Cluster Analysis of Top 200 Universities in Mathematics

Kathiresan Gopal
Institute for Mathematical Research (INSPEM),
Universiti Putra Malaysia (UPM)
43400 UPM Serdang, Selangor, Malaysia.
kathiresan@upm.edu.my

Mahendran Shitan
Institute for Mathematical Research (INSPEM),
Universiti Putra Malaysia (UPM)
43400 UPM Serdang, Selangor, Malaysia.
mahen@upm.edu.my

***Abstract*—University rankings are becoming a vital performance assessment for higher learning institutions worldwide. Besides the overall rankings, the universities are also ranked by subjects serving as comprehensive guide to discover the specialist strengths of universities worldwide by highlighting top 200 universities for a range of 30 individual popular subjects. Data for this ranking purpose consist four variables namely the academic reputation, employer reputation, citation per paper and H-index citations. In this ranking, universities are ranked according to their overall score calculated from linear combination of the aforementioned variables and their respective weightings. As the existing ranking technique based on overall score appears to be simple and since the rankings data are of multivariate nature, therefore it is possible to use multivariate statistical technique like cluster analysis. Agglomerative hierarchical cluster analysis of top 200 QS ranked universities by Mathematics subject area 2014 has been performed to obtain natural clustering of the universities in an objective manner. The agreement between cluster analysis and existing QS rankings is verified and it is suggested that the distance between universities can be used as an alternative measure to rank universities. Cluster analysis applied on the same variables would serve as an alternative way to rank universities and to look at the rankings in a different perspective.**

Keywords-Hierarchical cluster analysis; Rankings by Mathematics subject; QS World University Rankings

I. INTRODUCTION

The annual university rankings have been around since 2003 with the Academic Ranking of World Universities (ARWU) published by Institute of Higher Education, Shanghai Jiao Tong University as its pioneer. Ever since then, it is observed that there is a significant rise in the interest in university rankings as many universities are using it as a platform not only to promote their excellence but also to evaluate their annual performance and to improve in many aspects such as core activities of learning and teaching, research and knowledge transfer and recruitment of employees as well as to gauge progress based on the rankings [1-2]. This can be witnessed by the growing numbers of annual university rankings being published each year. Further, there is significantly higher competition among them due to the larger pool of universities every year which in turn makes the rankings more vital.

In accordance with the public demand of the rankings, several organisations are actively considering over 2,000 universities across the globe to evaluate and rank at least the top 800 universities annually. The list includes the popular Quacquarelli Symonds (QS) World University Rankings (QS-WUR), Times Higher Education World University Rankings (THE), Academic Ranking of World Universities (ARWU), Webometrics Ranking of World Universities and others in which the rankings are focused in either the academic or research performance or as in most of the cases, both criteria are equally considered. ARWU, QS-WUR and THE are recognised as the big three ranking systems as quoted by The Economist (The Economist, Oct 10, 2011).

As the World University rankings are becoming a global phenomenon and it is here to stay, Downing [2] asserted it is important to understand that the university rankings should not be regarded as an absolute measure of performance for a certain university. It is rather a relative comparison with other universities in the set to be ranked and the ranks should simply depict how a university fares with the remaining members in the ranking set [2].

One of the big three rankings, the annual QS-WUR have been in existence since 2004 and were originally published as Times Higher Education-QS World University Rankings by the QS and Times Higher Education. In 2009, this collaboration was terminated resulting in two separate and independent ranking publications [1]. QS-WUR is claimed to be the most established ranking that provides useful insights to cater the increasing public demand to obtain precise and relevant information in order to make useful comparisons between the world universities [3].

Beside the overall rankings, QS also provides the QS-WUR by subjects ranking which serves as comprehensive guide to discover the specialist strengths of universities worldwide by highlighting top 200 universities for a range of 30 individual popular subjects [3]. From a student's perspective, there is a clear demand for ways to compare the effectiveness of universities in narrower subject disciplines because they generally know first what they want to study before choosing where to study [3].

Four key indicator measures (variables) identified within this ranking methodology are *academic reputation* derived from academic reputation survey responses, *employer reputation* derived from employer reputation survey responses, *citation per paper* and *H-index citations*

978-1-4799-7896-0/15/$31.00 ©2015 IEEE

(introduced in 2013) both calculated using the data from Scopus [4]. Academic reputation (**weightings: 0.4**) is measured using a global survey, in which academics are asked to identify the institutions where they believe the best work is currently taking place within their field of expertise [4]. The employer reputation (**weightings: 0.2**) indicator is also based on a global survey, asks employers to identify the universities they perceive as producing the best graduates [4]. The citation per paper and H-index citations (**weightings: 0.2 each**) indicators aim to assess universities' research output using the information collected from Scopus [4]. Universities are ranked according to the overall score calculated using the equation defined in (1). Complete information regarding the ranking methodology can be obtained from [4].

Overall Score = 0.4(Academic Reputation) + 0.2(Employer Reputation) + 0.2(Citation per Paper) + 0.2(H-Index Citations). (1)

It appears that the existing technique using the overall score to rank the universities seems to be simple as the overall score is just a linear combination of the variables with their respective weightings. Since the rankings data are of multivariate nature, therefore it is possible to use multivariate statistical techniques like cluster analysis. For instance, if we wish to group the 200 universities into two groups; ordinarily we would place the first 100 in the first group and the remaining 100 in the second group. In this idea, we are predetermining the number of groups and fixing the number of universities in each group arbitrarily. The validity of such grouping is questionable as we have grouped them in a subjective manner. The advantages of cluster analysis are that it allows the formation of natural clustering of the universities in which we may observe that the number of universities in each groups will not be fixed. This means that we do not need to predetermine neither the number of groups nor the number of universities in each group. Eventually, we are able to remove this subjective element and able to obtain the groups in objective manner.

In view of the above discussion, the objective of this study is to rank the universities using multivariate statistical technique, namely cluster analysis. Further, we would like to compare the rankings in cluster analysis with that of in the existing QS by subject rankings. Ultimately, we are achieving the same purpose in a different perspective.

II. Methodology

A. Data Description

The data used in this study are the 2014 QS-WUR by Mathematics obtained from QS's official website for by subjects' rankings, available at [5].

The dataset is comprised of the aforementioned four variables and the overall score of the universities. Following the usual multivariate statistics notation, we shall let the random vector $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4)^T$ with the variables $\mathbf{X}_i$ as listed out as follows:

$\mathbf{X}_1$: Academic Reputation

$\mathbf{X}_2$: Employer Reputation

$\mathbf{X}_3$: Citation per Paper

$\mathbf{X}_4$: H-index Citations.

B. Cluster Analysis

Clustering (or cluster analysis) is an important multivariate statistical technique used in understanding the complex nature of multivariate relationships [6]. Clustering can provide an informal means for assessing dimensionality, identifying outliers and most importantly suggesting interesting hypotheses concerning relationships [6]. Clustering is performed on the basis of similarities or dissimilarities (distances). Cluster analysis in this study was performed using the Computational Statistics Toolbox in MATLAB [7-8].

In clustering a set of observations, proximity is usually indicated by a distance metric. Distances as dissimilarity measure provide an insight of how close observations are to each other in a set. Generally, longer distance indicates that observations are more dissimilar and vice versa. Common distance metrics available on most literatures are *Euclidean distance*, *Standardized Euclidean distance*, *City block distance* and *Mahalanobis distance* [7]. The Euclidean distance has been used in this analysis. The choice for the Euclidean distance used is detailed out in the Results and Discussions section. Suppose $\boldsymbol{w}$ and $\boldsymbol{y}$ are two observations (universities) from the data $\boldsymbol{X_i}$. Then, the Euclidean distance between universities $\mathbf{w} = (w_1, w_2, w_3, w_4)^T$ and $\mathbf{y} = (y_1, y_2, y_3, y_4)^T$ is defined as in (2).

$$d(\mathbf{w},\mathbf{y}) = \sqrt{(\mathbf{w}-\mathbf{y})^T(\mathbf{w}-\mathbf{y})} \quad (2)$$

The hierarchical clustering procedure consists of two methods, namely the *agglomerative* and *divisive* methods. In agglomerative method, we start with all the individual observations as a cluster and proceed by successive joining of observations and clusters until all the observations are fused into a single cluster [6]. The latter works in the opposite way; we start with all the observations in one cluster and proceed by successive divisions of clusters until all the observations are in separate clusters [6]. In this analysis, we have employed the agglomerative procedure.

Linkage methods are suitable for clustering observations in most of the agglomerative procedures. The conventional linkage methods are *single linkage* (minimum distance or nearest neighbour), *complete linkage* (maximum distance or

farthest neighbour) and *average linkage* (average distance) [6] (see [7-8] for more linkage methods). Given clusters r and s, the number of observations (universities) in each cluster is denoted by n_r and n_s; we can define the distance between cluster r and cluster s, $d_c(r,s)$ as follows:

Single linkage: $d_c(r,s) = \min\{d(x_{rw}, x_{sy})\}$

Complete linkage: $d_c(r,s) = \max\{d(x_{rw}, x_{sy})\}$

Average linkage: $d_c(r,s) = \frac{1}{n_r n_s} \sum_{w=1}^{n_r} \sum_{y=1}^{n_s} d(x_{rw}, x_{sy})$

where $d(x_{rw}, x_{sy})$ is the distance between university (observation) y from cluster r and university w from cluster s [7-8]. This is in fact the interpoint distance i.e. Euclidean distance in our case. We have opted for average linkage in this study and this choice is justified in the following section.

The results of agglomerative clustering can be well visualized in the form of a tree diagram known as a *dendrogram* [6]. A dendrogram shows the nested structure of the partitions (fusions) and how the various clusters are linked at each stage [7]. Additionally, it can be presented in both vertical and horizontal forms; however the vertical version is often preferred with the inter-cluster distances displayed on the vertical axis [7-8].

The fact that there are several hierarchical clustering techniques available, and that several of them can be used with many distance measures, means that there are a large number of options open to us. Consequently, we are faced with an arbitrary choice both of method and of distance metric and it is usual to try all or as many as possible choices of these and then compare the results. In order to ensure a good clustering, we can employ the *cophenetic* correlation coefficient technique introduced by Sokal and Rohlf [9] for comparing dendrograms.

The cophenetic correlation coefficient serves as a measure of degree of fit of a clustering to a dataset and as well as a criterion for evaluating the efficiency of various clustering techniques [9]. Suppose we have produced a dendrogram, T_i (a model of the original data $\mathbf{X}_i$) and we have the distance between observations, $d(\mathbf{w},\mathbf{y})$ from (2); we can denote the *dendrogrammatic* distance between the model points T_w and T_y as $t(\mathbf{w},\mathbf{y})$. This distance is the height of the node at which these two model points are joined together. Next, let d be the average $d(\mathbf{w},\mathbf{y})$ of and t be the average of $t(\mathbf{w},\mathbf{y})$, then the cophenetic correlation coefficient c is defined as in (3) [9].

$$c = \frac{\sum_{w<y}[d(\mathbf{w},\mathbf{y}) - d][t(\mathbf{w},\mathbf{y}) - t]}{\sqrt{\sum_{w<y}[d(\mathbf{w},\mathbf{y}) - d)^2] \sum_{w<y}[t(\mathbf{w},\mathbf{y}) - t]^2}} \quad (3)$$

Further evaluation on goodness of fit of the dendrograms can be measured by means of the Goodman-Kruskal gamma coefficient [9].

The agreement between cluster analysis and the existing QS rankings is then verified using the correlation coefficient between the overall scores and the distances between universities. This is plausible since the universities are being ranked using the overall score and the fact that distances are used for clustering the universities.

III. RESULTS AND DISCUSSIONS

The values of c for all possible combinations of available agglomerative techniques are presented in Table I.

TABLE I. COPHENETIC CORRELATIONS COEFFICIENT VALUES

Linkage Method	Distance Metric			
	Euclidean	Standardized Euclidean	City Block	Mahalanobis
Average	0.7483	0.7453	0.7365	0.6542
Complete	0.5082	0.501	0.515	0.4975
Single	0.5778	0.5797	0.6417	0.5825

The highest value of c is **0.7483** produced by average linkage with Euclidean distance indicating the best combination in relative to others. Hence, we proceeded to generate the dendrogram for this combination.

In most dendrograms, usually the leaf nodes (displayed as numbers on horizontal axis) contain a single observation [7]. In contrast to this, the nodes in MATLAB's dendrograms can be internal or terminal. The internal nodes contain or represent all observations that are grouped together based on the type of linkage and distance used [7]. They do not necessarily represent one of the original observations; they likely contain several observations [7]. Full information regarding this implementation is available in [7].

However when n is large (in this case, n=200), it is difficult to visualize the clusters, hence MATLAB's implementation becomes very handy (default number of leaf nodes is 30). In order to determine which universities are in the leaf nodes, we can make use of a built-in function to list them down [7]. The revised dendrogram with 30 leaf nodes is presented in Fig. 1 whereas the full scale dendrogram is given in Fig. 2.

Remark: MATLAB's built-in function to list the members in a leaf node.

Leaf node numbers for each data point in the original data set, returned as a column vector of length M, where M is the number of data points in the original data set. When there are fewer than P data points in the original data (P is 30, by default), all data points are displayed in the dendrogram, with each node containing a single data point. In this case, T is the identity map, T = (1:M)'.T is useful when P is less than the total number of data points. That is, when some leaf nodes in the dendrogram display correspond to multiple data points. For example, to find out which data points are contained in leaf node k of the dendrogram plot, use the specified function: ***find(T==k)*** *in MATLAB's command.[7].*

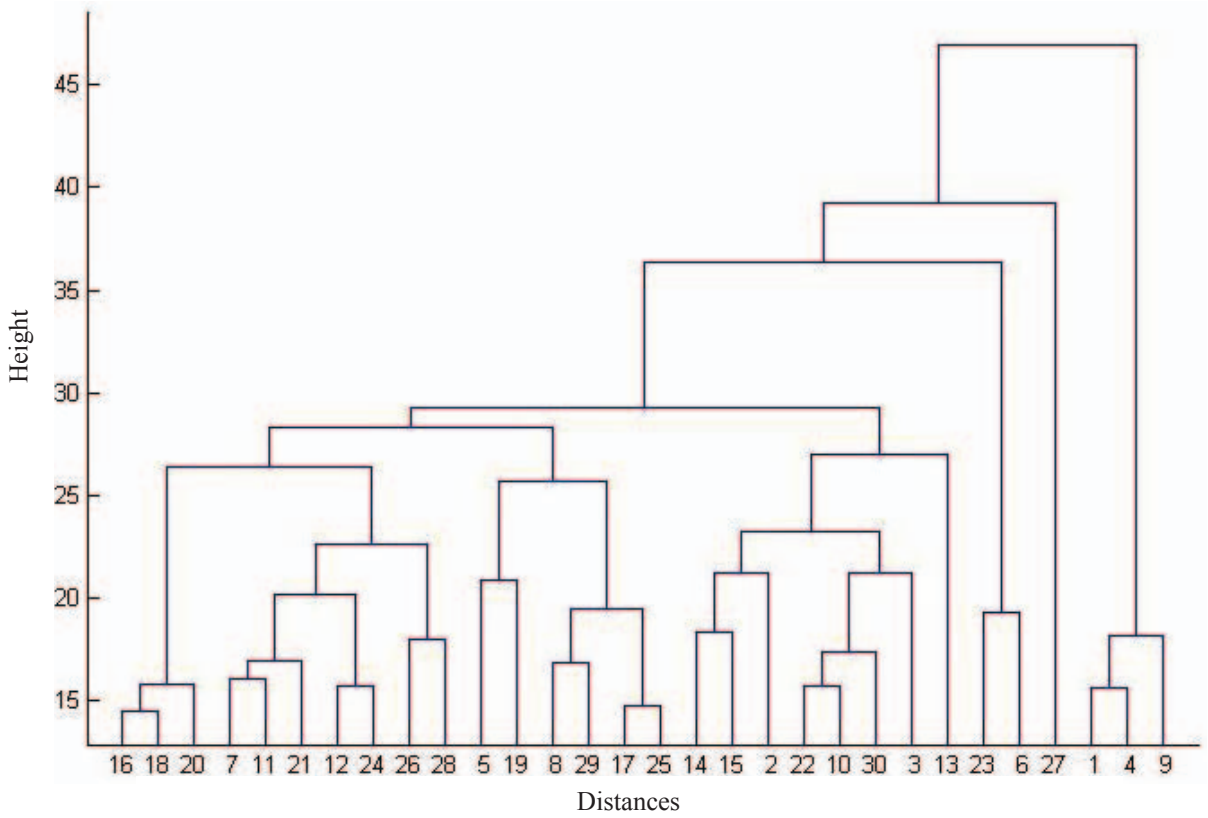

Figure 1. Dendrogram generated with 30 leaf nodes for average linkage with Euclidean distance.

To be specific, the numeral labels of the nodes on the above dendrogram (Fig. 1) are not the ranks directly but instead group labels consisting of several universities. The universities are identified by their respective ranks (available at [5]) and the corresponding leaf nodes containing them are displayed in Table II.

TABLE II. LEAF NODES AND CORRESPONDING UNIVERSITIES

Leaf nodes label	Universities with rank
1,4,9	1,2,3,4,5,6,7,8,9,10,11,12,13
27	167
23,6	23,49
13	74
22,10,30,3	22,29,30,31,35,36,38,39,40,44,46,47,48,53,56,62,67,70,73, 76,79,80,81,82,97
14,15,2	14,15,16,17,18,19,20,21,24,25,26,27,28,32,33,34,37,41,43, 50,57,90,99,100
8,29,17, 25	55,84,106,107,128,129,156,159,160,161,170,174,180
5,19	42,45,54,85,89,103,183,187
26,28	165,169,177,186,199
7,11,21,12, 24	51,52,58,59,60,61,63,64,65,66,68,69,71,72,77,78,83,86,87, 88,91,92,93,94,95,98,101,102,104,108,109,110,111, 112, 113,114,115,116,117,118,119,120,121,122,124,125,126, 127,130,131,132,133,134,135,136,138,139,140,141,143, 144,145,146,147,148,150,151,152,153,154,155,157,158, 162,163,164,166,168,171,172,175,176,178,179,181,182, 184,185,188,189,190,191,192,193,194,195,196,197,198, 200
16,18,20	75,96,105,123,137,142,149,173

The dendrogram provides a visual output of the formation of the natural clusters of the 200 universities. This aids us to group up the universities in an objective manner in which we can determine the number of groups based on the dendrogram and note that the number of universities in each groups is not fixed anymore and it depends on how we would like to group up them. Recall the idea of grouping up the 200 universities into two groups as mentioned in the Introduction section; we may notice that there would not be 100 universities in each of the groups as assumed in the arbitrary grouping but instead we have only the first 13 universities in the first group and the remaining universities in the second group. It is also clear that first 13 universities are in a compact group and more similar to each other in Mathematics in relative to the rest. Whereas, if say for instance in the arbitrary grouping we would have placed university ranked 14 in the first group; however from the natural clusters it is evident that this university is in fact does not belong to the first group as it is more similar to the rest of the universities compared to the first 13 universities. This depicts the advantages of using natural clusters rather than arbitrary grouping which could result in discrepancies.

Referring to Table III, we define D_i as the distance between university at rank i, (i = 1,2,…,200) and university ranked one (distance from origin) which is essentially the entries in the first column of the Euclidean

distance matrix. The correlation coefficient between the overall score and D_i is found to be **-0.98** which indicates a strong linear relationship between the two. In order to obtain positive correlation coefficient, we define another quantity D_i^* as the difference between the maximum D_i and the rest of the D_i 's. The correlation coefficient between the overall score and D_i^* is **0.98** which is a clear indication of positive linear relationship between the overall score and distances. This relationship verifies the agreement between cluster analysis and the existing QS rankings.

TABLE III. FIRST TEN UNIVERSITIES

QS Rank	School Name	Overall Score	D_i	D_i^*
1	Harvard University	97.7	0	77.09578458
2	Massachusetts Institute of Technology (MIT)	95.9	9.536771	67.55901364
3	University of Oxford	95.8	6.765353	70.33043167
4	University of California, Berkeley (UCB)	94.8	11.1866	65.90918545
4	University of Cambridge	94.8	13.50481	63.59097062
6	Princeton University	94.2	15.65567	61.44011468
6	Stanford University	94.2	7.828793	69.26699157
8	University of California, Los Angeles (UCLA)	91.6	20.39191	56.70387442
9	ETH Zurich (Swiss Federal Institute of Technology)	89	18.44858	58.6472074
10	New York University (NYU)	88.8	23.64487	53.45091199

As we can see from Table III, there exist ties in the ranks which is due to the equal overall score (for complete reference of such ties in ranks, please refer to [5]) which may mislead to an impression that two or more such universities are sharing exactly the same attributes for Mathematics subject. However, this is not true as they can be similar but not exactly the same. Recall that the overall score is calculated using (1); an example is illustrated below to show why this is not true in general:

i) *Overall score of University of California, Berkeley (UCB)*: **94.8** = 0.4(***98***) + 0.2(***88.3***) + 0.2(***94.1***) + 0.2(***95.8***)
ii) *Overall score of University of Cambridge*: **94.8** = 0.4(***98.9***) + 0.2(***100***) + 0.2(***89.2***) + 0.2(***87.1***)

From the above example, we can certainly see that the individual values for the variables are not identical but the use of weightings in the calculation gives rise to equal overall scores and finally to a situation of tied ranks.

Such an impression would be removed (and eventually no tied ranks) if distances (D_i) are used to rank up the universities instead of using overall score. Distances can be used as alternative measure since the cluster analysis and the existing QS rankings have a good agreement. Distances are unique for each of the universities and it measures the proximity between universities without the influence or affect of any form of weightings. Moreover, distances provide a clearer idea of the rankings as it resembles the dispersion of a university from the topmost position i.e. university at rank one.

If distances are used to rank, we would observe several shifts in the existing ranks. From Table III, it can be seen that University of Oxford at rank three would move up to the second rank as it is closer to Harvard University at rank one since the distance is shorter. Also, Stanford University should move up to rank three. Thus, there will be several shifts like these for the whole list of 200 universities if distances are used to rank them. An important assumption to initiate this ranking process is that the university at rank one in QS rankings will remain at its position in this technique as well due to the need of an origin or topmost position.

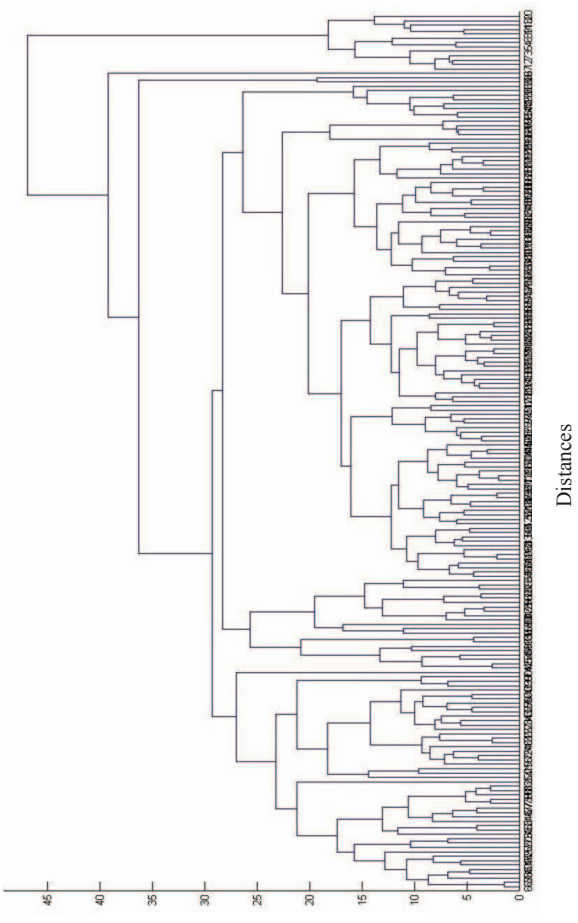

Figure 2. Full scale dendrogram

CONCLUSIONS

The objective of this study was to rank the universities using multivariate statistical technique, namely cluster analysis by using the 2014 QS-WUR by Mathematics rankings data. The resulting dendrogram and the illustration of the use of distances to rank the universities have been presented in the Results and Discussion section.

It is obvious that the cluster analysis agrees well with the existing QS ranking system indicated by the reasonably strong correlation coefficient of **0.98**. Some useful information in terms of similarities in Mathematics are able to be deduced through cluster analysis such as the distinction of the first 13 universities from the rest. It is suggested that the distance between universities can be used as an alternative measure to rank the universities provided that the assumption about the first ranked university in QS rankings is not violated. Besides the existing ranking technique employed by QS, cluster analysis applied to the same set of variables used in the rankings would serve as an alternative way to rank universities and to look at the rankings in a different perspective.

The main advantage of cluster analysis is the formation of natural clustering of the universities which would avoid any discrepancies rising from arbitrarily or subjectively grouping up the universities and enables the distance between universities to be used as an alternative measure besides the overall score.

REFERENCES

[1] I. F. Aguillo, 'Comparing university rankings', *Scientometrics*, vol. 85, no. 1, pp. 243-256, 2010.

[2] K. Downing, 'What's the use of rankings?', *Rankings and Accountability in Higher Education*, vol. 197, 2013.

[3] Top Universities, 'QS World University Rankings by Subject', 2014. [Online]. Available: http://www.topuniversities.com/subject-rankings/. [Accessed: 16- Jul- 2014].

[4] Iu.qs.com, 'QS Intelligence Unit | University Rankings', 2014. [Online]. Available: http://www.iu.qs.com/university-rankings/. [Accessed: 16- Jul- 2014].

[5] Top Universities, 'QS World University Rankings by Subject 2014 - Mathematics', 2014. [Online]. Available: http://www.topuniversities.com/university-rankings/university-subject-rankings/2014/mathematics. [Accessed: 16- Jul- 2014].

[6] R. Johnson and D. Wichern, Applied multivariate statistical analysis. Englewood Cliffs, N.J.: Prentice-Hall, 1988, pp. 543-554.

[7] W. Martinez, A. Martinez and J. Solka, *Exploratory Data Analysis with MATLAB, Second Edition*. Hoboken: Taylor and Francis, 2010, pp. 157-420.

[8] W. Martinez and A. Martinez, *Computational statistics handbook with MATLAB*. Boca Raton: Chapman & Hall/CRC, 2002, pp. 367-372.

[9] S. Saraçli, N. Doğan and İ. Doğan, 'Comparison of hierarchical cluster analysis methods by cophenetic correlation', *Journal of Inequalities and Applications*, vol. 2013, no. 1, p. 203, 2013.