Title: "On the Necessity of Multiple University Rankings"

Authors:

1. Lefteris Angelis (corresponding author):
School of Informatics, Aristotle University of Thessaloniki,
54124 Thessaloniki, Greece
email: lef@csd.auth.gr
phone: +302310998230

2. Nick Bassiliades:
School of Informatics, Aristotle University of Thessaloniki,
54124 Thessaloniki, Greece
email: nbassili@csd.auth.gr

3. Yannis Manolopoulos:
Faculty of Pure and Applied Sciences, Open University of Cyprus,
2220 Nicosia, Cyprus
email: yannis.manolopoulos@ouc.ac.cy

On the Necessity of Multiple University Rankings

Abstract

Nowadays university rankings are ubiquitous commodities; a plethora of them is published every year by private enterprises, state authorities and universities. University rankings are very popular to governments, journalists, university administrations and families as well. At the same time, they are heavily criticized as being very subjective and contradictory to each other. University rankings have been studied with respect to political, educational and data management aspects. In this paper, we focus on a specific research question regarding the alignment of some well-known such rankings, ultimately targeting to investigate the usefulness of the variety of all these rankings. First, we describe in detail the methodology to collect and homogenize the data and, second, we statistically analyze these data to examine the correlation among the different rankings. The results show that despite their statistically significant correlation, there are many cases of high divergence and instability, which can be reduced by ordered categorization. Our conclusion is that if, in principle, someone accepts the reliability of university rankings, the necessity and the usefulness, of all of them is questionable since only few of them could be sufficient representatives of the whole set. The overabundance of university rankings is especially conspicuous for the top universities.

Keywords: University rankings; Data acquisition; Statistical Analysis; Score categorization

1 Introduction and Motivation

Higher education has witnessed a massification during the last decades. More and more universities appear worldwide, more students seek for university education, more academic staff is necessary. Thus, between universities there is an increasing competition to attract a larger portion of human and other resources. In this context, university rankings play an important role as higher education institutes try to improve their position and appear higher in these lists.

University rankings exist for many years, mainly originating from United States back at the beginning of the 20th century [19]. However, during the last decade, there are plenty of such rankings; some are processed by private enterprises, whereas others by university research centers or by even national research institutions. The Wikipedia lemma on "university rankings" is extensive and contains a list of 24 such "global" rankings[1] as well as another lengthy list of regional and national rankings.

Out of this list, a few ones are well-known from newspapers, mass media, social/market actors and so on. Among them, we notice the following ones (in alphabetical order), which will be more-or-less mentioned later:

- ARWU (Shanghai)[2],
- CWTS (Leiden)[3],
- EduRoute[4],
- HEEACT / NTU (Taiwan)[5],

1 https://en.wikipedia.org/wiki/College_and_university_rankings
2 www.shanghairanking.com
3 www.leidenranking.com
4 http://www.eduroute.info/

- QS (Quacquarelli Symonds)[6],
- THE (Times)[7], split after 2009 from QS,
- USNWR (US News)[8],
- Webometrics (WR)[9].

A "paradox" phenomenon has been described in the literature [6], [37]; although these rankings are widely used and mentioned (they are here to stay [28]), at the same time they are heavily criticized for a number of reasons [9], [32]. In particular:

- they are not statistically robust [8], [26], [25],
- they are not reproducible as they are based on questionnaires to a great extent [34],
- they are not stable but show inconsistent fluctuations from year to year [27],
- they are not objective but select arbitrarily what/how they measure [31], [34],
- they use shallow proxies as correlates of quality [2], [29],
- they use different data sources (Web of Science or Scopus) [24],
- they depend on the citation counting method, which is an open bibliometric issue that may impact rankings [17], [18], [21],
- they depend on Impact Factor, which is a wrong metric to evaluate research performance [28],
- they favor universities of English-speaking countries [13], [33] focusing in hard science [26],
- they propagate errors which appear at indexing services [16], [30],
- they tweak their results to show movement and attract commercial interest [34],
- they compare apples-to-oranges (e.g. teaching vs. research institutions) as they use an ''one-size-fits-all'' approach, which leads to controversial results [31],
- they ignore the teaching dimension [6].

In this paper, we do not consider any methodological or policy issues about university rankings. Instead, we provide a "side by side" study of six rankings from the above list in order to infer on their simultaneous usefulness. The remainder of this paper is organized as follows. The next section includes a literature review on papers related to university rankings. Section 3 describes the methodology of the dataset collection and pre-processing. Section 4 describes the numerical transformations for homogenizing the data while Section 5 contains the main findings of our statistical analysis. Section 6 discusses the statistical results whereas the final Section 7 concludes the paper.

2 Related Work

In this section, we briefly review papers which deal with university rankings. The topic is largely discussed both by a world-wide audience and in the relevant scientific literature.

In most of the studies, ARWU is compared to some other rankings, most frequently with THE and/or QS, since these are supposed to be the most "prominent" ones. For example, ARWU and THE have been compared across various dimensions in [2], [5], [11], [23], [26], [27], ARWU, THE, QS and EduRoute in [37], ARWU, HEEACT and QS in [10], ARWU, QS, NTU and CWTS in [24], ARWU, THE and QS in [35], ARWU, THE and CWTS in [4], [17], ARWU, THE-QS, WR, HEEACT and CWTS [1].

From another point of view, we can separate the above references in two categories: the ones with a qualitative approach and those with a quantitative approach. Since here we provide a statistical comparison of

5 http://nturanking.lis.ntu.edu.tw/
6 www.topuniversities.com
7 www.timeshighereducation.co.uk
8 www.usnews.com/education/best-global-universities
9 www.webometrics.info

six rankings from the above list, we focus in the second category – i.e. the references with some technical merit and an experimental part by using any particular dataset – and present them in chronological order.

The study by [11] examines ARWU and THE, and states that they are not reliable (at the time of their research, 2007) due to non-transparency, no normalization and errors. For example, according to ARWU, credits are given to the university where a Nobel winner is affiliated although his pioneering work may have been done in another university. Aguilo et al. [1] have made a comparison of ARWU, THE, WR, HEEACT and CWTS by using a set of similarity metrics. Their findings demonstrate that there are reasonable similarities between them. In particular, the largest similarity has been observed between HEEACT and CWTS, whereas the smallest similarity has been observed between QS-THE and WR.

The work of Huang [10] reports a high-level comparison of ARWU, HEEACT and QS. Results are reported per country and differences in the ranking lists are mentioned. It is mentioned that QS favors British universities. In [17] the dataset are the publications of the period 2005-2007 and citations of the year 2009 of 7 Korean research universities; fractional counting according to co-authorship and normalization is applied. It is shown that these factors affect the ranking of the universities.

The study of [26] examines ARWU and THE and performs a robustness analysis to test the validity of the rankings. Interestingly, they conclude that: "apart from the top 10 universities, neither ARWU nor the THE should be used to compare the performance of individual universities". The above study goes on proposing a new framework to alleviate the influence of methodological choices (e.g. special weights). ARWU, THE and QS are examined in [35]; the conclusion is the same with that of reference [10]: British universities are favored in THE and QS rankings.

The authors of [24] present a national ranking system for Spain and compare the results with those of ARWU, QS, NTU and CWTS. It is stated that such comparisons should be used with caution due to the different methodologies, which need special interpretation. The study of [27] examines ARWU and THE aiming at investigating their reliability. It is interesting that they find inconsistent fluctuations from year to year with respect to the rankings of universities below position 50 (especially in THE). By using the ARWU and THE rankings, the study of [23] compares the performance of Chinese and Indian universities. It is concluded that worldwide, in general, Chinese universities perform second better (after USA and before UK), whereas Indian universities take the 9^{th} position.

Additionally, a study on ARWU without comparing it to other rankings is the work by Docampo [7], who certifies the credibility of the particular ranking system by matching the public perception to the outcome of a principal component analysis. In a recent study, ARWU is compared with a novel PageRank-based methodology to rank universities by examining links and entries about universities from 24 Wikipedia language editions. The finding is that for the top-100 universities of the Wikipedia ranking coincide 60% to the ARWU ranking [15]. A recent study performs a ranking of universities by using LinkedIn data about career paths of over 400 million professional from around the world [12]. Finally, another recent study examines country-specific factors that affect ARWU, QS and THE rankings, and reports that the position of universities in any ranking is determined by several political, financial and governmental variables [22].

In this work, we mainly focus on the association among the different University rankings instead of dealing with the evaluation criteria themselves that lead to the rankings. Our main purpose is not to compare the ranking lists and decide on the optimal ones, but rather to find their alignment especially for the "usually" top-ranked Universities. Our work is quite related to [1] in the sense that similarity and alignment are synonymous notions. However, our work presents a framework focused on the categorization of Universities in groups and the aggregation of scores. This framework helps us to locate and understand better the diversities between rankings.

3 Collecting University Rankings

To compare rankings so that safe conclusions about their correlation are drawn, data from their websites must be collected. In this section, we present the methodology of data collection from the various university ranking websites using a Prolog application we have developed, called URank [3]. URank:

a. extracts data from the ranking list sites using web data extraction techniques;
b. uniquely identifies the university entities within the ranking lists by linking them to the equivalent University entities in DBpedia;
c. constructs a complete dataset by merging the different ranking lists using Universities' DBpedia URIs as a primary key.

Table 1 contains the university ranking lists used in this study, performed during academic year 2015-2016[10]. We have chosen these 6 ranking lists, since they are considered by media as the most "prominent" ones, something that has been verified by counting the web search results from Google for each of the ranking lists mentioned in Section 1. Search result counts are reported in **Table 2**. Several technical challenges must be faced in order to collect the data from the sites of the ranking list web sites. Initially, data must be extracted by scraping the HTML pages of the ranking list sites, since there are no downloadable data. In this case, we have used the DEiXTo web data extraction tool [14], which is based on the W3C DOM. With DEiXTo highly accurate "extraction rules" (wrappers) can be created; these rules dictate what information to extract from a site. We utilized the graphical user interface of DEiXTo in order to build, test, fine-tune, save and modify the extraction rules, which were used for the wrapper component of URank to extract data from the ranking list sites at run-time.

Table 1. University ranking lists for data collection.

Ranking list	Acronym	Collected universities
Academic Ranking of World Universities	ARWU	500
CWTS Leiden Ranking	Leiden	842
Quacquarelli Symonds	QS	600
Times Higher Education	THE	800
U.S. News Best Global Universities	USNWR	500
Ranking Web of Universities	Webometrics	600

Table 2. Popularity of ranking lists counted by web search results.

Ranking list	Google Search Result count
US News ranking	798,000,000
THE ranking	374,000,000
ARWU | Shanghai ranking	64,400,000
QS ranking	9,010,000
CWTS | Leiden ranking	3,920,000
Webometrics	1,150,000
NTU | HEAACT ranking	519,000
EduRoute ranking	2,070

Another problem with data acquisition is the heterogeneity of the data schemata of the various websites. This heterogeneity problem has been resolved by a small OWL ontology we developed that provides a common schema for all ranked universities, regardless the ranking list site. Furthermore, we have used

10 For the CWTS ranking we have used the Impact indicator P (number of publications).

Prolog to customize the extraction rules in order to map the data from each site into this common schema, using site-specific transformations. Each extracted data set is in RDF format, so it can be published as Linked Open Data (LOD), individually.

There are two more challenges that depend on each other. More specifically, in order to merge different ranking lists into a single table a unique identifier for the universities is needed across all ranking lists. This task is difficult, since different ranking lists usually use different names for the same University. For example, in the QS list (**Table 1**) Imperial College[11] is mentioned as "Imperial College London", while in the ARWU list it is referred to as "The Imperial College of Science, Technology and Medicine". In order to find a unique key for each University to safely use it across datasets, we used should DBpedia[12] as a source of unique immutable identifiers for university entities. DBpedia is a community-run project that extracts structured information from Wikipedia and makes this information available on the Web as linked open data. Linking entities extracted from each ranking dataset to DBpedia serves two goals: (a) to link the data extracted in the first step with a popular linked open dataset, and (b) to use the DBpedia URI as a unique primary key across datasets to allow dataset merging.

Linking entities to DBpedia is also a difficult task. DBpedia and Wikipedia contain crowd-sourced data, which are sometimes inaccurate or incomplete. There are cases where a DBpedia university entity is wrongly classified under a higher level class of the DBpedia ontology (e.g. *owl*:*Thing*) instead of the relevant classes University or Educational Institution. Furthermore, there may exist at different places on earth universities with very similar names, such as Newcastle University[13] in the UK and University of Newcastle[14] in Australia. Another case is university splits or mergers along time; historical names of Universities still appear in Wikipedia and DBpedia. For example, University of Paris[15] was split in 1970 into 13 universities which have very similar normative names, such as "University of Paris I, II, …".

These issues cannot possibly be resolved with 100% accuracy using general purpose entity linking software (e.g. DBpedia Spotlight [20], SILK [36]). Domain-dependent knowledge on university names, geospatial reasoning and temporal reasoning must be deployed to disambiguate university entities in DBpedia. Furthermore, sometimes DBpedia does not contain up-to-date information because of revised Wikipedia articles; these late updates can be found at DBpedia Live[16] instead. Finally, when DBpedia cannot disambiguate an entity in satisfactory manner, our tool uses Wikipedia text search and web extraction techniques to locate better candidate entities, starting from Wikipedia lemmas and then moving to the corresponding DBpedia entities.

The architecture and data flow of our extraction tool (called URank) is shown in **Figure 1**. The main components are:

a. *Entity Extractor*: extracts university entities from the ranking list sites;
b. *Entity Linker*: links extracted university entities to DBpedia entities;
c. *Entity Merger*: merges all the ranking list datasets into a single dataset by creating a single entity for each university using its DBpedia URI as a primary key.

The *Entity Extractor* component uses extraction rules for each ranking list site defined by human users through the GUI of the DeiXTo tool [14]. Extraction rules are exported from DeiXTo into XML files fed to the Entity Extractor. Various libraries of SWI-Prolog [38] are used, involving e.g. XML, XPath and HTTP, to extract data from the web sites using a sophisticated algorithm in Prolog. The extraction rules

11 http://www3.imperial.ac.uk
12 http://dbpedia.org
13 http://www.ncl.ac.uk
14 http://www.newcastle.edu.au
15 http://en.wikipedia.org/wiki/University_of_Paris
16 http://wiki.dbpedia.org/DBpediaLive

are quite different for each site. From each site we collect names of universities, their global rank and their country; the latter is used for name disambiguation purposes. Additionally, we extract the URL that contains details about each university; sometimes the data transformation component needs to access the latter for disambiguation purposes, as well. Then, our tool applies website-specific transformations in order to clear and homogenize the extracted raw data and to generate the LOD datasets for each ranking list, in RDF format. These transformations are mostly about converting retrieved country data into proper country names, common across all ranking list sites. A common schema for all ranking list datasets is ensured through a lightweight university ranking ontology[17] we have developed that consists of the following classes: *RankingOrganization* and *RankedInstitution*. The first one has as instances the ranking lists, while the latter has as instances the university entities extracted from each ranking list site.

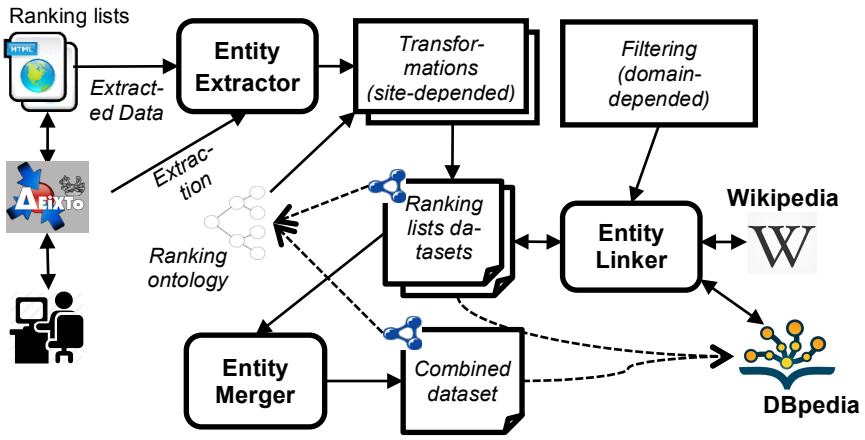

Figure 1. Architecture and dataflow of data collection.

The *Entity Linker* component links the dataset entities to DBpedia via a matching algorithm that consists of two main loops, for each ranking list and for each university entry. Inside the latter, three approaches (steps) are used in order to retrieve matching DBpedia entries. At each step substring matching is used to find a satisfactory match. If it is found, the algorithm terminates immediately; otherwise matching DBpedia entries are collected into a candidate set, scored according to a heuristic scoring function we have developed, and finally, the candidate with the highest score is returned. The two main methods for retrieving DBpedia entities is: (a) the DBpedia lookup service[18], and (b) DBpedia's SPARQL endpoint[19], using a template query we have derived from the Faceted Browser and Search & Find Service[20]. When these two methods cannot retrieve a match with a high score, then the keyword search engine of Wikipedia is used to retrieve Wikipedia lemmas as candidates for alternative (and possibly better) names for the universities.

In all the above methods, the retrieved entities are filtered using spatiotemporal domain-dependent constraints. For example, the DBpedia university entity retrieved by any of the search methods must be located in the same country as the one retrieved from the ranking list site. Furthermore, the University must still be in operation. Checking for location/country is sometimes difficult because DBpedia entries do not always keep that information, simple because the corresponding Wikipedia lemma is incomplete. For example, instead of a country-related property, a DBpedia entry might have City- or State-related information (USA and Spanish universities, mainly). In these cases, we use SPARQL queries to find in which

17 http://lpis.csd.auth.gr/ontologies/2013/university-rankings.owl
18 http://wiki.dbpedia.org/lookup
19 http://dbpedia.org/sparql
20 http://dbpedia.org/fct

country a city or a State is located. Moreover, sometimes the country name may differ from the country information extracted from the ranking list site. For example, the country property of *dbpedia*:*Harvard_University* is "U.S.", while the country data extracted from the ARWU list is "USA". To resolve this, we have created a compatibility matrix for country names. There are more, less important domain-specific filtering criteria to consider, such as using Roman or Arabic numbers in university names (e.g. "University of Montpellier II" vs. "Montpellier 2 University") or using translations of the word "university" in other languages (e.g. "University of Freiburg" vs. "Universität Freiburg").

Finally, *Entity Merger* generates a single dataset by merging together the datasets of the six ranking lists. This dataset contains all the universities with all corresponding rankings from each ranking list in a single entity. The merged dataset has RDF properties for the ranks of each ranking list. Merging is achieved using a straightforward algorithm, which creates a new university instance in the merged graph and copies its property values, for each ranking dataset RDF graph and for each university instance inside each dataset.

The algorithm checks whether a university instance already exists in the merged graph before creating a new one; in the former case, only the corresponding rank property is copied. The check is based on the *owl*:*sameAs* property; this is actually a link to the DBpedia university entry, discovered by the *Entity Linker*. Thus, it is really important to ensure a unique DBpedia URI for each university. This is not always the case, because sometimes there are many Wikipedia lemmas for the same topic, which are being redirected to a single Wikipedia page. This multiplicity of lemmas is also mapped to corresponding DBpedia entries; for the same real-world entity there can exist several DBpedia instances that (possibly) re-direct to a single DBpedia entry. The *Entity Linker* component ensures using a pointer-chasing algorithm that when a DBpedia URI is matched to a university name, the component will return the URI at the end of the chain of the re-direction links. Sometimes there are many instances with the above property; there are many different paths with re-direction links that lead to several different instances for the same real-world entity. This is checked by a recursive algorithm that follows the re-direction links until it finds instances that do not re-direct anywhere. In this case, our tool selects the most "informative" entity, i.e. the RDF instance which is the subject of the most RDF triples. Another problematic case is when the re-direction subgraph is cyclic; this is very rare, and it is usually temporary until the next DBpedia update. Nevertheless, this case is also handled using a closed set search. Using URank, we have managed to retrieve 1121 distinct universities, ranked along the 6 ranking lists of **Table 1**.

4 Methodology for Homogenizing the Data

For the statistical analysis, the retrieved raw dataset consisted of 1121 cases (*Universities* with their name), and 7 variables: the *Country* of the university (nominal variable with 78 values) and 6 more variables (*THE*, *ARWU*, *Webometrics*, *QS*, *CWTS* and *USNWR*) containing the rankings of the universities in the different systems. The first encountered problem concerned the values of the ranking variables, which do not use the same representations. More specifically, the variables contain mixed type values for their rankings:

- ***THE*:** 1-200 (numerical values explicitly), 201-250 (category), 251-300 (category), 301-350 (category), 351-400 (category), 401-500 (category), 501-600 (category), 601-800 (category), >800 (category).
- ***ARWU***: 1-100 (numerical values explicitly), 101-150 (category), 151-200 (category), 201-300 (category), 301-400 (category), 401-500 (category), >500 (category).
- ***Webometrics***: 1-599 (numerical values explicitly), >599 (category).
- ***QS***: 1-400 (numerical values explicitly), 401-410 (category), 411-420 (category), ..., 491-500 (category), 501-550 (category), 551-600 (category), >600 (category).
- ***CWTS*:** 1-842 (numerical values explicitly), >842 (category).
- ***USNWR***: 1-494 (numerical values explicitly), >494 (category).

Given that statistical analysis requires the values of each variable to be of the same type, we followed two approaches, which were subsequently compared.

The first approach involved transformation of the original values to numerical scores. Specifically, the following transformations were used, separately for each ranking:

- All ranking values represented as explicit numerical values were not transformed.
- All categories representing bounded intervals were transformed to numerical values represented by the mid-point of the interval. For example, the interval 201-250 was transformed to the value 225.5 while the interval 301-400 was transformed to the value 350.5.
- All categories representing intervals with only the lower bound, thus declaring that a university does not appear in the first predefined (separately for each ranking) positions, were transformed arbitrarily to the midpoint between the lower bound and the number of total universities in the list. For example, for the *THE* ranking system, a university categorized as >800 appears in the list of 1121 due to its explicit or interval position in another list of the dataset. For all these universities, comprising the "scree" of *THE* ranking, we decided to assign the value 961, which is the midpoint between 801 and 1121. Similarly, the category >500 of the *ARWU* system receives the value 811, the category >599 of *Webometrics* is transformed to 860.5, the category >600 of *QS* to 861, the category >842 of *CWTS* to 982 and the category >494 of *USNWR* to 808. Note that this approach gives a "bonus" to some universities with ranking >1121 in any ranking system. However, it is their appearance in the list of 1121 that is "rewarded" in a reasonable sense.

The analysis of the numerically transformed dataset involved descriptive statistics. We first analyzed the distribution of the variable *Country* (**Table 3** and **Figure 2**). Most of the universities in the dataset (19%) are from USA and next (9.8%) from China. Almost half of the universities of our dataset (49.6%) belong to 6 countries: USA, China, UK, Germany, Japan and France. Overall, 78 countries appear in the dataset.

Table 3. Distribution of countries.

Country	Frequency	Percent	Cumulative Percent	Country	Frequency	Percent	Cumulative Percent
USA	213	19.0	19.0	Saudi Arabia	6	.5	93.6
China	110	9.8	28.8	Colombia	5	.4	94.0
UK	82	7.3	36.1	Kazakhstan	4	.4	94.4
Germany	54	4.8	40.9	Mexico	4	.4	94.7
Japan	49	4.4	45.3	Norway	4	.4	95.1
France	48	4.3	49.6	Romania	4	.4	95.5
Italy	46	4.1	53.7	Ukraine	4	.4	95.8
South Korea	35	3.1	56.8	United Arab Emirates	4	.4	96.2
Spain	35	3.1	59.9	Indonesia	3	.3	96.4
Australia	32	2.9	62.8	Singapore	3	.3	96.7
Canada	31	2.8	65.6	Estonia	2	.2	96.9
Taiwan	27	2.4	68.0	Jordan	2	.2	97.1
Brazil	24	2.1	70.1	Lebanon	2	.2	97.2
India	23	2.1	72.2	Pakistan	2	.2	97.4
Turkey	18	1.6	73.8	Philippines	2	.2	97.6
Russia	16	1.4	75.2	Serbia	2	.2	97.8
Iran	14	1.2	76.4	Slovakia	2	.2	97.9
Poland	14	1.2	77.7	Slovenia	2	.2	98.1
Netherlands	13	1.2	78.9	Bahrain	1	.1	98.2
Sweden	12	1.1	79.9	Bangladesh	1	.1	98.3
Austria	10	.9	80.8	Belarus	1	.1	98.4
Finland	10	.9	81.7	Costa Rica	1	.1	98.5
Switzerland	10	.9	82.6	Croatia	1	.1	98.6
Argentina	9	.8	83.4	Cyprus	1	.1	98.7
Belgium	9	.8	84.2	Ghana	1	.1	98.8
Czech Republic	9	.8	85.0	Iceland	1	.1	98.8
Republic of Ireland	9	.8	85.8	Kenya	1	.1	98.9
Greece	8	.7	86.5	Latvia	1	.1	99.0
New Zealand	8	.7	87.2	Lithuania	1	.1	99.1
Thailand	8	.7	88.0	Luxembourg	1	.1	99.2
Israel	7	.6	88.6	Macau	1	.1	99.3
Portugal	7	.6	89.2	Morocco	1	.1	99.4
South Africa	7	.6	89.8	Nigeria	1	.1	99.5
Chile	6	.5	90.4	Oman	1	.1	99.6
Denmark	6	.5	90.9	Peru	1	.1	99.6
Egypt	6	.5	91.4	Puerto Rico	1	.1	99.7
Hong Kong	6	.5	92.0	Qatar	1	.1	99.8
Hungary	6	.5	92.5	Tunisia	1	.1	99.9
Malaysia	6	.5	93.0	Uganda	1	.1	100.0
Total	*1121*	*100.0*					

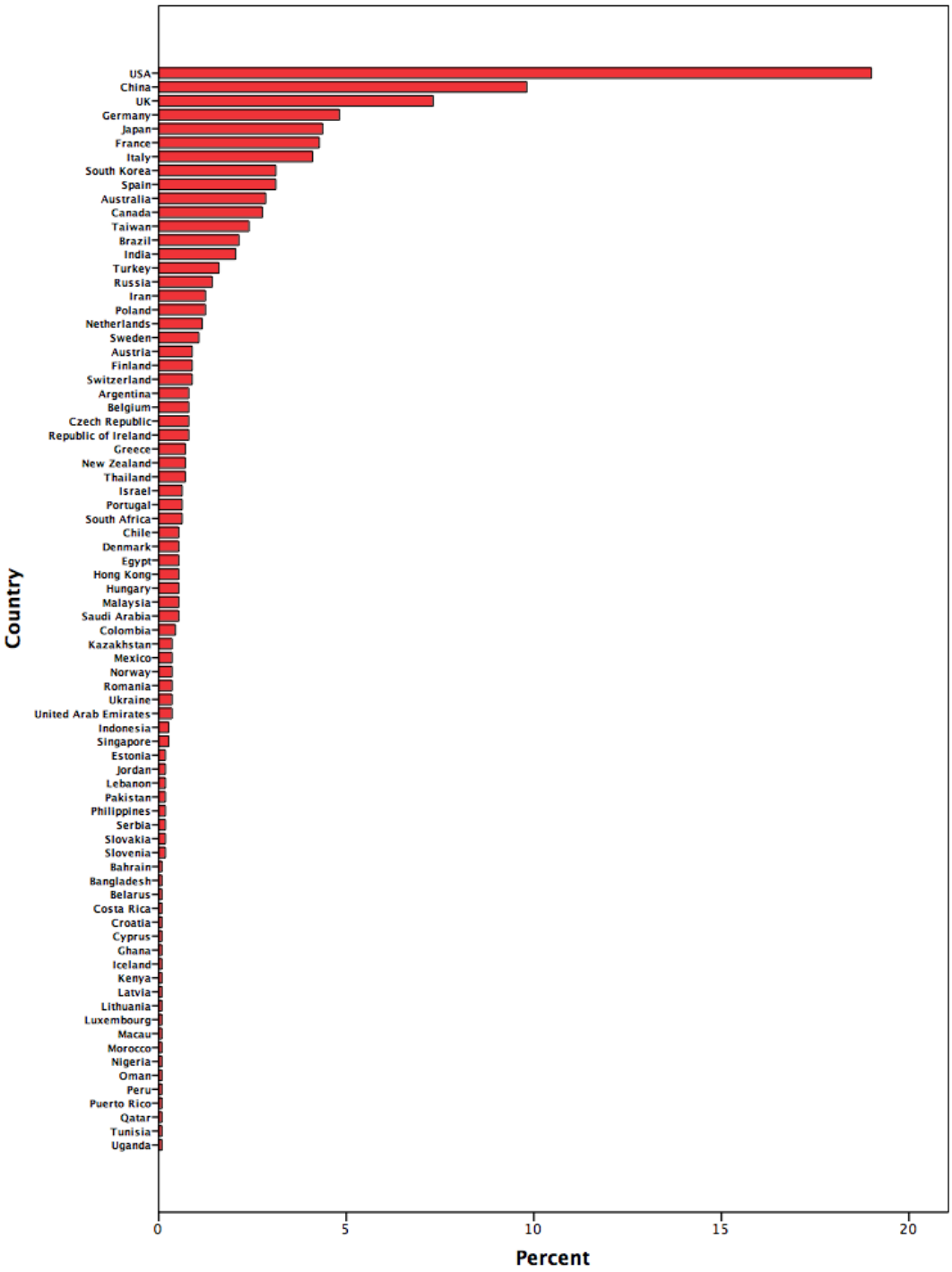

Figure 2. Bar chart of the distribution of countries.

5 Ranking by Aggregated Scores and Statistics

5.1 Statistical Aggregation of the Homogenized Data

Since one of our goals was to examine the deviation among the rankings after the homogenization described above, an overall score was computed for each university, based on the average of all its numerical rankings (*Av_score*). A small score indicates a university ranked in the first positions. According to this score, in the first position we find *Harvard University* with *Av_score*=2.17.

Along with this average score, we calculated the standard deviation of the 6 rankings for each university (*SdDev_score*). A small deviation score indicates agreement of rankings, whereas a large one indicates disagreement of rankings. However, to use a comparable metric of deviation among universities, we used the standardization of deviation known as the Coefficient of Variation (*CV_score* = *SdDev_score/Av_svore*).

Table 4 presents the list of the first 50 universities according to the *Av_score* in ascending order, showing also *SdDev_score* and *CV_score*. We notice that *California Institute of Technology* has large *CV_score*=1.73, since it is 1st in the *THE* ranking, 8th in the *ARWU* ranking, 35th in *Webometrics*, 5th in *QS*, 162nd in *CWTS* and 7th in *USNWR*. On the other hand, *Duke* is quite "stable" with respect to all rankings (20th, 25th, 24th, 24th, 26th and 20th respectively); this is reflected by its low *CV_score*=0.11.

The distribution of *CV_score* is presented in the histogram of **Figure 3** and is obviously skewed. This means that there are universities with high variability of their rankings across different systems. To depict how *CV_score* is distributed with respect to the 6 rankings, we provide the scatter plots in the panels of **Figure 4**. It is clear that large values of *CV_score* are observed in the first-ranked universities of all ranking systems, which is quite interesting and unexpected. This is an indication of instability and disagreement between ranking systems even for the first-ranked universities.

The aggregated scores *Av_score* and *CV_score* can be used for comparisons between countries. For instance, we compared these two metrics for the two highly represented countries in the dataset, i.e. USA and China. The comparison was made by the non-parametric Mann-Whitney test, which gave for both metrics $p<0.001$, i.e. a statistically significant difference. The boxplots of the metrics for these two countries are depicted in **Figure 5**. We can clearly see that USA universities tend to appear in the first places according to the *AV_score*. However, on the other hand their rankings under different systems present larger variability (*CV_score*) in comparison to the Chinese universities.

Apart from the *CV_score* metric, another method to investigate the agreement among different ranking systems is the calculation of correlations between rankings. For this purpose, we calculated the non-parametric Kendall's tau correlation coefficient between all pairs of rankings. These are presented in the correlation matrix of **Table 5**. The coefficient, which takes values in the interval [-1,1] and shows either positive or negative correlation, is accompanied by the significance (p-value). A correlation is considered statistically significant if $p<0.05$. For our data, all correlations have $p<0.001$ therefore they are all statistically significant and they all are positive.

However, the strength of correlation is not always that high. For example, the strongest correlation is between *ARWU* and *USNWR* with coefficient 0.769. The weakest one is between *THE* and *CWTS* with coefficient 0.386. A scatter plot of all pairs of rankings is given in **Figure 6**. The wide spread of the swarm of points shows the low degree of alignment (correlation) between ranking systems.

The correlations between *Av_score* and all rankings are given in **Table 6**. All p-values are <0.001. We can clearly see that the ranking that is best aligned with the "average score" is USNWR with correlation coefficient equal to 0.728.

Table 4. Average score and deviation metrics for all rankings (50 first universities).

	University	Country	Av_score	SdDev_score	CV_score
1	Harvard University	USA	2.17	2.04	.94
2	Stanford University	USA	3.50	1.87	.53
3	University of Oxford	UK	7.17	3.87	.54
4	University of Cambridge	UK	8.33	6.12	.73
5	Massachusetts Institute of Technology	USA	9.50	16.01	1.69
6	University of California, Berkeley	USA	11.17	9.95	.89
7	Johns Hopkins University	USA	12.17	4.26	.35
8	Columbia University	USA	13.33	5.39	.40
9	University of California, Los Angeles	USA	15.00	8.25	.55
10	University of Pennsylvania	USA	15.00	3.46	.23
11	University of Michigan	USA	15.33	9.07	.59
12	Yale University	USA	17.00	9.90	.58
13	University College London	UK	17.00	6.90	.41
14	Cornell University	USA	17.17	7.73	.45
15	University of Toronto	Canada	19.33	10.31	.53
16	University of Washington	USA	22.33	20.04	.90
17	Duke University	USA	23.17	2.56	.11
18	University of Chicago	USA	23.33	28.92	1.24
19	University of California, San Diego	USA	25.83	11.05	.43
20	ETH Zurich	Switzerland	26.83	20.27	.76
21	Imperial College London	UK	29.67	27.99	.94
22	Northwestern University	USA	31.00	8.74	.28
23	University of Tokyo	Japan	32.00	16.86	.53
24	University of Wisconsin-Madison	USA	32.33	15.53	.48
25	University of British Columbia	Canada	32.67	7.81	.24
26	Princeton University	USA	36.00	57.77	1.60
27	California Institute of Technology	USA	36.33	62.75	1.73
28	New York University	USA	38.50	16.79	.44
29	University of Illinois at Urbana-Champaign	USA	41.00	15.49	.38
30	Tsinghua University	China	41.17	17.41	.42
31	Peking University	China	42.33	15.67	.37
32	University of Melbourne	Australia	43.00	13.24	.31
33	University of Edinburgh	UK	43.67	26.12	.60
34	National University of Singapore	Singapore	44.17	27.42	.62
35	University of North Carolina at Chapel Hill	USA	45.83	21.15	.46
36	McGill University	Canada	47.33	13.26	.28
37	University of Texas at Austin	USA	47.50	12.03	.25
38	University of Manchester	UK	52.00	19.04	.37
39	University of California, Davis	USA	53.00	21.29	.40
40	The University of Queensland	Australia	55.33	10.67	.19
41	Washington University in St Louis	USA	56.33	28.82	.51
42	King's College London	UK	58.17	32.22	.55
43	Heidelberg University	Germany	59.17	21.29	.36
44	Ohio State University	USA	59.83	28.79	.48
45	Pennsylvania State University	USA	60.00	27.91	.47
46	University of Copenhagen	Denmark	60.67	26.21	.43
47	University of Sydney	Australia	62.33	27.34	.44
48	Kyoto University	Japan	63.17	33.40	.53
49	LMU Munich	Germany	63.33	24.14	.38
50	Nanyang Technological University	Singapore	67.25	36.23	.54

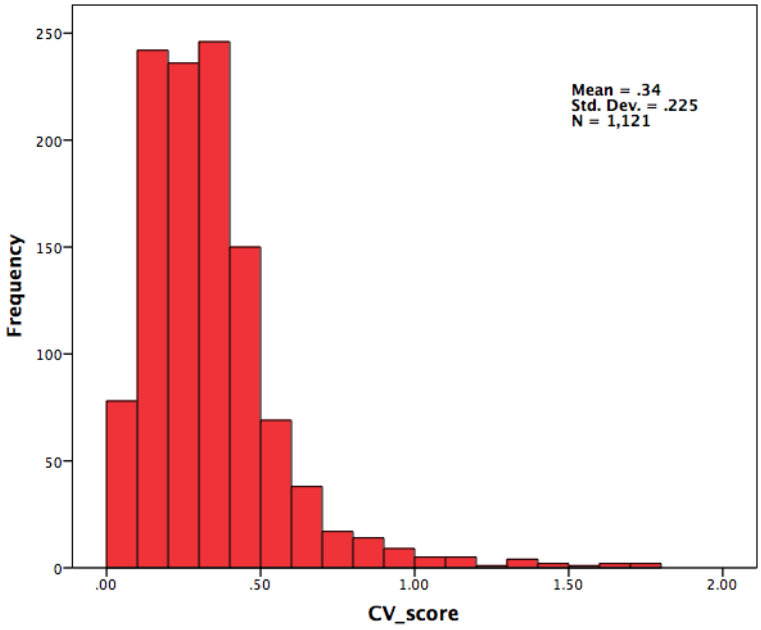

Figure 3. Distribution of CV_score for all universities.

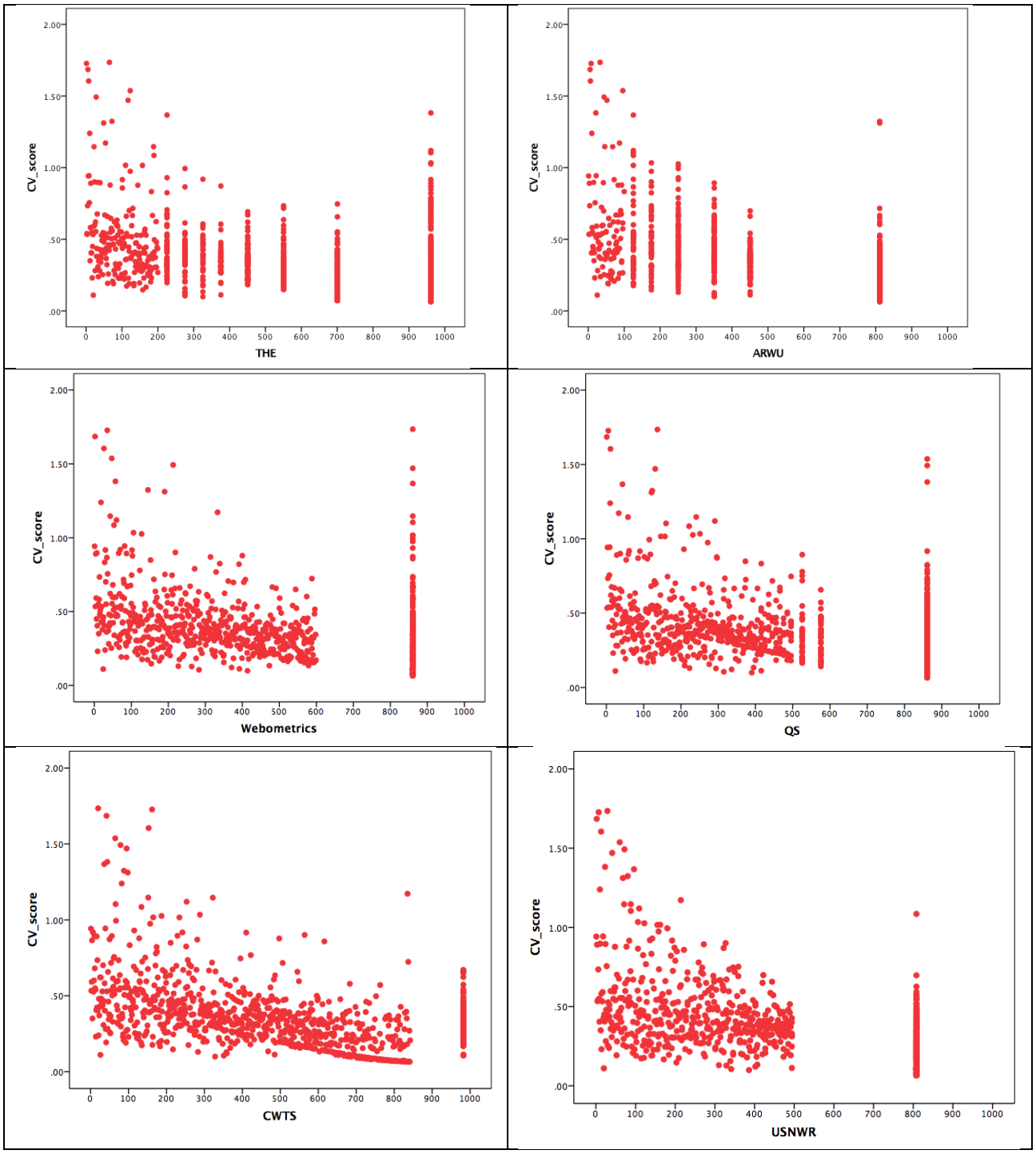

Figure 4. Distribution of CV_score with respect to all ranking.

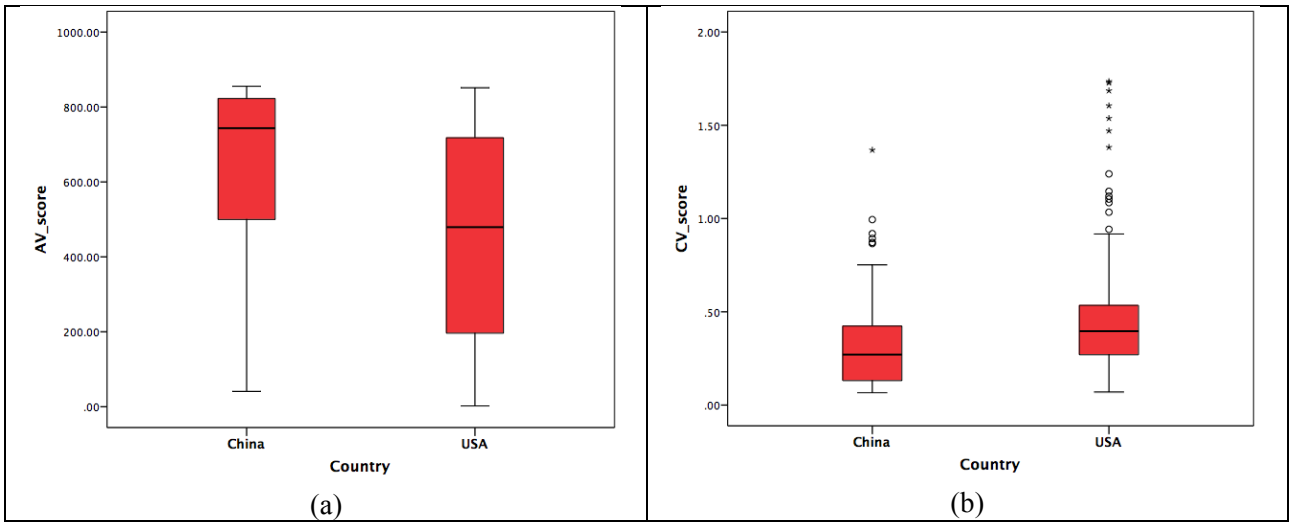

Figure 5. Comparison of (a) AV_Score and (b) CV_Score between USA and China.

Table 5. Correlation matrix for the Kendall coefficient among all pairs of rankings (first numerical transformation approach).

	THE	ARWU	Webometrics	QS	CWTS	USNWR
THE	1.000	.539**	.474**	.564**	.386**	.586**
ARWU	.539**	1.000	.612**	.568**	.646**	.769**
Webometrics	.474**	.612**	1.000	.471**	.581**	.646**
QS	.564**	.568**	.471**	1.000	.442**	.580**
CWTS	.386**	.646**	.581**	.442**	1.000	.617**
USNWR	.586**	.769**	.646**	.580**	.617**	1.000

**. Correlation is significant at the 0.01 level (2-tailed).

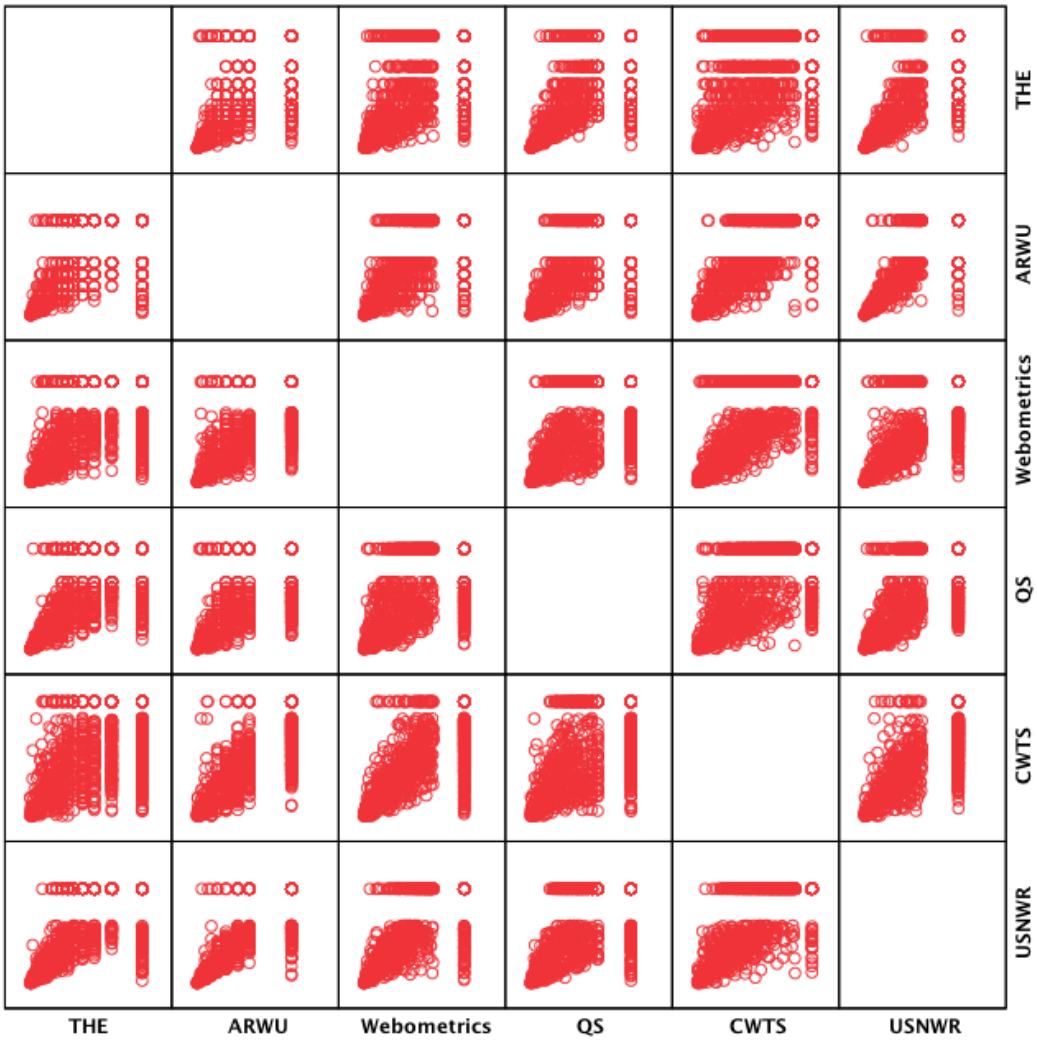

Figure 6. Scatterplot matrix of all pairs of rankings.

Table 6. Correlations of the Average Score with all rankings.

	THE	ARWU	Webometrics	QS	CWTS	USNWR
Kendall's tau_b AV_score	.629**	.718**	.677**	.632**	.650**	.728**
**. Correlation is significant at the 0.01 level (2-tailed).						

5.2 Statistical Aggregation after Ordinal Categorization

As a second approach for representing the original dataset, we recoded numerical values to ordinal values according to the following mapping for each ranking variable: 1-100 to "1", 101-200 to "2", 201-300 to "3", 301-400 to "4", 401-500 to "5" and >500 to "6". Thus, the new variables are: *THE (R)*, *ARWU (R)*, *Webometrics (R)*, *QS (R)*, *CWTS (R)* and *USNWR (R),* where (R) stands for "recoded".

Then we calculated the *CV_score (R)* for the new values. The distribution of this metric for the new values is shown in **Figure 7**. If we compare it with **Figure 3**, we can clearly see that the deviation values have been reduced significantly. This reduction can be concluded from the Wilcoxon test for paired samples between *CV_score* and *CV_score (R)*. The test gives $p<0.001$ and the reduction can be clearly seen in the boxplots of **Figure 8**.

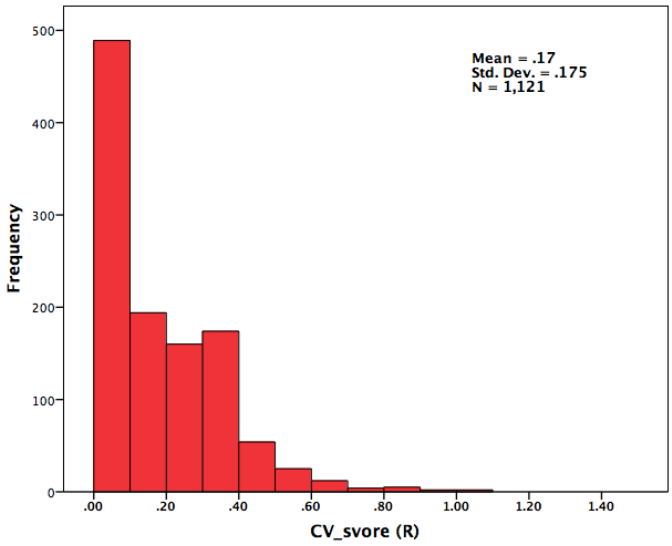

Figure 7. Distribution of CV_score for all universities. (R) means that the ordinal values of the second approach were used.

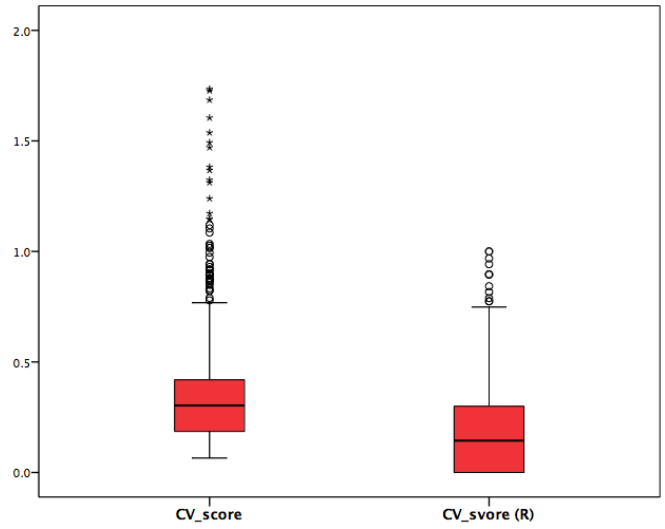

Figure 8. Comparison between CV_score computed from the first numerical approach and the second ordinal approach (R).

We also repeated the correlation analysis between all pairs of the new rankings. The new Kendall coefficients appear in the correlation matrix of **Table 7**. Comparing it with **Table 5** we can see that the ordering approach provides stronger correlations.

Table 7. Correlation matrix for the Kendall coefficient among all pairs of rankings (second ordinal transformation approach).

	THE (R)	ARWU (R)	Webometrics (R)	QS (R)	CWTS (R)	USNWR (R)
THE (R)	1.000	.617**	.546**	.645**	.502**	.663**
ARWU (R)	.617**	1.000	.644**	.590**	.716**	.776**
Webometrics (R)	.546**	.644**	1.000	.522**	.644**	.687**
QS (R)	.645**	.590**	.522**	1.000	.548**	.612**
CWTS (R)	.502**	.716**	.644**	.548**	1.000	.696**
USNWR (R)	.663**	.776**	.687**	.612**	.696**	1.000

**. Correlation is significant at the 0.01 level (2-tailed).

Finally, for this approach we applied another statistical method (appropriate for categories), the Multiple Correspondence Analysis (MCA) which treated the values 1, 2, ..., 6 of the new variables as separate categories. MCA based on the association of the variables represents their closeness in a two-dimensional space. The resulted representation is shown in **Figure 9**. We can clearly see that "1" values of all rankings form a clearly discrete, isolated and "compact" cluster. This shows the high agreement of all rankings regarding the first 100 universities. Then, there is a clear cluster of "2"s. However, the distance from the rest values is smaller and becomes even smaller as we move in the middle values "3", "4" and "5". The "6" values form also a clear cluster.

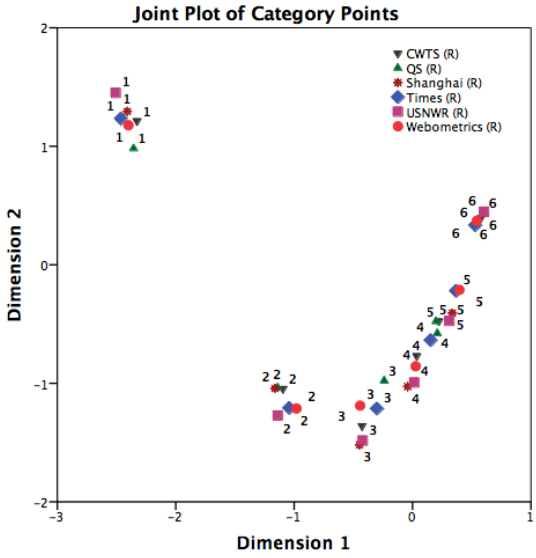

Figure 9. Result of Multiple Correspondence Analysis for the 6 rankings.

6 Discussion of Results

We chose to work with six university rankings, based on their popularity, using the information publicly announced; we extracted the ranking position of all universities appearing in at least one ranking system. The announced ranking position of each university is either an explicit number, for top universities, or a category representing a bounded or unbounded interval. Different systems define intervals in a different way. Therefore, from the beginning we pointed out an existing methodological problem, i.e. the difficulty to compare the position of a university in different rankings, if it does not belong in the set of 100-200 first universities. We believe that if we had the explicit ranking numbers of all universities in our dataset we could find larger deviations and inconsistencies, especially for universities tending to occupy last positions. In any case, the partial and "ad hoc" use of intervals is confusing and problematic for analysis and

inference. We addressed this problem by a quantification approach, assigning to each university falling in an interval, its middle value. This way we transformed the rankings to be comparable so as to study the degree of their agreement. Furthermore, we followed one more approach, the ordinal categorization of the rankings so as to define only few categories.

The transformed data of both approaches were statistically analyzed with measures and methods related to the study of correlation. Although correlations among different ranking systems were found statistically significant, their strength was not as high as one might intuitively expect. Of course, the causes behind high or low correlation strength between different rankings should be searched in the underlying criteria that are combined to produce the rankings. For example, ranking lists that use the same bibliographic data source (**Table 8**) tend to be higher correlated to each other (e.g. ARWU, CWTS and USNWR use Web of Science and their Kendall coefficient is close or higher than 0.7, whereas QS and THE that use Scopus are less correlated with e.g. CWTS). Furthermore, the total weight of the bibliographic-related criteria plays a role, since QS and THE have less than 40% weight for such criteria, whereas Webometrics that has 45% total weight for bibliographic-related criteria is more correlated to CWTS, even though they use different bibliographic databases.

Table 8. Bibliographic criteria of the ranking lists.

	Total weight of bibliographic criteria	**Source of bibliographic data**
ARWU	70%	Web of Science
CWTS	100%	Web of Science
USNWR	65%	Web of Science
QS	20%	Scopus
THE	38.5%	Scopus
Webometrics	45%	Scimago (Scopus), Google Scholar

Aggregated scores for each university are useful since not only they combine the rankings but also because they can show specific individual cases of universities with large deviations of rankings across different systems. For example, the case of California Institute of Technology shows that normalization plays a significant role, since ranking lists that signify research production per person or research impact per paper (e.g. THE, QS, USNWR, ARWU) rank Caltech in the first 10 positions, whereas ranking lists that use unnormalized criteria (e.g. Webometrics and the ranking indicator we have used in this study for CWTS), tend to rank Caltech in much lower positions.

Since correlation is a measure of what the public audience perceives as alignment or agreement of the rankings, the results of our analysis show that there are several cases where the different rankings may cause confusion. A byproduct of this statistical study is that if we were asked to choose a single representative, then this could be USNWR, as it is closer to an “average” ranking. Actually, this does not come as a surprise since USNWR uses criteria that are covered by all other ranking sites, whereas all other ranking lists have a number of unique criteria, not met in any other ranking list (**Table 9**). The CWTS ranking list criteria are also met by other ranking list; however, CWTS uses only a single dimension (for each ranking), that is why is not close to an “average” ranking.

Table 9. Unique ranking criteria per ranking list.

	Unique ranking criteria
ARWU	number of alumni / staff winning Nobel Prizes and Fields Medals
QS	Academic / Employer Reputation
THE	Doctorate-to-bachelor's ratio, Doctorates-awarded- to-academic-staff ratio, Institutional / Research / Industry income
Webometrics	Number of pages of the main web domain of the institution, Number of external networks originating backlinks to the institutions webpages
CWTS	-
USNWR	-

Another interesting finding is the effect of categorization. We have already discussed the confusing partial use of intervals in the announcement of rankings. That is why we proceeded in the ordinal categorization of all rankings and their correlation analysis. What we found is that the categorization increases the consistency and the perception of agreement, especially for the top universities.

7 Conclusions

There is a practical need for a global, well defined and widely accepted ranking system that takes into account different criteria used already in different rankings, which, however, have not been aggregated or combined so far, except for some early, superficial attempts at the level of newspapers[21]. Statistical analysis methodologies in the definition of indexes and scores can be proved beneficial to this end. New systems should satisfy certain criteria of consistency and stability.

Rankings of the type: "University X is #4, University Y is #7" and subsequent conclusions like "X is better than Y" are too arbitrary and prone to inaccuracy and instability. Explicit numerical rankings (1st, 2nd, etc.) although significantly correlated, exhibit high divergence from one ranking system to another even for "top" universities. If explicit rankings of all universities and all systems were available, it is our belief that inconsistencies would be much more striking.

Merging explicit rankings that fall in certain intervals in few ordered categories can reduce the problem of instability and the subsequent confusion. According to this perspective, the aforementioned example could be stated: "X and Y are both 1st class Universities". In this regard, an interesting debate could be based on whether the high granularity and, therefore, the instability are providing essential information about universities. From one perspective, different rankings represent different aspects of performance and therefore are useful. On the other hand, the instability can be interpreted as unnecessary confusion and noise caused by an artificial and "market-driven" need for so many ranking systems. Especially for the top universities, it seems superfluous to be evaluated by so many ranking systems. The use of few ordered categories has potentials to minimize the need and use of multiple rankings.

As a future work, we plan to study the stability of rankings across time, i.e. to test whether rankings are stable from year to year and to find out which ranking is more stable. Furthermore, we would like to integrate in our research alternative sources of information about Universities and/or national educational systems, such as research funding from governmental, international (e.g. EU) or private sources.

21 https://www.washingtonpost.com/news/grade-point/wp/2016/10/20/heres-a-new-college-ranking-based-entirely-on-other-college-rankings/?utm_term=.e8e65a3d50ac

References

[1] Aguillo, I. F., Bar-Ilan, J., Levene, M., & Ortega, J. L. (2010). Comparing university rankings. *Scientometrics*, 85(1), 243-256.

[2] Badat, S. (2010). Global rankings of universities: A perverse and present burden. Chapter in book "*Global inequalities and higher education: Whose interests are we serving*, by Unterhalter, E., & Carpentier, V., (eds.), 117-141, New York: Palgrave.

[3] Bassiliades, N. (2014). Collecting University Rankings for Comparison Using Web Extraction and Entity Linking Techniques. *Proceedings 10th International Conference on Information & Communication Technologies in Education, Research & Industrial Applications (ICTERI),* pp.23-46, Kherson, Ukraine.

[4] Buela-Casal, G., Gutiérrez-Martínez, O., Bermúdez-Sánchez, M. P., & Vadillo-Muñoz, O. (2007). Comparative study of international academic rankings of universities. *Scientometrics*, 71(3), 349-365.

[5] Da Hsuan Feng, V. (2005). World University Rankings – Generic and Intangible Feature of Universities?, *Proceedings 1st International Conference on World Class Universities*, Shanghai, China.

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

[7] Docampo, D. (2010). On using the Shanghai ranking to assess the research performance of university systems. *Scientometrics*, 86(1), 77-92.

[8] Freyer, L. (2014). Robust rankings. *Scientometrics*, 100(2), 391-406.

[9] Holmes, R.: blogspot http://rankingwatch.blogspot.gr/ Accessed: 8/12/2016

[10] Huang, M. H. (2011). A comparison of three major academic rankings for world universities: From a research evaluation perspective. *Journal of Library & Information Studies*, Vol.9(1), 1-25.

[11] Ioannidis, J. P., Patsopoulos, N. A., Kavvoura, F. K., Tatsioni, A., Evangelou, E., Kouri, I., Gontopoulos-Ioannidis, D. & Liberopoulos, G. (2007). *International ranking systems for universities and institutions: a critical appraisal*. BMC medicine, 5(1), 30.

[12] Kapur, N., Lytkin, N. I., Chen, B. C., Agarwal, D., & Perisic, I. (2016). Ranking Universities Based on Career Outcomes of Graduates. *Proceedings 22nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD),* pp.137-144, San Francisco, CA.

[13] Kivinen, O., Hedman, J., & Kaipainen, P. (2015) Reputation of universities in global rankings and top university research in six fields of research, *Proceedings 9th International Multi-Conference on Society, Cybernetics & Informatics (IMSCI)*, pp.157-161, Orlando, FL.

[14] Kokkoras, F., Ntonas, K., & Bassiliades, N. (2013). DEiXTo: a web data extraction suite. *Proceedings 6th Balkan Conference on Informatics (BCI)*, pp.9-12, Thessaloniki, Greece.

[15] Lages, J., Patt, A., & Shepelyansky, D. L. (2016). Wikipedia ranking of world universities. *The European Physical Journal B*, Vol.89(3), 69.

[16] Lee, D., Kang, J., Mitra, P., Giles, C. L., & On, B. W. (2007). Are your citations clean?. *Communications of the ACM*, Vol.50(12), 33-38.

[17] Leydesdorff, L., & Shin, J. C. (2011). How to evaluate universities in terms of their relative citation impacts: Fractional counting of citations and the normalization of differences among disciplines. *Journal of the Association of Information Science & Technology*, Vol.62(6), 1146-1155.

[18] Lin, C. S., Huang, M. H., & Chen, D. Z. (2013). The influences of counting methods on university rankings based on paper count and citation count. *Journal of Informetrics*, Vol.7(3), 611-621.

[19] Maclean Alick H. H. (1900) Where we get our best men: Some statistics showing their nationalities, counties, towns, schools, universities and other antecedents: pp.1837-1897. London: Simpkin, Marshall, Hamilton, Kent and Co.

[20] Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011). DBpedia spotlight: shedding light on the web of documents. *Proceedings 7th International Conference on Semantic Systems (I-Semantics),* pp.1-8, Graz, Austria.

[21] Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2015). Multiplicative versus fractional counting methods for co-authored publications. *Journal of Informetrics*, Vol.9(4), 974-989.

[22] Pietrucha, J. (2018). Country-specific determinants of world university rankings. *Scientometrics*, Vol.114(3), 1129-1139.

[23] Reddy, K. S., Xie, E., & Tang, Q. (2016). Higher education, high-impact research, and world university rankings: A case of India and comparison with China. *Pacific Science Review B: Humanities & Social Sciences*, Vol.2(1), 1-21.

[24] Robinson-García, N., Torres-Salinas, D., López-Cózar, E. D., & Herrera, F. (2014). An insight into the importance of national university rankings in an international context: the case of the I-UGR rankings of Spanish universities. *Scientometrics*, Vol.101(2), 1309-1324.

[25] Saisana, M., & D'Hombres, B. (2008). Higher education rankings: Robustness issues and critical assessment. JRC Scientific and Technical Reports.

[26] Saisana, M., d'Hombres, B., & Saltelli, A. (2011). Rickety numbers: Volatility of university rankings and policy implications. *Research Policy*, Vol.40(1), 165-177.

[27] Sorz, J., Wallner, B., Seidler, H., & Fieder, M. (2015). Inconsistent year-to-year fluctuations limit the conclusiveness of global higher education rankings for university management. *PeerJ*, 3, e1217. DOI 10.7717/Peerj.1217.

[28] Stergiou, K. I., & Lessenich, S. (2013). On impact factors and university rankings: from birth to boycott. *Ethics in Science & Environmental Politics*, Vol.13(2), 6-11.

[29] Thamm, M., & Mayr, P. (2011). Comparing webometric with web-independent rankings: a case study with German universities. *Proceedings 3rd ACM International Conference on Web Science (WebSci),* Koblenz, Germany.

[30] Tüür-Fröhlich, T. (2016). The non-trivial effects of trivial errors in scientific communication and evaluation. vwh Verlag Werner Hülsbusch.

[31] Usher, A., & Savino, M. (2007). A global survey of university ranking and league tables. *Higher Education in Europe*, Vol.32(1), 5-15.

[32] Van Raan, A. F. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, Vol.62(1), 133-143.

[33] Van Raan, A. F., Van Leeuwen, T. N., & Visser, M. S. (2011). Severe language effect in university rankings: particularly Germany and France are wronged in citation-based rankings. *Scientometrics*, Vol.88(2), 495-498.

[34] Vardi, M. Y. (2016). Academic rankings considered harmful. *Communications of the ACM*, Vol.59(9), 5.

[35] Vidal, P., & Filliatreau, G. (2014). Graphical comparison of world university rankings. *Higher Education Evaluation & Development*, Vol.8(1), 1-14.

[36] Volz, J., Bizer, C., Gaedke, M., & Kobilarov, G. (2009). Discovering and maintaining links on the web of data. *Proceedings 8th International Semantic Web Conference (ISWC)*, pp.650-665, Chantilly, VA.

[37] Wedlin, L. (2014). How global comparisons matter: The 'truth' of international rankings. Chapter in book "*Bibliometrics: Use and abuse in the review of research performance*" by Blockmans, W., Engwall, L. & Weaire, D. (eds.), Portland Press, London.

[38] Wielemaker, J., Schrijvers, T., Triska, M., & Lager, T. (2012). SWI-Prolog. *Theory & Practice of Logic Programming*, Vol.12(1-2), 67-96.