Paper SDA-010

Using SAS® PROC CLUSTER to Determine University Benchmarking Peers

Elayne Reiss, Sandra Archer, Robert Armacost, Ying Sun, and Yun (Helen) Fu
University of Central Florida, Orlando, FL

ABSTRACT

This paper will explore the steps taken by a large public research university to develop a list of peer institutions for consistent use in future benchmarking studies. At the heart of this process is the use of two tools available in SAS/STAT® – PROC CLUSTER and PROC FASTCLUS. Both of these procedures address clustering in differing ways, but this paper will demonstrate how we utilized the strengths of both procedures to strengthen our analysis. In describing the analytical process, we will focus on data selection and preparation concerns, the specifics of using these particular SAS® procedures for our benchmarking application, and the ways by which this application of clustering fits into benchmarking activities as a whole.

INTRODUCTION

In portions of the business world, the word “benchmarking” may suggest comparisons to sales and profit figures. However, in higher education, where the most important output is the latent variable of student learning, benchmarking is still an important process but is much more subjective in nature. When institutions engage in the benchmarking process, peer institutions with similar (or slightly improved) characteristics are identified so that programs within the institution can gain insight into the more tangible aspects of student learning and how they measure up over time.

The approaches taken by universities to identify benchmarking peers are as varied as their philosophies on the appropriate use of these lists of institutions. The preferred outcome in preparing the peer list is to attain the right mix of data, statistics, and judgment to ensure that all viewpoints are considered. The clustering methods present here, specifically those made available through the PROC CLUSTER and PROC FASTCLUS procedures in SAS/STAT®, provide an unbiased, data-driven approach to recommend a fair list of “comparison” peers to university leadership. The structure and development of the clustering models will be the main focus, with additional discussion on data preparation, model outcomes, and the ways by which input from key stakeholders was incorporated into the overall process.

WHY CLUSTER ANALYSIS?

The use of cluster analysis to develop a list of comparison peer institutions provides university leadership with an unbiased data-driven peer selection process for benchmarking purposes. The approach taken by this large metropolitan research institution is to create two sets of benchmarking peers: “comparison” peers and “aspirational” peers. The lists are intended to be held static over the course of the university’s strategic plan and provide a means by which progress toward the plan may be evaluated.

Cluster analysis is an exercise in exploratory data analysis that serves as an ideal tool for solving segmentation problems. This analysis involves classifying objects (in our case, universities) into groups, or clusters. There are a number of mathematical variants in determining the distance between groups to calculate the makeup of a cluster, but all methods involve the common goal of detecting groups within the data. Statistical significance of groupings is achieved when there is a degree of homogeneity within a cluster and heterogeneity between clusters of observations.

VARIABLES

In order to appropriately provide similarly-measured statistics to all institutions, it is necessary to begin with uniformly collected data. For this activity, all measures were collected from the datasets available as a part of the National Center for Educational Statistics (NCES) Integrated Postsecondary Education Data System (IPEDS), which holds information about every postsecondary institution that is eligible for federal aid. Since the IPEDS database maintains a quality standard for all of its contained information, there were no concerns regarding the accuracy of the data. However, having accurate data does not ensure a lack of missing values in some variables, which is an issue that will be addressed during the discussion of the clustering analysis.

SCOPE

Once all of the applicable datasets from IPEDS were obtained and brought into Base SAS®, management judgment was used to further filter the observations (institutions) to yield a more straightforward, applicable clustering analysis. The University of Central Florida (UCF) is a doctoral-granting university classified as a high research institution by the Carnegie Foundation for the Advancement of Teaching, a commonly-used standard for grouping postsecondary institutions by function. Therefore, only institutions fitting in the doctoral-granting research categories were considered. Additionally, UCF is an exceptionally large school with over 53,000 students enrolled as of fall 2009. Very small schools would not have the same culture and resources as UCF. After running basic frequency distributions, only those schools with enrollment over 18,000 were selected. Finally, because an institution's status as public or private influences major factors such as funding sources, all private institutions were removed from the dataset so UCF would be compared only to other public universities. These initial screening criteria reduced the number of observations in the dataset to 94.

VARIABLE SELECTION

For this benchmarking exercise, it was desirable to collect variables that would address a wide array of institutional descriptors, including student characteristics, degree offerings and productivity, financial issues, and faculty qualities. The specific variables used are listed in Table 1.

Table 1: List of Variables Used in Cluster Analysis

Category	Variable Name	Variable Type
Student Body Characteristics	Ratio of undergraduate to graduate students	Continuous
	Ratio of full-time to part-time students	Continuous
	Ratio of undergraduate degrees awarded to undergraduate enrollment	Continuous
	Ratio of graduate degrees awarded to graduate enrollment	Continuous
	Number of bachelor's degrees awarded	Continuous
	Number of master's degrees awarded	Continuous
	Number of doctoral degrees awarded	Continuous
Faculty Characteristics	Number of full-time faculty members	Continuous
	Ratio of publications to full-time faculty members	Continuous
Financial Characteristics	Expenditures per student	Continuous
	End of year value of endowment assets	Continuous
	NSF-funded research and development expenditures	Continuous
Other Characteristics	Institution grants a medical degree (Yes/No)	Binary
	Degree of urbanization of surrounding area	Ordinal

Numerous factors were considered in the decision of producing a final variable dataset.

Representation: Stakeholder input was taken into consideration when selecting variables that would jointly represent issues of importance to UCF as leaders assessed areas for continued growth. For example, with the opening of a new medical school at UCF, it was important to place some weight upon those institutions that also had this offering.

Uniqueness: A prior benchmarking exercise examined additional variables that were highly correlated with each other. When variables are not highly correlated with each other, they better represent unique qualities to measure that will not lead to redundancy when attempting to cluster observations.

Variable Structure: As will be described in a later section, optimal variables for clustering are typically continuous or binary in nature, not ordinal. Therefore, it was desirable to keep ordinal variables in the dataset to a minimum.

With this list of variables in place, the clustering of observations could begin.

PROC FASTCLUS AND PROC CLUSTER

This analysis employed two different clustering procedures in SAS/STAT®: PROC FASTCLUS and PROC CLUSTER. The two methods have advantages and disadvantages, so a method was developed to use them in conjunction with one another to maximize their potential.

PROC FASTCLUS

The PROC FASTCLUS method utilizes nearest centroid sorting to identify clusters. Means of clusters are first predicted through a set of points called cluster seeds. Temporary clusters are then formed by assigning each observation to the nearest seed. The means of these temporary clusters are then calculated to replace the seeds. This iterative process continues until the composition of the cluster no longer changes. This non-hierarchical method only utilizes nearest centroid sorting; other calculation methods are handled through PROC CLUSTER.

PROC CLUSTER

In PROC CLUSTER, the analyst can select from eleven different mathematical methods to hierarchically cluster observations in the SAS® data set. This process is hierarchical in that each individual observation begins as its own cluster. Distances between clusters are then calculated using the selected method; single-observation clusters are then turned into two-observation clusters to replace the old single-observation clusters. This iterative process continues until the process groups all of the observations into a single large cluster.

SUMMARY OF DIFFERENCES

Table 2 summarizes some of the advantages, disadvantages, and overall differences of both clustering methods as applied to this project.

Table 2: PROC FASTCLUS vs. PROC CLUSTER

	PROC FASTCLUS	PROC CLUSTER
Method	distance-based disjoint	hierarchical
Steps	• Arbitrarily choose *k* observations as the "seeds" • Assign all other observations to their closest "seed • Update the cluster mean and re-define the center	• Each observation begins in a cluster by itself • The two closest clusters are merged to form a new cluster • Merging continues until only one cluster is left
Advantages	• Faster • Can handle large datasets	• Produces a tree visualization • Provides more process details
Disadvantages	• Must specify number of clusters • Different seeds = different results	• Slower • Not suitable for larger datasets • Missing value imputation necessary

DATA PREPARATION

Before beginning the clustering analysis, one must ensure that the data are ready for the procedures. Several techniques were used to make the data appropriate for PROC FASTCLUS and PROC CLUSTER processing.

MISSING VALUES

Missing data are handled differently by PROC FASTCLUS and PROC CLUSTER. If there are missing values in PROC FASTCLUS, an adjusted distance is computed using the non-missing values. On the other hand, while using PROC CLUSTER, observations with missing values are excluded from the analysis and therefore will remain unclustered.

In this project, there were many missing values regarding institutional finances. During data collection, it was found that two reporting standards for expenses were used by institutions, GASB and FASB. GASB was used mainly by public institutions, which represented approximately two-thirds of the collected institutions. The FASB accounting system was primarily used by private institutions, representing approximately one-third of the collected institutions. As the expenses under these two systems were not comparable, only GASB data was used in cluster analysis. Therefore, financial data under FASB were left missing.

Since we wanted to evaluate all the universities regardless of the selected accounting system, PROC FASTCLUS and PROC CLUSTER were used in conjunction to handle the missing values. PROC FASTCLUS was used for a preliminary cluster analysis, producing a large number of clusters. PROC CLUSTER was then used to cluster the preliminary clusters hierarchically. This method was noted within the help files for SAS/STAT® as a solution to handle large datasets with PROC CLUSTER. In this example, a similar method was employed to handle missing data.

ORDINAL VALUES

Cluster analysis works most appropriately with binary or continuous data. Therefore, we were faced with a necessary decision when degree of urbanization, an ordinal variable describing the population level of the metropolitan area surrounding an institution, was selected as a variable of interest. As ordinal, or ranked, data are generally not appropriate to use in cluster analysis, it was transformed into a binary variable. A new variable was formed with a value of "1" for city-large, city-midsize, city-small and suburb-large and "0" otherwise. The final cut points were determined by descriptive statistics and management judgment and are located in Table 3**.**

Table 3: Levels of Degree of Urbanization to Transformed Binary Variable

Large Area (Urban = 1)	Small Area (Urban = 0)
City: Large	Suburb: Midsize
City: Midsize	Suburb: Small
City: Small	Town: Fringe
Suburb: Large	Town: Distant
	Town: Remote
	Rural: Fringe
	Rural: Distant
	Rural: Remote

VARIABLE STANDARDIZATION

In cluster analysis, variables with large values contribute more to the distance calculations, the basis for determining the cluster into which each observation is assigned. In order to avoid this occurrence from potentially spoiling the results of an analysis due to improper variable treatment, each variable needs to be transformed, or standardized. SAS provides the STANDARD procedure to handle this. In this example we standardize all the analytical variables to a mean of zero and standard deviation of one. In the example code below, the procedure creates the output data set *peer2006std* to contain the transformed variables.

```
/*standardize all the related variables, remove the effects of variable range, after
standardization, all the variables should have mean=0 standard deviation=1*/

proc standard data=peer2006 out=peer2006std mean=0 std=1;
    var V1-V14;
run;
```

ANALYSIS

After preparing the data, both PROC FASTCLUS and PROC CLUSTER were run multiple times and the results were combined and ranked by a scoring method.

USING PROC FASTCLUS TO DETERMINE PEERS

This section will address how PROC FASTCLUS was used to form clusters within a large list of institutions using multiple runs of the procedure. Each procedure had a different numbers of clusters identified.

PROC FASTCLUS uses a method called the k-means model, in which the user defines the number of clusters (*k*) and the observations are divided into *k* disjoint clusters. The process of this procedure is constructed in a straightforward manner. First, *k* cluster seeds are selected as the first guess of the centers of each cluster. By default, the first *k* observations with no missing values are selected as the initial seeds. The rest of the observations are then assigned to the cluster to which it is closest, and the cluster center is then updated. An iterative process subsequently begins where cases are grouped into their closest clusters and centers are recomputed until no further changes occur in the cluster centers. The final output contains a set of observations each belonging to a single cluster.

The following is the macro ***Clus*** used to cluster the institutions and demonstrate the results. Three procedures, PROC FASTCLUS, PROC CANDISC AND PROC SGPLOT, were used in this macro. The macro variable **N** represents the number of clusters defined in the PROC FASTCLUS procedure.

```
/*Run the fastclus first, no tree structure is required*/

%macro clus(N);

proc fastclus data=peer2006std out=peer2006clus&N.
     maxclusters= &N. maxiter=100;
     var v1-v14;
run;


 /*Use proc candisc along with proc gplot to get the visualized picture of the
clusters*/

proc candisc data=peer2006clus&N. out=peer2006cluscan;
    class cluster;
    var v1-v14;
run;

proc sgplot data=peer2006cluscan;
    scatter y=can2 x=can1 / group=cluster;
run; %mend;
%clus(3);
%clus(4);
%clus(5);
……
```

The FASTCLUS procedure uses the standardized data set *peer2006std* as input and creates the data set *peer2006clus&N*. Options REPLACE and RADIUS are set at their respective default values (REPLACE=FULL, RADIUS=0). REPLACE specifies the seed replacement method. RADIUS specifies the minimum distance for selecting new seeds. Both the MAXCLUSTERS and MAXITER options are set for our specific demands. MAXITER specifies the maximum number of iterations. We chose to set this as 100 to prevent a case of the procedure running for a long time when the data do not converge. MAXCLUSTERS specifies the maximum number of clusters. In this application, in order to obtain the institutions that were closest to UCF, we tried several values of the MAXCLUSTERS option and focused on the members in UCF's cluster in each data run. The VAR statement specifies the variables used in the cluster analysis.

The output data set *peer2006clus&N* contains the original variables and two new variables, **cluster** and **distance**. The variable **cluster** contains the cluster identification number to which each observation has been assigned. The variable **distance** contains the distance from the observation to its cluster seed. An example of this output is contained in Figure 1.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	UnitID	Institution	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	v11	v12	V13	V14	CLUSTER	DISTANCE
2	122409	San Diego State University	1.13	0.92	-0.4	0.64	0.11	-0.73	-0.12	-0.58	-1.31	-1	-0.48	-0.52	-0.88	0.44	4	1.433559
3	130943	University of Delaware	1.13	0.81	0.13	0.33	-0.62	-0.54	0.59	-0.47	-0.54			0.43	-0.72	0.44	4	1.803644
4	132903	University of Central Florida	1.13	1.45	-0.6	0.13	-0.55	-0.57	-0.45	-0.54	-0.92	-0.99	-0.49	-0.43	-0.57	0.44	4	1.241407
5	150136	Ball State University	1.13	1.53	0.01	-0.06	0.02	-1.03	-0.79	-0.47	-1.1	-0.75	-0.6	-1.39	-0.91	0.44	4	1.708454
6	145813	Illinois State University	1.13	2.34	0.21	1.09	-0.93	-1.04	0.43	-0.2	-0.92	-0.79	-0.55	-1.2	-1.02	0.44	4	2.706618
7	133951	Florida International University	1.13	1.18	-0.76	-0.8	0.63	-0.76	-0.31	-0.1	-0.85	-0.94		-0.29	-1.01	0.44	4	1.600977
8	206084	University of Toledo-Main Campus	1.13	1.88	-0.4	-0.81	1.04	-0.92	1.05	0.45	-1.03	-0.72	-0.55	-0.86	-1.01	0.44	4	2.669791
9	200800	University of Akron Main Campus	1.13	1.22	-0.59	-2.13	-0.11	-0.93	0.59	0	-1.03	-0.79	-0.54	-0.5	-1.14	0.44	4	2.184211
10	126614	University of Colorado at Boulder	1.13	1.29	-0.19	0.15	-0.41	1.28	-1.37	-1.12	-0.18	-0.35		0.33	0.31	0.44	4	3.213122
11	133669	Florida Atlantic University	1.13	0.85	-0.83	0.23	-0.87	-0.93	-0.88	-1.02	-1.07	-0.9	-0.4	-0.94	-1.01	0.44	4	1.392484
12	203517	Kent State University-Kent Campus	1.13	0.04	-0.28	0.05	-0.2	-1.01	-0.24	-0.34	-0.36	-0.74	-0.54	-0.79	-0.97	0.44	4	0.910124
13	229115	Texas Tech University	1.13	0.99	0.25	0.11	0.17	-0.8	0.24	0.69	0.35	-0.67	-0.21	0.08	-0.73	0.44	4	2.183661
14	209551	University of Oregon	1.13	0.9	0.35	0.95	1.22	-0.8	0.24	-0.23	-0.5	-0.39	-0.15	-0.44	-0.74	0.44	4	2.483656
15	225511	University of Houston	1.13	1.22	-0.6	-0.96	1.91	-0.72	-0.91	0.24	-0.39	-0.64	-0.33	-0.12	-0.65	0.44	4	2.857069
16	104151	Arizona State University at the Tempe Campus	1.13	0.67	-0.54	-0.84	-0.27	-0.14	-0.06	-0.3	-0.15	-0.59	-0.13	0.29	0.08		4	1.745183

Figure 1. Standardized Variable Values with Cluster Assignments and Distances to Cluster Seed

In this project, we used the PROC CANDISC and PROC SGPLOT procedures to check the distribution of the clusters graphically.

First, the CANDISC procedure performs a canonical discriminant analysis which finds linear combinations of variables, known as canonical variables, that provide maximum separation between classes or groups. The CLASS statement specifies the variable **cluster** to define groups for the analysis. The VAR statement specifies the variables used in the analysis.

Next, the SGPLOT procedure plots the two canonical variables generated from PROC CANDISC, **can1** and **can2**. The PLOT statement specifies the variable **cluster** as the identification variable. The resulting plot in Figure 2 illustrates the distribution of clusters from the FASTCLUS procedure.

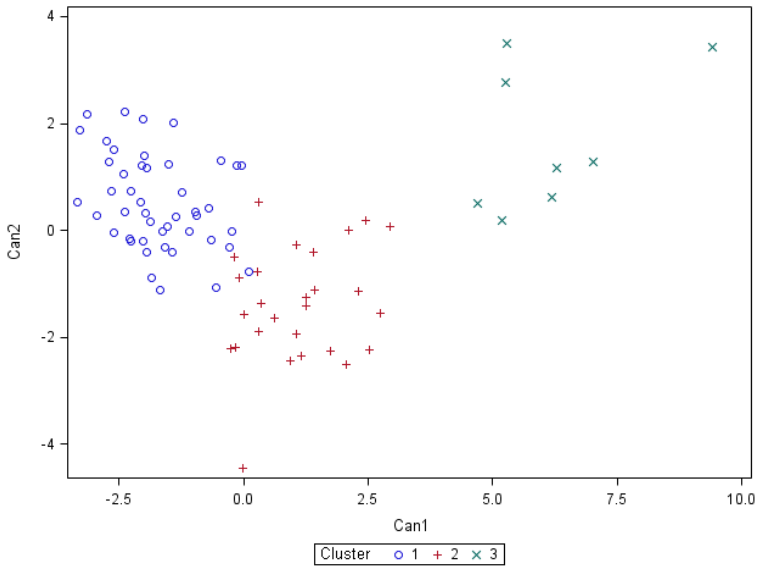

Figure 2. A Sample of a Resulting Plot of Clusters from PROC FASTCLUS

Multiple runs specifying different number of MAXCLUSTERS were conducted using 14 variables. The results of the cluster analysis were summarized and the members in the same cluster with UCF were counted. Only those data runs in which there were 10-20 institutions in UCF's cluster were discussed here. For each institution, the number of times they fell into UCF's cluster of 10-20 members were summarized and counted.

The results showed that there were 11 institutions falling into UCF's cluster at least 11 times out of the 18 runs and 10 of them were included in the final "comparison" peer list of UCF after combined results from both PROC FASTCLUS and PROC CLUSTER, a method that will be discussed in a subsequent section.

USING PROC CLUSTER TO ANALYZE PEERS

PROC CLUSTER was used to analyze the 94 institutions as well. PROC CLUSTER displays a history of the clustering process and creates an output data set that can be used to draw a tree diagram of the cluster hierarchy or to output the cluster membership at any desired level.

Unlike PROC FASTCLUS, PROC CLUSTER uses a hierarchical clustering method. The process begins by joining the closest pair of individual observations into a cluster and in a stepwise fashion, joins a pair of observations or pair of clusters until all observations are joined in the final cluster. Thus, with this method, once an observation joins a cluster it remains with the others in that cluster through to the final step. However, this method removes observations if any variable is left blank. Eleven different methods are available in PROC CLUSTER to carry out this analysis.

As previously discussed, this hierarchical method does not allow for missing observations. Therefore, those institutions with missing values would be excluded from the analysis. In the final data set of the 94 institutions, 14 of them had missing values. To include the 14 institutions in the analysis, PROC FASTCLUS was first conducted defining the maximum number of clusters to be 80. This was the largest possible number of clusters to yield both an absence of missing data and the lowest number of members in each cluster, which would minimize the effect of this additional use of PROC FASTCLUS on the final results.

```
/*Use Proc Fastclus for a preliminary cluster analysis producing a large number of
clusters and then use PROC CLUSTER to hierarchically cluster the preliminary
clusters*/

/*Run the fastclus first for preliminary cluster*/

proc fastclus data=peer2006std summary out=prelim_94
     maxclusters=80 maxiter=100 converge=0 mean=mean cluster=preclus;
     var v1-v14;
run;
```

Within PROC FASTCLUS, the SUMMARY statement suppresses the display of certain extra results, such as the statistics for variables. CONVERGE specifies the convergence criterion. The statement MEAN=[*SAS-data-set*] creates an output data set *mean* that contains the cluster means and other statistics for each cluster, which is then automatically used in the PROC CLUSTER process. CLUSTER=[**variable-name**] specifies a name for the variable that indicates cluster membership.

PROC CLUSTER was then utilized to analyze the preliminary clusters as shown in the following macro. PROC TREE was also run to produce the tree diagram, as well as PROC CANDISC and PROC SGPLOT to demonstrate the distribution of the final clusters as described previously.

```
/*The following macro, Clus, clusters the preliminary clusters*/

%macro clus(method);
    proc cluster data=mean method=&method;
       var v1-v14;
       copy preclus;
    run;

    proc tree ncl=6 out=out_94_&method;
       copy v1-v14 preclus;
    run;


    proc candisc data=out_94_&method out=peer2006cluscan&method distance anova;
       class cluster;
       var v1-v14;
    run;

    proc sgplot data=peer2006cluscan&method;
       plot can2*can1=cluster/frame cframe=ligr
       legend=legend1 vaxis=axis1 haxis=axis2;
    run;
%mend;

%clus(ward);
%clus(average);
%clus(centroid);
…
```

The macro variable represents the method specified in the METHOD statement of the PROC CLUSTER analysis. Eleven methods are provided in this procedure, with option names provided in parentheses: average linkage (average), centroid, complete linkage (complete), density linkage (density), maximum-likelihood hierarchical (eml), Lance-Williams flexible-beta (flexible), McQuitty's similarity analysis (mcquitty), Gower's median method (median), single linkage (single), two-stage density linkage (twostage), and Ward's minimum-variance method (ward).

The TREE procedure produces high-resolution graphics by default; a sample graphic is provided in Figure 3. The COPY statement copies the variables **v1** to **v14** and **preclus** from the preliminary PROC FASTCLUS clustering result into the output SAS data set *out_94_&method*. The CANDISC and PLOT statements request a plot of the two canonical variables, using the value of the variable **cluster** as the identification variable.

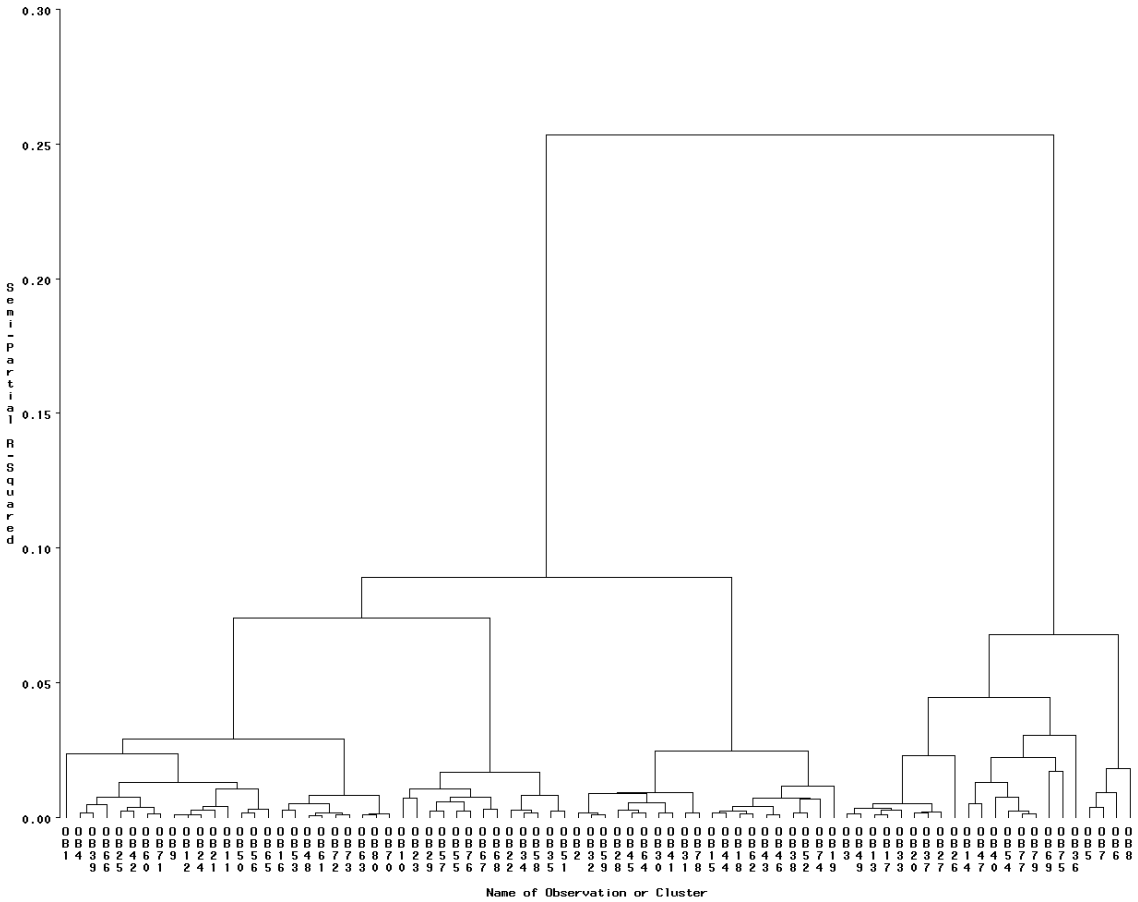

Figure 3: A Sample of the Tree Diagram Generated by PROC CLUSTER

Multiple data runs were applied using different linking methods and tree diagrams were drawn to demonstrate the hierarchical relations of the observations (the preliminary clusters). The tree diagrams were analyzed and the observations within the small branch of UCF of about 10-20 members were summarized. Similar to the method using PROC FASTCLUS, the number of times an institution fell into UCF's cluster of about 10-20 members was counted. The result showed that there were 14 institutions that fell into UCF's cluster at least 4 times of the 11 runs.

COMBINING METHODS

The results from PROC FASTCLUS and PROC CLUSTER were combined for institutions that appeared at least once in UCF's group of about 10-20 in both of the analyses. For each institution, the ratio of comparison peers under the two methods were generated using the number of runs that this specific institution fell into UCF's cluster divided by the total number of runs. A comparison peer ("CP") index was subsequently calculated as the sum of the ratios for each institution. This score can range from zero (did not appear in UCF's cluster at all) to two (appeared in UCF cluster during every run of both PROC FASTCLUS and PROC CLUSTER analyses).

Table 4 displays the list of UCF's comparison peers after this combined analysis, complete with the number of times the institution appeared in UCF's cluster group in both the PROC FASTCLUS and PROC CLUSTER analyses. The CP index is shown in the rightmost column. The 14 institutions with the highest CP index score were identified as potential comparison peers for UCF.

Table 4: Summary of UCF's Comparison Peer List after Combined Analyses

Institution	FASTCLUS (18 Runs)	CLUSTER (11 Runs)	CP Index
University of Central Florida	*18*	*11*	*2.00*
San Diego State University	18	11	2.00
University of Delaware	18	11	2.00
Ball State University	16	11	1.89
Illinois State University	18	6	1.55
Florida International University	14	8	1.51
University of Toledo – Main Campus	14	8	1.51
University of Akron – Main Campus	13	8	1.45
University of Colorado at Boulder	6	11	1.33
Florida Atlantic University	5	11	1.28
Kent State University – Kent Campus	16	4	1.25
Texas Tech University	13	5	1.18
University of Oregon	12	5	1.12
University of Houston	5	8	1.01

ADDITIONAL ANALYSIS AND STAKEHOLDER INPUT

As described throughout this paper, cluster analysis was performed using SAS® on 14 key variables that describe student population, degree productivity, scope of degree offerings, institutional finances, research productivity, faculty, and the degree of urbanization of the surrounding area.

The cluster analysis results provided the first suggestion for a list of potential peers, but several iterations of refinement were required in addition to this quantitative analysis to determine a final set of benchmarking peers. This refinement phase considered several factors that were not included in the cluster analysis.

The types of programs offered at an institution of higher education have a very large effect on the institution as a whole. Therefore, one factor carefully considered was program mix as measured by the proportions of graduate degrees awarded by discipline. The analysis of program mix was conducted in Excel using degree awarded data from IPEDS. Another important factor considered was the demographics of the surrounding metropolitan area of each potential peer institution. This analysis was conducted with a "ring" study of surrounding metropolitan areas by using mapping software. Areas that had similar diversity, income levels, and population growth projections as Orlando were considered. All these various aspects of analysis were provided to key stakeholders for consideration. Additional qualitative factors were discussed and carefully considered, including institutional mission and reputation. Finally, all aspects of analysis were synthesized to determine the final list of benchmarking peers, listed in Table 5.

Table 5: Final List of Benchmarking Peers for UCF

Comparison Peers	Aspirational Peers
Florida Atlantic University Florida International University Georgia State University Kent State University Portland State University San Diego State University University of Akron University of Delaware University of Houston University of New Mexico University of Texas – Arlington University of North Carolina – Charlotte University of South Florida Virginia Commonwealth University	Arizona State University – Tempe Auburn University North Carolina State University – Raleigh Oregon State University University of Cincinnati University of Colorado – Boulder University of Nebraska – Lincoln University of South Carolina – Columbia

CONCLUSION

Benchmarking is an important process in nearly any industry, yet its success depends upon the careful selection of a comparison list. Selections performed solely on the basis of management judgment can lead to disappointing and inaccurate results. The example discussed in this paper demonstrated how the clustering techniques available in SAS/STAT® can assist analysts in providing a data-driven approach to creating a benchmarking comparison list. Although this approach is limited to quantifiable variables and works best with continuous or binary measures, the combination of PROC FASTCLUS and PROC CLUSTER can overcome some other common data anomalies, such as missing observations. These procedures allow analysts to generate organized output, both tabular and graphical in nature, to share with others responsible for determining the members of a comparison list. It is critical to still apply some degree of management judgment in a benchmarking situation; however, with the use of a multi-faceted approach such as cluster analysis, decision-makers can begin with a much shorter list and benefit from a process with enhanced efficiency.

REFERENCES

[1] Robert L. Armacost, Alicia L. Wilson, Deivanayak Balakrishnan, Peer analysis, Technical Report, 2002: UAPS-TR-01-002.
[2] Sandra Archer, Shuxin Li, Identifying UCF Benchmarking Peers 2007 Phase I: Study of Methods, Technical Report, 2008: UAPS-TR-07-001.
[3] Sandra Archer, Ying Sun, Identifying UCF Benchmarking Peers 2007 Phase II: Updated Data Results, Technical Report, 2009: UAPS-TR-09-001.
[4] SAS® 9.1.3 Documentation, http://support.sas.com/onlinedoc/913/docMainpage.jsp.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Dr. Elayne Reiss (ereiss@mail.ucf.edu)
Dr. Sandra Archer (archer@mail.ucf.edu)
Dr. Yun (Helen) Fu (helenfu@mail.ucf.edu)
Ms. Ying Sun
University of Central Florida
Office of University Analysis & Planning Support
12424 Research Pkwy, Suite 215
Orlando FL 32826-3207
Phone: (407) 882-0285
Web: http://uaps.ucf.edu

Dr. Robert Armacost (armacost@mail.ucf.edu)
UCF College of Medicine
6850 Lake Nona Blvd, Suite 312C
Orlando FL 32827
Phone: (407) 266-1000