High Educ (2012) 63:1–18
DOI 10.1007/s10734-011-9417-z

A different approach to university rankings

Chris Tofallis

Published online: 31 March 2011

Abstract Educationalists are well able to find fault with rankings on numerous grounds and may reject them outright. However, given that they are here to stay, we could also try to improve them wherever possible. All currently published university rankings combine various measures to produce an overall score using an additive approach. The individual measures are first *normalized* to make the figures 'comparable' before they are combined. Various normalization procedures exist but, unfortunately, they lead to different results when applied to the same data: hence the compiler's choice of normalization actually affects the order in which universities are ranked. Other difficulties associated with the additive approach include differing treatments of the student to staff ratio, and unexpected rank reversals associated with the removal or inclusion of institutions. We show that a multiplicative approach to aggregation overcomes all of these difficulties. It also provides a transparent interpretation for the weights. The proposed approach is very general and can be applied to many other types of ranking problem.

Keywords League tables · Performance measure · University rankings

Introduction

> "Governments are swayed by them, universities fall out over them and vice-chancellors have even lost their jobs because of them" (Attwood 2009)
>
> "Our rankings have become hugely influential, and we recognise our responsibility to produce the most rigorous and transparent table we can" said Ann Mroz, editor of Times Higher Education. The rankings are used by governments and institutions worldwide to benchmark performance in higher education, but their methodology has been criticised. "We acknowledge the criticism, and now want to work with the

C. Tofallis (✉)
Statistical Services and Consultancy Unit, University of Hertfordshire Business School, College Lane, Hatfield, Hertfordshire AL10 9AB, UK
e-mail: c.tofallis@herts.ac.uk

sector to produce a legitimate and robust research tool for academics and university administrators." (Baty 2009)

University rankings (commonly known as 'league tables' in the United Kingdom) have been roundly criticised on many fronts but they continue to attract much attention and discussion every time they are released. Despite their weaknesses it is clear that they are here to stay. In a paper presented to, and discussed at The Royal Statistical Society, Goldstein and Spiegelhalter (1996) stated that "Our principal aim is to open up a discussion of the issues rather than prescribe specific solutions to what are clearly complex problems". They did however, "offer suggestions about appropriate ways of modelling and interpreting performance indicator data". In the same vein we do not claim to have answers to all the criticisms, but we do present serious anomalies which have not been widely appreciated and we offer a suggested way forward.

University league tables are used in making decisions by various groups of people. Each such ('stakeholder') group uses the tables for different purposes:

- For students intending to go to university they provide a collection of useful data in one convenient place. It is therefore likely that they influence which institutions they will apply to. Prior to the appearance of league tables applicants would have had to contact each institution individually to obtain its prospectus. Data which did not show a university in a good light would likely not be mentioned in the prospectus, and so the applicant would be faced with a collection of non-comparable and selective pieces of information.
- Employers faced with selecting from applicants with degrees of the same classification may also be influenced by university rankings in making their recruitment decisions.
- Principals and other directors of higher education institutions tend to find fault with league tables. Yet they and their marketing departments find it difficult to resist quoting them if there has been an upward shift in their ranking. Perhaps this is to be expected in a competitive environment. Given the influence of such tables on employers and prospective students, it is not surprising that directors might want to take steps to improve their ranking. This could be by focusing effort on those criteria they can most easily improve, or where a given expenditure would have the most impact. So even their strategic decisions are likely to be influenced by the expected impact on their ranking. A detailed and comprehensive report commissioned by the Higher Education Funding Council for England (CHERI et al. 2008) included an investigation into "how higher education institutions respond to league tables generally and the extent to which they influence institutional decision-making and actions".

Another possible effect of performance tables is on the tuition fees that universities feel they can charge. Prestigious institutions with high positions in the world rankings can point to this in support of asking for higher fees. Currently British universities are restricted to charging a fixed annual fee to undergraduate students from the European Union. This fixed fee is to be replaced by variable fees (within a given range) which will be chosen by each institution. It is likely that rankings will play a part in influencing the level of tuition fees.

This paper does not deal with the problem of selection of valid criteria. Neither does it deal with the vexing question of which weight values should be used. In our view, both of these issues depend on the intended purpose of the table as well as the intended audience. Ultimately, it could be argued that these choices are really a matter of personal preference and so should be chosen by the user in an interactive online table. Rather, the purpose of this paper is to look at how the criteria are combined. We shall explain the current

approaches and highlight their flaws–flaws which are not appreciated by the vast majority of users. We shall then present a different approach which does not suffer from these flaws.

Normalization

The usual approach to constructing a ranking from multiple criteria or measures includes the following three steps:

1. Normalize the data.
2. Attach weights to the criteria.
3. Add together the weighted values to produce an overall score.

The first step makes the magnitudes of the values 'comparable' or similar across criteria. There are various ways in which this can be achieved, including the following:

(1) Dividing by the largest value. This converts the largest value on each criterion to unity and all others convert to a proportion of the highest achieved value. A variation is to also multiply by 100 so that the normalized values are percentages of the highest achieved score for that measure.

(2) Range normalization. The largest value is given a value of 1 or 100 as above, but in addition the lowest value is converted to zero. Thus there will be an actual observation at each end of the range, and all criteria will have an equal range of observations, from 0 to 1, or 0 to 100. The formula for achieving this for criteria where 'more is better' is:

$$\text{Range normalized score} = (x - x_{\min})/(x_{\max} - x_{\min})$$

where $x_{\max}$ and $x_{\min}$ are the largest and smallest observed values for the given measure.

Notice that this conversion ruins proportionality: If, say, one cost is twice that of another, then this ratio will no longer be maintained for the range normalized scores.

(3) z-scores (statistical standardisation). These are obtained by subtracting the mean and then dividing by the standard deviation. So a value of $z = -2$ indicates a value that is two standard deviations below the mean. As with range normalization, proportionality is lost: consider doubling the x-value (e.g., a cost) associated with $z = -1$, this will not correspond to the cost associated with $z = -2$. Thus an institution which spends twice as much per student as another will not have twice the z-score.

(4) Dividing by the sum. For each criterion we sum all the values, and divide each value by this sum, thus giving a proportion of the total.

Compilers of league tables have, on occasion, switched from one normalization to another. This in itself seems to indicate that the choice of what to do at this stage of the computations is not clear cut. We shall see that this choice is far from innocuous, and indeed can have a dramatic effect on rankings. For example, in 2007 The Times Higher/QS World University Rankings stopped using scores out of 100 (the first type of normalization above), on the grounds that their new 'approach gives fairer results and is used by other ranking organisations' (Ince 2007a). The new approach involved z-scores 'with the top mark set at 100' (Times Higher Education 2008). How the z-values are converted to values out of 100 with no zero or negative values is not explained, and the compiler did not respond to an inquiry about this question. The *Times Higher Education* noted that

'The adjustments in our statistical methods mean substantial change in the results between 2006 and 2007'. One of these was that the London School of Economics plummeted from 17th position to 59th. Even more dramatic was the University of Zurich which soared from 112th to 9th. (These moves may have also been influenced by the fact that a larger citation database was used, but this would not explain such large shifts in position because it only affected one of the indicator variables.)

Because shifts from one year to the next may be explained by factors other than a change in normalization methodology, it would be better to work with a single set of data for a particular year and compare normalizations with that. This is precisely what Yorke and Longden (2005) did. They used the Sunday Times 2004 league table of 119 UK institutions and considered what would happen if the data had been standardised using normalization of type 3 above, compared to the type 1 normalization used by the newspaper. They found that 21 institutions moved by more than 10 places in their rank position, with a few moving by more than 20 places! Thus the choice of normalization can clearly make a substantial difference.

In a very extensive investigation carried out for HEFCE (Higher Education Funding Council for England), it was observed that:

> The weightings "do not necessarily ensure that institutions that perform well on indicators with high weightings have this reflected in their rankings. This is because other aspects of the calculations performed, such as standardising and 'normalising' scores, can have a bigger influence on the overall rankings than the nominal weighting given to each variable" (CHERI et al. 2008, p. 55).
>
> The report also noted that "Compilers are not always clear about their methods for standardising the individual variables, despite this potentially having a major impact on the rankings." (CHERI et al. 2008, p. 21).

The staff:student ratio

The number of students per member of staff is unusual in that it is a measure where high values are worse than low ones. One way of dealing with this is to subtract this variable in the overall weighted score. Another approach (used by *The Guardian* newspaper) is to take the reciprocal, which gives the staff:student ratio, so that it can now be added together with the other variables in the weighted average. It should be noted that these two approaches do not lead to equivalent rankings, since for this to be true the effect of subtracting X would have to be equivalent to adding $1/X$. We shall see later that the approach we shall propose has the benefit of being able to accommodate either the staff:student ratio or the student:staff ratio with identical results. There is thus no longer a need for the table compiler to make an ad hoc choice on this issue.

The unexpected effect of excluded institutions: rank reversal

In any comparison one must decide which institutions are to be included in the analysis. Typically this might consist of all universities in the country in question, although worldwide rankings also exist. The list might be reduced for a particular user, for example they might only be interested in institutions within a given geographical region or city. Some institutions may not appear in the tables because they have requested to be excluded, or they

have declined to provide data, or their data did not arrive in time and will have to be inserted later. Starting from a 'complete' set of data and institutions one might expect that the effect of removing some of them and repeating the computations on the remainder would merely be that those ranked below those excluded would simply shift up the rankings en bloc. Surprisingly, what in fact happens is that the remaining institutions will be re-arranged, with some pairs having their relative positions reversed. A simple numerical illustration of this effect is provided by Filinov and Ruchkina (2002), involving just four universities. When one of these is removed the ranks of the remaining three are completely reversed!

This absurd result means that the 'best' in any set of institutions may not be the best in any subset in which it appears, thus breaking a rule in decision theory known as Sen's alpha condition (also known as the Chernoff condition). The same effect has also been demonstrated by Holder (1998) using the data in *The Times* university ranking. Let us see why this paradoxical effect occurs. The problem is directly related to the normalization step. Each type of normalization involves transforming individual data values using statistics derived from the data set as a whole; for example, dividing by the maximum, or the range, or the standard deviation in each criterion column. If the data for certain institutions is removed (or inserted) then these derived statistics will be altered. Each criterion column will be affected in a different way. The result is that each criterion will now make a different contribution to the total score. This in turn affects the rankings. The way in which the raw data is transformed imputes a weighting. We shall discuss this implied weighting further in the next section.

Filinov and Ruchkina (2002) therefore argue that "it is necessary to exclude the use of the various normalizations". They propose that any methodology should satisfy a requirement that "if some universities refuse to participate in the ranking, the relative positions of those institutions that remain in the ranking should not be changed". Our proposed method satisfies both of these recommendations.

Understanding the interpretation of weights

All published league tables make use of weights in their scoring system. None of them explain what these weights represent. Perhaps it is felt that this is intuitively obvious. It is rather easy to elicit criteria weights from most people. Indeed *The Independent* newspaper's website allows users to create their own 'customised ranking' by choosing weights in their interactive league table at www.thecompleteuniversityguide.co.uk. This allows weights in the range 0–2.5 in steps of 0.5. (The zero weight is particularly useful for eliminating those criteria which might be deemed irrelevant by particular stakeholders.) But when one asks people to explain precisely what these weights represent, they find great difficulty in doing so.

To see what the weights represent in an additive score function let us begin with a simple case of just two criteria: facilities expenditure per student, and entry points of incoming students. The latter is used as a measure of the academic calibre of the students enrolling at an institution. In the UK most school leavers have qualifications called A-levels: an A grade is deemed to be worth 120 points, a B is 100 points, etc. down to an E which is worth 40 points. The average A-level point score is a widely used criterion in league tables. A typical value for this would be of the order of 300. Suppose our score function is simply the sum of these two variables: A-level score + facilities expenditure per student in pounds sterling i.e., suppose we use equal weights. These weights could then be interpreted thus: every extra pound (£) of expenditure on academic facilities per student

would add one unit to the score, as would one extra point in the A-level average (mean entry standard). It also implies an equal 'exchange rate' in the sense that one could retain the same score by 'trading' average A-level points for expenditure per student on a one-to-one basis. Thus the weightings represent substitution rates or trade-off rates.

Now suppose that our data for expenditure is measured in *thousands* of pounds. In that case, using equal weights would mean that one extra A-level point in the mean entry score would contribute the same as £1,000 additional expenditure per student to the total score. Thus a vital point to appreciate is that the interpretation of weights depends on the units in which the variables are measured.

If we attach unequal weights W to the two variables (X, Y) such that:

$$\text{Weighted sum score} = W_X X + W_Y Y \tag{1}$$

then the interpretation would be that an extra unit of the y-variable would be equivalent to W_Y/W_X units of the x-variable in its contribution to the total score. Alternatively, W_Y units of X are equivalent to W_X units of Y.

The meaning of weights becomes more complicated once the data has been normalized. Suppose type 1 normalization has been applied, i.e., a simple rescaling to make the highest score equal to 100 for each criterion. What do equal weights mean in this context? The score function can now be written thus:

$$100X/X_{\max} + 100Y/Y_{\max} \tag{2}$$

Comparing this with Eq. 1 the normalization has effectively introduced (imputed) weights of $100/X_{\max}$ and $100/Y_{\max}$. It now follows that one extra unit of the y-variable contributes the same as $X_{\max}/Y_{\max}$ units of the x-variable, which is the ratio of the best score on the x-variable to the best score on the y-variable. Another way of expressing this is to say that $Y_{\max}$ units of Y are worth $X_{\max}$ units of X. Once explicit weights are attached this will change to: $W_Y\, X_{\max}$ units of X are equivalent to $W_X\, Y_{\max}$ units of Y.

A number of the published league tables, including The Times and The Guardian, use the z-score normalization and then adjust this in some arbitrary way to make the best score 100 and the worst score some positive number. The details are never made entirely clear. For example, according to *The Complete University Guide* (published in association with *The Independent* newspaper): "The Z-scores on each measure were weighted by 1.5 for student satisfaction and research assessment and 1.0 for the rest and summed to give a total score for the university. Finally, these total scores were transformed to a scale where the top score was set at 1,000 with the remainder being a proportion of the top score. This scaling does not affect the overall ranking but it avoids giving any university a negative overall score." The claim that the scores are true proportions of the top score is suspect since some component Z-scores will have been negative (indicating they were below the mean); thus some constant must have been added into remove negative Z-values. Personal communication with the compiler revealed that when the weighted z-scores are summed, if the lowest value is $-L$ then $2L$ is added to make them all positive. This is an ad-hoc approach and does not maintain proportionality.

One more point to notice is that for each type of normalization a particular weighting of the criteria is introduced, albeit inadvertently, even if the user then attaches equal weights. For example Eq. 2 shows that an exchange rate involving the ratio $X_{\max}/Y_{\max}$ has been introduced regarding the relative worth of these two criteria. Since this ratio will change from year to year, the effective exchange rate will also change, and also

therefore the relative weighting. This reduces the comparability of tables across different years.

The multiplicative approach and its advantages

We require a formula to calculate a score based on the criteria values. It is apparent that an additive aggregation formula leads to various difficulties and anomalies. So we must be prepared to open up our minds to alternative aggregation schemes. Let us draw some inspiration from the field of physics where physical laws expressed as equations represent the combined impact of variables measured in different units. For example the momentum of an object in motion is equal to its mass multiplied by its velocity. Other examples: the force of an object equals its mass multiplied by its acceleration (Newton's second law of motion); the impulse of a force is equal to the size of the force multiplied by the time for which it acts. The key thing to notice regarding equations in physics is that the physical laws are generally in the form of a product of variables (i.e. multiplication), and not a sum. This immediately overcomes the issue of incommensurability—adding together quantities measured in different units—the 'adding apples and oranges' problem. It also allows for 'more is worse' variables to be included—one simply divides by them so that increased values lead to a reduction in the impact. This is illustrated by the universal law of gravitation, which states that the force between masses M and m is GMm/d^2 where G is a constant and d is the distance between the masses. The greater the distance—the weaker the force. This also shows us how weighting can be incorporated—one simply raises a variable to a power.

Thus the overall score from a set of indicators X_1, X_2 etc. under the multiplicative approach is found by using the formula:

$$\text{Multiplicative score} = X_1^{W1} X_2^{W2} \ldots X_n^{Wn} \tag{3}$$

If there are indicator variables where 'less is better', then these are divided into the above expression. All approaches have their limitations. In this case, to ensure scale-invariance (or units-invariance—invariance of the rankings to the units of measurement) the indicator variables must be measured on a ratio scale. Also, we cannot really use this type of scoring if an indicator has a zero value. This is unlikely to be a problem in the case of university data although it is conceivable that a particular institution might, for example, not carry out any research activity whatsoever in any department and hence have no rating on that indicator. In such an instance one would be dealing with an organisation that is materially different from the others, and, some would argue, does not constitute a university. One would have to assess such organisations separately.

Let us now turn to the anomalies and flaws associated with the weighted sum scoring approach and see how the multiplicative approach fares. We began by discussing normalization, which had the effect of adjusting the numerical magnitude of variables prior to them being weighted and added together. When the multiplicative approach is used the fact that some variables are numerically much greater than others does not matter since a rescaling of any variable (by multiplying by a positive constant) would have no effect on the results. For example, consider a switch from measuring expenditures in thousands of pounds to pounds; this would simply lead to a multiplication of the score function by 1000 for every institution. The relative scores (ratio of one to another) would remain the same for the institutions, and therefore so would their ranks. This remains true if the adjusted variable has a weight (W) associated with it: the scores would then all be multiplied by the same factor of 1000^{w}.

Avoiding the need for normalization simplifies the procedure and makes it more transparent. It also means that the compiler no longer has to make a choice regarding which type of normalization to apply to the data, with its consequent effect on relative rankings. This is a big step forward.

We previously noted that dealing with the student:staff ratio was problematic in the additive scoring scenario. Under the multiplicative approach it becomes straightforward: criteria where 'more is worse' are divided into the score function. Furthermore, if we instead use the staff:student ratio we would simply multiply this with the other criteria. Whichever of these two approaches is used, the result is identical (since dividing by a quantity is equivalent to multiplying by its reciprocal).

Next we turn to the issue of institutions that are excluded from the analysis causing rank reversals. An institution's score under the multiplicative scheme is not affected in any way by the data of other universities. Since there is no longer any normalization (which would be based on the rest of the data), the relative order of the remaining institutions is unaffected i.e., there are no rank reversals. All that happens is that those universities which were below the excluded ones merely shift up the rankings. This is as it should be.

We also noted that with the additive scoring model comparability across time is reduced because the precise form of the normalizations will change as the data changes from one year to the next. This problem does not arise with multiplicative scoring and is therefore another benefit.

Weights in the multiplicative approach

Since weightings now appear as powers (exponents) of the criteria, their interpretation is now in terms of percentage: a weighting of W means that a 1% improvement in a performance indicator will lead to a W% change in the score. A useful benefit is that this remains true even if the unit of measurement is re-scaled, e.g., from pounds to thousands of pounds. It also means that a given weight has the same effect irrespective of which criterion it is applied to—this was not true in the additive case because the weight interpretation depended on the units of measurement as well as on the type of normalization chosen. So weights are much simpler to comprehend under the proposed scheme.

Another interesting consequence is that the meaning of 'equal weights' is clear and unambiguous in multiplicative scoring. For example if all weights equal unity, it means that a 1% change in *any* indicator measure will lead to a 1% change in the overall score. By contrast, in additive scoring, the precise interpretation of 'equal weights' depends firstly on the units of measurement used for each indicator, and secondly on the normalization that is to be applied. The vast majority of users of existing tables would not be aware of such complications.

When considering weights in the proposed scheme it is as well to understand the difference in effect between values less than unity and values greater than unity. A weight exceeding unity always has the effect of stretching out differences at the upper end of the scale for that criterion. (Think of the graph of the function x^2, where a unit change for large values of x leads to a bigger change in x^2 than occurs for a unit change at smaller values of x.) Weights below unity lead to differences at the upper end of the scale being muted or compressed relative to those at the lower end of the scale. (Think of the graph of the square root of x.) In deciding how to place limits on the weights one might note the general tendency for diminishing marginal utility: When one already has a very large amount of a given resource, one extra unit will make less of a difference to that institution than for an

institution which is at the other end of the scale. To model this effect one would choose a weight less than unity; this is in fact what one would normally expect to use and one can ensure this by arranging for the sum of weights to equal 1 or 100%. The assignment of weights could then be compared to slicing up or allocating segments of a pie chart. Note that in this case, we have the useful property that if all criteria are increased by the same percentage then the overall score will change by that same percentage.

A comparative illustration

Let us compare results between the additive and multiplicative scoring models. We cannot simply transfer the weights chosen by any particular publisher of tables to the multiplicative case because the weights operate in a different way for each of these approaches. Given this caveat, perhaps a prima facie comparison might be to take some published data and use equal weights under both schemes. We have used data published in *The Complete University Guide 2008* and we are grateful to Dr. Bernard Kingston of "*The Complete University Guide*" for granting permission. We omitted institutions for which there was missing data. (We do not follow the practice of dealing with missing data by inserting the mean of that criterion across the other universities.) This resulted in data for 99 institutions. The criteria employed were: Student Satisfaction, Research quality, Student-staff ratio, Expenditure on academic services and facilities, Completion rate, Percentage gaining a good honours degree (at least upper second class), Graduate prospects (percentage in graduate level employment or further study 6 months after graduation), and Entry standards. We shall compare results with the additive approach to the limited extent that this is possible. For additive scoring we used the simplest normalization: allocating a score of 100 for the best observed value on each criterion; this retains proportionality and a ratio scale. For the sake of comparison equal weights will be used in both approaches. As a result of these steps the ranks using our additive scheme are not the same as the published ones. The Sect. "Appendix" displays the results with the institutions listed in the order of their published ranking, though we emphasise that this ranking cannot be compared with our own results here—apart from other differences, the published version applied unequal weights to the criteria.

To begin with let us compare the distribution of scores. Figures 1 and 2 show the results for additive and multiplicative scores respectively. In each case we have arranged for the arithmetic mean score to equal one, this was achieved by simply dividing the scores by the

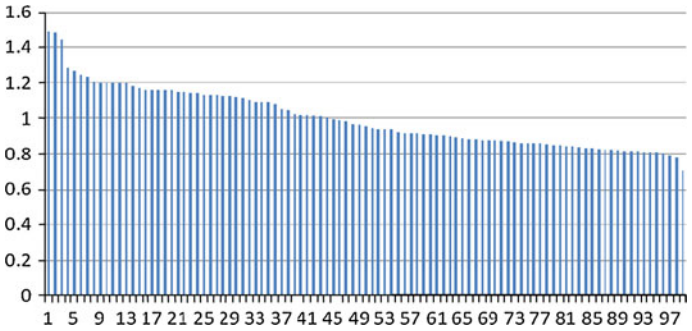

Fig. 1 Distribution of scores using the additive approach. Rank on horizontal axis

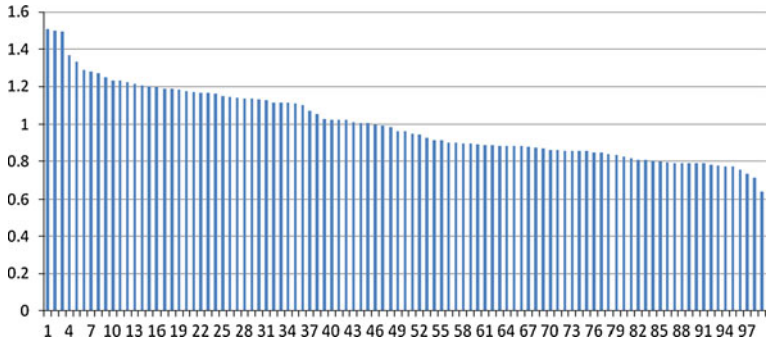

Fig. 2 Distribution of scores using the multiplicative approach. Rank on horizontal axis

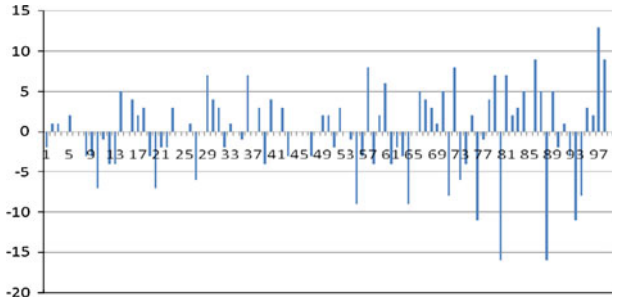

Fig. 3 Difference in ranks between additive and multiplicative approaches (equal weights). A positive value implies that the institution was placed higher according to the multiplicative scheme. The horizontal axis shows the ordering according to *The Complete University Guide*

arithmetic mean. The distributions are not dissimilar and both display a sharp drop in scores after the top three institutions.

If we look at the differences in the ranks (out of 99) between the two approaches for any given institution we find at one extreme a gain of 13 places using the multiplicative approach (for London South Bank), and a fall of 16 places for both Bolton and York St. John. The mean change in rank is of course zero since any gain in rank is always balanced by a loss elsewhere. A more useful statistic is the mean absolute change in rank; we found this to be 3.6 rank positions. There were 18 institutions that maintained exactly the same rank. Figure 3 displays the difference in rankings: additive minus multiplicative. One notices that the larger changes tend to occur at the lower performance end.

Let us delve deeper and try to understand what gives rise to the differences. We begin by comparing two institutions which have an almost identical additive score but a very different multiplicative score: Bolton and Liverpool John Moores are ranked 78th and 79th under the additive scheme. Under multiplicative scoring however, Liverpool John Moores rises by 7 places whilst Bolton falls by 16. Given that we have used equal weights, there must be something in the composition of the individual measures that is causing this difference. It is difficult to perceive what is going on across criteria by referring to actual scores so we shall refer to the ranks instead. For Bolton the ranks on individual criteria are: 30, 39, 47, 70, 92, 94, 99, 99. Notice that four of these are near the bottom whilst two of these (30, 39) are much better than its other ranks. Thus on the additive model the two

good performances are helping to cover up or compensate for the very poor performance on four other measures. Whereas under the multiplicative scheme it has not been able to do this. The individual criteria ranks for Liverpool John Moores are: 54, 58, 69, 71, 72, 75, 79, 95. Notice that these are more closely spaced. We can measure this scatter using the standard deviation of the ranks; for Bolton the figure is 28.8 whereas for Liverpool John Moores is only 12.6.

Next we consider institutions with a similar multiplicative score but a very different additive score. Bedfordshire and Leeds Metropolitan are next to each other on the multiplicative ranking with geometric mean scores of 23.07 and 23.05 respectively, but differ by 17 places on the additive ranking with Bedfordshire being ahead. The individual criteria rankings for Bedfordshire are: 4, 64, 71, 87, 89, 91, 94, 98; whilst for Leeds Metropolitan they are: 55, 58, 62, 67, 77, 78, 89, 90. The median of these ranks for Bedfordshire is 88 whilst for Leeds Metropolitan it's higher, at 72, yet Bedfordshire wins out in the additive ranking. When we look at the scatter of the individual criteria ranks there is a marked difference, with the standard deviation for Bedfordshire being 30.9 whereas for Leeds Metropolitan it is only 13.5. Once again we observe an advantage in the additive model when there is a wide scatter in results because poor performance can be covered up or compensated by good performance elsewhere to a far greater extent than in the multiplicative model. In contrast, the multiplicative model rewards consistency to a greater extent than does the additive model.

Understanding the differences in ranking between the additive and multiplicative approaches

To put the above observations on a more solid footing let us consider general cases. Let's begin with just two criteria (X, Y) with strictly positive scores between zero and 100 with equal weighting. Suppose two institutions have the same total additive score $X + Y$ and therefore the same ranking. In general their multiplicative score (XY) need not be the same and so one will rank higher than the other in this scheme. One can prove that the multiplicative score is maximised when $X = Y$. For example if the total score is 20, individual scores of 10 and 10 give the highest multiplicative score, beating 11 and 9, 12 and 8 etc. This is a consequence of a well known mathematical result known as the arithmetic mean-geometric mean inequality which states that for any given positive numbers their geometric mean is always less than or equal to their arithmetic mean, with equality occurring when the component numbers are equal. Applied to our example the arithmetic mean is fixed (total score = 20) but by adjusting its components we can raise the geometric mean value until it reaches the value of the arithmetic mean. This famous inequality applies for any set of n non-negative numbers. Applied to our situation this means that for a given additive score, the multiplicative score is maximised when the components are equal, and as the components diverge away from equality the multiplicative score declines. This is illustrated in Fig. 4 where we now assume that individual component scores have a maximum of unity; the straight line represents all points having the same aggregate score of 1 under an additive scheme—all but one of these points lie below the curved contour for the multiplicative scheme—indicating that their multiplicative score is not high enough to reach that contour; the exception is when the components are equal. Notice that points near the ends of the straight line—which have a very high score on one attribute and a very low score on the other—will be much further

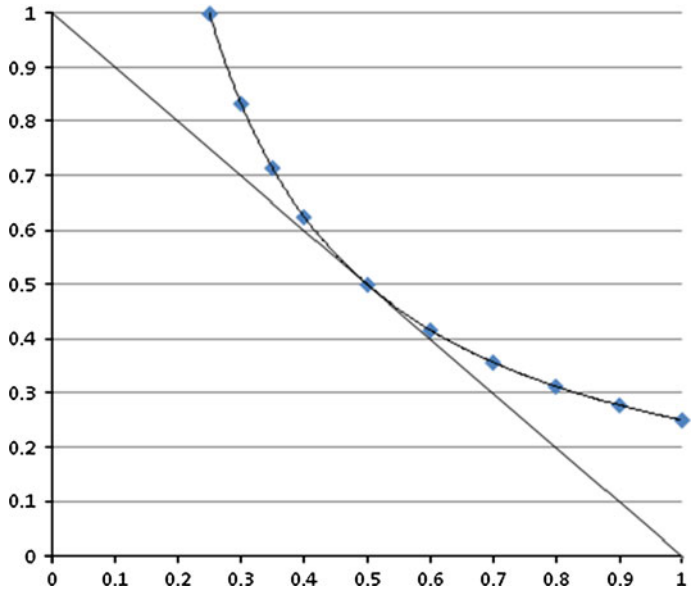

Fig. 4 *Contour lines* for additive (*straight*) and multiplicative (*curved*) aggregations. Each point on the contour has the same score under its respective aggregation scheme

from the curved contour, indicating they will be strongly disadvantaged under a multiplicative scheme.

Our inequality states that

Geometric mean $\leq$ *Arithmetic Mean*

where the equality occurs when the components are the same. Let us now look at this inequality from the other direction: consider institutions with the same multiplicative score and hence the same value for the geometric mean—these are points on the curve in Fig. 4.

The geometric mean value now acts as a lower bound and the lowest score for the arithmetic mean occurs when the component scores are the same. If we change the component values—make them diverge from each other—the arithmetic mean will increase from its lowest value, and hence so will the total additive score. We thus see that an institution rated under additive scoring will benefit from a wide spread of scores whereas one which has more uniform or less scattered scores will do well under the multiplicative scheme.

The above discussion assumed equal weights, but it extends to the case where unequal weights are applied. Suppose we have non-negative weights W_i which sum to unity, then the general arithmetic mean-geometric mean inequality states that

$$\sum W_i X_i \geq X_1^{W1} X_2^{W2} \ldots X_n^{Wn} \tag{4}$$

with equality occurring when the components x_i are equal. In our context the left hand side is the additive score and the right hand side is the multiplicative score. As before, this demonstrates that for two institutions with the same additive score (and thus rank), the one having equal or similar component scores will have a higher multiplicative score and rank. If the components diverge in value then the score is reduced. Conversely, for institutions

with the same multiplicative score and rank, the additive score will be lowest when the components are equal, and this score will be higher for institutions which have components that diverge from each other.

In summary, consistency of performance across the various indicators is rewarded to a greater extent under the multiplicative scheme, and excellence in a few fields will not automatically imply a high ranking. Under the additive scheme very poor performance in some criteria can be compensated to a greater extent by good performance elsewhere.

Conclusion

There are a great many issues associated with rankings. It is important not just to point to their weaknesses, but to suggest ways in which they can be improved. This paper has focused on the issue of how measures are aggregated. All current publishers of league tables use an additive approach which includes a normalization step to make the individual performance indicators 'comparable' before they are combined to produce a single value score. The problem is that there are different ways of achieving this comparability. A choice therefore needs to be made. If the result of this choice did not affect the final rankings there would be no problem—but it does affect them. The attraction of the multiplicative approach is that a normalization step is not required and so the problem is avoided from the start.

Rank reversal is an anomalous feature of additive scoring which arises when the chosen normalisation is data-dependent. This is where the inclusion or exclusion of a particular institution (C) reverses the relative ranking of other institutions (A ranked above B changing to B ranked above A). It was for this reason that Filinov and Ruchkina (2002) proposed that "it is necessary to exclude the use of the various normalizations", and that any ranking approach should satisfy the principle that "if some universities refuse to participate in the ranking, the relative positions of those institutions that remain in the ranking should not be changed". The multiplicative approach satisfies this principle as it does not involve normalization.

One indicator that is commonly included in league tables is that involving the number of academic staff relative to the number of students. If an additive methodology is employed the compiler has to make a choice: Either use the ratio staff/students and add this into the total, or use the ratio students/staff and subtract from the total. The effect of adding a quantity is not equivalent to subtracting its reciprocal, and this in turn affects the overall ranking. One again the compiler is forced into making a choice which will affect the final results. Under the multiplicative methodology this issue does not arise and so is avoided. This is because if the student/staff ratio is used then it is divided into the aggregate, whereas if the staff/student ratio is used it is multiplied with the other indicators; the result is the same under a multiplicative scheme, and this remains true if unequal weights are applied.

Choosing weight values has always been a difficult cognitive exercise. This is perhaps in part due to the fact that most people do not have a clear understanding of what these numbers represent. If the movement towards interactive league tables on the internet expands then it is important that the end-user appreciates the effect and meaning of the weights that he or she is selecting. The interpretation of weights under the multiplicative scheme is simply this: applying a weight W means that if a criterion improves by 1% then this will lead to a W% change in the final score. Under the additive scheme using a weight *W* implies that that if the *normalized* score on that criterion is increased by one unit then

the overall score will increase by W units. Since the effect of a weight depends on the form of normalization selected the weight interpretation is less straightforward under the additive scheme. A typical user presented with data adjusted using two completely different normalisations would most likely submit the same weight values. They would then be perplexed as to why the rankings were not the same.

Even the simplest case of equal weights can be problematic, even though this notion is one that people feel intuitively comfortable with and think they understand. But again, using equal weights and adding together data that has been normalised in different ways leads to different results. By contrast, under a multiplicative scheme, 'equal weights' has a simple and unique interpretation and leads to a unique result; under equal weights if *any* one performance measure changes by a certain percentage then the final score is affected in the same way. For example, if the weights are all equal to unity, then an improvement of 1% in any one criterion implies a 1% improvement in the overall score.

We need to remind ourselves that a single figure cannot possibly represent all the activities that take place in any large organisation. It is more preferable by far to have a number of separate scores for each activity or function, for example: research, postgraduate teaching, undergraduate teaching etc. This would replace the single overall score by a performance *profile*, which would make it easier to see where the strengths and weaknesses lie. The selection and grouping of measures for such profiles would, in general, depend on the interests of particular users or stakeholders as well as the purpose of the analysis. One approach for grouping together the various attributes is according to whether they are inputs, outputs, process measures etc. For example, efficiency is the ratio of output to input, and so if an efficiency measure were being sought the outputs would be multiplied together and divided by the product of the inputs. This is equivalent to multiplying the outputs by the reciprocals of the inputs, and so the multiplicative approach can be applied.

Even with such multi-dimensional performance profiles it is still the case that we would need to combine measures for each of these general headings or functions, and so a method of aggregation would still be required. Hence the content of this paper remains relevant.

While this paper was being revised in accordance with helpful comments made by this journal's referees, an important and relevant development occurred in another field. Each year the United Nations publishes the Human Development Index. This is a ranking of countries based on an aggregate of three dimensions using an arithmetic mean, i.e., additive aggregation. At the end of 2010, after 20 years of using this approach the United Nations decided that it would be an improvement to switch to a geometric mean, i.e., multiplicative aggregation.

> "We reconsidered how to aggregate the three dimensions. A key change was to shift to a geometric mean, thus in 2010 the HDI is the geometric mean of the three dimension indices. Poor performance in any dimension is now directly reflected in the HDI, and there is no longer perfect substitutability across dimensions. This method captures how well rounded a country's performance is across the three dimensions. As a basis for comparisons of achievement, this method is also more respectful of the intrinsic differences in the dimensions than a simple average is. It recognizes that health, education and income are all important, but also that it is hard to compare these different dimensions of well-being." (United Nations Development Programme 2010, p. 15)

Their stated reason is that the old scheme allowed for 'perfect substitution', meaning that poor performance in one dimension could be substituted or covered up by better performance in another. The new Human Development Index "thus addresses one of the most serious criticisms of the linear aggregation formula, which allowed for perfect substitution across dimensions." (UNDP 2010, p. 216).

Their findings also confirm our discussion regarding the effect of such a change:

"Adopting the geometric mean produces lower index values, with the largest changes occurring in countries with uneven development across dimensions." (UNDP, p. 217)

We are convinced that the multiplicative approach has benefits, not just for university ranking but for many other applications where criteria are aggregated (e.g. Tofallis 2008). The fact that the United Nations Development Programme has adopted multiplicative aggregation should encourage others to consider its advantages. We therefore commend it for serious consideration and implementation.

Appendix

See Table 1.

Table 1 Comparison of scores and rankings

Institution	Additive score	Multiplicative score	Rank using additive approach	Rank using multiplicative approach
Cambridge	100.0	99.3	1	3
Oxford	97.1	99.6	3	2
Imperial College	99.3	100.0	2	1
LSE	85.1	88.4	5	5
St Andrews	80.3	81.7	13	11
UCL	86.2	90.8	4	4
Bristol	83.5	85.5	6	6
Bath	83.0	81.7	7	10
Durham	80.7	79.9	11	14
Loughborough	77.1	75.4	21	28
Aston	79.5	79.6	14	15
Royal Holloway	78.0	77.7	17	21
Nottingham	81.0	81.1	8	12
York	78.0	80.4	18	13
Edinburgh	80.8	83.0	9	9
King's College London	80.4	84.4	12	8
Exeter	75.6	75.8	28	26
Lancaster	75.4	75.6	30	27
East Anglia	75.4	73.9	29	32
Leicester	76.5	74.6	24	31
Southampton	78.1	78.9	16	18
Newcastle	78.6	78.9	15	17
SOAS	80.8	84.8	10	7
Sheffield	77.9	78.6	19	19

Table 1 continued

Institution	Additive score	Multiplicative score	Rank using additive approach	Rank using multiplicative approach
Sussex	77.0	77.4	22	22
Cardiff	76.0	77.1	25	24
Queens—Belfast	76.5	75.2	23	29
Reading	73.1	73.5	35	35
Glasgow	75.8	78.0	27	20
Manchester	77.9	79.5	20	16
Birmingham	76.0	77.3	26	23
Essex	73.1	72.9	34	36
Surrey	74.8	75.0	31	30
Kent	70.3	71.2	37	37
Leeds	73.2	73.7	33	34
Queen Mary	73.7	76.0	32	25
Hull	68.2	68.1	41	41
Liverpool	72.5	73.8	36	33
Aberystwyth	68.3	66.9	40	44
Bangor	67.4	68.1	44	40
Swansea	67.6	67.1	43	43
City	68.1	68.1	42	39
Bradford	68.8	68.0	39	42
Keele	66.0	65.9	47	47
Goldsmiths College	65.4	65.3	48	48
Brunel	69.9	70.1	38	38
Oxford Brookes	66.3	64.0	46	49
Ulster	66.6	66.8	45	45
Nottingham Trent	63.0	62.6	54	52
Plymouth	63.1	62.9	53	51
Lampeter	62.1	59.7	55	57
University of the Arts	64.8	66.3	49	46
Salford	64.3	63.8	50	50
Roehampton	63.3	61.6	52	53
Central Lancashire	61.3	58.4	58	67
UWCN—Newport	60.4	58.5	63	66
Bournemouth	59.0	58.9	70	62
Central England	63.6	60.6	51	55
Glamorgan	60.8	59.5	61	59
Brighton	61.0	60.8	60	54
Bath Spa	57.6	54.9	76	80
Winchester	59.0	57.1	69	71
Gloucestershire	61.4	59.1	57	60
UWIC—Cardiff	61.5	58.6	56	65
Northumbria	60.2	58.6	64	64
Portsmouth	59.3	59.0	66	61

Table 1 continued

Institution	Additive score	Multiplicative score	Rank using additive approach	Rank using multiplicative approach
West of England	60.6	59.5	62	58
Sheffield Hallam	61.1	59.8	59	56
Chichester	56.0	53.4	85	84
Staffordshire	57.7	57.1	75	70
Coventry	59.2	56.9	67	75
Kingston	58.7	58.7	71	63
Worcester	56.0	52.5	84	90
Chester	55.1	51.9	88	92
Canterbury Christ Church	54.6	52.5	91	89
Bedfordshire	59.7	56.3	65	76
Sunderland	58.2	56.9	72	73
De Montfort	58.1	57.9	73	69
Liverpool John Moores	56.9	57.0	79	72
Bolton	57.2	51.3	78	94
Huddersfield	54.9	53.7	89	82
Hertfordshire	56.8	55.7	80	78
Northampton	55.6	53.6	86	83
Leeds Metropolitan	56.5	56.3	82	77
Westminster	57.8	56.9	74	74
Manchester Metropolitan	57.5	58.0	77	68
Teesside	54.6	53.1	90	85
York St John	56.7	48.7	81	97
Derby	54.5	52.6	92	87
Anglia Ruskin	54.2	51.2	93	95
Cumbria	55.3	52.8	87	86
Southampton Solent	54.0	47.6	95	98
Middlesex	59.1	55.5	68	79
Wolverhampton	56.3	52.5	83	91
Lincoln	53.8	51.7	96	93
Liverpool Hope	52.4	50.3	98	96
London South Bank	54.2	54.3	94	81
Greenwich	53.2	52.6	97	88
Edge Hill	47.4	42.5	99	99

References

Attwood, R. (2009). Redrawing ranking rules for clarity, reliability and sense. *Times Higher Education*, December 10, 2009.

Baty, P. (2009). New era for the world rankings. *Times Higher Education,* November 5, 2009.

Centre for Higher Education Research and Information (CHERI), Open University, and Hobsons Research. (2008). *Counting what is measured or measuring what counts? League tables and their impact on higher education institutions in England.* Report to HEFCE.

Filinov, N. B., & Ruchkina, S. (2002). Ranking of higher education institutions in Russia—Some methodological problems. *Higher Education in Europe, 27*(4), 407–421.

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance (with discussion). *Journal of Royal Statistical Society A, 159*(3), 385–443.

Holder, R. D. (1998). The good university guide and rank reversal. *Journal of Operational Research Soc, 49*(9), 1019–1020.

Ince, M. (2007a). Fine tuning reveals distinctions. *Times Higher Education Supplement*, 9th November supplement.

Ince, M. (2007b). Ideas without borders as excellence goes global. *Times Higher Education Supplement*, 9th November supplement.

Kingston, B. (2007). *The Complete University Guide 2008*.

Times Higher Education. (2008). Strong measures. *Times Higher Education,* October 9th, supplement.

Tofallis, C. (2008). Selecting the best statistical distribution. *Computers and Industrial Engineering, 54*, 690–694.

United Nations Development Programme. (2010). *Human Development Report 2010. The Real Wealth of Nations: Pathways to human development.*

Yorke, M., & Longden,B. (2005). Significant figures: Performance indicators and league tables. SCOP (Standing Conference of Principals). Available from: http://www.scop.ac.uk/UploadFolder/SCOPsigfigfinalprint2.pdf.