High Educ (2015) 69:859–866
DOI 10.1007/s10734-014-9809-y

Technical pitfalls in university rankings

Marie-Laure Bougnol · Jose H. Dulá

Received: 14 May 2014 / Accepted: 23 August 2014 / Published online: 10 September 2014

Abstract Academicians, experts, and other stakeholders have contributed extensively to the literature on university rankings also known as "league tables". Often the tone is critical usually focused on the subjective aspects of the process; e.g., the list of the universities' attributes used in the rankings, their respective weights, and the size and composition of the comparison group. These aspects of a ranking are an easy target since, after all, they are based on someone's opinion even if this person is considered an expert. There are other, purely technical, reasons why ranking schemes are problematic. In this paper we discuss these aspects of rankings by studying the handling of the data, exposing logical mistakes, and raising interpretation issues. We present these as a list of four "pitfalls" invoking in each case an example from an actual rankings. Each case also results in recommendations that address the technical issues involved.

Keywords Colleges and universities · Rankings · US News and World Report · Principal component analysis

Introduction

Rankings, especially when applied to higher education, are exposed to all sorts of criticisms from many different sides. There are three aspects of rankings that are the target of the criticisms: attributes, weights, and domain (Casper 1996; Gladwell 2011). The choice of attributes used in the comparison are supposed to capture the essence of the entities being compared. Inevitably, however, this choice involves subjective decisions and

M.-L. Bougnol (✉)
Jacksonville University, Jacksonville, FL 32211, USA
e-mail: mbougno@ju.edu

J. H. Dulá
Virginia Commonwealth University, Richmond, VA 23284, USA
e-mail: jdula@vcu.edu

different experts are likely to come up with different lists especially when dealing with such highly complex entities as universities. An attribute commonly used in rankings measures the institutions' reputation. The data for this attribute is usually collected via surveys. Even though this attribute appears frequently it remains controversial (Jaschik 2007). Weights also have a direct effect on rankings. Most mainstream rankings report the weights used in their calculations without much more of a rationale than that their experts believe them to be appropriate (Webster 2001). The domain of a rankings is the set of entities to be compared. There are competing forces for both large and small sets. Large sets give the appearance of detail and completeness but smaller sets are better defined and the comparisons more intimate. This issue is at the core of the debate as to whether public and private institutions should be compared in the same ranking as is done in "Best National Universities" from *USNWR* (De Vise 2011). All three of these aspects of a ranking are, to a great degree, subjective.

Higher education rankings are also exposed to different, more technical, criticisms. The study of the technical issues relating to rankings in higher education is not entirely new in the literature. For example, Tofallis (2012) presents a detailed technical discussion of the different ways data can be normalized and how this affects rankings. Others who have taken a technical approach to the issue of rankings in higher education are Goldstein and Spiegelhalter (1996) and Filinov and Ruchkina (2002). The current work focuses on specific technical problems appearing in actual rankings in higher education that have not yet been addressed. The rankings used range from highly commercial enterprises such as *US News and World Report* (*USNWR*) "Best National Universities" (*USNWR*2014) and sister venture "QS World University Rankings" (Quacquarelli and Symonds 2012) to more academic ones such as the "Academic Ranking of World Universities" (ARWU) also known as the "Shanghai Ranking" (Liu and Cheng 2005; ARWU 2013). The Leiden University "Green Ranking" (Enserink 2007) is also discussed but for other reasons.

We avoid any discussion that has to do with the subjective nature of the design and implementation of a ranking. This means that we are not interested in issues related to attribute and weight selection as well as the determination of the domain of the study.

The next section presents the list of pitfalls that were selected for discussion. Each pitfall is discussed from a technical viewpoint and connected with an actual practice by some rankings provider in higher education. For every pitfall we present a recommendation as to how to proceed and avoid it or correct it.

Pitfalls and recommendations

Pitfall 1: Anti-isotonic attributes

A weighting scheme that uses positive weights for the attributes' values rewards larger magnitudes. University rankings typically apply positive weights to the values of the attributes in the model. Therefore, an attribute is indicated for use in calculating a university's scores in a ranking if it is *isotonic* in the sense of Dyson et al. (2001); that is, it is reasonable to expect the better universities in the ranking to have higher values for this attribute. A pitfall in creating a model for a ranking scheme based on positive weights is to include attributes that are not clearly isotonic.

Consider the case of *USNWR* 2014 and the attribute "% of Classes w/ 50 or More". It was classified as a subcategory within "Faculty Resources" and contributed 2 % to the total score. An attempt at explaining how the attribute could be isotonic is proposed by

Webster (2001). The explanation as to why it would be reasonable to reward higher values of this attribute when calculating scores for a ranking is that larger sections result in more effective use of faculty. From the students' and parents' perspective, ostensibly the target constituencies of this enterprise, it is hard to see how this is an isotonic attribute. Moreover, this attribute stands in direct contrast with its similar counterpart "% of Classes under 20" which gets a weight of 6 %.

Recommendations

The solution to the problem in the case of "% of Classes w/ 50 or More" depends on the constituency for the ranking. Upper level university administrators may be interested in rewarding this attribute with a positive weight but if the constituency is prospective students and their parents, or even faculty, the solution is to omit this attribute altogether.

Pitfall 2: Rewarding inefficiency

An issue related to isotonicity of attributes is whether an attribute is an input or an outcome and whether these two types of attributes should be treated differently. An entity's efficiency increases if it can lower its inputs and still have the same outcomes, or, conversely, if more outcomes are obtained while maintaining the same inputs. From this viewpoint if two entities, X and Y, have identical values for all the attributes except for one input, then X is considered more efficient than Y if the value for this input is smaller. A pitfall occurs when inputs and outcomes in a rankings scheme are treated in the same way by assigning them positive weights. It then becomes a problem from a strictly efficiency perspective that one way to climb in the rankings is to increase the values for input attributes. In other words, one way to improve in a rankings is by becoming less efficient. *USNWR* provides some guidance as to the classification of the attributes in their model as inputs and outcomes (USNWR 2014 Ed., p. 73). For example, in their narrative they state (p. 74):

> The indicators include input measures that reflect a school's student body, its faculty, and its financial resources, along with outcome measures that signal how well the institution does its job of educating students.

An example of an input is "SAT/ACT Score" which is an attribute in the "Student Selectivity" category. If we compare two imaginary universities in a ranking using the USNWR model with identical values for all attributes except "SAT/ACT Score" we may conclude that the one with the lower of these two values is more efficient; it produces the same outcomes with the same inputs but proportionately fewer top-testing students. Another example occurs in the *Academic Ranking of World Universities* (Liu and Cheng 2005, web site) from the Shanghai Jiao Tong University where the value "Staff of an Institution Winning Nobel Prizes and Fields Medals" is one of the model's attributes. From a strictly human resources point of view, this attribute is an input since it is easier for a university with more Nobel Laureates and Fields Medalists to generate the measured outcomes in the model: Alumni winning Nobel Prizes and Fields Medals, Highly Cited Researchers, Article Published in *Nature* and *Science*, and Articles Indexed in SCI, all other things being equal.

Treating inputs the same way as outcomes by giving both positive weights means that higher positions in a rankings cannot be interpreted as higher efficiency and the incentive to increase the values for input attributes to rise in the rankings can end up rewarding inefficiency.

Recommendation

One way to address the problems resulting from input measures in a rankings scheme is not to use them. This is in fact what is done in the Leiden Ranking (Waltman et al. 2012), more specifically, in the part dealing purely with bibliometric attributes. Another approach is to assign negative weights to inputs. Allowing negative weights carries with it its own problems. For example, it no longer makes sense to require the weights to add up to unity but other conditions may be imposed such that the score must be less than or equal to zero. Taking the notion of using negative weights beyond, leads to Data Envelopment Analysis (DEA) where the weights for inputs and outputs are determined by a linear program so as to maximize the efficiency score of the entity being evaluated (Bougnol and Dulá 2006).

Pitfall 3: Co-linearity in the data

A problem with ranking schemes may result from co-linearity in the data. Co-linearity among attributes' data is a manifestation of excess information. This means that, in theory, some of this information can be removed without affecting the rankings. Note that this does not mean that *any* of the information can be removed. The following result formalizes this.

Result 1 *Consider a ranking with m attributes and the following co-linearity relation among (w.l.o.g.) the first $k \leq m$ attributes*

$$\sum_{j=1}^{k} b_j x_j = b_0;\ b_j \neq 0;\ j = 1, \ldots, k; \tag{1}$$

then any one of the k attributes can be omitted from the model without affecting the ranking with an appropriately adjusted set of weights.

Proof W.L.O.G., consider removing the kth attribute. Apply the co-linearity relation to express x_k as follows:

$$x_k = b'_0 - \sum_{j=1}^{k-1} b'_j x_{i,j}; \tag{2}$$

where $b'_j = b_j / b_k;\ j = 0, \ldots, k-1$.

Suppose entity i has score $\mathcal{S}_i$ in this ranking scheme when the weights are $w_1, \ldots, w_m$. Then

$$\mathcal{S}_i = w_1 x_{i,1} + \cdots + w_k x_{i,k} + \cdots + w_m x_{i,m}. \tag{3}$$

We can use the expression for x_k above in the calculation of the entity's score:

$$\mathcal{S}_i = w_1 x_{i,1} + \cdots + w_k \left(b'_0 - \sum_{j=1}^{k-1} b'_j x_{i,j} \right) + \cdots + w_m x_{i,m}.$$

After canceling out the terms involving b'_0 we group based on the data:

$$\mathcal{S}_i = x_{i,1}(w_1 - w_k b'_1) + \cdots + x_{i,k-1}(w_{k-1} - w_k b'_k) + w_{k+1} x_{i,k+1} + \cdots + w_m x_{i,m}. \tag{4}$$

Define new weights:

$$\tilde{w}_j = \begin{cases} w_j - w_k b'_j, & j = 1, \ldots, k-1; \\ w_j, & j = k+1, \ldots, m; \end{cases} \tag{5}$$

and therefore

$$\mathcal{S}_i = \sum_{\substack{j=1 \\ j \neq k}}^{m} \tilde{w}_j x_{i,j}. \tag{6}$$

Since scores are unaffected by the removal of an attribute in a perfect co-linear relation among a group of attributes using the recalculated weights, the ranking is the same after this removal. □

Formula (5) can result in negative weights. Presumably, all attributes in a ranking scheme are isotonic selected, in part, for their positive impact on the entities' score. A positive weight for an attribute rewards higher values in its dimension. A negative weight means that entities are rewarded for attaining lower values of the attribute. Therefore, an attribute involved in a co-linearity relation cannot be removed if doing so generates a negative weight in Formula (5).

Rankings use real data and therefore are unlikely to exhibit perfect co-linearity. Co-linearity in these cases is a matter of degree. Principal Component Analysis (PCA) is used to assess the degree of co-linearity in data (see Joliffe 2002). An experiment can be used to illustrate the effect of co-linearity in rankings. PCA was applied to the data from QS World University Rankings (Quacquarelli and Symonds 2012). The order of magnitude difference between the largest (2.27) and smallest (0.211) eigenvalues indicates a high degree of co-linearity in this data. The coefficients $(0.747, -0.610, -0.043, -0.210, -0.051, 0.148)$ define a space which captures 95.05 % of the information contained in the data. Using ranking weights of (0.4, 0.05, 0.2, 0.2, 0.05, 0.1) the omission of either the first attribute ["Academic Reputation (AR) Score"] alone or second ["Employer Reputation (ER) Score"] attribute individually, result in positive recalculated weights when Formula (5) is used. The two plots in Fig. 1 illustrate the effect on the rankings when each of these attributes is the one omitted and the weights are adjusted accordingly for the remaining five attributes in the model.

The difference between the entities' ranking before and after the omission is seen by how points deviate from the 45° line. Under perfect co-linearity there would be no deviations. From the plots it is clearly visible that points are more scattered around the 45° line when the "AR Score" is removed (Fig. 1a) than when the "ER Score" is removed (Fig. 1b). Recall that "AR Score" is relatively much heavier than "ER Score" (0.4 vs 0.05), so the fact that removing "AR Score" has a greater impact on the ranking than removing "ER Score" may appear predictable. This however, is not as obvious as this result may indicate. The correlation between the attribute's values and the final rank plays a role. When an attribute and rank are strongly correlated, as the case of "AR Score" ($r^2 = 0.76$) or "ER Score" ($r^2 = 0.44$), then the weight has a proportional impact on the new ranking when it is removed. This may not apply when dealing with attributes that are not well correlated with the ranking. Another effect is a visible jump in the scatter after about the 50th entity in the plots in figures a and b, particularly evident in the first of these two. This can be explained by the relation between attributes' values and weights. At the top of the rankings, entities' scores are high across all attributes and gradually become smaller for entities with lower rankings. As attribute values decrease they become more

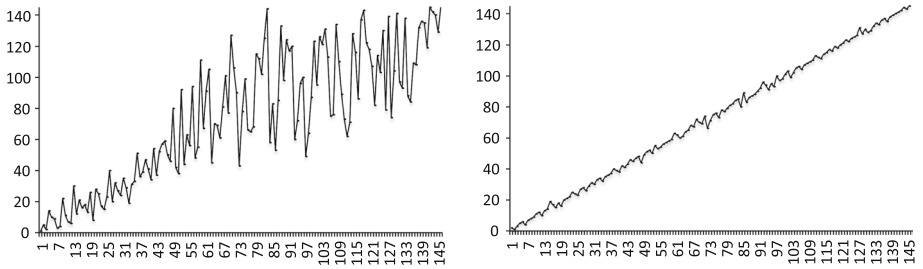

Fig. 1 Effect of co-linearity on rankings: **a** Heavily weighted attribute omitted ($w_{AR} = 0.4$). **b** Lightly weighted attribute omitted ($w_{ER} = 0.05$)

sensitive to the fixed weights so that when these change the deviations from the original ranking are accentuated.

Recommendation

The effect of co-linearity on rankings depends on the degree of this effect as well as on the weight of the attribute and its correlation with the original ranking. We have seen how, in the presence of strong co-linearity, for the case of light attributes that are highly correlated with the original ranking their removal will not have much of an impact on the new ranking especially for the first quarter of the entities where rankings are most important. Therefore, such attributes do not provide much information to the ranking and should be removed.

Pitfall 4: Transparency and reproducibility

From the rankings consumer's point of view, especially the scholars interested in the topic, reproducibility of ranking results is important and, essential to that, is transparency. Ideally, a ranking will provide both the data as it was used in the calculation of the scores and the weights. Unfortunately, not all ranking schemes live up to these ideals. An example of this occurs in the most recent "Best National Universities" rankings of *USNWR*. Although this publication provides information about the weights and their rationale in the model, it is not entirely forthright when it comes to the data and actually suppresses the values for some of the attributes. For example, SAT/ACT scores are clearly being reported in a format that is different from how this attribute's value is used in the calculations. First because they report two types of scores that use different scales and, second, because they report a range. It is not at all clear how this data is used in the actual calculation and any researcher wishing to duplicate the study needs to make assumptions that may not coincide with what the magazine does (Gnolek et al. 2014). The problem also applies to the attributes "Graduation Rate Performance", "Student Faculty Ratio", "National Resources Rank" and others. The lack of clarity about how much of the data is manipulated for the rankings calculations makes it nearly impossible to duplicate the ranking scores even if all the data were available, which is not the case. As of the 2014 edition of *USNWR*, the "Best National Universities" rankings (US News and World Report 2013) suppresses information for the three attributes "Faculty Salary", "Proportion of Professors with Highest Degree", and "Financial Resources per Students" (for which they only provide a ranking). The suppression of two or more of the attribute values for each of the entities in the study, makes it impossible to infer the missing values. Although this information is potentially

available and could be gathered through laborious research, collecting and processing it to obtain the same values used by *USNWR* in their calculations would be nearly impossible as Gnolek et al. (2014) frankly admit.

Recommendation

All the data, along with the weights, as used in the calculations should be revealed if the purpose of a ranking is to provide information about a group of complex entities defined by multiple attributes that are measured using magnitude values. This is indeed the spirit in the case of many of the rankings including the ARWU (Center for World-Class Universities 2013), QS (Quacquarelli and Symonds 2012), and Leiden (Enserink 2007) in that no weights are used.

Concluding remarks

There is a hierarchy of the rankings' consumers in terms of importance from the viewpoint of those who provide them. At the top of this hierarchy are the paying customers. Some of these pay for the rankings by directly buying products; for example, *USNWR*'s yearly print "Best Colleges", or subscription web services. Presumably many of these are prospective students and their parents although universities are eager enough to gain access to the databases available in some of the subscriptions. Another type of paying customers are the advertisers who buy ads promoting their products and services in the media through which the agents broadcast their rankings. Interestingly, the "*USNWR*'s 2014 Best Colleges" is replete with one-page spreads advertising specific universities many of which rank low or not at all in the "Best National Universities" list elsewhere in the publication. Low in this hierarchy are the university administrators who react to the rankings. Unfortunately, these individuals have limited power to affect how rankings are designed and implemented. Finally, at the bottom, are the legions of scholars and researchers who enhance their careers by studying rankings but who contribute nothing to the provider's bottom line. It is not surprising therefore, that ranking agents who design their product for commercial gain pay little attention to the researchers who so eagerly offer their science to improve them.

There are analogies between the rankings and entertainment industries; any scientist's criticisms of the bad science in a science fiction film will have little immediate effect on how a film maker will produce his/her next film especially if it was a commercial success. However, even science fiction evolves with the science. The entertainment providers in this and other genres respond with more realism and accuracy as audiences become more educated about technology and science.

So many of the popular rankings in higher education are like bad science fiction. It is up to the scientists to educate the producers and, more importantly, the consumers of rankings about the technical problems. We hope that the reaction from the rankings providers to criticisms of the technical defects in their products will be to provide better designed and implemented products that actually inform their audience and help them make better decisions.

Of all the ranking schemes analyzed in this study, the one that emerges as having successfully eschewed the pitfalls presented above is the CWTS Leiden Ranking (2013). To begin with, it largely avoids the controversies of the three subjective aspects of a ranking: attributes, weights, and domain. The rankings are created by sorting a single attribute selected by the user. Therefore the model is not defined on a fixed, predetermined,

attribute mix and the use of weights is obviated. The issue of domain is left to the user to decide from among lists based on countries, geographical regions, or disciplines. The design and spirit of the Leiden Ranking definitely avoids the pitfalls we have identified. The user can pick and choose individual attributes on which to sort along with a vast array of parameters. Since attributes are not combined, there are no issues of isotonicity, rewarding inefficiency, data magnitude compatibility, and co-linearity. All the data is published so there is complete transparency and reproducibility. By the standards set in this work, the Leiden Ranking applies the best science.

Acknowledgments We would like to acknowledge that this paper has been inspired and patterned after Dyson et al. (2001).

References

Bougnol, M. L., & Dulá, J. H. (2006). Validating DEA as a ranking tool: An application of DEA to assess performance in higher education. *Annals of Operations Research*, *145*, 339–365.

Bougnol, M. L., & Dulá, J. H. (2013). A mathematical model to optimize decisions for climbing in multi-attribute rankings. *Scientometrics*, *95*, 785–796.

Casper, G., Criticism of College Rankings. (1996). *Private letter to the editor of USNWR*. http://www.stanford.edu/dept/pres-provost/president/speeches/961206gcfallow.html. Accessed March 2014.

Center for World-Class Universities (CWCU), Graduate School of Education of Shanghai Jiao Tong University. (2013). *Academic Ranking of World Universities*. http://www.shanghairanking.com/ARWU2013.html. Accessed April 2014.

CWTS Leiden ranking. (2013). http://www.leidenranking.com/ranking. Accessed April 2014.

De Vise, D. (2011). The ups and downs of U.S. news rankings. *The Washington Post*.

Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S., & Shale, E. A. (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research*, *132*, 245–259.

Enserink, M. (2007). Who ranks the university rankers? *Science*, *317*, 1026–1028.

Filinov, N. B., & Ruchkina, S. (2002). Ranking of higher education institutions in Russia—Some methodological problems. *Higher Education in Europe*, *27*, 407–421.

Gladwell, M. (2011). The order of things. *The New Yorker*.

Gnolek, S. L., Falciano, V. T., & Kuncl, R. W. (2014). Modeling change and variation in U.S. News & World Report college rankings: What would it really take to be in the top 20? *Research in Higher Education*. doi:10.1007/s11162-014-9336-9.

Goldstein, H. D., & Spiegelhalter, J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Series*, *159*, 385–443.

Jaschik, S. (2007). Battle lines on 'U.S. News'. http://www.insidehighered.com/news/2007/05/07/usnews. Accessed March 2014.

Joliffe, I. T. (2002). *Principal component analysis* (2nd ed). Berlin: Springer Series in Statistics.

Liu, N.-C., & Cheng, Y. (2005). The academic ranking of world universities. *Higher Education in Europe*, *30*, 127–136.

Quacquarelli and Symonds. (2012). *World university rankings*. http://www.iu.qs.com; Accessed March 2014.

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

US News and World Report. (2013). *Best colleges 2014 edition*, Published by US News and World Report L.P., Washington, D.C.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., et al. (2012). The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the Association for Information Science and Technology*, *63*, 2419–2432.

Webster, T. J. (2001). A principal component analysis of the U.S. News & World Report tier rankings of colleges and universities. *Economics of Education Review*, *20*, 235–244.