University Ranking Analysis:

Comparing and Identifying Biases Among Different Ranking Systems

| Bingjie Chen  University of Washington  [cbjmandy@uw.edu](mailto:cbjmandy@uw.edu) | **Nelson D’souza**  University of Washington  [nelsonds@uw.edu](mailto:nelsonds@uw.edu) | Maria George  University of Washington  [gmaria@uw.edu](mailto:gmaria@uw.edu) | Cory Shyu  University of Washington  [coryshyu@uw.edu](mailto:coryshyu@uw.edu) |
| --- | --- | --- | --- |

# ABSTRACT

## Ranking is a key decision-making factor for students in their university selection process. Ranking methodology and criteria can vary significantly across different ranking systems and it can be difficult to understand and compare various ranking systems. Our study focuses on comparing university rankings by three reputable ranking systems and identifying potential biases among them. There are three major findings in our study. Firstly, volatility in the Times ranking systems is consistently much higher than others, which may indicate systematic biases. Secondly, we found strong evidence for biases in certain input factors, which exhibit similar pattern as the divergence in ranking between two ranking systems. Thirdly, we implemented our own clustering model to visualize university clusters based on difference in ranking and found evidence for potential biases in ranking universities in certain countries across all three ranking systems. Overall, we also found that top 40 universities tend to have the least bias in terms of magnitude as well as the number of disagreement, and universities beyond top 40 are subject to much larger biases.

## Author Keywords

University ranking, volatility in university ranks, input bias, K-Means clustering

# INTRODUCTION

University ranking has always been an important issue for schools, students, and academia worldwide. This can be observed from the increasing number of annual rankings published (Aguillo et al, 2010) as well as the number of research in the topic.  Rankings are used by the administrations to receive funding from the government. With increased research grants, the universities can then improve their research performance, which in turn can attract eminent scholars and researchers to the institutions. Most importantly, university ranking plays a key role in student’s application and enrollment decision-making process (Claassen, 2015).  Thus, universities also leverage their rankings as an advertising strategy in order to withstand the strong competition in the ‘market’ (Huang, 2011).

While there is little dispute on the importance of university ranking, there is no consensus in ranking criteria and methodology across different ranking systems. While some ranking systems resort to calculation of ranking through conducting surveys, other ranking systems utilize different procedures such as collecting data from independent reputable institutions. Moreover, there is discrepancy regarding the weightage given to similar criteria like citations and publications. Thus, it is difficult to analyze and compare university rankings across different ranking systems. How do we know that a particular ranking is genuine? How do we know that a particular ranking has not willfully promoted certain universities? How can we check for inherent bias in the rankings? Having relied on the university ranking systems ourselves when applying to a particular university inspired us to probe further into the bias among the various ranking systems.  We believe that such an analysis will help students to make informed decisions.

Through this research study, we intend to analyze potential biases in the rankings published by the following different ranking systems: Times Higher Education World ranking system (“Times”), Academic Ranking of World Universities, also known as the Shanghai ranking (“Shanghai”), and the Center for World University Rankings (“CWUR”). We performed volatility analysis to identify high variation in ranking within each ranking system between 2012 and 2015. Additionally, we compared similar ranking criteria between different ranking systems to identify potential biases in the ranking input factors. Lastly, we implemented similarity measures and performed clustering analysis to identify and visualize favoritism toward certain universities as well as universities in certain countries.

# Relevant works

In the field of university rankings, the number of qualitative studies overwhelmingly exceeds the quantitative studies (Claassen, 2015). In order to limit the scope of the study, we performed extensive research on the existing research studies which compare and contrast different ranking systems. The study by Huang (2011) compared the University rankings by three different ranking systems. The research method identified the common universities among all the ranking lists and took into account the difference in rankings. The difference was then normalized to the range [0,1]. The researchers then devised a new mathematical approach, similarity measure, for comparison. The study by Aguillo, Ilan, Levene and Ortega (2010), compares the annual rankings of five different systems using three complementary measures, the size of the overlap, the Speraman’s footrule, and the M measure.

In the recent study by Claassen (2015), Bayesian hierarchical latent trait model was used to measure university quality. By measuring the accuracy of each ranking systems as well as degree of bias towards universities in particular countries, the model developed in the study computes a common ranking system. Furthermore, the study by Soo & Dill (2005) pointed out that it is necessary to have an international census for the cross-national university ranking, which can minimize the cultural and political bias and help students as well as policy makers make decisions wisely. Specifically, the article pointed out that some ranking systems provide platforms for institutions to manipulate data to improve their reputations.

The similarity measure computed in this research study is inspired from the study by Huang (2011) and Aguillo et al (2010). However, we have combined the idea of difference in rankings and size of overlap to calculate the impartiality score. Also, the aforementioned studies compare the university rankings at the country level. In this study, in addition to comparing the rankings at the country level, we have performed a volatility analysis at the university level.

# Data

The university ranking dataset was obtained from kaggle.com, “a platform for predictive modeling and analytics competitions[[1]](#footnote-1).” For our analysis, we have used the following three datasets:

1. Times Higher Education World University Ranking

The dataset contains world university rankings by Times Higher Education. Founded in the UK in 2010, it is considered to be one of the most reputable ranking institution. It has the top 200 university rankings from 2011, the top 400 rankings from 2012 to 2015 as well as the top 800 rankings from 2016.

1. Academic Ranking of World Universities

This dataset contains world university rankings by ARWU also known as the Shanghai Rankings. Founded in China in 2003, it has the top 500 rankings from 2005 to 2015.

1. Center for World University Rankings

The CWUR ranking comes from Saudi Arabia, founded in 2012 and it contains the top 100 rankings from 2012 and 2013 as well as stop 1000 rankings from 2014 and 2015.

For further detail in ranking criteria and methodology, please refer to Appendix I.

These particular data sets require significant effort in data wrangling. There are two major challenges: matching university names across the three ranking systems and handling missing ranking information.

1. Matching University names

We decided to consider only the universities the universities within the top 100 ranking across all the ranking systems for the years 2012 to 2015. This decision was made based on the size of overlapping measure described in Aguillo et al (2010). In order to achieve this, the main challenge was matching the name of the university. The name of the same university was recorded in three different ways across different universities. For instance, Texas A&M University was shown as Texas A&M University, College Station in one ranking system, Texas A&M university in another system and Texas A & M university (note, space between A & M) in another system. Moreover, the same university was recorded differently within a ranking system as well because of the prevalence of special characters (e.g. University of Michigan, Ann Arbor, University of Wisconsin-Madison). We have made extensive effort in replacing university names so that we can perform join operations later on, which mostly required manual sorting and eyeballing. To ensure that no duplicate university names are used, we assigned a unique identification to each unique university for all three ranking systems.

1. Handling missing ranking information

As mentioned previously, CWUR provides all ranking information from 1 to 1000, while only a range of ranking is provided for rankings beyond top 100 and top 200 for Shanghai and Times, respectively. Additionally, Shanghai only provides ranking information up to 500 while Times provides up to 400 between 2012 and 2015. However, in order to compare rankings across three systems, it is necessary to impute ranks when the exact rank is missing. We used the midpoint when given a range for both Shanghai and Times. For universities that are present in one ranking system but absent in another, for the purpose of distance calculation we assigned a ranking of 700. Please refer to the discussion in the distance clustering section.

# volatility analysis

Ranks of individual universities often change, but the volatility in ranking can be quite different across different ranking systems. To answer the question whether bias exists within each ranking system, we analyzed the volatility in each ranking system for all top 100 universities between 2012 and 2015. We compared the volatility in rank across the three ranking systems. The variance in ranking for a ranking system is computed as the average variance of all the top 100 universities in the ranking system. This requires the assumption that rank of one university is independent from all other universities. It can be argued that a change in ranking for one university necessarily impacts ranking of some other universities from a ranking perspective. However, the criteria used to compute the ranking are independent from one institution to another. Ultimately it is the criteria that drive changes in rankings. Therefore, we will use the independent assumption and ignore the interaction terms between universities. The table below summarizes the volatility for each ranking system.

**Table 1. Volatility in Ranks**

| **Ranking System** | **Volatility** |
| --- | --- |
| Times | 20.178 |
| CWUR | 11.900 |
| Shanghai | 6.926 |

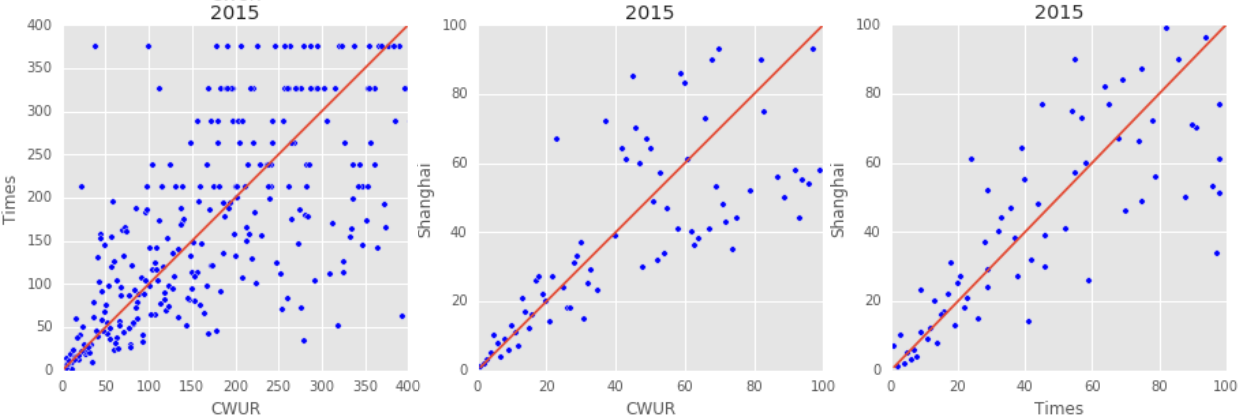
University ranks in Shanghai are clearly more stable over time relative to the other two, whereas Times shows high volatility in ranks among the three and CWUR is somewhere in the middle. Volatility is defined as the standard deviation in ranking. This implies that on average, we expect to see more than 30% of the top 100 universities change by more than 20 ranks from one year to another in the Times ranking system, versus 7 in Shanghai and 12 in CWUR.

One reason universities in “Shanghai” have relatively stable rankings is that this ranking system gives one third of the ranking weight to numbers of alumni and staff winning Nobel Prizes and Fields Medals. Because it is very rare to win Nobel Prizes and Fields Medals, we should expect very little change in the alumni and award criteria (Saisana, d’Hombres, Saltelli, 2010). In contrast, CWUR includes “quality of education” which represents faculty members who have won major international awards, prizes, and medals. The range of relevant awards, prizes and medals is broader than Nobel prizes and therefore, we should expect higher variation in the factor. Additionally, “quality of education” for CWUR is adjusted to the university's size, which allows the accomplishment of outstanding faculties in small-sized schools to have a significant influence on their university ranking.

For Shanghai, the research criterion focuses on natural and social science (Saisana, d’Hombres, Saltelli, 2010). In the science field, it can take much longer to produce meaningful research results. Therefore, we expect low throughput rate in scientific research, which leads to less ranking volatility for Shanghai. On the contrary, both CWUR and Times give more weight to an institution’s influence in industry, such as number of alumni assuming executive positions in top companies. And we expected those related indicators will change more frequently.

# pairwise ranking comparison

To visualize how the 3 ranking systems differ, we have examined the pairwise ranking scatterplot by year and found a common pattern across all years and all pairs. For universities that were ranked within the top 40, variations in ranking tend to be much smaller. As ranking increases, the variation in rankings widens.



**Figure 1: Pairwise Ranking**

Table 2: P-values of pairwise T-Test for university rankings beyond top 40

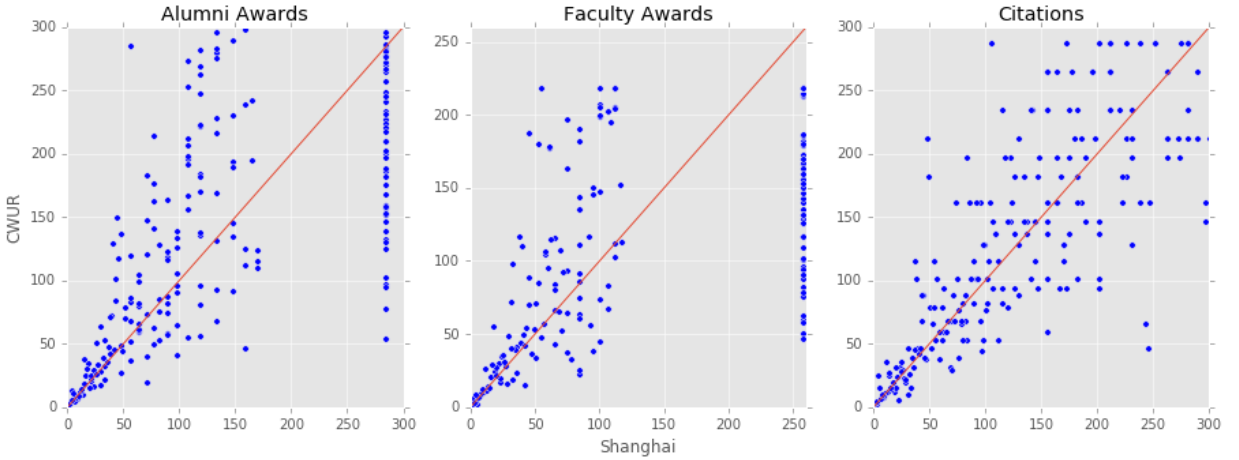
|  | CWUR/  Times | CWUR/  Shanghai | Times/  CWUR |
| --- | --- | --- | --- |
| 2013 | 0.0237 | 0.4515\* | 0.4702 |
| 2014 | 0.0038 | 0.1663\* | 0.0000\* |
| 2015 | 0.0000 | 0.0035 | 0.1768 |
| 2016 | 0.0000 | 0.0056 | 0.2407 |

## \*Number of records <100

None of the pairs shows statistically significant difference for top 40 universities. The T-test table above shows the p-values associated with the T-test for difference in ranking beyond top 40 between each pair. The ranking difference between CWUR and Time has mostly been statistically significant and also has deepened progressively between 2012 and 2015. In contrast, the difference in ranking between Shanghai and Times has mostly not been statistically significant. The difference between CWUR and Shanghai has been statistically significant in 2014 and 2015 but not in 2012 and 2013. CWUR’s ranking seems to deviate more from Shanghai and Times. However, given that the Shanghai ranking system only provides a range for all rankings beyond 100 and that the Times ranking system provides a range for all rankings beyond 200, for visualization purposes we have filled in the midpoint of ranking if given by a range. Therefore, the T-test is likely to underestimate the difference.

To explain the possible reasons for the above observation and also test our initial hypothesis that biases exist across different ranking systems, we compared similar input criteria (alumni award, faculty award, and citation) between Shanghai and CWUR in 2015. Please refer to the ranking criteria for a complete description in Appendix I. While these criteria are based on measures such as the number of alumni winning major awards, the number of faculty members winning major awards, and the number of citations of important publications, surprisingly, there is significant divergence in these input criteria between the two. To our surprise, the seemingly objective measures also show similar patterns as the ranking variable. Variation for the input data within the top universities are much smaller than the those ranked beyond the top 40.

The T-test for each of the factor has a corresponding p-value of close to 0, which indicates highly statistically significant differences between the two systems on similar ranking criteria. This finding may indicate potential biases in certain input variables, which ultimately drive the difference in ranking across all ranking systems.



**Figure 2: Input Factor Comparison Between Shanghai and CWUR**

# Clustering Analysis

We have seen in the previously section that there is statistically significant difference between some of the ranking systems. The next step is to develop a metric to visualize the distance between each ranking system and analyze the different clusters of universities based on distance. We chose the Manhattan distance as the distance measure. While there are multiple options to measure distance such as Euclidean, L-Norm (multi-dimensional), and Chebyshev, we chose Manhattan because it is more intuitive in the context of ranking.

There is a total of 142 universities present in the top 100 list of any of the 3 ranking systems. 20 universities within the top 100 list were not common to at least 2 of the ranking systems. For these universities, we imputed a dummy rank of 700. The rationale is to ensure that the ones which were not common to other ranking systems would be considered in the distance calculation and that we are able to visualize differences for such instances. Given that Shanghai only ranks universities up to 500 and Time ranks up to 400, our best guess for universities that are not present should be somewhere between 400 and 1000. Therefore, we chose 700 as the imputed rank.

The distribution of distance between each ranking system pair over 2014 and 2015 is shown in Appendix II.

All of the pairwise difference in ranking for 2014 and 2015 are almost similarly distributed, which suggests that each ranking system has remained stable over the 1-year period and so has the difference in ranks between each pairs. As shown in the histograms, the majority of the difference is between 0 and 100, while there is fat tail on the right indicating extreme differences.

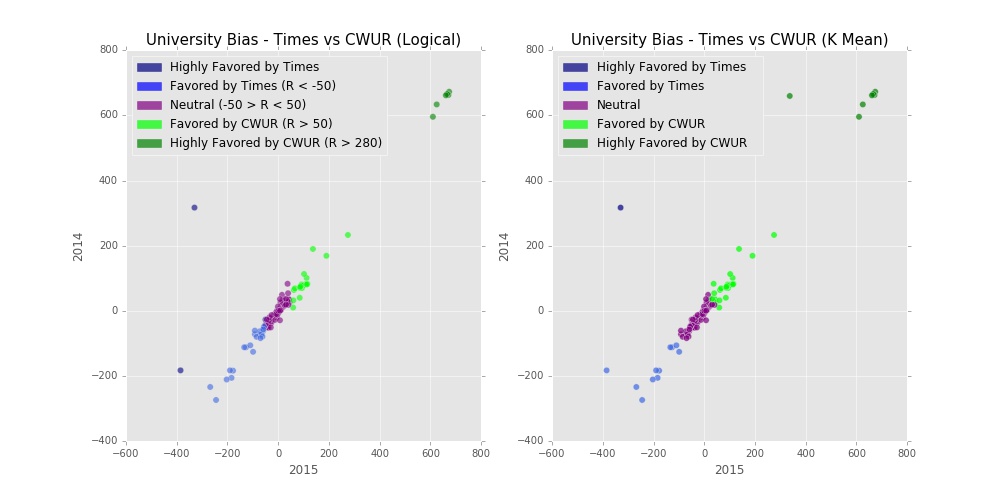
Given the above distribution, we hypothesized that 50 would be a good threshold for classification of bias. In other words, the universities which had a rank difference of greater than 50 would be considered to be biased (either towards or against) by a ranking system.

There is total of 4 types of disagreement when considering 3 ranking pairs, as described below:

1. **Consistent ranking** - This category represents universities which have consistent ranking across all the 3 ranking systems and none of the pairs shows significant difference.
2. **Somewhat consistent ranking** - This category represents universities which have consistent rankings across 2 pairs of rankings but not for the third one. These universities would thus have some bias (either towards or against) owing to one of the rankings systems.
3. **Somewhat inconsistent ranking** - This category represents universities which have inconsistent rankings across 2 pairs of rankings, but not for the third pair. These universities would thus have bias (either towards or against) owing to two of the ranking systems.
4. **Inconsistent ranking** - This category represents universities which have inconsistent ranking across all the 3 ranking systems

Before categorizing the universities, we need to first validate our hypothesis for difference threshold. We first examined the difference in ranking between each pair, then clustered the university based on the difference. For each pair there are 5 possible clusters. One cluster contains universities in which difference between each pair is below certain threshold. The second and third clusters represent universities in which difference between either of the other two pairs exceeds the threshold. Ideally we would only have these 3 clusters, however the impact of extreme distance due to absence of university in another system can influence the clusters significantly. Therefore, in the case wherein a university is missing in one ranking system in the pair, there will be 2 more clusters. The fourth cluster contains universities missing in one ranking system and the fifth cluster contains universities missing in other ranking system.

Next, we employed a semi-supervised learning method because we knew what our clusters should be but we did not know the optimal threshold for each of the clusters. Thus, before running a completely unsupervised K-Means algorithm, we used a validation set to judge how K-Means algorithm performed in comparison to our initial hypothesis. We first formed logical clusters for the pair Times and CWUR by assigning codes for each university such that universities with similar pairwise difference were assigned similar codes. In forming the logical clusters, we assigned all universities with rank difference within range of -50 and 50 to one cluster, the universities with rank between -50 to -316 and 50 to 316 to two other clusters respectively and finally we had two more clusters for universities with rank difference of greater than 316 and lesser than -316 respectively. 316 was chosen as the threshold because we have assigned 700 for missing data, which requires a larger threshold compared to 50. Thus, in all, we formed 5 logical clusters. All the clusters except the one which had universities in the range -50 to 50 is considered as biased. Then, we applied the general K-Means algorithm to cluster universities from Times and CWUR by keeping K=5 so that we could test our logical model. The results of both our logical clusters and the K-Mean clusters are shown as below. We performed the analysis for both 2014 and 2015 to control for volatility in a University ranking.

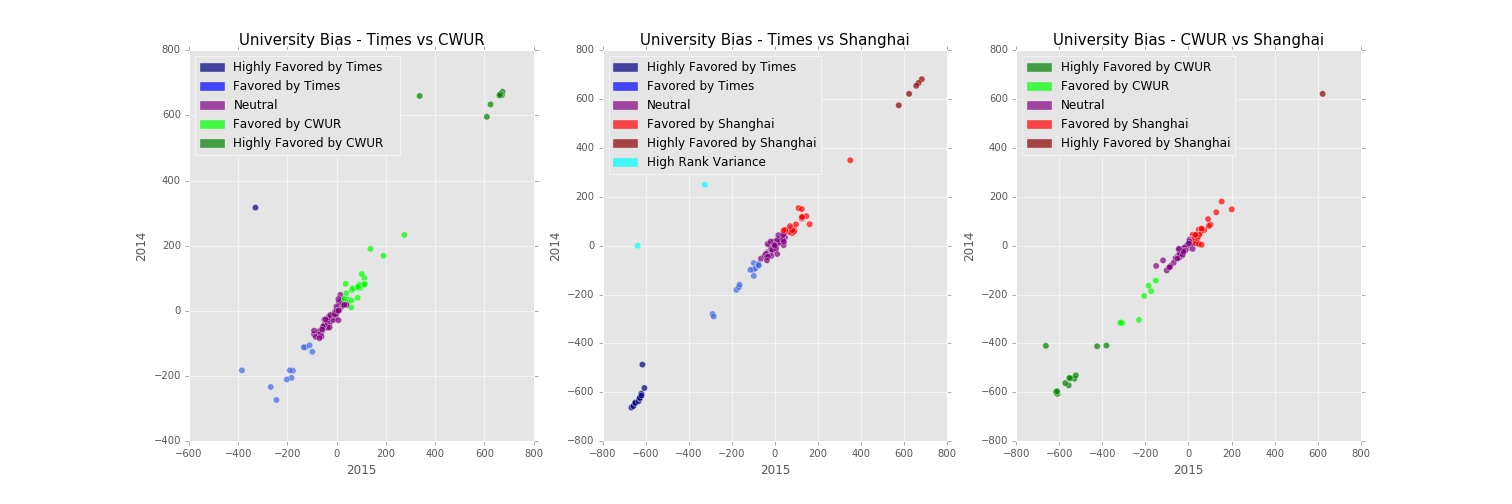


**Figure 3: Logical Model and K-Means Clustering Algorithm Comparison**

As shown in the above graph, our logical cluster is fairly similar to the K-Means algorithm. In order to check the accuracy of our logical model, we computed a confusion matrix:

We found via the confusion matrix that the K-Means algorithm had **82.23%** accuracy. The data in both cases were clustered into 5 groups. Firstly, we had a neutral cluster which contains universities which did not exhibit a high difference in ranking for both systems. We had two other clusters which had universities favored by Times and CWUR respectively and finally two more clusters which had universities missing in either of the ranking system pairs.

Then, we implemented the K-Means clustering process for all the 3 pairs and the results are as below:



**Figure 4: Pairwise Clustering**

These plots help us visualize the count of universities as well as the magnitude of inconsistency between any two ranking systems. They also help us understand the form in which significant difference exists. For all the 3 graphs, the universities contained in the clusters at the diagonally far end of the chart are the ones which are highly inconsistent with respect to a particular ranking system. The ones nearer to the center are universities that are fairly consistent with respect to a particular ranking system.

After putting universities into clusters, we aggregated the results to understand which of the pairs were clustered similarly and which ones are clustered differently across the ranking systems. We first assigned binary values to each of the 5 clusters present in each pair-wise ranking. A value of 1 denotes that the university was in a cluster which did not have significant inconsistencies among the ranking system pairs and a value of 0 signified that the university belonged to one of the other 4 clusters.

The binary values for all universities across the 3 pairs were then summed to calculate the Impartiality Score as given below. These scores correspond to the categories we had defined above wherein 3 signifies score of group of universities with consistent ranking and 0 signifies score of group of universities with inconsistent ranking.

**Table 3. Impartiality Scores**

| **Impartiality Score** | **Number of Universities** |
| --- | --- |
| 3 | 46 |
| 2 | 45 |
| 1 | 31 |
| 0 | 20 |

Universities having score of 3 are consistent because none of the ranking systems disagrees on their rank whereas universities having a score of 0 were inconsistent because none of the ranking systems agreed on their rank. Either way, the ranking systems showed systematic bias either against or towards a particular university. For visualization of university clusters, please refer to Appendix III.

The next logical question would be: what factors drive biases among the different systems? Given the country information, we analyzed for systematic biases against or towards universities from any particular country by a ranking system using the same threshold of 50. If a ranking system pair exhibits a difference in rank of 51 or more, we would consider it to be biased either towards or against (based on the sign of the rank difference) universities from that country. We also considered only those countries which had 3 universities or more in the aggregated top 100 rankings for the 3 ranking systems to eliminate false positives owing to small sample size. The results are shown in Appendix IV.

We concluded that Times (a ranking system from the UK) exhibits extremely high bias for universities from the Australia and South Korea relative to the other two systems, whereas CWUR (a ranking system from Saudi Arabia) exhibits extremely high bias for universities from Japan, Israel and South Korea and Shanghai (a ranking system from China) exhibits extreme high bias for universities from France, Israel and Sweden. On the other hand, Times exhibits extremely high bias against universities from Japan, France and Israel relative to the other two systems, whereas CWUR exhibits extremely high bias against universities from Germany, Australia and Sweden, and Shanghai exhibits extremely high bias against universities from South Korea.

# Conclusion

University ranking can significantly influence a student’s decision to select a university and thus play a critical part in the student’s life. With such high stakes, our aim from the beginning of this project was to provide students with a way of interpreting the major ranking systems and alert them against potential bias. In order to make informed decisions about a university, all the 3 ranking systems should be considered in aggregation rather than separately given that biases exist in at least one ranking system. Our major findings are as follows:

1. **Bias in input criteria** - We tend to attribute differences in ranking to different methodologies and different criteria. However, we found similar divergence even within similar criteria such alumni awards, faculty awards, and citation, which should be based on facts. This finding provides a warning against taking measuring criteria of ranking systems at face value.
2. **Volatility in Ranks** - We analyzed the volatility of each university to determine how stable the university rankings are for each ranking system over a given time period. This would help in identifying and guarding against universities with highly volatile rankings, or ranking systems which are inherently unstable. We found that Shanghai is the most stable among all three, while Times is almost three times as volatile compared to Shanghai and CWUR is close to 2 times as volatile.
3. **University clusters** - The pairwise university clusters we derived help us visualize the universities which are most similar or different to each other for a particular ranking system pair based on the Manhattan distance metric. We also used the Impartiality Score to visualize the number of instances in which the 3 ranking system pairs agree or disagree. This serves as another dimension which is complementary to the magnitude of distance. Through the visualization of university clusters, we have reaffirmed that universities within the top 40 tend to be more consistent, and as ranking increases, the number as well as the magnitude of disagreement increase.
4. **Country-Specific Bias** - We have explored how each ranking system tends to either favor or disapprove of universities from a particular country. We found that all three systems exhibit strong biases with respect to different set of countries.

# Limitations and Future Work

We have clustered top 100 universities from the 3 ranking systems, however, we could not fully utilize these clusters due to absence of common features across the 3 rankings. Country was the only common feature available and because of this, the scope of the bias analysis is rather limited.

The next steps for this study are increasing the number of common features and also incorporating a greater number of ranking systems. From Wikipedia, we could scrape a host of additional features such as university type (public/private), number of undergraduate students, founding year of university, etc. We can incorporate multiple dimensions in the distance calculation and moreover, identify which critical features drive the differences in ranking across different system pairs. By considering a greater number of ranking systems, we can validate or adjust our model accordingly.

With our study, we tried to develop a methodology to compare rankings from disparate ranking systems and identify potential biases. We have made our findings available on our website[[2]](#footnote-2).  The website also contains interactive visualizations to explore our results in further detail. We hope people find it useful, especially those in the critical phase of school search process.

**REFERENCES**

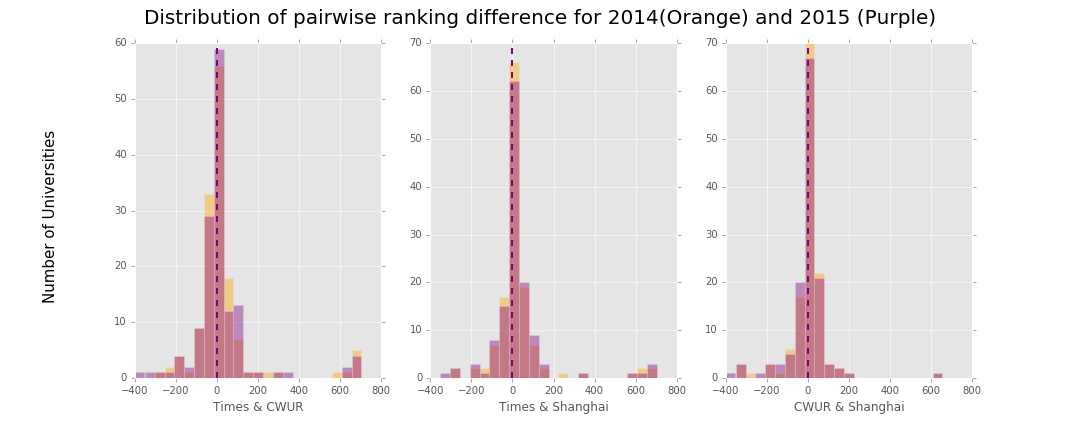
1. Aguillo, I. F., Bar-llan, J., Levene, M., & Ortega, J. L. (2009, July 17). Comparing University Rankings. <https://www.researchgate.net/publication/220365497_Comparing_university_rankings>
2. Claassen, C. (2015). Research Repository. <http://repository.essex.ac.uk/14726/>
3. Huang, M. (2011, June). A Comparison of Three Major Academic Rankings for World Universities: From a Research Evaluation Perspective. <http://jlis.lis.ntu.edu.tw/article/v9-1-1.pdf>
4. Soo, M. & Dill, D.D. (2005). Academic quality, league tables, and public policy: A cross-national analysis of university ranking systems. Higher Education. 49: 495- 533.
5. Saisana, M., d’Hombres, B., & Saltelli, A. (2010) Rickety numbers: Volatility of University rankings and policy. Research Policy. 40. 165 - 177. Elsevier B.V.

**Appendix**

1. **Criteria and methodologies**

|  |  |  |  |
| --- | --- | --- | --- |
| Ranking System | Ranking Criteria | Weight | Description |
| Times | Teaching | 30% | University score for teaching (the learning environment) |
|  | Research | 30% | University score for research (volume, income and reputation) |
|  | Citations  International | 30%  7.5% | University score for citations (research influence)  University score international outlook (staff, students, research) |
|  | Income | 2.5% | University score for industry income (knowledge transfer) |
| Shanghai | Award | 20% | Staff of an institution winning Nobel Prizes and Fields Medals |
|  | HiCi | 20% | Highly cited researchers in 21 broad subject categories |
|  | N&S | 20% | Papers published in Nature and Science |
|  | PUB  Alumni | 20%  10% | Papers indexed in Science Citation Index-expanded and Social Science  Citation Index  Alumni of an institution winning Nobel Prizes and Fields Medals |
|  | PCP | 10% | Per capita academic performance of an institution |
| CWUR | Quality of Education | 25% | The number of a university's alumni who have won major international awards, prizes, and medals relative to the university's size |
|  | Alumni Employment | 25% | The number of a university's alumni who have held CEO positions at the world's top companies relative to the university's size |
|  | Quality of Faculty | 25% | The number of academics who have won major international awards, prizes, and medals |
|  | Publications | 5% | The number of research papers appearing in reputable journals |
|  | Influence | 5% | The number of research papers appearing in highly-influential journals |
|  | Citations | 5% | The number of highly-cited research papers |
|  | Broad Impact | 5% | The university's h-index |
|  | Patents | 5% | The number of international patent filings |

1. **Distribution of ranking difference**



1. **University Clusters**



1. **Biases against or for specific countries**

1. <https://www.kaggle.com/mylesoneill/world-university-rankings> [↑](#footnote-ref-1)
2. <http://univrankinganalysis.github.io/> [↑](#footnote-ref-2)