CSE564 Visualization Project Report

Zhenxun Zhuang, 111131765

May 14, 2017

1 Introduction

Although the credibleness and validity of university rankings have been widely debated for years, those rankings are still considered to be an indispensable and important indicator when one is applying for undergraduate or graduate schools. Most of the time, we only care about the final rank, instead of how this number is computed. However, the ranking mechanism under the hoods turns out to be a hard, political, and controversial practice. Different ranking systems usually emphasize different parts and use different ranking mechanisms. Thus, it is not surprising that there are hundreds of different national and international university ranking systems, many of which disagree with each other. Therefore, before referencing a rank, we should be aware of which factors are included and how they affect the results.

In this project, I do not intend to develop a novel ranking system; instead, I try to discover interesting patterns from an existing system, which is one of the most influential and widely observed university measures – the Times Higher Education World University Ranking[1]. Founded in the United Kingdom in 2010, it is praised for having a new improved methodology; nevertheless, it has also been criticized for its commercialization and for undermining non-English-instructing institutions.

According to the official explanation of their methodology[2], the rankings are calculated as follows:

- **Teaching (the learning environment): 30%**
  - reputation survey: 15%
  - staff-to-student ratio: 4.5%
  - doctorate-to-bachelor's ratio: 2.25%
  - doctorates awarded-to-academic staff ratio: 6%
  - institutional income: 2.25% (*It indicates an institution's general status and gives a broad sense of the infrastructure and facilities available to students and staff.*)
- **Research (volume, income and reputation): 30%**
  - reputation survey: 18%
  - research income: 6%

[1] `https://www.timeshighereducation.com/world-university-rankings`
[2] `https://www.timeshighereducation.com/news/ranking-methodology-2016`

  - research productivity: 6% (*It counts the number of papers published in the academic journals indexed by Elsevier's Scopus database per scholar, thus reflects the university's ability to publish papers in quality peer-reviewed journals.*)
- **Citations (research influence): 30%**
- **International outlook (staff, students and research): 7.5%**
  - International-to-domestic-student ratio: 2.5%
  - International-to-domestic-staff ratio: 2.5%
  - International collaboration: 2.5% (*It calculates the proportion of a university's total research journal publications that have at least one international co-author and reward higher volumes.*)
- **Industry income (knowledge transfer): 2.5%** (*This category looks at how much research income an institution earns from industry, thus suggests the extent to which businesses are willing to pay for research and a university's ability to attract funding in the commercial marketplace.*)

So, to check whether the actual rankings follow the above methodology, we can utilize Principal Component Analysis.

Apart from verifying the methodology, there are many other things to do, among which I plan to do the following:

1. Classify different universities according to their respective attributes using t-distributed Stochastic Neighbor Embedding, Multi-Dimensional Scaling, and Parallel Coordinate Plot.
2. Study the change in rankings and specific attributes of different colleges over time.
3. Compare the number of top universities between different countries.

This problem looks very interesting to me, as I am always wondering what makes a good university, in other words, what do good univerisities have in common. Actually, when I was applying for graduate schools, I felt really upset about all those rankings, how can a school be characterized by a single number? If I had been given a detailed visualized chart like Parallel Coordinate Plot, I could have obtained a better knowledge of which are my dream schools. I am pretty sure I am not the only one frustrated by those rankings, so I think this work shall prove instructive and helpful.

2 Data Preparation

2.1 Dataset Description

The dataset is obtained from Kaggle[3]. It contains the Times Higher Education World University Ranking from 2011 to 2016. There are 2603 data items, one per row representing the statistics of a university in one year, along with 14 columns, each represents an attribute. Columns are described below:

1. **world_rank** - world rank for the university.

[3] `https://www.kaggle.com/mylesoneill/world-university-rankings`

2. **university_name** - name of university.
3. **country** - country of each university.
4. **teaching** - university score for teaching (the learning environment).
5. **international** - university score international outlook (staff, students, research).
6. **research** - university score for research (volume, income and reputation).
7. **citations** - university score for citations (research influence).
8. **income** - university score for industry income (knowledge transfer).
9. **total_score** - total score for university, used to determine rank.
10. **num_students** - number of students at the university.
11. **student_staff_ratio** - Number of students divided by number of staff.
12. **international_students** - Percentage of students who are international.
13. **female_male_ratio** - Female student to Male student ratio.
14. **year** - year of the ranking (2011 to 2016 included).

Attribute 1 will be coded into visual variables like size or color.

Attribute 2 and 3 are text data, thus it's more appropriate to use them as tool tips.

Attribute 4 - 8 are major factors influencing rankings.

Attribute 9 is redundant as it's just a linear combination of attribute 4 - 8 following predefined rules. Moreover, its sole purpose is to generate the rankings, namely attribute 1. Thus, it is discarded.

Attribute 10 - 13 are additional factors which may be related to ranking. Initially, they are in non-numeric forms, hence I transformed them to make them computable.

Attribute 14 shows that the number of universities included in this ranking system is increasing, from only 200 in 2011, to 400 in 2012, and finally 800 in 2016.

2.2 Filling Missing Values

In the original dataset, there are a few missing values. They should be considered and filled before applying any data analytic method.

For those data items which miss a single value among attribute 4 - 9, we can approximate that value by reversely using the ranking generation rules. As said above, the *total_score* is a linear combination of attribute 4 - 8; thus, it is like a parity bit which can reconstruct one missing value. However, for those data items which have more than one missing values among attribute 4 - 9, it is impossible to estimate those values, and I have to discard them.

For those data items which miss a single value among attribute 10 - 13, we can fill in using the average value of all other data items. Also, if one data item missed more than two values, I would remove them since they have lost too much information.

3 Anticipated Work

3.1 Methodology Verification

Principal Component Analysis is the perfect tool to verify whether they have used the methodology that was published in their website. By applying PCA, we can find out which attribute is the most influential, and which is the least. Also, we can come up with an optimal combination of all attributes which can account for most variances in the data; this can be done by selecting the principal component with the highest eigenvalue. Moreover, we can also examine the relationship between rankings and other unaccounted attributes like *female_male_ratio*.

3.2 Clustering

Which universities are alike? Do good universities have similar patterns? These questions can be solved by using following clustering techniques:

- **t-distributed Stochastic Neighbor Embedding**
  t-SNE converts affinities of data points to probabilities. The affinities in the original space are represented by Gaussian joint probabilities and the affinities in the embedded space are represented by Students t-distributions. The Kullback-Leibler (KL) divergence of the joint probabilities in the original space and the embedded space will be minimized. This allows t-SNE to be particularly sensitive to local structure and has a few other advantages over existing techniques:
  - Revealing the structure at many scales on a single map
  - Revealing data that lie in multiple, different, manifolds or clusters
  - Reducing the tendency to crowd points together at the center
- **Multi-Dimensional Scaling**
  Using MDS, we can map high dimensional data into two dimensions, thus obtain a straightforward and immediate view of how similar one university is to others. And by applying different similarity metrics, we can generate different clusters and gain different views.
- **Parallel Coordinate Plot**
  PCP is a popular way of analyzing multivariate data as well as visualizing high-dimensional geometry. By drawing n parallel, vertical, and equally spaced lines, a point in n-dimensional space can be represented as a polyline with vertices on each parallel axis, and the position of the vertex on the i-th axis corresponds to the i-th coordinate of the point. We can explore more relationships through reordering axes.

3.3 Variations over time

Since the dataset contains statistics ranging from 2011 to 2016, it is of great interest to reveal the changing trends of the rankings and specific attribute values of different universities

over time. We may find out what leads to the increase or decrease of the ranking of a specific university.

3.4 Comparison of Top Universities among Countries

The following questions can be answered in this section:

- Which country contributes most top universities? (Though the answer is quite simple, the United States of course...)
- Which country grows most rapidly in the number of top universities, which indicates the fast development of education and economy?
- Reversely, which country is facing the severest education recession?

4 Preliminary Results

4.1 Homepage

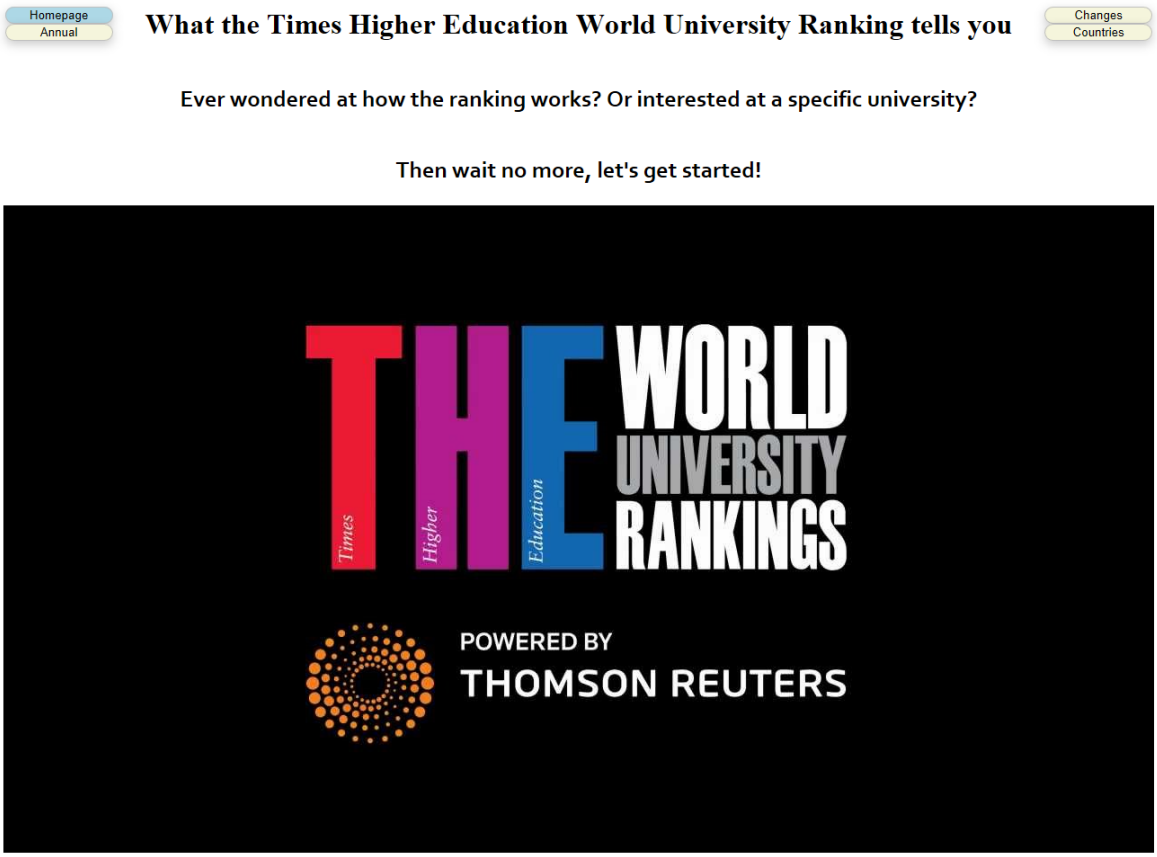


Figure 1: Homepage

4.2 Methodology Verification

From Figure 2, we can see that the first component alone is able to discriminate good universities from average ones quite satisfactorily. This means the ranking is exactly based on the linear combination of those afore-mentioned attributes.

Combined with Figure 3, it is clear that they are using something very similar to the methodology published on their website. In the first component, teaching and research play much more important roles than any other attribute. The only discrepancy is that, according to the methodology, citations should be as important as teaching and research whereas it's not the case in the results. This shows that they propably changed the methodology a bit when calculating rankings.

It is not surprising that rankings are positively correlated with the international students percentage, because it is actually part of international outlook. What's interesting is that a higher female male ratio should make a university score less. It is something to be clarified by the Times system. Finally, the number of students seems to be not very important to a university's ranking, what really matters is the student to staff ratio.

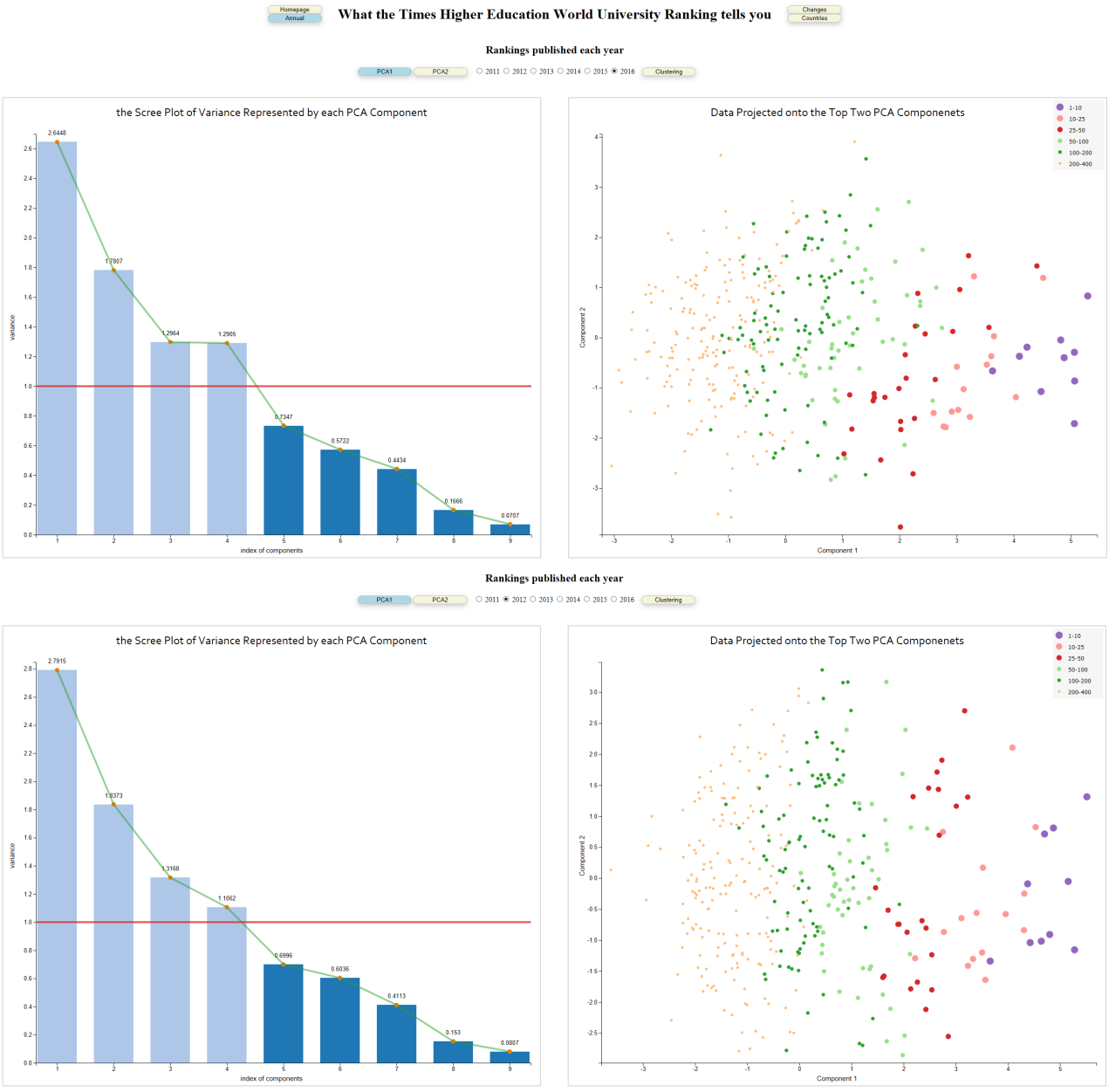
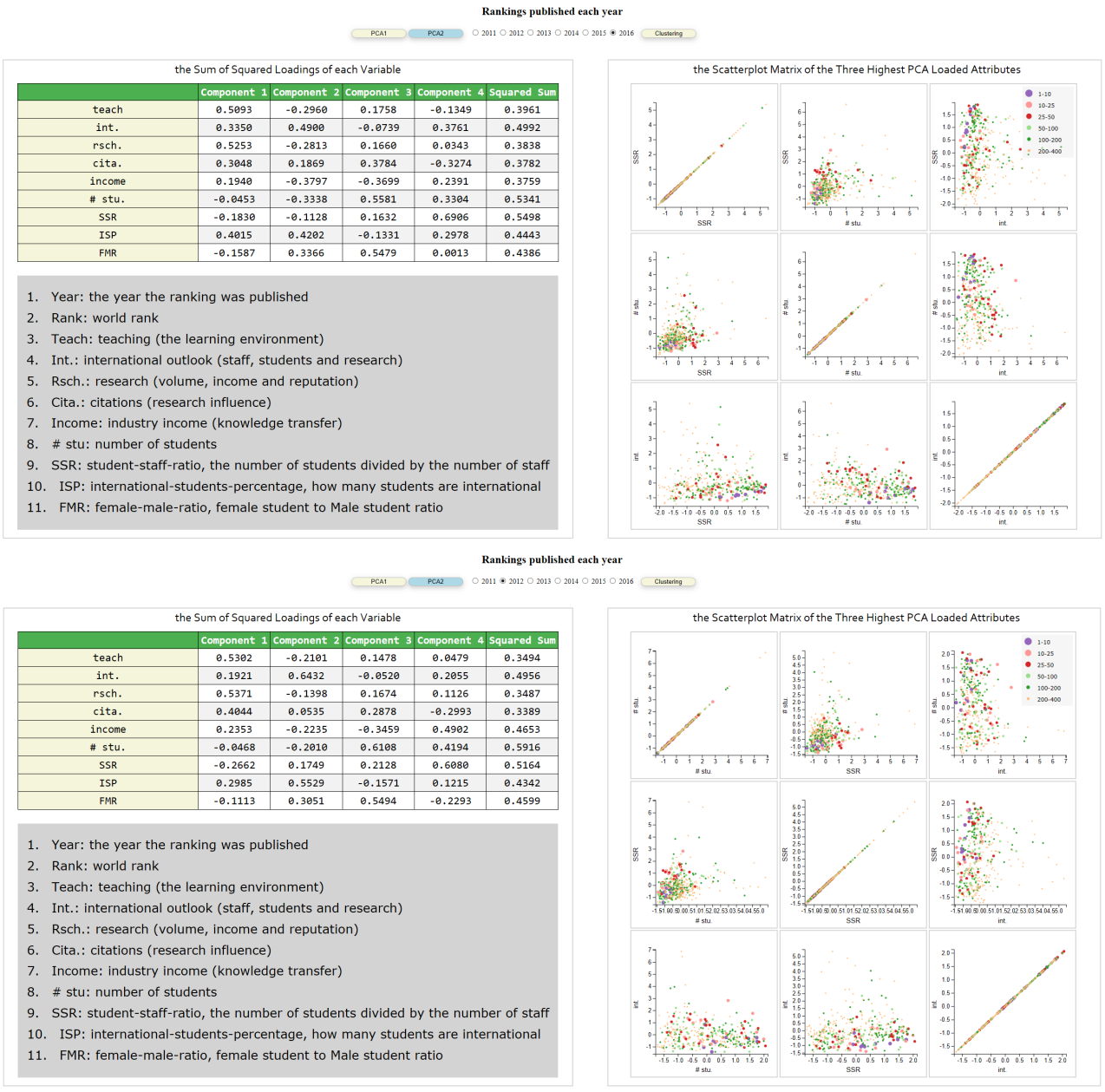


Figure 2: PCA components scree plot and projecting data onto top two PCA component.



	Component 1	Component 2	Component 3	Component 4	Squared Sum
teach	0.5093	-0.2960	0.1758	-0.1349	0.3961
int.	0.3350	0.4900	-0.0739	0.3761	0.4992
rsch.	0.5253	-0.2813	0.1660	0.0343	0.3838
cita.	0.3048	0.1869	0.3784	-0.3274	0.3782
income	0.1940	-0.3797	-0.3699	0.2391	0.3759
# stu.	-0.0453	-0.3338	0.5581	0.3304	0.5341
SSR	-0.1830	-0.1128	0.1632	0.6906	0.5498
ISP	0.4015	0.4202	-0.1331	0.2978	0.4443
FMR	-0.1587	0.3366	0.5479	0.0013	0.4386

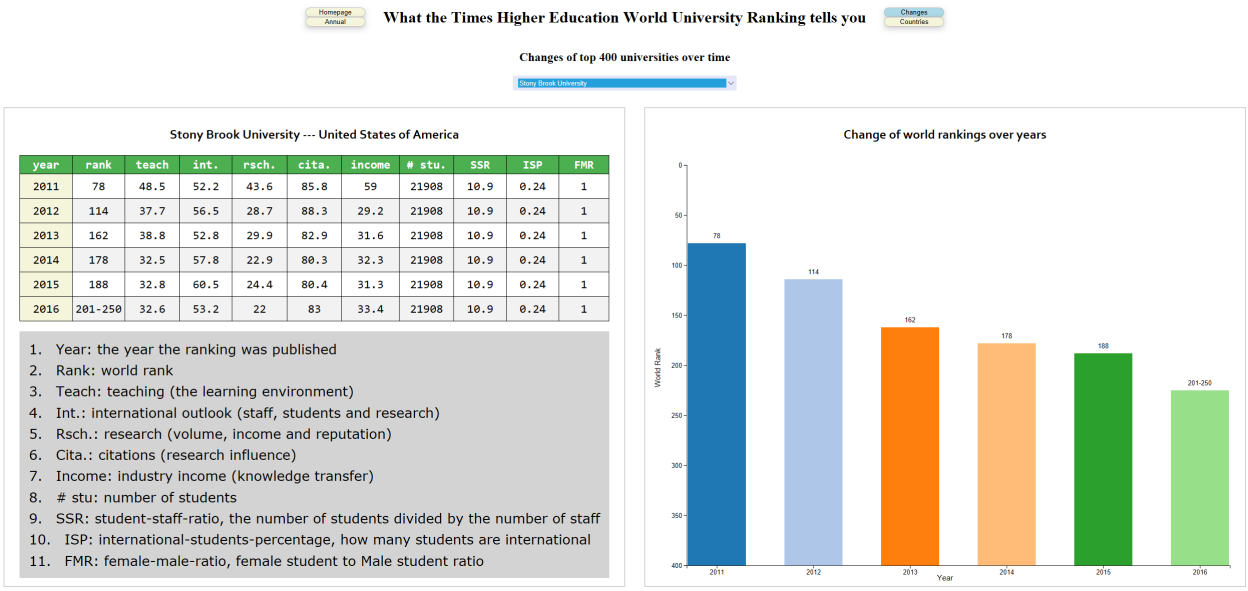
1. Year: the year the ranking was published
2. Rank: world rank
3. Teach: teaching (the learning environment)
4. Int.: international outlook (staff, students and research)
5. Rsch.: research (volume, income and reputation)
6. Cita.: citations (research influence)
7. Income: industry income (knowledge transfer)
8. # stu: number of students
9. SSR: student-staff-ratio, the number of students divided by the number of staff
10. ISP: international-students-percentage, how many students are international
11. FMR: female-male-ratio, female student to Male student ratio

	Component 1	Component 2	Component 3	Component 4	Squared Sum
teach	0.5302	-0.2101	0.1478	0.0479	0.3494
int.	0.1921	0.6432	-0.0520	0.2055	0.4956
rsch.	0.5371	-0.1398	0.1674	0.1126	0.3487
cita.	0.4044	0.0535	0.2878	-0.2993	0.3389
income	0.2353	-0.2235	-0.3459	0.4902	0.4653
# stu.	-0.0468	-0.2010	0.6108	0.4194	0.5916
SSR	-0.2662	0.1749	0.2128	0.6080	0.5164
ISP	0.2985	0.5529	-0.1571	0.1215	0.4342
FMR	-0.1113	0.3051	0.5494	-0.2293	0.4599

1. Year: the year the ranking was published
2. Rank: world rank
3. Teach: teaching (the learning environment)
4. Int.: international outlook (staff, students and research)
5. Rsch.: research (volume, income and reputation)
6. Cita.: citations (research influence)
7. Income: industry income (knowledge transfer)
8. # stu: number of students
9. SSR: student-staff-ratio, the number of students divided by the number of staff
10. ISP: international-students-percentage, how many students are international
11. FMR: female-male-ratio, female student to Male student ratio

Figure 3: The scatter plot matrix of the three highest PCA loaded attributes..

4.3 Variations Over Time



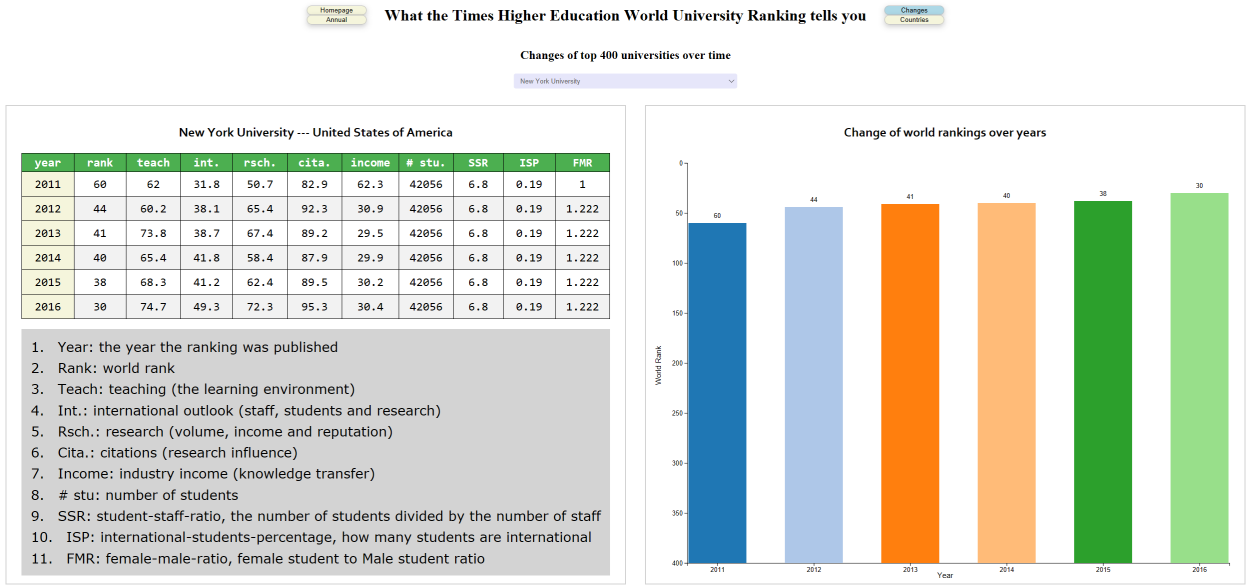


Figure 4: Changes of the performance of Stony Brook University VS. NYU over years.

From Figure 4, we may see that the difference between Stony Brook University and New York University is increasing dramatically. The major indicators in which SBU is getting lower and lower scores are teaching and research, so it's time to find out the reason and try to do better.

4.4 Comparison of Top Universities among Countries

From Figure 5, the United States of American takes up the largest share, contributing around 25% of the total number. Then comes the United Kindom, Germany, Australia, and Canada. These five countries are home to more than half of the top 400 universities in the world. Interestingly, four of them use English as official, or a de facto official language. No wonder the Times ranking system is suffering from the criticism of undermining non-English-instructing institutions. Frankly speaking, though, these countries had and are still having most of greatest universities in this planet. However, the Times ranking system ought to pay more attention to non-English-speaking countries where excellent colleges are emerging rapidly these years.

Homepage Annual

What the Times Higher Education World University Ranking tells you

Changes Countries

Comparison of the number of top universities in each country

○ 2011 ○ 2012 ○ 2013 ○ 2014 ○ 2015 ○ 2016 ◉ Summary

country	2011	2012	2013	2014	2015	2016	country	2011	2012	2013	2014	2015	2016
Australia	7	21	19	19	20	22	Italy	0	14	14	15	17	18
Austria	2	5	5	6	5	5	Japan	5	16	13	11	12	6
Belgium	2	7	7	7	7	7	Netherlands	10	13	13	13	13	13
Brazil	0	2	2	2	2	2	New Zealand	1	5	6	5	5	3
Canada	6	17	18	18	18	18	Norway	1	4	4	4	4	4
Chile	0	1	0	0	1	0	Poland	0	1	2	1	1	0
China	6	10	9	10	11	8	Portugal	0	4	3	2	2	1
Colombia	0	0	1	1	1	0	Republic of Ireland	2	5	5	5	5	6
Czech Republic	0	1	1	1	1	1	Russian Federation	0	2	2	1	2	5
Denmark	3	5	5	5	5	6	Saudi Arabia	0	0	1	2	0	1
Egpyt	1	1	0	0	0	0	Singapore	2	2	2	2	2	2
Estonia	0	1	1	1	0	1	South Africa	1	3	4	3	3	3
Finland	1	5	5	4	6	6	South Korea	4	7	6	7	9	8
France	4	8	11	11	11	15	Spain	2	8	7	9	6	5
Germany	13	21	25	24	27	34	Sweden	5	10	10	9	9	11
Greece	0	1	1	1	1	1	Switzerland	6	7	7	8	8	9
Hong Kong	4	6	6	6	6	6	Taiwan	4	8	7	8	6	4
Iceland	0	1	1	1	1	1	Thailand	0	1	1	1	1	0
India	0	1	3	3	4	2	Turkey	1	3	4	4	5	2
Iran	0	1	1	1	2	0	United Kingdom	27	49	46	47	42	44
Israel	0	4	3	3	3	3	United States of America	69	100	96	98	99	109

Note: the ranking for 2011 only includes 200 universities, whereas in other years there are 400.

Figure 5: The summary of the number of top 400 universities each country has from 2011-2016.

From Figure 6 and 7, we can see the number of top univerisities located in France is increasing significantly, nearly doubled; whereas Japan and Portugal are possibly experiencing educational recessions. Apart from those three countries, however, there is no other big change from 2012 to 2016. This shows that education quality of a country is quite stable. It takes time to develop a good education system, but once established, it will last a long time unless some radical social changes happen. For example, before World War II, Germany had generated more Nobel laureates in scientific fields than any other nation, and compelled as the best country in the natural sciences. However, the tensions heralding the onset of WWII spurred sporadic but steady scientific emigration, or Brain Drain, in Europe. Many of these emigrants were Jewish scientists, fearing the repercussions of anti-Semitism, especially in Germany and Italy, and sought sanctuary in the United States. One of the first to do so was Albert Einstein in 1933. At his urging, and often with his support, a good percentage of Germany's theoretical physics community, previously the best in the world, left for the US. As a result, in the post-war era the US was left in a position of unchallenged scientific leadership. By the mid-1950s

the research facilities in the US were second to none, and scientists were drawn to the US for this reason alone. The changing pattern can be seen in the winners of the Nobel Prize in physics and chemistry. During the first half-century of Nobel Prizes, from 1901 to 1950, American winners were in a distinct minority in the science categories. Since 1950, Americans have won approximately half of the Nobel Prizes awarded in the sciences.

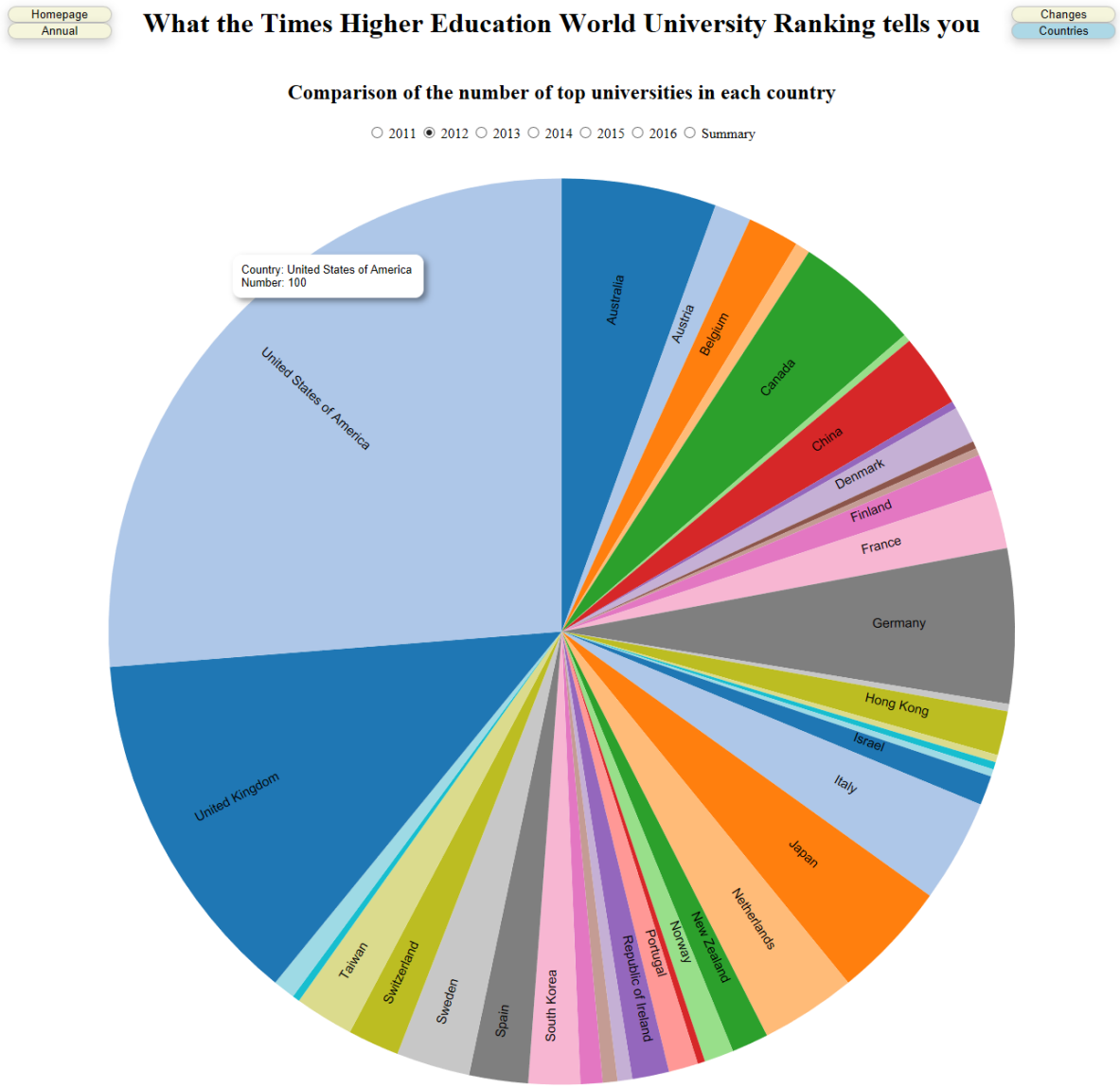


Figure 6: Pie chart showing the constitution of top 400 universities in 2012.

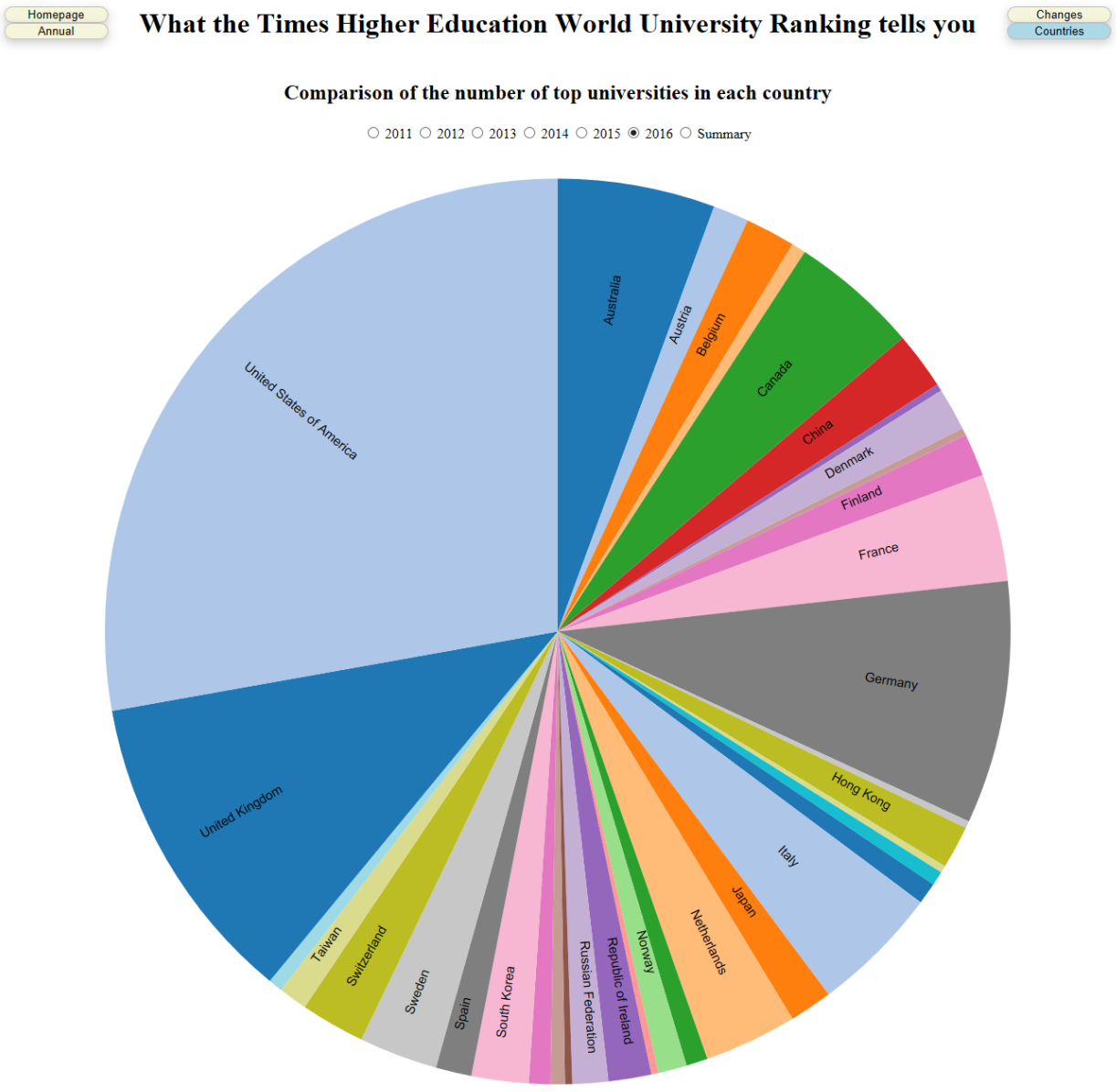


Figure 7: Pie chart showing the constitution of top 400 universities in 2016.

5 Supplementary results since the prelim report

5.1 Comparison of Top Universities among Countries: World Map

Figure 8 gives a more straight forward view of the distribution of top universities over the world. As can be seen easily, most top universities are located in Europe and North America.

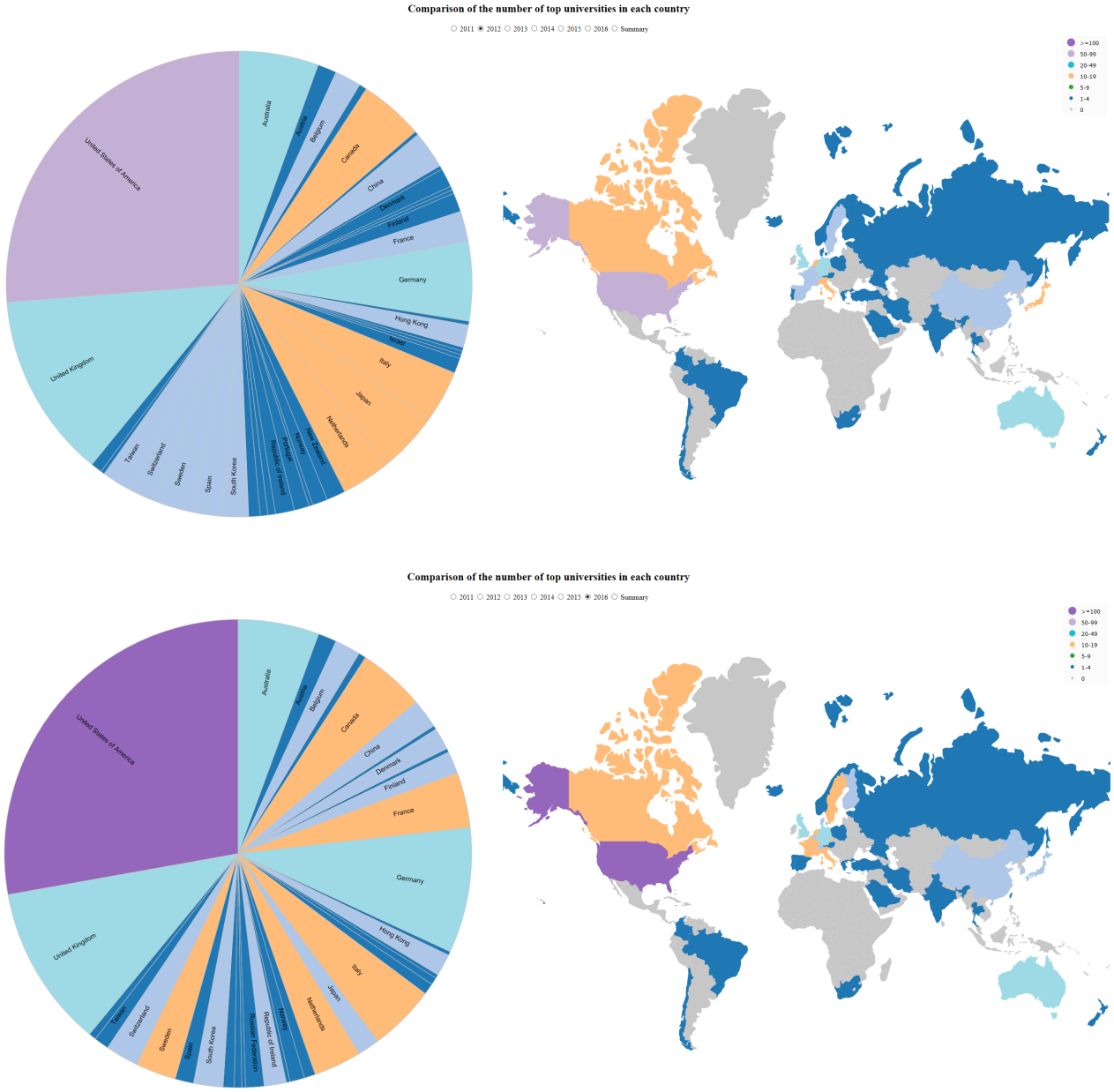


Figure 8: Pie chart and world map showing the distribution of top 400 universities.

5.2 Finding your dream universities using Parallel Coordinates

By applying the brushing technique on the parallel coordinate plot, it is very convenient to find your dream universities. You only need to set the range of each indicator, and those qualified universities will be highlighted. Moreover, by linking the parallel coordinate plot with the results of MDS and Isomap, you can also obtain a view of where those universities locate with respect to other universities. In addition, you can rearrange the relative order of these axes to attain a better view of data.

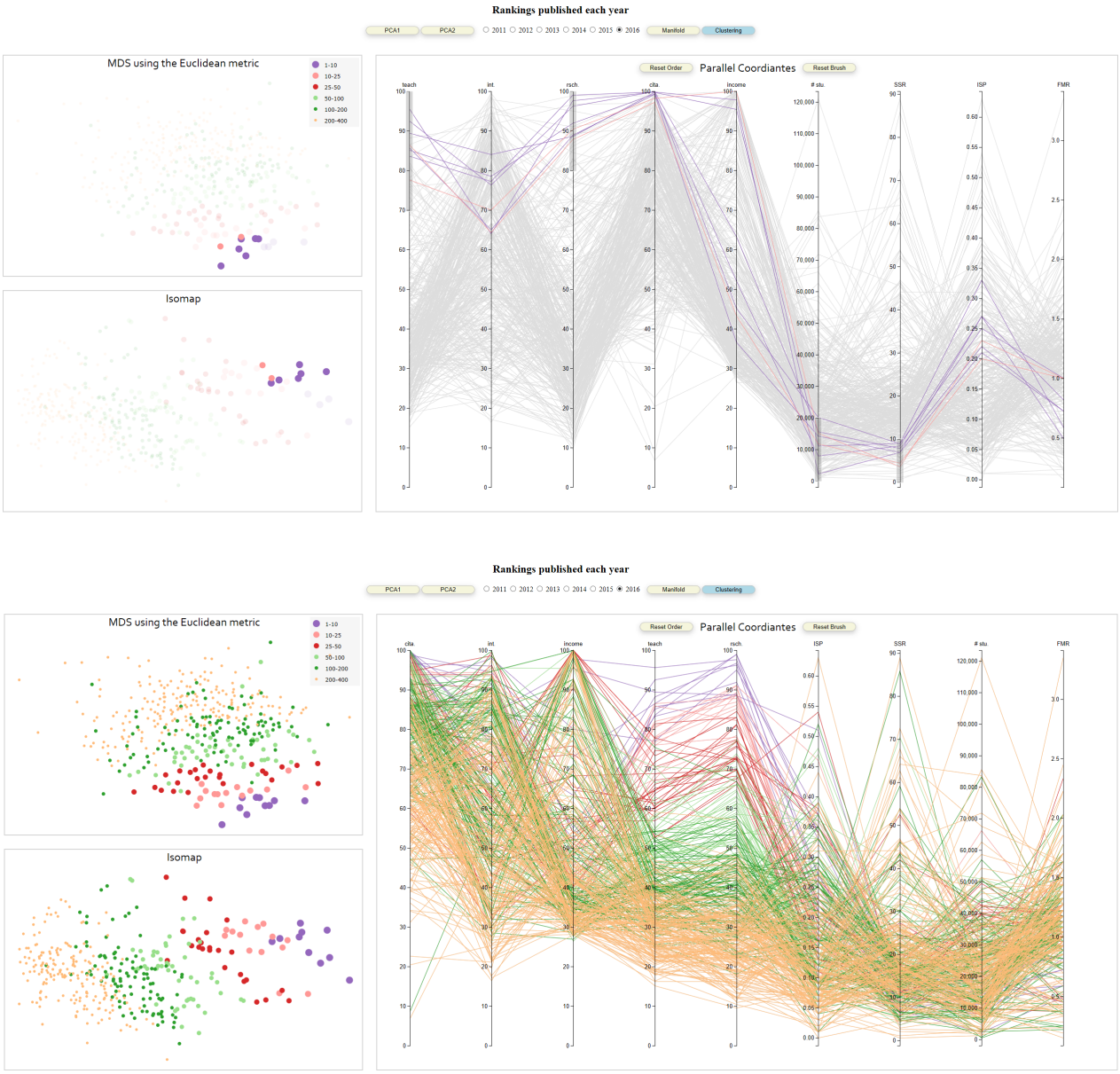


Figure 9: Filtering schools using parallel coordinate plot.

5.3 Finding the best clustering method

5.3.1 Which metric of MDS is the best for this dataset

From Figure 10, we can see that the Euclidean metric is much more suitable for our case than either the Pearson correlation metric or the Cosine metric, as data are clustered best under the Euclidean metric. Consequently, the Euclidean metric is chosen.

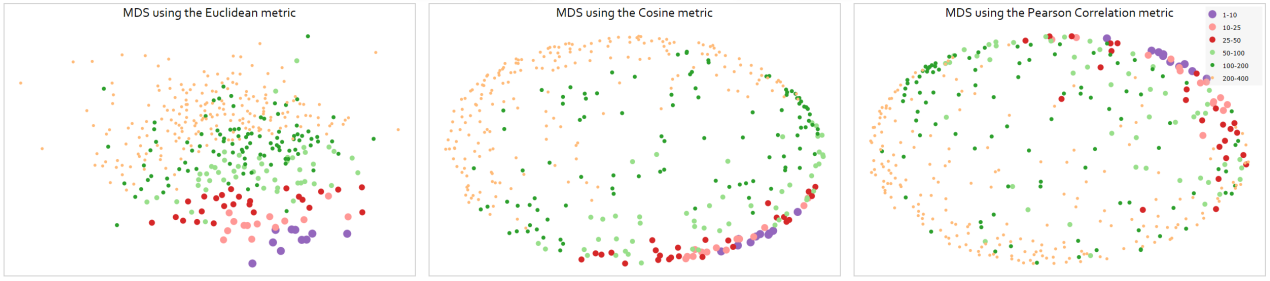


Figure 10: MDS results of 2016 data under different metrics.

To see why, we must bear in mind that the ranking is based on the weighted linear combination of several indicators. If two universities have small Euclidean distance, they must have similar scores in most indicators, thus their rankings must also be similar. On the contrary, two universities with similar distribution of scores over indicators may be quite different in the final rankings. For example, under the Pearson correlation metric or the Cosine metric, Cal Tech is most similar to John Hopkins University instead of Oxford.

world_rank	university_name	teaching	international	research	citations	income	# students	stu_staff_ratio	int_stu	f_m_ratio
1	California Institute of Technology	95.6	64	97.6	99.8	97.8	2243	6.9	0.27	0.493
2	University of Oxford	86.5	94.4	98.9	98.8	73.1	19919	11.6	0.34	0.852
11	Johns Hopkins University	77.6	70	90.4	98.2	100	15128	3.6	0.23	1

Figure 11: Comparison of three universities.

5.3.2 How many neighbors should be used to compute Standard LLE

Figure 12 shows that Standard Locally Linear Embedding is not able to satisfactorily cluster all universities, though things become better as the number of neighbors used increases.

Figure 12: Standard Locally Linear Embedding.

5.3.3 How many neighbors should be used to compute Modified LLE

Figure 13 shows that Modified Locally Linear Embedding performs really terribly on this dataset. Most points entangle together, while a couple of points locate so far away from the main cluster that they can be safely considered as outliers. Although it suggests the use of log scale, after looking into the actual coordinate values, it turns out that those outliers are not that apart compared with the variance among points of the main cluster.

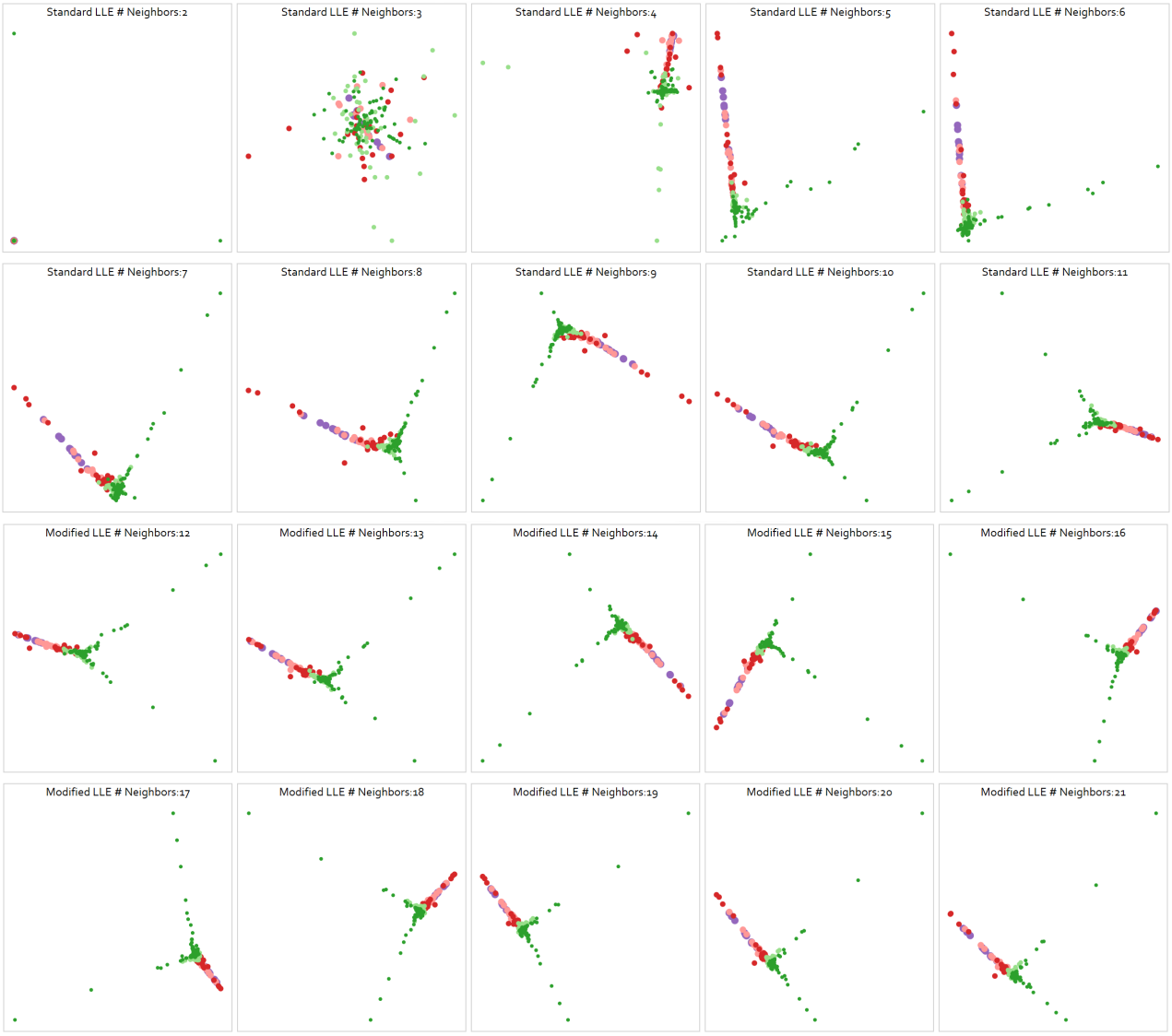


Figure 13: Modified Locally Linear Embedding.

5.3.4 How many neighbors should be used to compute Isomap

Isomap behaves much better than Locally Linear Embedding, but the number of neighbors used should not be too small.

Figure 14: Isomap.

5.3.5 How many neighbors should be used to compute Spectral Embedding

When the number of neighbors used exceeds 5, the results of Spectral Embedding do not change a lot. Its performance is acceptable.

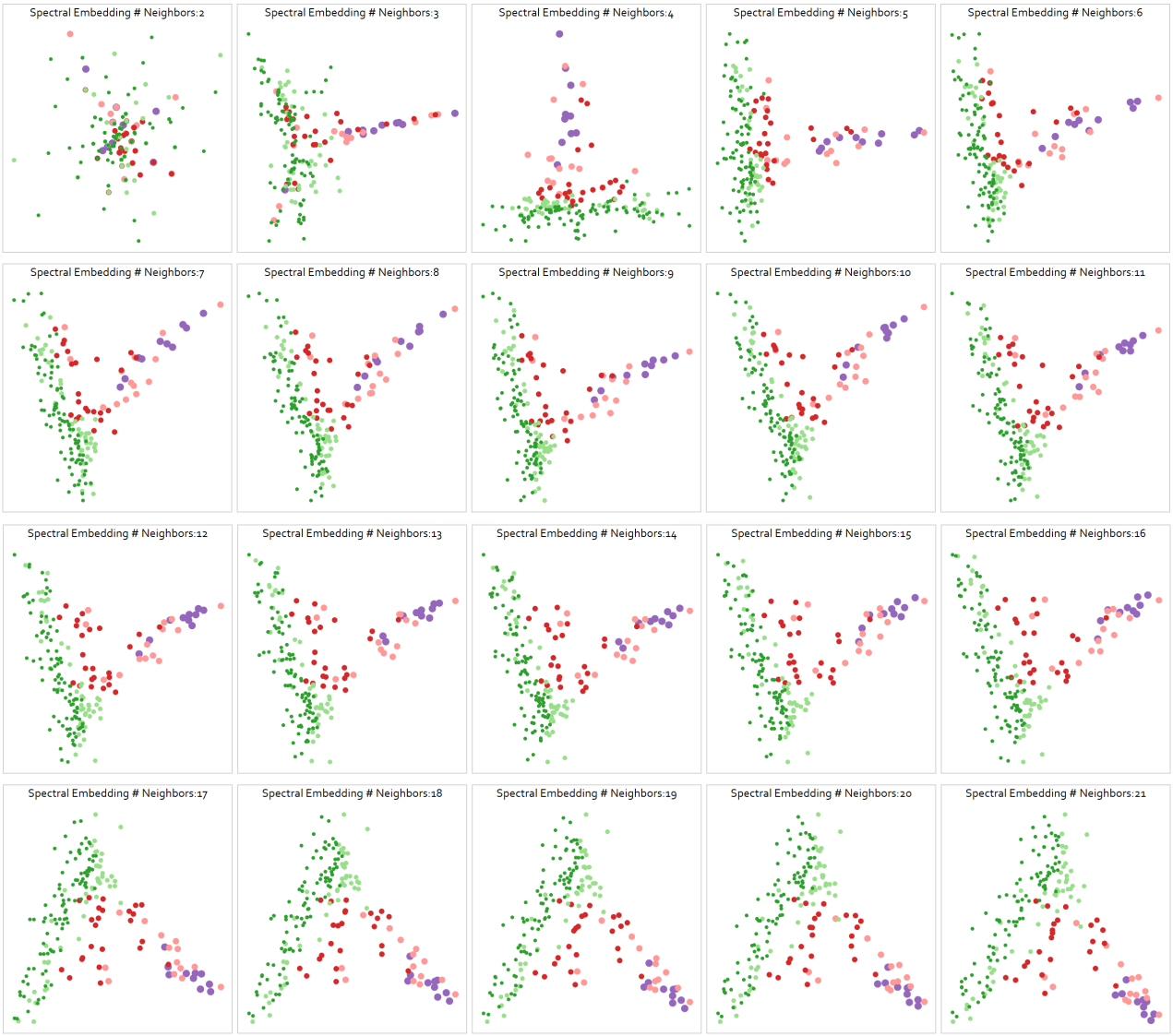


Figure 15: Spectral Embedding.

5.3.6 Comparison of all these methods

As can be seen from Figure 16, MDS using the Euclidean metric achieves the best clustering result. In addition, from Figure 17, we can see that two closely located points in one plot are also near each other in another one. That is because all these 6 manifold learning method are more or less depending on nearest neighbors.

Figure 16: Comparison of different manifold learning methods.

Figure 17: Brushing.