White Paper

Selecting Peer Institutions Using Cluster Analysis - Summer, 2014

Institutional Research, Planning, and Assessment

Build the Pride

About the Author

Dr. Andrew L. Luna, is Director of Institutional Research, Planning, and Assessment. He has served over 28 years in higher education, with 19 of those years in institutional research. He has published research studies on many topics including salary studies, assessment, market research, and quality improvement. Dr. Luna received his Ph.D and M.A. degrees in higher education administration and his M.A. and B.A. degrees in journalism, all from the University of Alabama.

Table of Contents

EXECUTIVE SUMMARY

Sensing a need to update the University of North Alabama's peer institution list, the Vice President for Academic Affairs and Provost charged the Office of Institutional Research, Planning, and Assessment with the task of creating a more scientific and reliable method for selecting UNA's peers.

The method used is referred to as cluster analysis, which is defined as an exploratory data analysis technique for classifying and organizing data into meaningful clusters, groups, or taxonomies by maximizing the similarity between observations within each cluster. The purpose of cluster analysis is to discover a system of organizing observations into groups where members of the groups share properties in common.

The process required the designation of an initial group that shared a similar role, scope, and mission to UNA; identification of variables to be used in the analysis; and the determination of the fit of the clusters in relationship to UNA. After the analysis was completed, it was determined that two cluster groups overlapped and that UNA could use peers from either cluster. Taking geographical and accreditation considerations into account, the Office of Institutional Research recommended the following as its new peers:

- Nicholls State University (Louisiana)
- Auburn University at Montgomery
- NcNeese State University (Louisiana)
- Northwestern State University of Louisiana
- Midwestern State University (Texas)
- Pittsburg State University (Kansas)
- Radford University (Virginia)
- University of South Florida - St. Petersburg
- Western Carolina University (North Carolina)

Out of these recommend peers, Nicholls State University, Auburn University at Montgomery, Northwestern State University, and Pittsburg State University are among UNA's current peer group.

INTRODUCTION

Within the current state of higher education, colleges and universities must strive to be competitive in both the quality of education they offer as well as the cost of attendance. At the same time, higher education is being held more accountable by federal and state governments, as well as by the communities they serve. This accountability varies broadly by legislative bodies, governors' offices, faculty committees, federal mandates, students and other constituencies. Therefore, the use of comparator institutions as a reference point within higher education has become common practice.

The use of peer comparator institutions allows administrators to compare both the quality and quantity of academic programs and delivery methods, as well as institutional expenditures and revenues. Comparisons like these allow for more focused strategic and long-range planning strategies in order to meet goals and objectives.

When identifying peers, it is important to understand the focus for the comparison group, as more than one set of peer groups may be utilized by an institution. There are various kinds of peers, such as:

- **Comparable:** Similar institutional level (two-year vs. four-year), control (e.g. private not-for-profit vs. public) and enrollment profile characteristics.
- **Aspirational:** Institutions with similar institutional characteristics yet are significantly different in several key performance indicators, such as significantly higher graduation rates or endowments.
- **Competitors:** Based on cross applications, institutions may have significantly different institutional characteristics, yet a significant percentage of the institution's applicants choose to attend another institution.
- **Consortium:** Institutions belonging to a consortium for a common purpose and/or to share data may be another peer group for review.

These peer institutions tend to share the same basic Carnegie Classification (e.g. Master's Institution vs. Associate of Arts), in addition to similar graduation rates and enrollment mix (e.g. percent full-time vs. part-time).

In 2009, the University of North Alabama updated its list of peer institutions through a series of discussions and recommendations by the President's Executive Council as well as the Council on Academic Deans. This peer list was created solely on the experience and understanding that the administration had towards each one of the institutions chosen, the relative close proximity to UNA, as well as certain academic programs that the institutions offered. The current list of peer institutions for UNA is:

1. Auburn University at Montgomery
2. Austin Peay State University (Tennessee)
3. Jacksonville State University
4. Morehead State University (Kentucky)
5. Murray State University (Kentucky)
6. Nicholls State University (Louisiana)
7. Northwestern State University of Louisiana
8. Pittsburg State University (Kansas)
9. University of West Georgia
10. Western Carolina University (North Carolina)

Sensing a need to update this list, the Vice President for Academic Affairs and Provost charged the Office of Institutional Research, Planning, and Assessment with the task of creating a more scientific and reliable method for selecting UNA's peers. The process of utilizing statistical methodologies in the identification of peer institutions began more than 20 years ago (Terenzini, et al., 1980; Teeter & Brinkman, 1987; and McLaughlin &McLaughlin, 2007). The overall goal during this time has been to identify appropriate methods for comparing the performance of a reference institution relative to a group of similar institutions, and to make goal and outcome decisions concerning the reference institution based on the performance of the comparator institutions.

While the use of statistical methodologies supports scientific objectivity, their complexity often makes them difficult to understand by the end user. Other studies have also indicated that these types of methodologies inherently contain statistical error due to the additive and multiplicative attributes of the procedures used (McLaughlin & McLaughlin, 2007). It is, therefore, recommended that the institution not rely solely on the outcome of a statistical peer analysis. Rather, the data from the analysis should be used in conjunction with other knowledge gained.

"The process of utilizing statistical methodologies in the identification of peer institutions began more than 20 years ago."

This study used cluster analysis, which is defined as an exploratory data analysis technique for classifying and organizing data into meaningful clusters, groups, or taxonomies by maximizing the similarity between observations within each cluster. The purpose of cluster analysis is to discover a system of organizing observations into groups where members of the groups share properties in common. The goal of this analysis, therefore, is to sort variables into groups or clusters so that the degree of association or relationship is strong between members of the same cluster and weaker between members of different clusters.

The appropriate cluster algorithm and parameter settings depend on the individual data set and intended use of the results. Furthermore, cluster analysis is an iterative process of knowledge discovery and optimization to modify data processing and model parameters until the result achieves both the preferred as well as appropriate properties.

The choice of methods used for cluster analysis depends on the size of the data set as well as the types of variables used. In this study, hierarchical clustering is more appropriate because the data set is small. The steps in obtaining and preparing the data for cluster analysis are as follows:

- Screen institutions to determine what type and size of institution will be used in the analysis based upon the IPEDS data service
- Choose variables to download from IPEDS that will be used in the analysis
- Standardize all quantifiable variables that will be used in the analysis
- Run the cluster analysis procedure
- Determine the fit and reliability of the model
- Identify those institutions that are within the same cluster as UNA

"...cluster analysis, [is] defined as an exploratory data analysis technique for classifying and organizating data into meaningful cluster, groups, or taxonomies..."

IPEDS INITIAL INSTITUTIONAL SCREENING

To start the process of determining institutional peers, an initial reference group was established. Larger research institutions, two-year colleges, and specialty institutions with a significantly different role, scope, and mission than UNA were screened out. This screening process was generated through the Grouping procedure found within the IPEDS Data Center. Below are listed the screening criteria within the Grouping procedure as well as what was chosen for this study:

1. **Select:** "First Look University" which included institutions currently within the IPEDS universe, those that were open to the public, and those that participated in federal financial aid programs.
2. **Special Missions:** This criterion was left null because UNA is not an Historically Black College or University, tribal institution, or land-grant institution.
3. **State Or Other Jurisdiction:** All 50 states within the US.
4. **Geographic Region:** Since all 50 states were chosen above, there was no need to choose a specific geographic region. Therefore, this criterion was left null.
5. **Sector:** Public, 4-year or above.
6. **Degree-Granting Status:** Degree-Granting.
7. **Highest Degree Offered:** Doctor's Degree (Other) and Master's Degree.
8. **Institutional Category:** Degree-Granting, Primarily Baccalaureate or Above.
9. **Carnegie Classification:** Master's Colleges and Universities (Larger Programs), Master's Colleges and Universities (Medium Programs).
10. **Degree of Urbanization:** City (Medium), City (Small), Suburban (Large), Suburban (Medium), Suburban (Small).
11. **Institutional Size:** 5,000 – 9,999 and 10,000- 19,999.
12. **Reporting Method:** Student charges for full academic year and fall Graduate/Student Financial Aid/Retention rate cohort.
13. **Has Full-Time First-Time Undergraduates:** Yes
14. **All Programs Offered Completely Via Distance Education:** No

Based on this initial screening, a total of 61 institutions were chosen through the IPEDS system. From these institutions, specific variables were chosen to be used in the cluster analysis procedure.

"Larger research institutions, two-year colleges, and specialty institutions with a significantly different role, scope, and mission were screened out."

Choosing Variables to Use in the Analysis

Once the initial 61 institutions were selected, a total of 12 selected variables were downloaded from the IPEDS Data Center for each institution. These variables were selected by the OIRPA office and the Vice President for Academic Affairs and Provost following both a discussion and a literature review process. The variables selected are listed below:

1. Undergraduate enrollment for latest fall semester
2. Graduate enrollment for latest fall semester
3. FTE for latest fall semester
4. Six-year graduation rate based on the IPEDS defined freshman cohort
5. Total core revenues
6. Tuition and fees as a percent of core revenues
7. State appropriations as a percent of core revenues
8. Total core expenditures
9. Instructional costs as a percent of core expenditures
10. Endowment Assets per FTE
11. In-state tuition and fees on-campus
12. Out-of-state tuition and fees on-campus

Standardizing all quantifiable variables used in the analysis

Many researchers have noted the importance of standardizing variables for multivariate analysis. Otherwise, variables measured at different scales may not contribute equally to the analysis. This practice holds true for cluster analysis. Because of the sensitivity of most cluster models, raw values used for the variables may significantly alter the outcomes.

For example, in selecting peer institutions, a variable that ranges between $5 million and $10 million will influence significantly and have more weight in the analysis than a variable that ranges between 20 and 50. Therefore, transforming the data to comparable scales can prevent this problem. Typical data standardization procedures equalize the range and/or data variability. In the case of this study, variable values were standardized using z-scores with a mean of zero and a standard deviation of 1.

The z-score is a very useful statistic because it allows researchers to calculate the probability of a score occurring within the normal distribution and it enables researchers to compare two scores from different normal distributions. The standard

"Many researchers have noted the importance of standardizing variables for multivariate analysis. Otherwise, variables measured at different scales may not contributes equally to the analysis"

score does this by converting scores in a normal distribution to z-scores using the following formula:

$$z = \frac{x - \overline{x}}{S}$$

where x represents an individual score or observation in a set of scores, $\overline{x}$ represents the average of all individual scores or observations, and S represents the standard deviation of the scores or observations.

The z-score is synonymous to the standard deviation. A z-score of 2 is essentially 2 standard deviations above and below the mean. A z-score of 1.5 is 1.5 standard deviations above and below the mean. A z-score of 0 is equal to the mean of the distribution.

Z-scores exist on both sides of the mean. For example, 1 standard deviation below the mean is a z-score of -1 and a z-score of 2.2 can be 2.2 standard deviations above the mean. A z-score of -3 is 3 standard deviations below the mean. Put another way, the standard deviation and z-scores are just the average distance that individual values are from the mean.

RUNNING THE CLUSTER ANALYSIS PROCEDURE

While there are numerous ways in which clusters may be formed, hierarchical clustering is one of the most straightforward methods. It can be either agglomerative or divisive. Agglomerative hierarchical clustering begins with each institution being a cluster unto itself. At successive steps, similar clusters are merged. The algorithm ends with all institutions in one, but useless, cluster. Divisive clustering starts with all institutions in one cluster and ends with each institution in its own cluster which, again, is not helpful. To find a good cluster solution, the researcher must look at the characteristics of the clusters at successive steps and decide when an interpretable solution is found that has a reasonable number of fairly homogeneous clusters.

"While there are numerious ways in which clusters may be formed, hierarchical clustering is one of the most straightforward methods."

This study used PROC FASTCLUS within SAS to determine the clusters. While the FASTCLUS procedure is intended for larger data sets, it can be used with smaller, although it can be sensitive to the order of the observations within the data set. This issue can be negated by standardizing the variables. PROC FASTCLUS also uses algorithms that place a large influence on variables with larger variance. Again, standardizing the variables before performing the analysis is highly recommended.

PROC FASTCLUS performs a disjoint cluster analysis on the basis of distances computed from one or more quantitative variables. The observations are divided into clusters so that every observation belongs to one cluster. By default, PROC FASTCLUS uses Euclidean distances, so the cluster centers are based on least squares estimation. The cluster centers are the means of the observations assigned to each cluster when the algorithm is run to complete convergence. PROC FASTCLUS is designed to find good clusters, not the best possible clusters, with only two or three iterations of the data set and changing the number of clusters requested. This procedure can be effective in detecting outliers which appear as clusters with only one institution.

To run the analysis a two-step process was used to determine the number of possible clusters. This process used the CLUSTER procedure within SAS in order to examine eigenvalues, differences, and proportions. According to **Table 1**, a large difference exists between the first (4.686) and second (2.755) eigenvalues, proportions go from .3905 to .2296, with the cumulative proportion for the second eigenvalue equal to .6201. While this seems significant, a total of 61 institutions within only two clus-

ters would be considerably underspecified and the cumulative proportion indicates more clusters could be formed.

Upon further examination of the table, there exists a moderate change from eigenvalues eight (.3475) and nine (.1159), proportions go from .0290 to .0097, with the cumulative proportion for the ninth eigenvalue equal to .9912. Therefore, further investigation of eight clusters will be examined with results from PROC FASTCLUS.

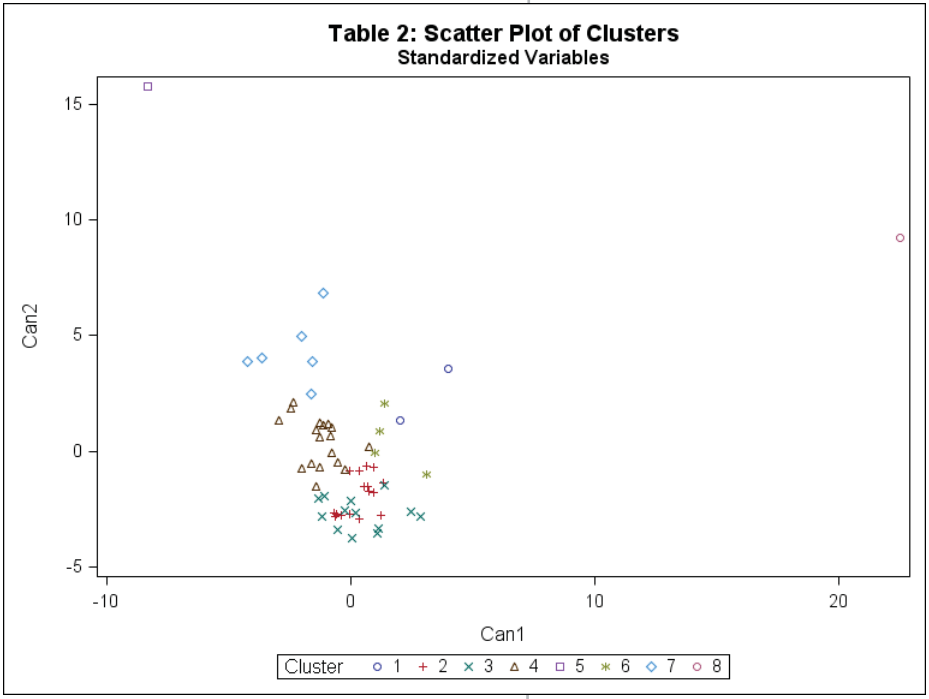
Table 1: Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.686	7.931	0.391	0.391
2	2.755	1.500	0.230	0.620
3	1.255	0.268	0.105	0.725
4	0.987	0.207	0.082	0.807
5	0.780	0.251	0.065	0.872
6	0.528	0.089	0.044	0.916
7	0.440	0.092	0.037	0.953
8	0.347	0.232	0.029	0.982
9	0.116	0.033	0.010	0.991
10	0.083	0.068	0.007	0.998
11	0.016	0.009	0.001	0.999
12	0.007		0.001	1

Running the FASTCLUS procedure on eight clusters generated a significant Pseudo F Statistic of 13.26 and an observed over-all R-Squared value of .64. The multivariate statistics and F approximations were then computed to test the fit of the model and the Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace, and Roy's Greatest Root all confirmed that the model was significant with eight clusters.

DETERMINING FIT AND RELIABILITY OF MODEL

After the cluster analysis procedure determined that the 61 institutions could be reduced into eight unique clusters, a canonical discriminant analysis was run to create grouped variables for use in a scatterplot in order to determine where each of the clusters fall. Canonical discriminant analysis is used to find a linear combination of features which characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier or, more commonly, for dimensionality reduction before later classification.



The first canonical correlation is the maximum correlation that can be obtained between a linear combination of one set of variables and a linear combination of another set of variables. The second canonical correlation is the maximum correlation that can be obtained between linear combinations of the two sets of variables subject to the constraint that these second linear combinations are orthogonal (independent/uncorrelated) to the first linear combinations. The second canonical variable provides the greatest difference between group means while being uncorrelated with the first canonical variable.

Within this study, the first two canonical correlations explain about 83% of the variation in the study, so plotting the first canonical correlation against the second should give a good indication where the clusters fall and how closely related they are to each other. Following the FASTCLUS procedure, the first canonical variable was plotted against the second canonical variable. Together, these variables indicate where the various clusters reside, how widely distributed they are, and how close they are to each other. As can be seen in **Table 2**, all of the clusters are distinct, albeit they are close together with clusters two (red t's) and three (green x's) overlapping. Clusters five and eight contain only one institution and, therefore, are considered outliers.

RESULTS

The FASTCLUS procedure within SAS indicated that the 61 institutions would best be divided into eight clusters. The results also indicate that clusters two and three are close together. Clusters five and eight only contain one institution each and should be considered outliers. For purposes of identification, the institution within cluster five was California State University – Fullerton, and the institution within cluster eight was Citadel Military College of South Carolina. According to the study, UNA fell within cluster two. Those institutions within cluster two were:

1. Augusta State University (Georgia)
2. East Stroudsburg University of Pennsylvania
3. Fitchburg State University (Massachusetts)
4. Framingham State University (Massachusetts)
5. Indiana University – South Bend
6. Indiana University – Southeast
7. Minnesota State University – Moorhead
8. Nicholls State University (Louisiana)
9. Purdue University – Calumet Campus (Indiana)
10. Radford University (Virginia)
11. Salisbury University (Maryland)
12. Southern Oregon University
13. State University of New York at New Paltz
14. University of North Alabama
15. Westfield State University (Massachusetts)
16. Worchester State University (Massachusetts)

Within this cluster, only three institutions reside in the same geographic region as UNA. Furthermore, Nicholls State University is the only institution that is listed among UNA's current peers. It should be noted that data from the IPEDS system is, in most cases, at least a year old. As a result of this lag, Augusta State University is no longer a stand-alone master's comprehenisve institution. Since the lastest IPEDS data collection, it has sinced merged with other institutions, including Georgia's medical college, to form a comprehensive research institution. Therefore, while the data used in this analysis was accurate, ASU can no longer be considered a peer to UNA.

With the cluster analysis indicating that cluster three was close to cluster two, further investigation as to which institutions reside wthin this cluster should be done. According to the analysis, those institutions within cluster three were:

"With the cluster analysis indicating that cluster three was close to cluster two, further investigations as to which institutions reside within this cluster should be done."

1. Albany State University (Georgia)
2. Auburn University at Montgomery
3. McNeese State University (Louisiana)
4. Midwestern State University (Texas)
5. Montana State University – Billings
6. Northwestern State University of Louisiana
7. Pittsburg State University (Kansas)
8. SUNY Institute of Technology at Utica – Rome (New York)
9. Southern Polytechnic State University (Georgia)
10. The University of Texas of the Permian Basin
11. University of South Florida – St. Petersburg
12. Western Carolina University (North Carolina)

Within this cluster, eight institutions reside in the same geographic region as UNA. Furthermore, Auburn University at Montgomery, Northwestern State University of Louisiana, Pittsburg State University, and Western Carolina University are also listed among UNA's current peers. While, initially, it looks like UNA may be a better fit with cluster three than with two, a look at the actual variables used in the analysis may shed some additional light. It should be noted that Albany State University is an Historically Black College and University (HBCU) with a different role, scope, and mission than UNA. While the initial screening did not include HBCU's as a sole critera source, these institutions were also not excluded from the study. Also, since the IPEDS data were released, Southern Polytechnic State University has merged with Kennesaw State University and is no longer a stand-along institution.

Table 3: UNA Values Compared to Primary and Secondary Clusters

Variables Used in Study	UNA Value	Primary Cluster Mean	Secondary Cluster Mean
Undergraduate enrollment for latest fall semester	6,098	6,371.75	5,017.15
Graduate enrollment for latest fall semester	934	959.69	788.62
FTE for latest fall semester	5,933	6,027.94	4,735.46
Six-year graduation rate based on the IPEDS defined freshman cohort	32	45.19	36.69
Total core revenues	82,986,521	89,108,260.56	78,797,441.46
Tuition and fee as percent of core revenues	44	41.63	32.00
State appropriations as a percent of core expenditures	32	33.38	34.69
Total core expenditures	77,588,308	81,075,721.88	70,838,115.31
Instructional costs as a percent of core expenditures	51	53.19	44.38
Endowment Assets per FTE	3,946	2,377.25	4,363.54
In-state tuition and fees on-campus	16,564	20,191.94	18,006.85
Out-of-state tuition and fees on-campus	21,892	28,818.75	27,592.54

The data listed in **Table 3** includes all of the twelve variables used in the study, UNA's value for each of these variables, and the mean values of each variable for cluster two (Primary), and cluster three (Secondary). From these data, it is clear that UNA's values more closely align with the means of cluster two than they do with cluster three. The exceptions are the six-year graduation rate where UNA is lower than both means but closer

to cluster three; endowment assets per FTE where UNA is more in line with cluster three; and both in- and out-of-state tuition where UNA is lower than both means but closer to cluster three.

According to this study, institutions currently belonging to UNA's peer group that were not present in either cluster two or three included Austin Peay State University, Jacksonville State University, Morehead State University, Murray State University, and the University of West Georgia. Based on the results of this study, all of these instutitons were placed into cluster four, which, according to the graph on **Table 2**, indicates a definitive cluster with no overlap on cluster two. Comparing to the average values of the 12 variables, UNA is significantly lower on most.

While the initial study did not narrow down peer selection by geographic area, the proximity of the institutions being compared to should also be considered concerning such factors as cost-of-living and regional accrediting associations. These factors could significantly affect comparability within any model.

Recommendations

Based on the cluster analysis outcomes from this study, along with external factors such as cost-of-living and accredition considerations, the Office of Institutional Research, Planning, and Assessment recommends eight institutions that were within both the second and third cluster. The data within **Table 4** indicates the institution chosen, which cluster the institution was grouped, and if the institution is in UNA's current cluster.

While cluster analysis is clearly an exploratory data analysis technique for classifying and organizing institutions into meaningful groups, the results of such analyses are not definitive and should be reviewed with other quantitative and qualitative criteria. These methods, however, can save time and resources as institutions seek to find peer institutions to match their benchmarking needs.

Table 4: Recommended Peer Institutions

Institution	Cluster No.	Current Peer
Nicholls State University (Louisiana)	2	Yes
Auburn University at Montgomery	3	Yes
McNeese State University (Louisiana)	3	No
Northwestern State University of Louisiana	3	Yes
Midwestern State University (Texas)	3	No
Pittsburg State University (Kansas)	3	Yes
Radford University (Virginia)	3	No
University of South Florida - St. Petersburg	3	No
Western Carolina University (North Carolina)	3	No

References

McLaughlin, G.W. and McLaughlin, J.S. (2007). The information mosaic: Strategic Decision making for universities and colleges. AGB: Washington, DC.

Teeter, D. J and Brinkman, P.T. (1987). Peer institutional studies/institutional comparisons. Primer for Institutional Research, J. Muffo and G. McLaughlin (eds), Association for Institutional Research: Tallahassee.

Terenzini, P. T., Hartmark, L., Lorang, W. G., & Shirley, R. C. (1980). A conceptual and methodological approach to the identification of peer institutions. Research in Higher Education,12, 347-364.