Data Analysis and Visualisatition
World University Rankings

Judith Corsel, Aleksej Horvat, Sybren Osinga, Tim Stolp

June 2018

Introduction:

Judging the quality of education can be an arduous task, with numerous differences between each one and a very different approach to communication their persuasion. A number of companies has taken up the task to rank different institutions, each basing the order on their own methodology. Do these methods of grading hold any ground? When just looking at the measured data, do the patterns suggest any flaws in the applied system?

These broader questions have lead to a central one: Will ranking universities provide additional insight in an institutions excellence, or does it merely underline the arbitrary nature of such lists? In this research a data-set provided by one of these companies is tested on consistency and analyzed for anomalies and interesting patterns. It is hypothesized that thorough analysis will reveal meaningful insights.

The Data Set:

A substantial part of our data was provided by Times Higher Education, further called THE. This company collects data from numerous universities. Universities are eligible for review if they meet the requirements for that year. In 2018, some of the requirements entailed the teaching of undergraduates and over 1000 published articles (between 2012 and 2016). On the rare occasions when a particular data point is not provided, they enter a conservative estimate for the affected metric. After review the data from around 1300 universities were published on their web-pages.

In order to analyse and visualize the data, understanding how the data is constructed is vital. The overall ranking of all universities in the data set is calculated with a set formula, allocating weight to each measured category (figure 1). Barring the slight variance over the different years, the formula allots 30% to each of the three main categories (teaching, research and citations). The categories "international outlook" and "industry income" share the remaining ten percent. The justification of each weight is based on a qualitative assessment of the categories values, and are quite well documented by THE. In general the distribution of weights is constructed to maximize value for students attending.

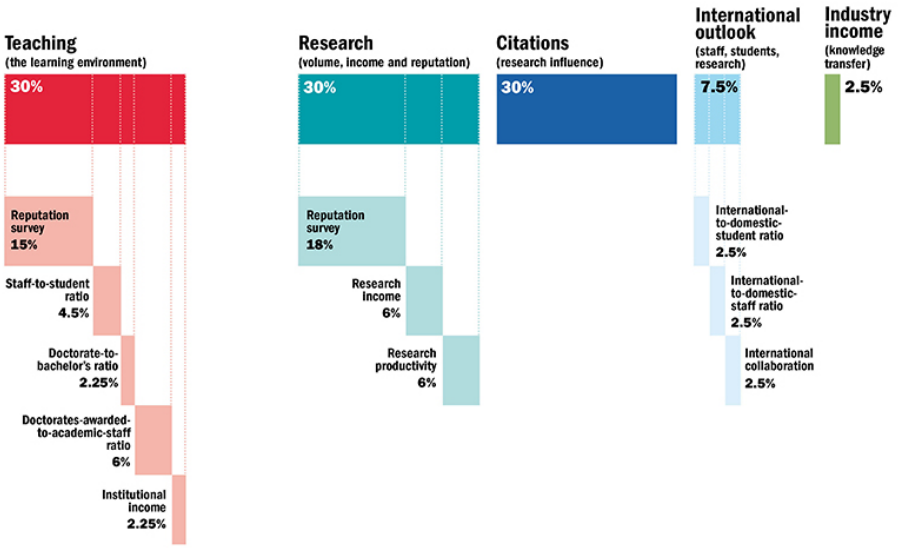

Figure 1: The assigned weight to each category in the 2018 survey by Time High Education, on wich the overall ranking is based

Attempts at scraping these figures did not yield usable data, since these were not present in the HTML itself. A web-driver, Selenium, was considered, tot have the JavaScript elements load, then scrape, but this did not work as intended. However, via the developer tools of Firefox a complete JSON file could be extracted for each year. This satisfied our demand of data from THE.

Cleaning the data, multiple columns were dropped for being duplicates or containing information only valuable to the website itself, such as "member level". Some data was reformatted for easier calculations, like the "female:male ratio" was replaced by "percentage male". Even after review by THE some entries for "percentage male" still displayed missing values. These values were satisfied by a mean of the university's country, as national regulations on gender equality can be a deciding factor in accessibility to education. All other cells showed no noise.

Anomalies:

The nature of the formula used effectively removes almost all anomalies from the top 200 universities. For the three main categories, in most years, the scores of the top 200 universities broadly follow the overall rankings. The subset "top few" derived from any formula should follow the characteristics dedicated by

the dominant factors of such formula.

Certain individual entries have their high score in teaching levelled by their lower score in research, and there subsequent even lower score in citations. The Lomonosov Moscow State University, ranked 194 in 2018, has a teaching score ranked 26Th in that category. But it has one of the lowest scores in citations and is placed 933rd. Other cases where teaching lowers the overall rank are less frequent in the top 200, no more then four entries each year.

The relation between research and citations is not linear (figure 2). It is the case that higher scores in research are predominantly paired with higher citation scores. However, many display a very high citation score to a much lower research score. The top 200 universities show a very high citation score, regardless of research. Top universities having a high citation score might not surprise many: the well known universities would be well published. Such attention could spark new research. But it could also be explained by the possibilities these universities might have to attract quality staff, improving the usability of the research done on these grounds.

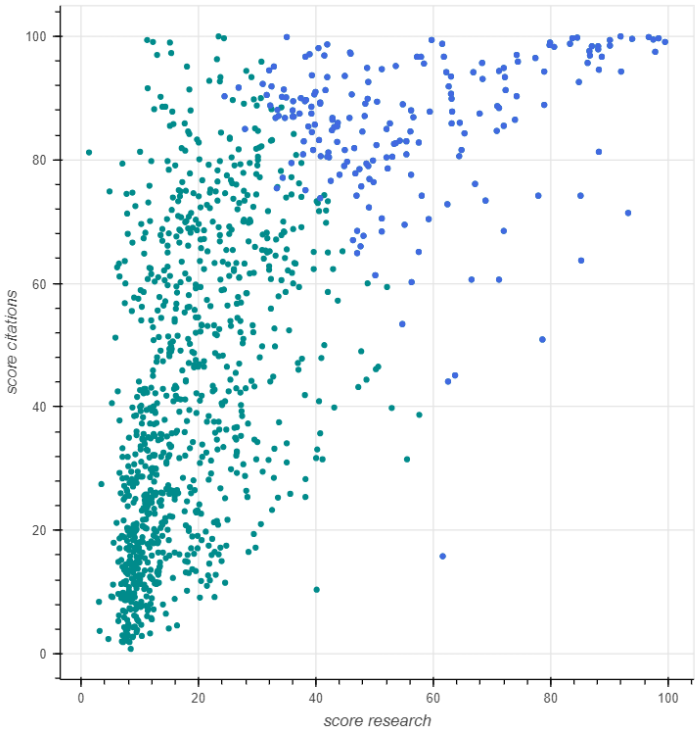

Figure 2: The research and citation scores of each university in the 2018 survey. Universities within the top 200 overall rank are shown in blue.

Looking at the countries that could provide the best education the total number of entries per country was collected. At a glance, certain countries dominated the charts. The United States occupies 157 places in the top 1000, against the meager single entry for Luxembourg. This isn't very peculiar, because of the size differences between the countries. Judging a countries quality of higher education, it is more insightful to look at the percentage of entries compared to the total number of universities in each country. In this case Luxembourg tops the charts, having only three universities. The United states plummets to the bottom 40%.

Consistency:

Of the roughly 1100 universities included in the 2018 survey, exactly 200 were already present in 2011. These institutions, being present in all years in between, served a good subset to analyze over a longer period.

It became apparent that for the vast majority of these universities, both the scores for industry income and for international outlook have been declining. In many cases by over 30 percent. This does not apply a hefty influence on their overall position, as discusses earlier, and this was shown in the majority of these not having dropped in the ranks. Comparing these figures to the extended list of later years however, shows a prevalence of higher scores by the newcomers. Still, there is no clear pattern in any other ranking of these universities, so no connection to their performance is apparent. It could very well be the nature of a company as THE that moves them to include these new entries on the basis of the high scores in international outlook or industry income. It might even make them more visible in the first place.

Comparing the subset of every year against their individual scores in years following, the data is very stable, almost identical over the three main categories. Even plotting these against additional provided measurements, like the ration of staff to students, did not show any noteworthy differences.

To further analyze the quality of universities their steadiness was taken into consideration. A heavily fluctuating rank for a given university shows a high possibility for a vastly different learning experience each year. For the 200 universities included in the 2011 survey their variance was calculated over the subsequent years, until 2018. The variance shows how stable the ranks were throughout the years, a high stability gives a low variance. When the stability decreases, the variance grows exponentially.

Example:

citation rank	2011	2012	2013	2014	2015	2016	2017	2018
Harvard University	8	7	10	9	12	5	8	8

$$Numberyears(n) : 8$$

Average:

$$\frac{8+7+10+9+12+5+8+8}{8} = 8,375$$

Variance:

$$\frac{\begin{array}{l}(8-8,375)^2+(7-8,375)^2+(10-8,375)^2+(9-8,375)^2\\ \quad +(12-8,375)^2+(5-8,375)^2+(8-8,375)^2+(8-8,375)^2\end{array}}{n-1}$$

$$= 4,2678571429$$

Clustering:

When evaluating data, it is interesting to look the different countries. All countries have a different culture, and have different morals and values. When taking decisions they take they take different aspects into consideration. This leads to the idea that this must also be true for universities. What will happen when plotting data points from countries? Will they cluster together per country, or will a top 1000 only attract one elite type of university?

Investigating this, a plot was made that interactively shows data of the 70 countries in the top 1000. A graph that is capable of selecting a country, and showing what data belongs to every university in that country. Using the multi_select feature of Bokeh, multiple locations can be viewed at the same time. This rises the opportunity for clear evaluation, without overcrowding.

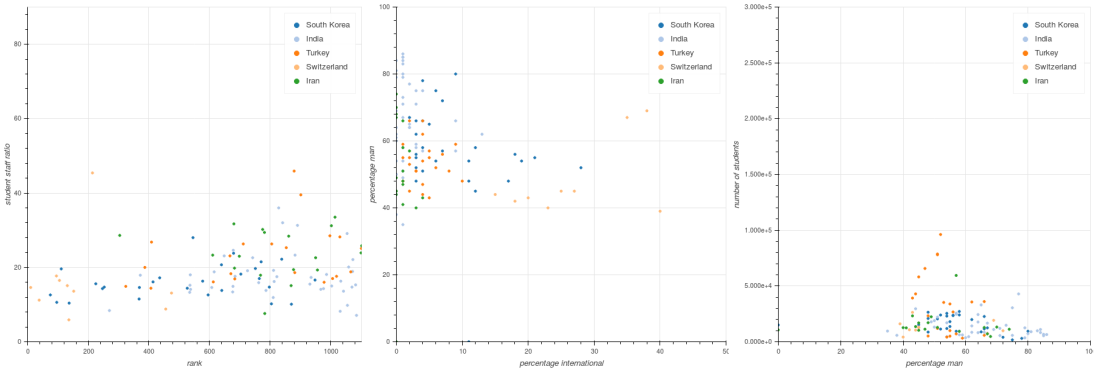

Figure 3: datapoints plotted in graphs, per country.

Evaluating different countries makes it apparent that most universities are quite alike in nature. Inspecting the figure above, there is no obvious clustering in the graphs. Ergo no reason to spend time programming machine learning algorithms regarding clustering. More so, because the dataset is very small, just 1103 datapoints, clustering is not a very favored technique.

In the middle graph, there is case of slight clustering. Not a lot, but enough to see a pattern. This graph plots the percentage of international students against percentage man. Taking India, for example. To some extent, all the light blue dots cluster in the top left corner. This means these universities have more men, and less international people. This stands opposed to Switzerland, which datapoints are located more in the middle of the graph. Remarkable, might be, that Iran is located on the middle left of the graph, whereas you might suspect it in the top left corner.

Hidden Meaning:

To investigate if you can estimate the appeal of universities to international students multivariate regression was applied, using data from the five categories used in the ranking formula and the percentage of international students. In the official ranking, 'international outlook' is not hugely taken into account. It is a mere 7.5% of the total ranking score. a high score for 'Teaching' and 'Research' would have a much bigger impact on the final ranking, each counting for 30%. Using international outlook, THE maps how involved a university is with international business, so not just the sheer amount of foreign students. Interesting would be, if you could still estimate the percentage of international students, considering all five elements of the overall score. Because in theory, this percentage is quite sidelined in the ranking.

A good way to examine this, is using regression. In particular logistic regression, not linear, because it is better when dealing with percentages. From the 2018 database, extract the target and a data frame of variables. In this case, these would respectively be the percentage of international students and the five scoring categories. Of course splitting both in a train set and a test set. A model fitted over this data predicts a percentage of international students, using the data of the ranking scores. In order to interpret how accurate the model is, the educated guesses from the algorithm with the actual percentages are compared. If the difference is small, the model is accurate.

After training the data on the training set, the time has come to check the test-set for accuracy. There are 2 ways this could be done; checking by how much the estimated value varies by percent from the original, or by checking whether the values fall within a threshold. Say the former is chosen, the value 2 would have a smaller accepted variance than 20 would. Hence why the latter check has been chosen, since bigger values do not have to be punished. MSE is not favourable, because there is nothing to compare to.

When comparing the calculated values with the estimated values the test set was used, to prevent issues like over-fitting. Taking a threshold of 5 percent, 57.6 percent of the test set is accurately predicted. This means more than half of the test set is estimated just 5 percent off the correct percentage. Looking at a threshold of 9%, 81.9 % is rightly predicted. Meaning there is definitely a trend in the ranking scores and the percentage of international students. The trend surely is not perfect, but considered percentage international students is not something actively assessed by THE, it is quite substantial. So yes, the international appeal is certainly visible in the ranking scores.

Conclusions:

It has proved difficult to deduce any explicit patterns from the used data-set that is not a direct result of the used formula. All categories within said formula showed patterns directly derived from their given weights. Furthermore, the size of the set also provided to little data to effectively carry out higher level analysis, like clustering. However, regression showed there is some value in the rankings regarding predicting external factors, like the attendance of, and perchance the appeal of international students.

The proposed hypotheses that analysis would reveal a treasure of insight is not necessarily deflated. However it does instigate the recommendation for a broader analyses of a much bigger data set.