A peer-reviewed version of this preprint was published in PeerJ on 25 August 2015.

View the peer-reviewed version (peerj.com/articles/1217), which is the preferred citable publication unless you specifically need to cite this preprint.

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
https://doi.org/10.7717/peerj.1217

High statistical noise limits conclusiveness of ranking results as a benchmarking tool for university management

Johannes Sorz, Martin Fieder, Bernard Wallner, Horst Seidler

Regression analyses of results from the Times Higher Education (THES)-Ranking and Shanghai University's Academic Ranking of World Universities (ARWU)-Ranking from 2010-2014 show fluctuations in the rank and score for lower scoring universities (below position 50) which lead to inconsistent "up and downs" in the total results, especially in the THES-Rankings. Furthermore year-to-year results do not correspond in THES- and ARWU-Rankings for universities below rank 50. We conclude that the observed fluctuations in the THES do not correspond to actual university performance and ranking results are thus of limited conclusiveness for the university management of lower scoring universities. We suggest that THE and ARWU alter their ranking procedure insofar as universities below position 50 should be ranked summarized only in groups of 25 or 50. The year to year changes in the ARWU scores are very small, so essential changes from year to year could not be expected, so therefore we argue to publish the ranking less frequently. Additionally, we argue for introducing a standardization process for ranking data in both rankings by using common suitable reference data to create calibration curves represented by non-linearity or linearity .

High statistical noise limits conclusiveness of ranking results as a benchmarking tool for university management

Johannes Sorz[1], Martin Fieder[2], Bernard Wallner[2] and Horst Seidler[2]

[1]University of Vienna, Office of the Rectorate, Universitätsring 1, A-1010 Vienna (Austria)
[2]University of Vienna, Department for Anthropology, Althanstrasse 14, A-1090 Vienna (Austria)

Abstract

Regression analyses of results from the Times Higher Education (THES)-Ranking and Shanghai University's Academic Ranking of World Universities (ARWU)-Ranking from 2010-2014 show fluctuations in the rank and score for lower scoring universities (below position 50) which lead to inconsistent "up and downs" in the total results, especially in the THES-Rankings. Furthermore year-to-year results do not correspond in THES- and ARWU-Rankings for universities below rank 50. We conclude that the observed fluctuations in the THES do not correspond to actual university performance and ranking results are thus of limited conclusiveness for the university management of lower scoring universities. We suggest that THE and ARWU alter their ranking procedure insofar as universities below position 50 should be ranked summarized only in groups of 25 or 50. The year to year changes in the ARWU scores are very small, so essential changes from year to year could not be expected, so therefore we argue to publish the ranking less frequently. Additionally, we argue for introducing a standardization process for ranking data in both rankings by using common suitable reference data to create calibration curves represented by non-linearity or linearity.

Key words: Times Higher Education Ranking, ARWU Ranking, Shanghai Ranking, Statistical fluctuations, Regression analysis

Introduction

Global higher education rankings have received much attention recently and, as can be witnessed by the growing number of rankings being published every year, this attention is not likely to subside. Besides the arguable use of ranking results as an instrument for rational university management, it is still a common practice in many universities to use rankings as an indicator for academic performance. Rankings became a big business and as of today a plethora of regional and national rankings exist, advocated by their publishers as potentially efficient and effective means of providing needed information to universities on areas needing improvement (Dill & Soo, 2005). Numerous studies have analyzed and criticized higher education rankings and their methodologies (van Raan, 2005; Buela-Casal et al., 2007; Ioannides et al., 2007; Hazelkorn, 2007; Aguillo et al., 2010; Benito and Romera 2011; Hazelkorn 2011; Rauhvargers, 2011; Tofallis, 2011; Saisana et al. 2011; Safon, 2013; Rauhvargers, 2013; Bougnol & Dulá, 2014). This casts justified doubt on a sensible comparison of universities hailing from different higher education systems and varying in size, mission and endowment based on mono-dimensional rankings and league tables and hence on the usability of such rankings for university management and policy making (O'Connell, 2013; Hazelkorn, 2014). Several studies have demonstrated that data used to calculate ranking scores can be inconsistent. Thus, bibliometric data from international databases (Web of Science, Scopus), used in most global rankings to calculate research output indicators, favor universities from English-speaking countries and institutions with a narrow focus on highly-cited fields, which are well covered in these databases. This puts universities from non-English-speaking countries, with a focus on the arts, humanities and social sciences, at a disadvantage when being compared in global rankings (Calero-Medina et al., 2008; van Raan et al., 2011; Waltman et al., 2012). Data submitted by universities to ranking agencies (e.g. personnel data, student numbers) can be problematic to compare due to different standards. These incompatibilities are being amplified because university managers have become increasingly aware of global rankings and try to boost their performance by "tweaking" the data they submit to the ranking agencies (Spiegel Online, 2014). Beyond all the data issues, there is the effect that universities with lower positions in the rankings often encounter volatile ups and downs in their consecutive year-to-year ranks. These effects make global university rankings an inconclusive benchmarking tool for university managers: the ranking results simply do not reflect the universities' actual performance or

their management strategies. Ranking results need to be consistent to be of use, so that long-term strategies (e.g. the hiring of high-calibre researchers from abroad or the establishment of a competitive doctoral education funding scheme) are reflected in year-to-year scores and ranks. Furthermore, results from various rankings should be concordant to allow a sort of meta-analysis of rankings.

Bookstein et al. (2010) found unacceptably high year-to-year variances in the score of lower ranked universities caused by statistical noise in the Times Higher Education World University Ranking (THES), Jovanovic et al. (2012) and Docampo (2013) found a large number of fluctuations and inconsistencies in the ranks of the ARWU-Ranking. We again observed puzzling variances in the THES- and the ARWU-Rankings 2014, that were both published recently. Accordingly, we here analyze the fluctuations in score and rank of the THES- and the ARWU-Ranking by calculating regression analyses for consecutive years for 2010-2014 to determine the random component of these fluctuations. Also we calculated a regression of the ranking positions of the first 100 universities in the THES- Ranking 2014 on the first 100 universities in the ARWU-Ranking 2014.

THES-Ranking

The methodology of the THES-Ranking was revised several times in varying scale, before and after the split with Quacquarelli Symonds (QS) in 2010 and the new partnership with Thompson Reuters. Times Higher Education (THE) calculates 13 performance indicators, grouped into the five areas Teaching (30%), Research (30%), Citations (30%), Industry income (2.5%) and International outlook (7.5%). However, THE does not publish the scores of individual indicators, only those of all five areas combined. Since 2010, the research output indicators are calculated based on Web of Science data. Most of the weight in the overall score is made up by the normalized average citations per published paper (30%), and by the results of an academic reputation survey (33%) assessing teaching and research reputation and influencing the scores of both areas (Rauhvargers, 2013; THE, 2014). In the past, criticism has been levied against this survey. Academic peers can choose universities in their field from a preselected list of institutions and, although universities can be added to the list, those present on the original list are more likely to be nominated. This leads to a distribution skewed in favor of the institutions at the top of the rankings (Rauhvargers, 2011 and 2013). THE allegedly addressed this

issue by adding an exponential component to increase differentiation between institutions, yet no information is available on its mode of calculation (Baty, 2011 and 2012).

ARWU-Ranking

ARWU ranks more than 1000 (of ca. 17,000 universities in the world) and publishes the best 500 on the web. In addition ARWU offers to field rankings that cover several subjects and subject rankings for Mathematics, Physics, Chemistry, Computer Science and Economics & Business. Universities are ranked according to their research performance, including alumni (10%) and staff (20%) winning Nobel Prizes and Fields Medals, highly cited researchers in in 21 broad subject categories in the Web of Science (20%), papers published in Nature and Science (20%), papers indexed in major citation indices (20%), and the per capita academic performance of an institution (10%). Calculation of indicators remained relatively constant since 2004. ARWU ranks universities individually or into bands by sorting on the total score which is the linearly weighted sum of the six research output indicator scores derived from the corresponding raw data by transformations. Institutional data (number of academic staff) is not provided by universities but obtained from national agencies such as ministries, national bureaus and university associations (ARWU, 2013). In contrast to the THES, there are no teaching/student related indicators or any peer survey component in the ARWU-Ranking. Due to reliance on ISI subject fields the areas of natural sciences, medicine and engineering dominate the citation indicator, putting universities with a focus on the arts, humanities and social sciences. The per capita performance is the only ARWU indicator that takes into account the size of the institution, thus small but excellent institutions have less of a chance to perform well in the ARWU-Ranking (Rauhvargers, 2011). Already several studies, i.e. Docampo (2011 and 2013) analyzed the ARWU-Ranking and its indicators and found inconsistencies and unwanted dynamical effects.

Methods

We used the publicly available data on scores and ranks from the THES- and ARWU-Rankings for the years 2010, 2011, 2012, 2013 and 2014, including only those universities ranked from 1 to 200. We performed the following analysis: i) we regressed the scores of the rankings of the year t-1 on the scores of the year t; ii) we regressed the ranks of the rankings of the year t-1 on the ranks of the year t; iii) we plotted the scores in descending order iv), we determined the random component of the fluctuations in the ranks from year to year and v) finally we investigated the concordance (ranking

position of the first 100 universities) of the THES ranking with the ARWU ranking. For this purpose we regressed the position of the first 100 universities in the THES-Ranking (2014-15) on the ranking position of the first 100 university in the ARWU-Ranking (2014).

Results

THES-Ranking:

Regression of the scores and ranks of two consecutive years

The regression of the scores – particularly of the ranking 2010-2011 regressing on the scores of the ranking of 2011-2012 – shows a very high fluctuation/noise (figure 1a), especially for the lower ranked universities. Moreover, the noise among the lower ranked universities seems to be higher compared to the already very noisy THES-Ranking performed by QS before 2010 (Bookstein et al. 2010, figure 1). Note that in the rankings in the years following 2010-2011, the noise in the THES-Ranking did improve (figure 1b-d).

Association between Scores and Ranks

Nonetheless, a general problem of the THES-Ranking remains: the difference in the scores among the 50 highest scoring universities is considerably higher compared to the difference among the lower scoring universities. This clearly suggests a non-linear relationship between scores and ranks (figure 2 a-e). The consequence is that the ranks of the high scoring universities are much more robust to deviations in the scores from year to year. In the lower ranking universities, however, even very small, more or less random deviations (around 0.5%) lead to unexpected "high jumps" in the ranks from year to year (figure 1e-h).

ARWU-Ranking:

While still on a high level, the regression of ranks and scores of the ARWU-Ranking, show much less fluctuations compared with the THES. This indicates a more robust set of indicators. Furthermore the

ARWU-Ranking shows a similar, but even a more extreme pattern of non-linearity between ranks and scores. Particularly the first ranked university, Harvard, scores far ahead of all the other universities in the ARWU-Ranking at each year. As in the THES-Ranking the association between ranks and scores flattens from rank of 50 on (figure 4a-e). As in the THES ranking the non-linear relationship between ranks and scores increases the "noise" in the position shifts of the universities ranked approx. below 50 from year to year (figure 3a-d). A really dramatic amount of noise reveals the regression of the ranks in the THES ranking on the ranks in the ARWU-Ranking: for the universities ranked approximately lower the 50th rank there is virtually no correlation between the THES and the Shanghai ranking (figure 5). Regression could only be plotted for universities ranked in both rankings among the first 100).

Discussion

High ranking positions achieved by a small group of universities are often self-perpetuating, especially due to the intensive use of peer review indicators, which improve chances of maintaining a high position for universities already near the top (Bowman & Bastedo, 2011; Rauhvargers, 2011). This phenomenon also corresponds to the Matthew effect, which was coined by Merton (1968) to describe how eminent scientists will often get more credit than a comparatively unknown researcher, even if their work is similar: credit will usually be given to researchers who are already famous. The intensive and exaggerated discussion in the media of the "up and downs" of universities in the THES-Rankings is particularly misleading for the lower scoring universities (below approximately a score of 65% and a rank of 50; above scores of 65%, the relationship between ranks and scores is steeper, and it flattens for scores below 65%). This is because the ranking positions suggest substantial shifts in university performance despite only very subtle changes in score. In fact, merely random deviations must be assumed. One reason lies in the weighing of indicators by THE, with the emphasis on citations and peer review (totalling more than 65% of the total score). For lower ranked universities, a few highly cited publications, or the lack thereof, or few points asserted by peers in the reputation survey, probably make a significant difference in total score and position.

Ranking results have a major influence on the public image of universities and can even impact their claim to resources (Espeland & Saunder, 2007; Hazelkorn, 2011). Accordingly, fluctuations can have serious implications for universities, especially when the media or stakeholders interpret them as direct

results of more or less successful university management. The use of monodimensional rankings for university management is generally doubtful. Our results show, that the THES-Rankings in their current form have very limited value for the management of universities ranked below 50. This is because the described fluctuations in rank and score probably do not reflect actual performance, whereby the results cannot be used to assess the impact of long-term strategies. “Rankings are here to stay, and it is therefore worth the time and effort to get them right” warns Gilbert (2007).What could be done to address the fluctuations in the THES-Rankings for universities below rank 50 to make it a more usable tool to assess actual performance for university management? THE has already addressed fluctuations to some extent by ranking universities only down to position 200, followed by groups of 25 from 201-300 and groups of 50 from 300 to 400. Nonetheless, based on our data we believe that this is not going far enough and suggest that universities should be summarized in groups of 25 or 50 below the position of 50. The analyzed curves of scores vs. ranking positions in figure 2 do have analogous characteristics for example to semi-logarithmic curves produced in analytic biochemistry. The accuracy of such curves is limited to the steepest slope of the curve, whereas asymptote areas deliver higher fuzziness (Chan, D.W. (ed), 1992). Thus, a further suggestion to avoid the blurring dilemma is the methodological approach of introducing a standardization process for THES- Ranking data. This would involve using common suitable reference data to create calibration curves represented by non-linearity or linearity.

Comparing the year to year fluctuation in the ARWU ranking with the THES ranking reveals, that fluctuation in the ARWU ranking is overall lower as in the THES ranking (figure 1 vs. figure 3), i. e. the ARWU ranking seems to be more stable. This is on one hand a good message: a smaller amount of „noise”, but on the other hand, it has to be asked if a yearly publication of the ARWU makes sense, if no “real” changes can be expected. However the same holds true for all rankings published on a yearly basis: no factual changes reflecting university strategies can be expected,

Furthermore, the astonishing low correlation between the ranks of the THES and the ARWU ranking, particularly for the universities ranked below 50 in both rankings, creates another serious doubt if rankings should be used for any management purposes at all. Maybe a “meta- analysis” of rankings could be reasonable to derivate consistent and reliable results from rankings. If done, such and meta-analysis should include as many rankings as possible to reduce random perturbations.

References

Aguillo IF, Bar-Ilan J, Levene, M, and Ortega L.J. 2010. Comparing university rankings. Scientometrics 85: 243–256

ARWU (2013) The Academic Ranking of World Universities – Methodology (http://www.shanghairanking.com/ARWU-Methodology-2013.html, Retrieved on 23.02. 2015)

Benito M, Romera N. 2011. Improving quality assessment of composite indicators in university rankings: a case study of French and German universities of excellence. Scientometrics (2011) 89: 153–176

Baty P. 2011. "THE Global Rankings: Change for the better", Times Higher Education, 6 October. Retrieved on 11 November 2014 from: http://www.timeshighereducation.co.uk/world-universityrankings/2011-12/world ranking/methodology

Baty P. 2012. "The essential elements in our world-leading formula", Times Higher Education, 4 October. Retrieved on 11. November 2014 from: http://www.timeshighereducation.co.uk/worlduniversity-rankings/2012-13/world-ranking/methodology

Billaut JC, Bouyssou D, Vincke P. 2010. Should you believe in the Shangai ranking: An MCDM view. Scientometrics, 84: 237–263.

Bowman AM, Bastedo NM 2011. Anchoring effects in world university rankings: exploring biases in reputation scores. Higher Education 61: 431–444

Bookstein FL, Seidler H, Fieder M, Winckler G. 2010. Too much noise in the Times Higher Education rankings. Scientometrics 85: 295–299

Buela-Casal G, Gutierrez-Martinez O, Bermudes-Sanchez M, Vadillo-Munoz O. 2007. Comparative study of international academic rankings of universities. Scientometrics 71: 349-365

Bougnol ML, Dula, JH 2014. Technical pitfalls in university rankings, Higher Education, DOI 10.1007/s10734-014-9809-y

Calero-Medina C, López-Illescas C, Visser MJ. Moed HF 2008. Important factors when interpreting bibliometric rankings of world universities: an example from oncology. Research Evaluation, 17: 71–81

Chan, D.W. (ed) 1992. Immunoassay automation: a practical guide. Academic Press: San Diego

Dill DD, Soo M 2005. Academic quality, league tables, and public policy: A cross-national analysis of university ranking systems. Higher Education, 49: 495–533

Docampo D. 2011. On using the Shanghai ranking to assess the research performance of university systems. Scientometrics, 86: 77–92.

Docampo, D. 2013. Reproducibility of the Shanghai academic ranking of world universities results. Scientometrics, 94: 515–524.

Espeland WN, Sauder M. 2007. Rankings and Reactivity: How Public Measures Recreate Social Worlds, American Journal of Sociology 113: 1–40

Gilbert A. 2007. Academics strike back at spurious rankings. Nature 447: 514–515

Hazelkorn E. 2007. Impact and influence of league tables and ranking systems on higher education decision-making. Higher Education Management and Policy 19: 87–110

Hazelkorn E. 2011. Rankings and the Reshaping of Higher Education: the battle for world class excellence (Basingstoke, Palgrave-MacMillan).

Hazelkorn E. 2014. Reflections on a Decade of Global Rankings: what we've learned and outstanding issues, European Journal of Education, 49: 12–28

Ioannidis JPA, Patsopoulos NA, Kavvoura FK, Tatsioni A, Evangelou E, Kouri , et al. 2007. International ranking system for universities and institutions: A critical appraisal. BMC Medicine, 5: 30.

Jovanovic M, Jeremic V, Savic G, Bulajic M, Martic M. 2012. How does the normalization of data affect the ARWU-ranking? Scientometrics 93: 319–327

Liu NC, Cheng Y. 2005. Academic ranking of world universities: Methodologies and problems, Higher Education in Europe 30: 127–136.

Merton RK 1968. The Matthew effect in science. Science 159: 56-63

O'Connell C. 2013. Research discourses surrounding global university rankings: exploring the relationship with policy and practice recommendations, Higher Education 65: 709–723

Rauhvargers A. 2011. EUA Report on Global Rankings and their Impact – Report I (European University Association)

Rauhvargers A. 2013. EUA Report on Global Rankings and their Impact – Report II (European University Association)

Safon V 2013. What do global university rankings really measure? The search for the X factor and the X entity. Scientometrics 97: 223–244

Saisana M, d'Hombres B, Saltelli X. 2011. A Rickety numbers: Volatility of university rankings and policy implications. Research Policy 40: 165–177

Spiegel Online 2014. Deutsche Unis im "THE"-Ranking: Das Wunder von Tübingen. 02.10.2014, Retrieved on 11. November 2014: http://www.spiegel.de/unispiegel/studium/uni-ranking-hochschulenim- the-ranking-a-994684.html

Times Higher Education 2014. World University Rankings 2014-2015 methodology http://www.timeshighereducation.co.uk/world-university-rankings/2014-15/worldranking/methodology

Tofallis C. 2012. A different approach to university rankings. Higher Education 63, 1 –18

van Raan, T. 2005. Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. Scientometrics, 62(1), 133–143

van Raan T, Leeuwen T, and Visser M. 2011. Severe language effect in university rankings: particularly Germany and France are wronged in citation-based rankings. Scientometrics 88: 495–498

Waltman L, Calero-Medina C, Kosten J, Noyons NC, Tijssen R.J, and Wouters P. 2012. The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. CWTS Working Paper Series. http://arxiv.org/abs/1202.3941

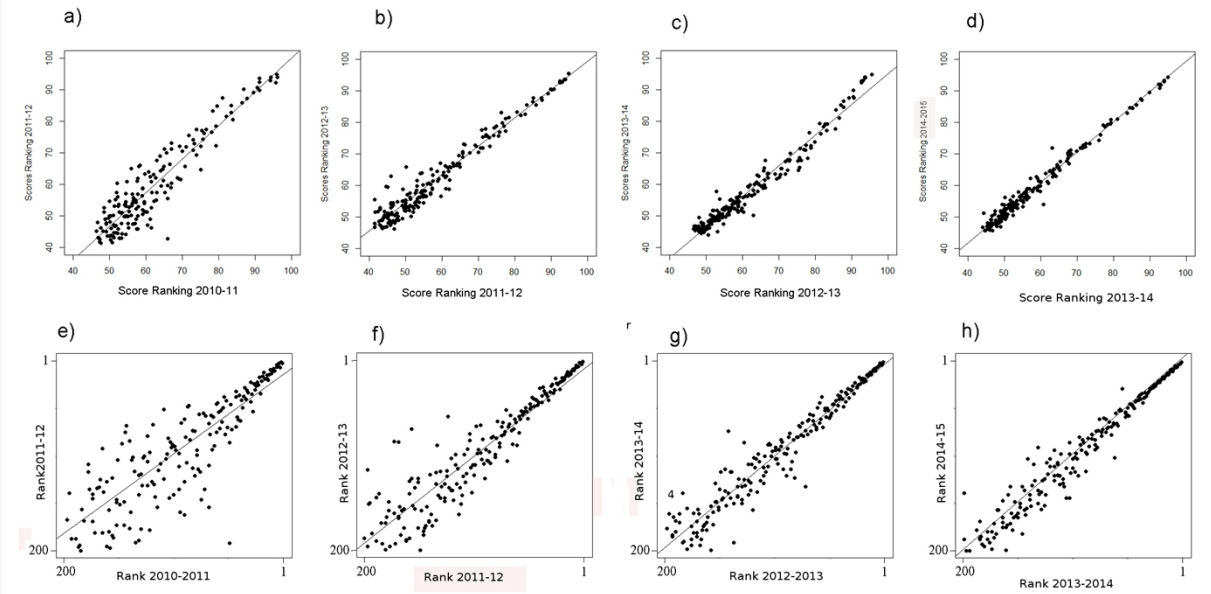

Figure 1a-1d) Scores of the year t-1 regressing on the score of the year t from the ranking 2010-11 on. Figure 1e-1h) Ranks of the year t-1 regressing on the ranks of the year t from the ranking 2010-11 on. Linear regression line indicates perfect association, e.g. no changes in ranks and scores between two consecutive rankings. Remark: THES denotes the rankings in academic years (figures 1 & 2), whereas ARWU denotes the rankings in years (figure 3 & 4).

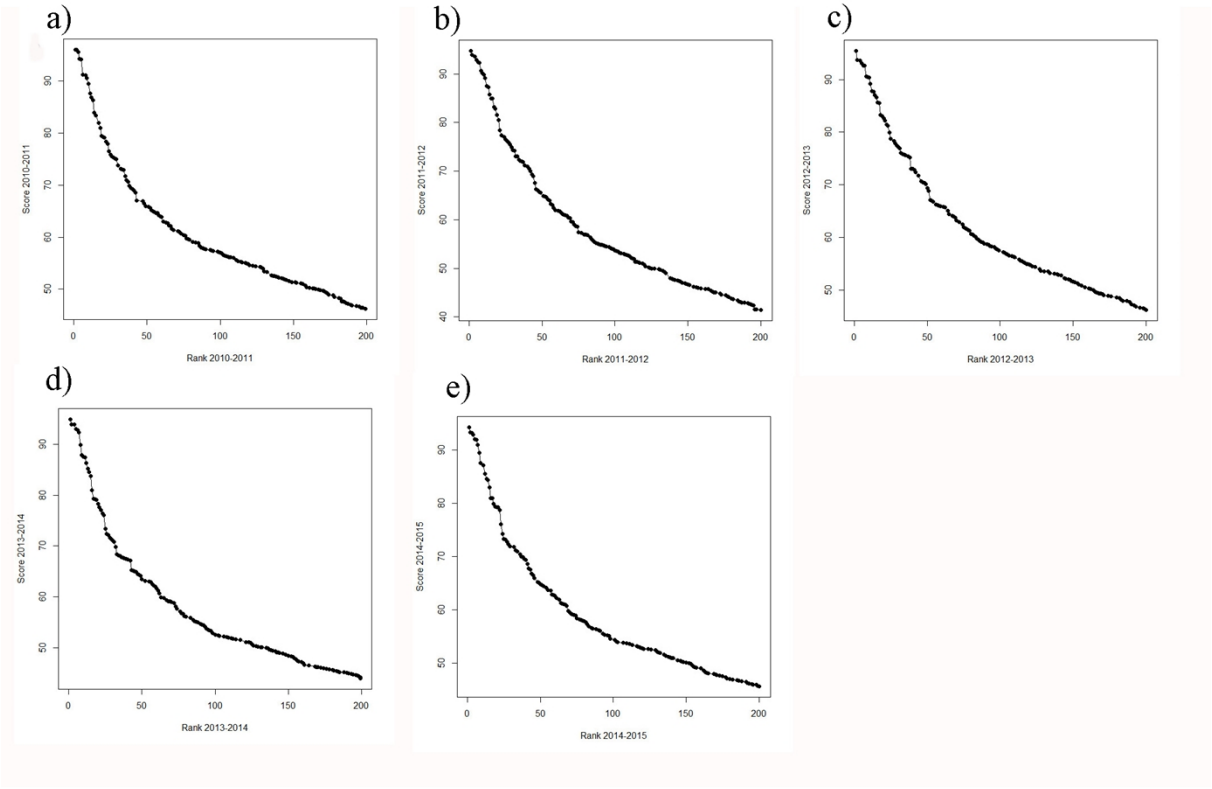

Figure 2 a-e). Ranks plotted against scores for the THES-Ranking a) 2010-11; b) 2011-12; c) 2012-13; d) 2013-2014; e) 2014-15

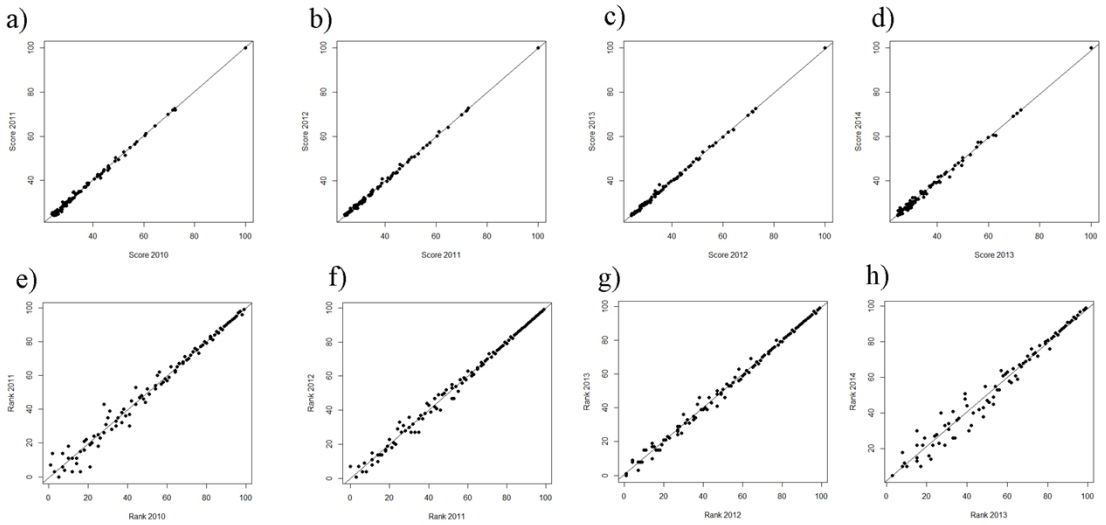

Figure 3 a-d). Scores vs scores (a-e) Ranks vs. ranks (e-h) for the ARWU-Rankings from the 2010 on. a) 2010; b) 2011; c) 1012; d) 2013; e) 2014.

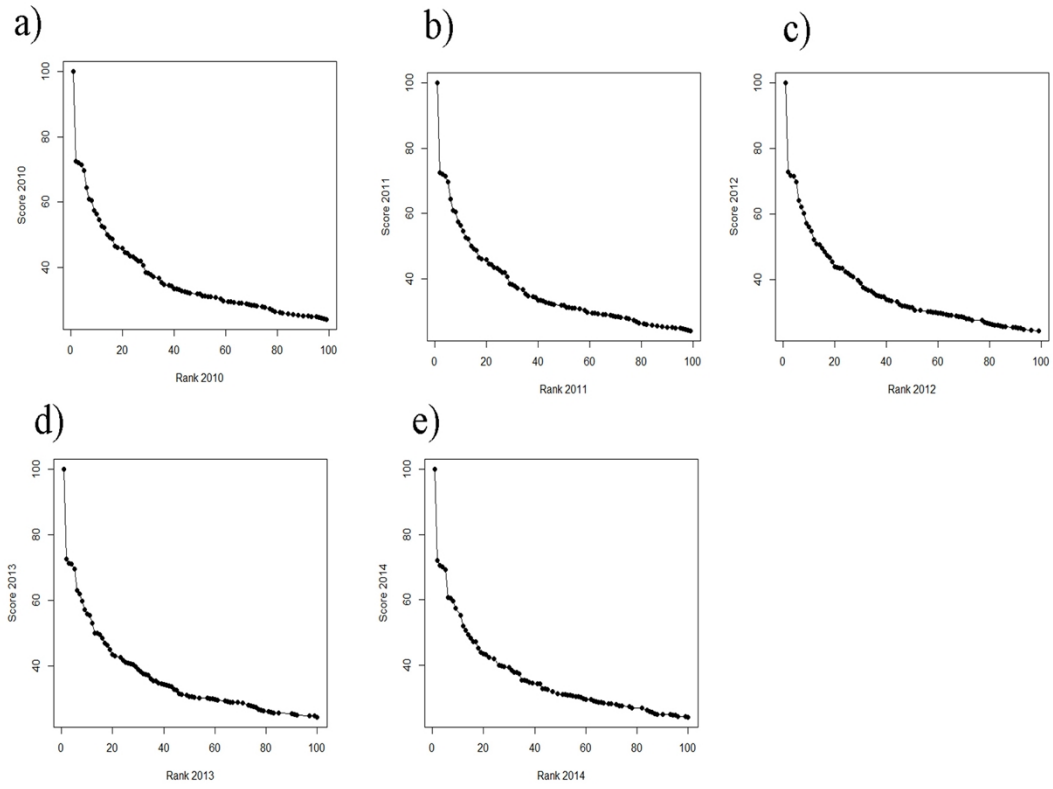

Figure 4 a-e). Scores vs. ranks for the ARWU-Rankings from 2010 on. a) 2010; b) 2011; c) 2012; d) 2013; e) 2014.

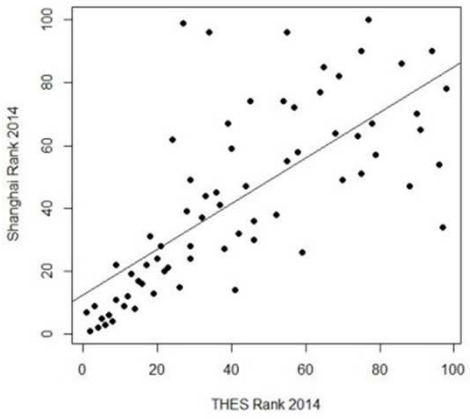

Figure 5) Regression of the ranking positions of the first 100 universities in the THES- Ranking 2014 on the first 100 universities in the ARWU-Ranking 2014.