reasonable criterion is to average scores across a country-size related number of institutions from each university system, a number that should enable us to make sensible comparisons among countries of very different sizes.

It is well known that research results are linked to the development of countries and their economic capacity, so we decided to analyze the performance of university systems in relation with the economic size of the country they belong to, as measured by its share of the World's GDP.

A key issue is how to choose the proper number of universities from each country in such a way that comparisons among different university systems are fair and meaningful. Fair comparisons among country's university systems mean that for two countries with the same GDP share we should average the same number of universities. For countries with different size we should work the proportions in relation with our economic yardstick: the GDP world's share of the country. Meaningful comparisons mean that we should round up a large enough multiple of the country's share of World's GDP, but not too large as to exclude too many countries from the analysis. By choosing 3 times the GDP share we can accommodate the majority of the countries with universities in the ranking. However, we should make sure that the resulting scores are decoupled from the GDP data, proving therefore that the criterion does not carry any bias into the analysis. In "Exploratory Analysis of the Data" section we will carry out an appropriate ANOVA to show how the impact of country sizes on the aggregated scores for university systems is not statistically significant.

Countries are included in the sample for further analysis when they fulfil the following conditions:

1. The country does show at least two universities in the ranking.
2. The number of universities from a country in the ranking does not fall below 75% of three times its share of the World's GDP.

Small countries are forced to enter with at least two universities. For some university systems (Brazil, China, Poland, Greece, Chile, Portugal, Hungary, Singapore) the number of institutions subject to averaging coincides exactly with the number of universities from those countries present in ARWU. Those are the best institutions from their country's higher education system. Hence, in order to make fair comparisons among different university systems, to compute the aggregates we should select for the rest of the countries their first institutions according to the ranking. When institutions are not precisely ranked in ARWU, we compute their scores using the ARWU weights and sub-rank them within the countries accordingly. We have made an exception in the case of Belgium that should appear with three universities, but since its first four universities in ARWU are so close together, taking the four of them into account provides smoother average scores for Belgium's higher education system. For China we should have averaged 21 universities, but only 18 are included in ARWU: we believe it is better to include China in the study. Poland should be represented by three universities, but only two found room in ARWU. We have included Poland, though its results would be a little worse should a third university be accounted for.

Insofar as the PCP indicator is only computed correctly in the case of a short number of countries, it is clear that it cannot be used in a study of aggregate effects by country university systems. In this we are in agreement with Dehon et al. (2010) who also left this indicator out of their analysis.

Table 1 Share of World's GDP, number of institutions used to compute university system scores, and averaged system scores on the five indicators

Country	Acronym	GdP	NoU	AlU	AwD	HiCi	NaS	PuB
Australia	AUS	1.66	5	15.24	8.16	22.56	19.44	50.92
Austria	AUT	0.68	2	15.15	7.75	8.80	13.10	35.00
Belgium	BEL	0.83	4	9.70	12.00	17.10	13.55	44.25
Brazil	BRA	2.19	6	0.00	0.00	4.15	7.08	42.68
Canada	CAN	2.46	8	13.60	7.14	23.81	20.85	55.00
Switzerland	CHE	0.82	3	22.80	26.73	27.80	29.90	45.90
Chile	CHL	0.28	2	4.75	0.00	3.65	6.60	28.65
China	CHN	7.10	18	0.68	0.00	0.41	5.81	43.18
Germany	DEU	6.03	18	18.11	12.21	15.13	16.89	39.06
Denmark	DNK	0.56	2	20.10	21.55	16.75	24.40	51.05
Spain	ESP	2.63	8	2.16	0.00	5.23	9.18	37.39
Finland	FIN	0.45	2	8.20	8.95	17.20	14.20	41.95
France	FRA	4.71	14	17.02	13.09	10.71	16.79	35.94
United Kingdom	GBR	4.40	13	27.60	27.39	29.97	29.50	52.63
Greece	GRC	0.59	2	0.00	0.00	9.95	4.25	39.75
Hong Kong SAR	HKG	0.35	2	0.00	0.00	14.50	7.55	43.15
Hungary	HUN	0.26	2	8.65	7.75	8.80	6.90	23.20
Ireland	IRL	0.44	2	6.70	7.05	7.30	8.25	31.65
Israel	ISR	0.33	2	22.65	21.55	19.30	17.45	41.20
Italy	ITA	3.80	12	6.58	4.55	11.84	9.03	39.53
Japan	JPN	8.06	24	6.20	3.46	14.57	14.10	38.55
Korea	KOR	1.53	5	0.00	0.00	2.92	10.12	47.32
Netherlands	NDL	1.44	5	11.28	11.28	23.64	20.40	47.82
Norway	NOR	0.74	3	12.97	11.13	11.80	16.30	38.93
New Zealand	NZL	0.21	2	7.75	0.00	10.30	12.15	34.80
Poland	POL	0.87	2	15.55	0.00	3.65	5.70	32.35
Portugal	PRT	0.40	2	0.00	4.45	0.00	5.60	29.75
Singapore	SGP	0.30	2	0.00	0.00	7.25	9.75	50.60
Sweden	SWE	0.79	4	24.97	29.70	20.13	18.73	44.33
Taiwan	TWN	0.64	2	6.70	0.00	10.30	6.00	53.45
United States	USA	23.71	71	21.07	23.12	40.06	33.56	51.94
South Africa	ZAF	0.45	2	21.90	0.00	5.15	11.80	30.20

Table 1 shows the list of the higher education systems that will be subjected to statistical analysis, along with the following data:

GdP: Percentage of World's GDP in current US$, IMF(2009).
NoU: Number of averaged universities to produce the system scores.
AlU: Averaged score of a university system on the indicator Alumni.
AwD: Averaged score on the indicator Award.
HiCi: Averaged score on the indicator HiCi.
NaS: Averaged score on the indicator N&S.
PuB: Averaged score on the indicator PUB.

Exploratory analysis of the data

A collection of scores on five variables from averages taken over 32 higher education systems are now available for further analysis. Well aware of the results from Dehon et al. (2010), who found an outstanding number of outliers in ARWU at the institutional level, we first conducted an exploratory analysis of the aggregated data in search of possible outliers. The results from SPSS indicate the presence of just one individual measure that could be deemed an outlier: the USA in the HiCi indicator. Removing the USA from the analysis is clearly not an option, since more than 30% of the universities in ARWU are American institutions. By keeping it, we are aware of the risk of introducing a slight bias. However, there are two facts that support the retention of the USA in the analysis. On one hand, the trimmed mean (12.72) of the HiCi indicator (i.e. the mean computed after removing the top and bottom 5% of the cases) and the actual mean (13.27) are not very different, so extreme scores (including the outlier) are not having a strong influence on the mean. On the second hand, and more importantly, when we searched for multidimensional outliers, using the Mahalanobis distance, we found none. The maximum value of the Mahalanobis distance in the sample (13.65) does not correspond to the United States but to South Africa, and when we compare it with the critical statistical value (see Tabachnick and Fidell 2007, Table C4) for five variables at the 0.01 level, 20.52, we realize that the maximum value lies well on the safe side, confirming the absence of multidimensional outliers in the sample. We thus decided to retain the whole sample for further analysis.

To check for a possible bias due to differences in size of the countries with presence in ARWU, a one-way between-groups analysis of variance was conducted to explore the impact of country sizes on levels of research performance on their university systems, as measured by the ARWU aggregated indicators. Countries were divided into three groups according to their size. Group 1 (10 countries): world's GDP share in excess of 2.0. Group 2 (11 countries): world's GDP share between 0.5 and 2.0. Group 3 (11 countries): world's GDP share below 0.5. The Levene's test confirms that we have not violated the homogeneity of variance assumption in any of the indicators. There was no statistically significant difference at the $p = 0.5$ level in the scores for the three groups. AlU: $F(2, 29) = 0.59$, $p = 0.943$; AwD: $F(2, 29) = 0.260$, $p = 0.773$; HiCi: $F(2, 29) = 0.149$, $p = 0.862$; NaS: $F(2, 29) = 0.220$, $p = 0.804$; PuB: $F(2, 29) = 1.19$, $p = 0.343$. Post-hoc comparisons using the Tukey HSD test (Hsu 1996) indicated that mean scores for the three groups and all the indicators were not statistically different. Indicator(p): AlU(0.942), AwD(0.760), HiCi(0.850), NaS(0.784), PuB(0.372), confirming therefore that the methodology used did not introduce any bias due to differences in size of the countries with presence in ARWU.

Collectively, the scores can be taken as five measures from a scale; hence, the question of how the variables group together arises naturally. Specifically, we are interested in the dimensionality of the so constructed scale, we want to know whether the five variables can be meaningfully combined into a single measure on which the countries significantly differ, and can explain a great deal of the variance from the original scores. We are bound to use an empirical statistical method, since we have not been able to formulate our hypotheses beforehand and consequently design our experiment. Principal Component Analysis (PCA) is the most effective and most widely used statistical technique for an empirical study of dimensional reduction. PCA uses the correlation among the variables to develop a small set of components that summarize the correlation among the variables (Tabachnick and Fidell 2007). PCA provides a description rather than a theoretical explanation, but it can help in understanding our data and can serve as a starting point for further conceptual analysis.

Table 2 ARWU 2009 Pearson correlation matrix

Indicator	AwD	HiCi	NaS	PuB
AlU	0.812**	0.649**	0.773**	0.195
	AwD	0.759**	0.839**	0.347*
		HiCi	0.893**	0.605**
			NaS	0.551**

(** and *): correlation significant at the 0.01 and 0.05 levels (1-tailed), respectively

Prior to the computation of the principal components, the suitability of the data for exploratory factor analysis was assessed. Table 2 shows the correlation matrix (with indication of statistical significance of the coefficients) for the higher education systems in Table 1.

Inspection of the correlation structure of the five aggregated scores from Table 2 revealed that the majority of the correlation coefficients were highly significant (at the 0.01 level 1-tailed), one of them was significant (at the 0.05 level 1-tailed), and only one was not statistically significant (the one corresponding to the pair AlU and PuB). The Kaiser-Meyer-Olkin value was 0.81 far exceeding the recommended value of 0.6 (Kaiser 1974). The Bartlett's Test of Sphericity (Bartlett 1954) reached statistical significance. Both results stand in support of the factorability of the correlation matrix.

The ARWU scale

Much in the same way as Principal Components and Factor Analysis are used as psychometric tools in Psychology, i.e. in the development of tests for measuring of personality traits or intelligence, we will use PCA in this paper to elucidate the following question regarding the ARWU based aggregated scores: whether there is a single coherent aggregate measure that accounts for a percentage of the population variance to such an extent as to be accepted as a measure of a unique underlying phenomenon that meaningfully accounts for the correlations among the variables. The specific goal of our PCA will thus be to try and summarize patterns of correlations among the measured variables and reduce them to one or possibly two components accordingly. Further investigation will be needed to asses the validity of the construct resulting from the interpretation of the component(s) arising from PCA. At this stage of our research our questions will be related to the number and nature of the principal components, their importance (as far as the explanation of variance is concerned), their possible interpretation, and the scores the subjects of the analysis receive on the resulting component(s).

It is worth pointing out that conceptually a scale is an instrument that makes use of a number of items to measure a single construct. The items within the scale should be interchangeable, in such a way that they can be taken as different ways to enquiry about the same theoretical property or, as it is our case, a characteristic difficult to measure directly. In a well constructed and meaningful scale, the magnitude of the correlations between its items should be relatively large, and that happens to be the case of the ARWU scale. Furthermore, the scores on the items of the scale should be aggregated in some way to produce the actual measurement, which is precisely the way the authors of the ranking chose to compute institutional scores. Hence, we can consider the five ARWU indicators selected at the aggregate level of countries as entries to a scale that attempts to measure an underlying variable, one that we associate with the research performance of a university system as a whole.

Table 3 Item-total statistics

Indicator	Corrected item-total correlation	Cronbach's Alpha if item deleted
AlU	0.71	0.89
AwD	0.82	0.86
HiCi	0.87	0.85
NaS	0.94	0.84
PuB	0.46	0.94

We are then trying to summarize the information provided by a set of indicators into a single measure that combines them all in such a way that a significant part of the variance gets statistically explained. That is precisely the rationale behind Principal Component Analysis, just a technique to reduce the dimensionality of a data set in search of the dimensions (underlying constructs) that account for most of the variance of our data set.

Reliability of the scale

To test the validity of a scale one should be able to answer the question of whether or not the scale properly represents the theoretical construct it is meant to measure. To do that we would need a different measure of the same theoretical construct to make the appropriate comparisons, something we do not have in this case. What we can do is to measure the reliability of the ARWU scale. Reliability is tantamount to precision (small variance) or consistency of measurement; in other words, reliability is based on the overall proportion of true score variance to total observed variance (Brown 2006). Taking into account that only a single set of measurements is presented for our analysis, we have to rely upon the computation of the Cronbach's alfa (Tabachnick and Fidell 2007). Using that parameter we will be measuring the internal consistency reliability, which concerns the homogeneity of the items comprising the scale (DeVellis 2003).

In the sample, the Cronbach's coefficient was 0.90, so we can conclude that the scale shows good internal consistency. Given that the variance of the PuB indicator is not adequately explained, we decided to also take a look at the information in the Item-Total Statistics, as shown in Table 3.

Again, the results from Table 3 point to a relatively weak, although noticeable, correlation between the indicator PuB ant the total score (i.e. the score computed by just adding all the items in the scale). The last entry to Table 3 provides the most valuable information: PuB is the only ARWU indicator that causes the Cronbach's Alpha of the scale to increase if the indicator is removed, a clear sign that PuB does not belong in the same pack as the other indicators.

Finally, we arrive at the same conclusion when we compute the adjusted R^2 for each item when it is predicted by means of a multidimensional stepwise linear regression model, using the rest of the items in the scale as predictors. R^2 measures the proportion of variance from a variable that gets explained by the regression model that uses the rest of the indicators as independent variables. In relation with the reliability of a scale, the larger the R^2, the more the item is contributing to the internal consistency of the scale. Table 4 shows the values of R^2 and the predictors for the five items in the scale. The results from Table 4 again point to the isolation of the indicator PuB among the five items from the ARWU scale.

Table 4 Proportion of variance (R^2) explained by the five indicators

Indicator	Predictor 1	Predictor 2	R^2
AlU	AwD		0.65
AwD	NaS	AlU	0.75
HiCi	NaS		0.79
NaS	HiCi	AlU	0.85
PuB	HiCi		0.34

Table 5 Loadings of the five indicators on the first principal component

Loadings		Communalities	
Indicator	Loading	Initial	Extraction
AlU	0.84	1.0	0.70
AwD	0.92	1.0	0.84
HiCi	0.93	1.0	0.86
NaS	0.96	1.0	0.92
PuB	0.57	1.0	0.33

ARWU as a one-dimensional scale

To asses the properties of the scale constructed with the five ARWU indicators from Table 1, the data from the 32 countries were examined for Principal Component Analysis using SPSS. Because no multidimensional outliers were detected, there was no need to employ robust principal components. Since the measures are reasonably commensurable, Morrison (2000) suggests that performing the component analysis on the covariance matrix is preferable for statistical reasons.

Principal Component Analysis on the covariance matrix revealed the presence of just one component with eigenvalue exceeding 1, explaining a great deal of the variance in the sample, almost 73%.

The loadings on this component, as shown in Table 5, are particularly relevant for all the variables, except for the indicator related to the scientific throughput (PuB). The loadings are related to the degree of correlation between each indicator and the component. The communalities explained by this first factor, also shown in Table 5, clearly reveal that unless the case of the PuB indicator, the variance of the other 4 indicators is adequately accounted for by the first principal component.

We are then entitled to conclude that the ARWU ranking could be taken, at the level of countries, as a one-dimensional scale that attempts to assess the research quality of their university system, as measured by the indicators of individual (AlU, AwD, HiCi) and collective (NaS) quality. It favors indicators of individual excellence, and does not particularly value the efforts to indiscriminately increase the research throughput.

Finally, Table 6 shows the scores of the 32 university systems under analysis on the first principal component, split by the median of the distribution of the scores. Although we have already mentioned that no other measure is available to test the validity of the scale, we think that the data shown in Table 6 enable us to say that our results are not in contradiction with the conventional wisdom around the research quality of the university systems around the world.

Table 6 Scores of higher education systems on the first principal component: columns above and below the median

Above median		Below median	
Country	Score	Country	Score
USA	2.28	ITA	−0.44
GBR	2.17	TWN	−0.47
CHE	1.85	ZAF	−0.48
SWE	1.43	NZL	−0.60
DNK	1.21	HKG	−0.65
ISR	1.00	IRL	−0.66
CAN	0.78	SGP	−0.69
NDL	0.71	HUN	−0.72
AUS	0.71	POL	−0.82
DEU	0.41	KOR	−0.87
BEL	0.24	ESP	−0.94
FRA	0.21	GRC	−0.94
NOR	0.12	BRA	−0.99
FIN	0.09	CHN	−1.12
AUT	−0.18	CHL	−1.14
JAP	−0.28	PRT	−1.25

ARWU as a two-dimensional scale

We have seen how the bulk of the scientific production of a university system does not exert a significant influence on the first component, since its variance remains largely unexplained by it. Besides, the second largest eigenvalue, 0.9, is to close to the threshold to be discarded without further inspection of the data. To arrive at a decision we have to select an appropriate stopping rule for the number of components. Although we are aware that the Scree test (Catell 1966) is only a graphical substitute for a significance test (Jackson 2003), we made use of it to see what the plot reveals about the slope of different segments connecting the five eigenvalues, shown in Fig. 1. After a quick inspection it is not difficult to conclude that eigenvalues 3, 4 and 5 do constitute the rubble at the bottom of the cliff, revealing a clear break after the second eigenvalue. The second principal component was therefore retained for further investigation.

Table 7 shows the proportion of variance of each indicator accounted for by the two components, that now jointly explain more than 90% of the variance in our sample. Entries to Table 7 enable us to connect the second principal component to the PuB indicator as indicated by the percentage of its variance that gets explained by that component. The explanation of variance also points to a clear association of the second component with the indicator AlU, and to a lesser extent with the indicator AwD.

Without rotation, the raw principal components loadings do not facilitate an easy interpretation of the axes; Varimax rotation (to preserve orthogonality of the components) was therefore performed. The loadings (only the significant ones) of the five indicators on the two rotated components are shown in Table 8.

In Fig. 2 we have represented the scores of the 32 higher education systems on the two principal components as produced by SPSS.

The rotated solution does not separate the set of indicators in a clear fashion, as expected given the large amount of correlation among the first four indicators. However,

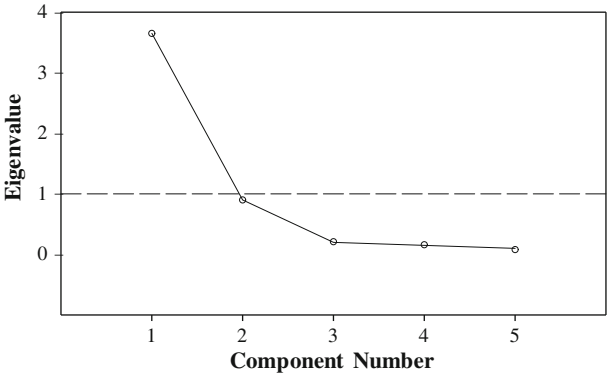

Fig. 1 Scree Plot

Table 7 Proportion of variance explained by the two first principal components

Indicator	Princ. comp.		Unexplained variation
	1	2	
AlU	0.70	0.19	0.11
AwD	0.84	0.06	0.10
HiCi	0.86	0.04	0.11
NaS	0.92	0.00	0.08
PuB	0.33	0.60	0.07

Table 8 Loadings on the two principal components

Indicator	Component 1	Component 2
AlU	0.94	
AwD	0.92	
HiCi	0.71	0.63
NaS	0.82	0.50
PuB		0.96
% of variance explained	58.35	32.40

taking into account that indicators HiCi and NaS appear to be neutral (they load on both components), we have a useful interpretation of the principal components in terms of research quality (related to all the indicators, but greatly to Nas and HiCi), individual excellence (linked to the first component) and research quantity (associated with the second component). The interpretation of the figure has been simplified by the rotation: going just East is tantamount to scoring in AlU and AwD. Going North-East is tantamount of High Quality. Going just North means scoring in PuB. We have labeled the figure in accordance with the interpretation of the two rotated components.

If we now want to cluster the subjects of our analysis, an information which is worth pursuing instead of trying to establish a precise ranking, we should take into account the fact that university systems with similar research performance should be aligned from

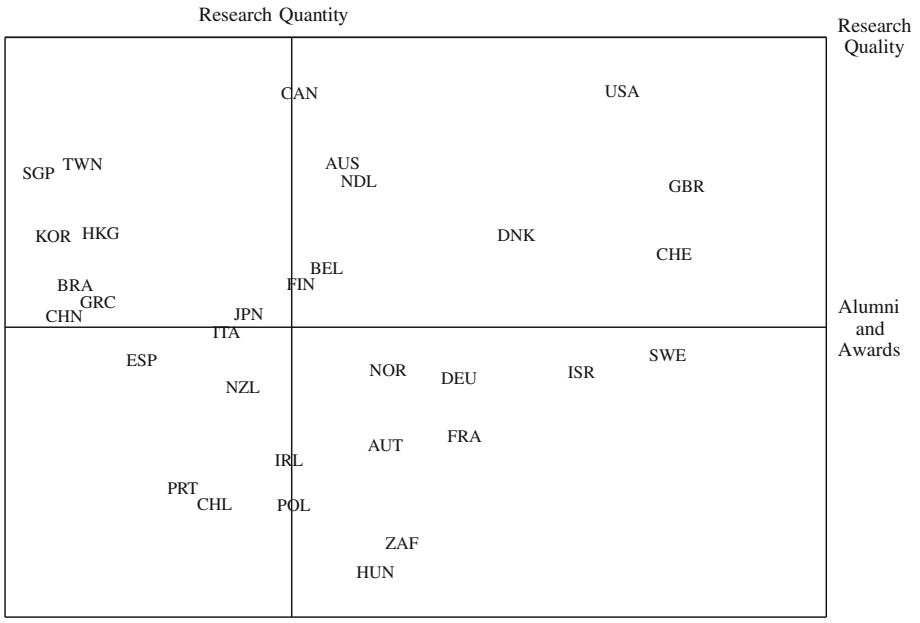

Fig. 2 Plot of the two principal components with factor interpretation

North-West to South-East, the direction orthogonal with the line of Research Quality (from South-West to North-East). Figure 3 provides the appropriate clustering, with a labeling that we deem self-explanatory.

Discussion

In spite of the many criticisms that the Shanghai's rankings have attracted, if ARWU is taken at face value, i.e. as a reasonable tool to measure the research quality of a university through some carefully selected indicators related to the quantity and quality of its scientific production and the excellence of its students and alumni, the information it provides, when properly used, allows us to gain a useful insight into the research performance of whole university systems.

A limitation of the study worth mentioning is associated with the length of the sample. The study was undertaken with a sample of 32 cases, a sample that cannot be enlarged at the country level. We know that the sample size (6.4 cases for each item) is not very large for a reliability analysis unless the items show a great deal of correlation among them. Hence, the value of communalities for the four indicators closely related to research quality (AlU, AwD, NaS and HiCi), and the large values of the correlations (as shown in Table 2) provide empirical support to the adequacy of the sample size for the analysis we have carried out. Nevertheless, the fact that the PuB indicator, i.e. the bulk of the scientific production, does show a poor communality, 0.35, and a weak correlation with some of the indicators (not even reaching the level of significance with AlU) points to a shortcoming of ARWU as a one-dimensional scale, at least at the aggregate level we are dealing with.

Another limitation of the study is the dependence of the reliability analysis on the Cronbach's alpha, a parameter that may either underestimate or overestimate scale

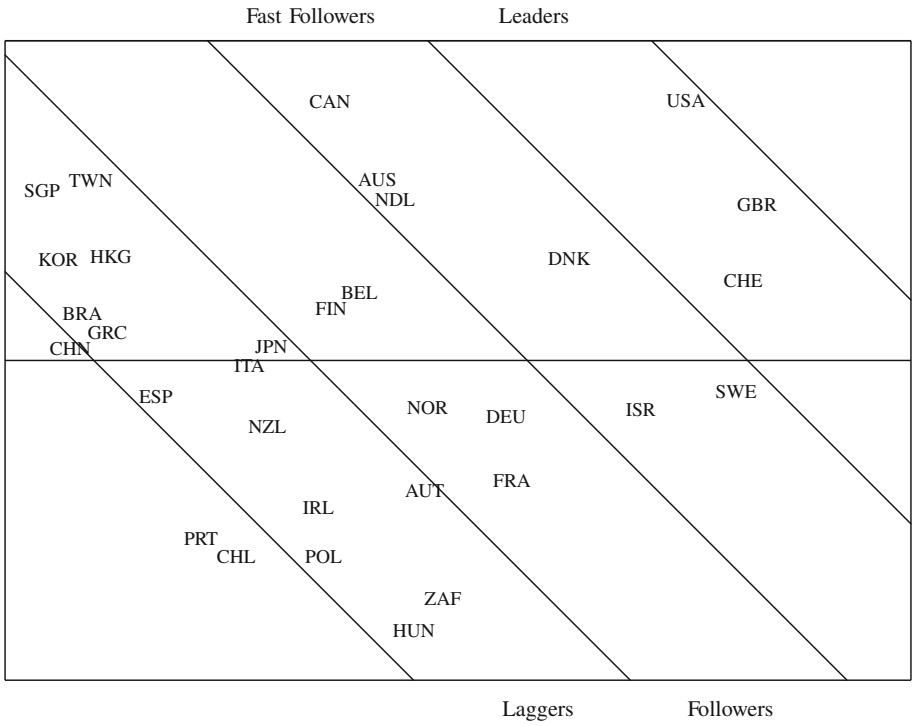

Fig. 3 Plot of the two principal components with cluster interpretation

reliability depending on the underlying measurement parameters (Brown 2006). A complete confirmatory factor analysis of the data set shall therefore be the subject of further work.

Furthermore, when we look at the loadings of the indicators on the second principal component before rotation, we realize that we may not have a reliable second component. Stevens (1996) has compiled a set of recommendations for the retention of principal components, chiefly based on the Monte Carlo study by Guadagnoli and Vellicer (1988). They recommend to retain components with four or more loadings above 0.60, as is the case of the first principal component, because they are reliable, regardless of sample size. That is not the case of the second principal component, another pointer to the opportunity of carrying out a confirmatory factor analysis on the data set.

However, we interpret the fact that only a small percentage of the variance of the PuB indicator is taken care of by the first principal component as a sign that the first dimension, in spite of the large percentage of the global variance explained, may not be sufficient to tell the whole story of ARWU. That, and a careful view of the scree plot did recommend the introduction of a second component in the analysis; as a result, the picture of ARWU is more colorful and helps in explaining the differences in performance of the higher education systems around the world.

Once the second component is introduced, the results from our analysis are not in disagreement with the ones reported by Dehon et al. (2010), in the sense that we have on one side of the second component "the research conducted at the highest level, as measured by alumni and faculty recipients of a Nobel Prize or Fields Medal" and the measured research in terms of output on the other one. As a matter of fact, the second rotated

component is associated with the indicators PuB and AlU, as Table 8 shows. The second component contributes to explain a great deal of the variance of the PuB indicator, 60%, almost twice the percentage of the variance explained by the first principal component. It also completes the explanation of the variance of the indicator AlU with an additional 20%. Besides, it complements the explanation of the variance of the indicator AwD with an additional 7%. At the aggregate level of the higher education systems, the indicators HiCi and NaS, which are clearly associated with research quality although not at the highest level, are not differentiated by the two principal components.

Conclusions

In this paper we have tried to enhance the understanding of the Shanghai ranking by suggesting some meaningful uses of the ARWU data to assess the research performance of higher university systems around the world. The analysis carried out on the aggregated data shows that, at the country's university system level, the Shanghai Jiao Tong Academic Ranking of World Universities indicators can be taken as entries to a one-dimensional scale that shows a high internal consistency.

The results from Table 6 point out the relative strength of each of the analyzed university systems in terms of the quality of their research performance. Those results are very much in agreement with the public perception of the relative position of university systems in the global research arena, and render credibility to the endeavor of the creators of the Shanghai ranking.

The picture obtained by clustering the scores on the two principal components contributes to refine the analysis and interpretation of the ARWU data. The association of the indicators with the two components enables us to perform a meaningful clustering of the subjects under analysis, one that explains the different strengths and shortcomings of university systems worldwide in relation with their research quality.

Our findings support the use of the Shanghai ranking at the aggregate level to monitor the research performance of the different university systems around the world. Future research directions for our work include a study of the evolution of the metrics along years of existence of ARWU and a reliability study based on a confirmatory factor analysis.

Methods summary

Data on academic institutions were gathered directly from the Shanghai Jiao Tong University ARWU website, http://www.arwu.org; data on World's GDP share by countries were downloaded from the International Monetary Fund website. Aggregated data for countries was computed using Excel by averaging scores over a number of country universities, as explained in "Country scores" section. ANOVA, Exploratory Analysis of the data, Reliability analysis of the ARWU scale using Cronbach's alpha and Principal Component Analysis through the covariance matrix were performed using SPSS Statistics 17.0. Data from SPSS output was transferred to the file to compose the figures that appear in the text using the epic.sty and eepic.sty packages.

Acknowledgments The author would like to thank the financial support from Xunta de Galicia through the IMAN Program. The author would also thank the anonymous reviewer for the insightful comments and suggestions which helped to improve the manuscript.

References

Bartlett, M. S. (1954). A note on the multiplying factors for various chi-square approximations. *Journal of the Royal Statistical Society, 16*(B), 296–98.

Billaut, J. C., Bouyssou, D., & Vincke, P. (2010). Should you believe in the Shangai ranking: An MCDM view. *Scientometrics, 84*(1), 237–263.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research* (3rd ed.). New York: The Guilford Press.

Catell, R. B. (1966). The scree test for number of factors. *Multivariate Behavioural Research, 1*, 245–276.

Dehon, C., McCathie, A., & Verardi, V. (2010). Uncovering excellence in academic rankings: A closer look at the Shanghai ranking. *Scientometrics, 83*(2), 515–524.

DeVellis, R. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, CA: Sage.

Docampo, D. (2008). Rankings internacionales y calidad institucional. *Revista de Educación, Número Extraordinario*, 149–176.

Florian, R. V. (2007). Irreproducibility of the results of the Shangai academic ranking of world universities. *Scientometrics, 72*(1), 25–32.

Guadagnoli, E., & Vellicer, W. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265–275.

Hsu, J. (1996). *Multiple comparisons: Theory and methods*. London: Chapman & Hall.

IMF. (2009). World Economic Outlook (WEO) Database: Downloaded from the International Monetary Fund server on November 19th 2009. http://www.imf.org/external/pubs/ft/weo/2009/02/index.htm..

Jackson, J. E. (2003). *A users guide to principal components*. Hoboken, New Jersey: Wiley.

Kaiser, H. (1974). An index of factorial simplicity. *Psychometrika, 39*, 31–36.

Liu, N. C., & Cheng, Y. (2005). Academic ranking of world universities: Methodologies and problems. *Higher Education in Europe, 30*(2), 127–136.

Marginson, S. (2005). There must be some way out of here. Tertiary Educ. Management Conference, Keynote address, Perth, Australia.

Morrison, D. (2000). *Multivariate statistical methods* (3rd ed.). New York: McGraw-Hill.

Stevens, J. (1996). Applied multivariate statistics for the social sciences. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson Education, Inc./Allyn and Bacon.

Zitt, M., & Filliatreau, G. (2007). *The world class universities and ranking: Aiming beyond status* (pp. 141–160), Romania: UNESCO-CEPES, Cluj University Press, chap Big is (made) beautiful: Some comments about the Shangai ranking of world-class universities, Part Two.