Paper 1279-2014

Selecting Peer Institutions with Cluster Analysis

Diana Suhr
University of Northern Colorado

Abstract
Universities strive to be competitive in the quality of education as well as cost of attendance. Peer institutions are selected to make comparisons pertaining to academics, costs, and revenues. These comparisons lead to strategic decisions and long-range planning to meet goals. The process of finding comparable institutions could be completed with cluster analysis, a statistical technique. Cluster analysis places universities with similar characteristics into groups or clusters. A process to determine peer universities will be illustrated using PROC STANDARD, PROC FASTCLUS, and PROC CLUSTER.

Introduction

Cluster analysis is an exploratory data analysis technique for classifying and organizing data into meaningful clusters, groups, or taxonomies by maximizing the similarity between observations in 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 cluster analysis is to sort cases (people, things, events) into groups, or clusters, so that the degree of association/relationship is strong between members of the same cluster and weak between members of different clusters. Cluster analysis creates groups without any preconceived notion.

The appropriate clustering algorithm and parameter settings (including values such as the distance function, density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis is an iterative process of knowledge discovery and optimization to modify data processing and model parameters until the result achieves preferred as well as appropriate properties.

Cluster analysis originated in anthropology through studies by Driver and Kroeber in 1932 and was introduced to psychology by Zubin in 1938 and Tryon in 1939. Cluster analysis techniques were used by Cattell beginning in 1943 for trait theory classification in personality psychology.

Applications

Clustering occurs in most areas of daily life. SAS® Global Forum clustered presentations by skill level, industry, and job role. The United States can be divided into a number of clusters according to lifestyle, buying habits, climate, job opportunities, cost of living, or any of a number of variables.

In business, consumer segment clusters are found to determine successful marketing strategies. Groups of individual are formed with similar buying habits, behaviors, and demographics. The characteristics of each group help to identify different marketing approaches that would be appropriate for each group. Clustering is used in software development to restructure and improve software functionality. When analyzing social networks, clustering is used to identify communities. Internet search engines use clustering to search more intelligently.

A medical diagnosis looks at symptoms and diseases to provide a taxonomy of illnesses. Psychological inventories cluster patients into subgroups and determine appropriate treatment while studying characteristics of diseases. Biologists organize and group different species of animals before identifying differences between animals. A well-known clustering classifies stars into a main sequence, white giants, and red dwarfs, according to temperature and luminosity.

A grocery store displays groups of similar products together, such as meats or vegetables or bakery items. Cluster analysis in the food industry could make recommendations for new items based on customer preferences. The military uses cluster analysis of body measurement data to reduce the number of different uniform sizes kept in inventory. Law enforcement resources can be managed more effectively by identifying areas of higher and similar types of crime. Education uses cluster analysis to identify and examine groups of students or schools with similar properties.

Cluster analysis vs. discriminant and factor analysis
Both cluster analysis and discriminant analysis deal with classification. Discriminant analysis requires prior knowledge of group membership to classify observations while cluster analysis requires no prior knowledge of which observations belong to which groups. Cluster analysis may be used in conjunction with discriminant analysis, discovering a linear structure of either the measures used in the cluster analysis and/or different measures.

Cluster analysis, like factor analysis, makes no distinction between independent and dependent variables. Factor analysis reduces the number of variables by grouping them into a smaller set of factors. Cluster analysis reduces the number of observations by grouping them into a smaller set of clusters.

Cluster Analysis Steps
Cluster analysis methods are not clearly established. There are many options one may select when performing cluster analysis. For example, each case could start as a separate cluster and continue to combine similar clusters until there is one cluster left. This method uses distances between data points. The most common and generally accepted way to calculate distances between objects in a multi-dimensional space is to compute Euclidean distances (an extension of Pythagoras' theorem).

The choice of methods used for cluster analysis could depend on the size of the data set as well as the types of variables. Hierarchical clustering is appropriate for a small data set. K-means clustering where the number of clusters is known works for a moderately large data set. A large data set or a mixture of continuous and categorical variables requires a two-step procedure.

Start with a number of cases to subdivide into homogeneous groups.

1) Choose variables
2) Standardize variables (if there is a wide range of values and/or different scales of measurement). Variables with larger values contribute more than variables with smaller values. To provide an optimal solution, variables should contribute equally to the distance or similarity between observations.
3) Decide which clustering procedure to use, based on number of case and variable types.
4) Determine how many clusters to represent the data

Hierarchical clustering

- Choose a statistic that quantifies how far apart (or similar) two cases are.
- Select a method for forming groups
- Determine how many clusters to represent data

K-means clustering

- Select the number of clusters
- Algorithm selects cluster means
- Assigns cases to the cluster where the smallest distance to the cluster mean.

Standardizing Variables
If variables are measured on different scales, variables with large values contribute more to the distance measure than variables with small values. The distance between two people based on their IQ and income in dollars would find that the differences in incomes dominate any distance measures. For example, if a person's monthly income is $7500 while their IQ is 150, the squared distances are 56,250,000 and 22,500 respectively. The income will contribute more to the distance measures than the IQ. By standardizing each variable with a mean of 0 and a standard deviation of 1, each variable will be weighted equally.

PROC STANDARD syntax

```
PROC STANDARD DATA=<input dataset> OUT=<output dataset>
  MEAN = <mean value>  (specifies mean value)
  STD = <std value>    (specifies standard deviation value)
  <options>:
```

```
VAR <variables>;
FREQ <variable>;
WEIGHT <variable>;
BY <variables>;
```

Clustering Methods

There are many practical problems involved in the use of cluster analysis. The selection of variables to be included in the analysis, the choice of distance measure and the criteria for combining cases into clusters are all critical. Because the selected clustering method can itself impose a certain amount of structure on the data, it is possible for spurious clusters to be obtained. In general, several different methods should be used (Collins Dictionary of Sociology, 2006).

Hierarchical clustering, also known as connectivity based clustering, is one of the most straightforward methods and is based on the concept of objects being more related to nearby objects than to objects farther away. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. Hierarchical clustering can be agglomerative or divisive.

Agglomerative hierarchical clustering (bottom up) begins with single observations and aggregates them into clusters. At successive steps, similar clusters are merged. The algorithm ends with all observations in one "useless" cluster. In agglomerative clustering, once a cluster is formed, it cannot be split; it can only be combined with other clusters. Agglomerative hierarchical clustering doesn't let observations separate from clusters that they've joined. Once in a cluster, always in that cluster. Divisive clustering (top down) begins with all observations in one cluster and ends with individual clusters.

To form clusters using a hierarchical cluster analysis, select:

- A criterion for determining similarity or distance between cases
- A criterion for determining which clusters are merged at successive steps
- The number of clusters to represent the data

To find an appropriate cluster solution, the characteristics of the clusters must be examined at successive steps, e.g. frequencies or means, with a decision on an interpretable solution which includes a reasonable number of clusters and appropriately represents the data. There is no right or wrong answer as to how many clusters to select.

Warning: Computations for the selected distance measure are based on all of the variables selected. If a mixture of nominal and continuous variables are selected, it is recommended to use the two-step cluster procedure because the distance measures in hierarchical clustering or *k*-means are unsuitable for use with both types of variables.

Hierarchical clustering requires a distance or similarity matrix between all pairs of cases, which could translate into longer computing time. A clustering method that doesn't require computation of all possible distances is *k*-means clustering. It differs from hierarchical clustering in several ways. The number of clusters must be known a priori. Solutions for a range of the number of clusters are obtained by rerunning the analysis for each different number of clusters. The algorithm repeatedly reassigns cases to clusters, so the same case can move from cluster to cluster during the analysis.

The algorithm is called *k*-means, where *k* is the number of clusters selected, since a case is assigned to the cluster for which its distance to the cluster mean is the smallest. The algorithm centers around finding the *k*-means. The analysis begins with an initial set of means and classifies cases based on their squared distances to the centers. Next, cluster means are computed again, using the cases that are assigned to the cluster. Then all cases are reclassified based on the new set of means. This step is repeated until cluster means are almost equal between successive steps. Finally, cluster means are calculated again and cases are assigned to their permanent clusters.

Note: If a data set is large, take a random sample of the data and try to determine a good number, or range of numbers, for a cluster solution based on the hierarchical clustering procedure. Hierarchical cluster analysis can be used to estimate starting values for the *k*-means algorithm.

In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points.

The clustering model most closely related to statistics is based on distribution models. Clusters can then easily be defined as objects belonging most likely to the same distribution. This approach closely resembles the way artificial data sets are generated: by sampling random objects from a distribution.

While the theoretical foundation of these methods is excellent, there is one key problem known as overfitting (unless constraints are put on the model complexity). A more complex model will usually be able to explain the data better, which makes choosing the appropriate number and type of variables a key factor in model development.

PROC CLUSTER

The CLUSTER procedure hierarchically clusters the observations in a SAS data set by using one of eleven methods. The data can be coordinates or distances. If the data are coordinates, PROC CLUSTER computes squared Euclidean distances. Clustering methods are

- average linkage
- centroid method
- complete linkage
- density linkage (including Wong's hybrid and th-nearest-neighbor methods)
- maximum likelihood for mixtures of spherical multivariate normal distributions
- flexible-beta method
- McQuitty's similarity analysis
- median method
- single linkage
- two-stage density linkage
- Ward's minimum-variance.

All methods are based on the usual agglomerative hierarchical clustering procedure. Each observation begins in a cluster by itself. The two closest clusters are merged to form a new cluster that replaces the two old clusters. Merging of the two closest clusters is repeated until only one cluster is left. The various clustering methods differ in how the distance between two clusters is computed.

The CLUSTER procedure is not practical for very large data sets because the CPU time is roughly proportional to the square or cube of the number of observations. The FASTCLUS procedure requires time proportional to the number of observations and can be used with much larger data sets than PROC CLUSTER. If you want to cluster a very large data set hierarchically, use PROC CLUSTER for a preliminary cluster analysis. Then use PROC FASTCLUS to determine cluster membership.

Plots of the pseudo *F* statistic and cubic clustering criterion (CCC) are options in PROC CLUSTER. Values of CCC, pseudo F, and observed overall r-squared, and approximate expected overall r-squared are shown with PROC FASTCLUS. Values of cubic clustering criterion greater than 2 or 3 indicate good clusters. Values between 0 and 2 indicate potential clusters, but they should be taken with caution; large negative values can indicate outliers. Note: approximate expected r-R square and CCC values are not valid for correlated variables. For correlated variables, cluster principal component scores from the PROC PRINCOMP procedure.

PROC CLUSTER displays a history of the clustering process, showing statistics useful for estimating the number of clusters in the population from which the data are sampled. PROC CLUSTER also creates an output data set that can be used by the TREE procedure to draw a tree diagram of the cluster hierarchy or to output the cluster membership at any desired level. For example, to obtain the five-cluster solution, you could first use PROC CLUSTER with the OUTTREE= option, and then use this output data set as the input data set to the TREE procedure. With PROC TREE, specify NCLUSTERS=5 and the OUT= options to obtain the five-cluster solution and draw a tree diagram.

Syntax PROC CLUSTER

```
PROC CLUSTER DATA=<input dataset>
    METHOD=<name>
    OUTTREE=<dataset>
    std                 (standardize, mean=0, std=1)
    ccc                 (cubic clustering criterion plot)
    pseudo              (pseudo F plot)
    <options>;
  BY <variables>;       (separate analysis)
  COPY <variables>;     (copied from input dataset to OUTTREE=dataset)
  FREQ <variable>;      (frequency of occurrence of variables)
  ID <variables>;       (to identify observations)
  RMSSTD <variables>;   (variable containing root mean squared deviations)
  VAR <variables>;      (variables in the analysis)
```

PROC CLUSTER <options> include specifying the input and output data sets and clustering method, control of data processing prior to clustering, density estimation, display of cluster history, and output. Plots of the pseudo *F* statistic and cubic clustering criterion (CCC) are options in PROC CLUSTER. Values of CCC, pseudo F, and observed overall r-squared, and approximate expected overall r-squared are shown with PROC FASTCLUS. Note approximate expected r-R square and CCC values are not valid for correlated variables. For correlated variables, cluster principal component scores from the PROC PRINCOMP procedure. Values of cubic clustering criterion greater than 2 or 3 indicate good clusters. Values between 0 and 2 indicate potential clusters, but they should be taken with caution; large negative values can indicate outliers.

PROC FASTCLUS

PROC FASTCLUS is intended for use with large data sets, 100 or more observations. With small data sets, the results can be highly sensitive to the order of the observations in the data set.PROC FASTCLUS uses algorithms that place a larger influence on variables with larger variance. Standardizing the variables before performing the cluster analysis is recommended. PROC FASTCLUS produces brief summaries of the clusters it finds. For more extensive examination of the clusters, you can request an output data set containing a cluster membership variable.

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 such that every observation belongs to one and only one cluster; the clusters do not form a tree structure as they do in the CLUSTER procedure. For an analysis for different numbers of clusters, run PROC FASTCLUS several times changing the number of clusters requested in each analysis.

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 (an a –means clustering approach). PROC FASTCLUS is designed to find good clusters, not the best possible clusters, with only two or three passes through the data set. PROC FASTCLUS can be an effective procedure for detecting outliers which appear as clusters with only one member.

Syntax PROC FASTCLUS

```
PROC FASTCLUS DATA=     <input data>
   INSTAT =             <dataset created with OUTSTAT=>
   OUT =                <output SAS dataset>
   OUTSTAT =            <output SAS dataset containing statistics>
   CLUSTER =            <specifies name for cluster membership variable>
   MEAN =               <output SAS dataset containing cluster centers>
   MAXCLUSTERS=n;       (if MAXCLUSTERS is not specified, n=100, by default)
 VAR <variables>;       (variables in the analysis)
 ID <variables>;        (to identify observations)
 FREQ <variable>;       (frequency of occurrence of variables)
 WEIGHT <variable>;     (compete weighted cluster means)
 BY <variables>;        (separate analysis)
```

Graphical Representation
The results of the application of the clustering technique can be described using a dendogram or binary tree. The objects are represented as nodes in the dendogram and the branches illustrate when the cluster method joins subgroups containing that object. The length of the branch indicates the distance between the subgroups when they are joined.

A dendogram that clearly differentiates groups of objects will have small distances in the far branches of the tree and large differences in the near branches. When the distances on the far branches are large relative to the near branches, then the grouping is not very effective. Dendograms are also useful in discovering "runts", or objects that are joined to a group in the near branches. A runt, at least in a single linkage dendogram, does not join the main group until the last step. Runts are exceptions to the grouping structure.

A plot or scatter plot provides a graphical representation of the clusters (see Figure 1 for an example). Canonical variables are created using the PROC CANDISC procedure and plotted with PROC SGPLOT.

Syntax for PROC TREE

```
PROC TREE <options>;
  NAME <variables>;
  HEIGHT <variables>;
  PARENT <variables>;
  BY <variables>;
  COPY <variables>;
  FREQ <variables>;
  ID <variables>;
run;
```

Syntax for PROC SGPLOT;

```
proc candisc data=clus7 out=can7
noprint;
  var TotalPriceInstate_OnCampus--
InstrExp_Percent;
  class cluster;
proc sgplot data=can7;
  scatter y=can2 x=can1 /
group=cluster;
run;
```

Example
A data set was created with data extracted from the IPEDS Data Center. Public doctoral granting institutions were selected (n=310). Continuous data were used in the cluster analysis. Categorical data were used to describe the final cluster including the target institution. Variables include

- institution name
- total price for in-state students living on campus
- total price for out-of-state students living on campus
- undergraduate enrollment
- graduate enrollment
- graduation rate
- total dollars of core revenues
- tuition and fees as a percent of core revenues
- state appropriations as a percent of core revenues
- revenues from tuition and fees per FTE
- revenues from state appropriations per FTE
- instructional expenses as a percent of total core expenses
- state
- geographic region
- IPEDS data feedback report comparison group category

Note: Data presented in the example is simulated using data from the IPEDS data center. Due to confidentiality, actual data analysis is not shown.

Standardizing variable values
Data was examined by running frequencies and means. Ranges and average values of the variables used in the analysis varied. Variables with a larger value will influence and have more weight in the analysis. Therefore, variable values were standardized with PROC STANDARD, mean of zero and standard deviation of 1.

```
proc standard data=rawsub1 mean=0std=1 out=stan9;
varTotalPriceInstate_OnCampus--InstrExp_Percent;
run;
```

Determining the number of clusters
A two-step process was used to determine the number of clusters, cluster membership, and peer institutions. Cluster membership can change depending on the intent or purpose of clustering, on the variables used and criterion for selecting the number of clusters.

The first step included a cluster analysis with PROC CLUSTER, examining eigenvalues, cubic clustering criterion, and pseudo F. Emphasis was placed on eigenvalues with selection of the number of clusters using cumulative proportion and difference in eigenvalues. SAS code used for step 1 was

```
proc cluster data=rawsub1x outtree=tree1x std method=average ccc pseudo;
 var TotalPriceInstate_OnCampus--InstrExp_Percent;
 id unitid;
run;
```

Eigenvalues, differences, proportions, and cumulative proportions are shown in Table 1. A large difference in eigenvalues is shown between the fourth (1.0917) and fifth (0.4263) eigenvalues, proportions go from 0.0992 to 0.0388, with cumulative proportion for the fourth eigenvalue equal to 0.8600. Further investigation of four clusters will be examined with results from PROC FASTCLUS.

Further examination of Table 1 shows a moderate change between eigenvalues for the seventh (0.3076) and eighth (0.1552) eigenvalues. Differences are 0.1524 compared to 0.0338 with a cumulative proportion for the seventh eigenvalue of 0.9590. Further investigation of seven clusters will be examined with results from PROC FASTCLUS.

Table 1. Eigenvalues (step 1)

	Eigenvalue	Difference	Proportion	Cumulative
1	4.7601	2.3437	0.4327	0.4327
2	2.4164	1.2249	0.2197	0.6524
3	1.1915	0.0999	0.1083	0.7607
4	1.0917	0.6654	0.0992	0.8600
5	0.4263	0.0709	0.0388	0.8987
6	0.3554	0.0478	0.0323	0.9310
7	0.3076	0.1524	0.0280	0.9590
8	0.1552	0.0338	0.0141	0.9731
9	0.1214	0.0171	0.0110	0.9842
10	0.1043	0.0343	0.0095	0.9936
11	0.0699		0.0064	1.0000

PROC FASTCLUS
The code below standardizes data with a mean of 0 and a standard deviation of 1 (PROC STANDARD), uses PROC FASTCLUS to extract 7 clusters, uses PROC CANDISC to create variables to produce a scatter plot with PROC SGPLOT. Similar code can be used to extract 4 factors, using maxclusters=4.

```
proc standard data=rawsub1x mean=0 std=1 out=stan7;
  var TotalPriceInstate_OnCampus--InstrExp_Percent;
proc fastclus data=stan7 out=clus7 maxclusters=7 summary;
  var TotalPriceInstate_OnCampus--InstrExp_Percent;
title2'standardized variables';
proc candisc data=clus7 out=can7 noprint;
  var TotalPriceInstate_OnCampus--InstrExp_Percent;
  class cluster;
proc sgplot data=can7;
  scatter y=can2 x=can1 / group=cluster;
run;
```

Results from a seven clusters solution found an observed over-all r-squared of 0.5816, cubic clustering criterion equal to 72.197, and pseudo F equal to 70.19. Results from a four clusters solution found an observed over-all r-squared equal to 0.4126, cubic clustering criterion equal to 56.174, and pseudo F equal to 71.66. Table 2 shows frequencies and nearest clusters. Figure 1 illustrates a scatterplot of seven clusters while Figure 2 illustrates a scatterplot of four clusters.

Table 2. Seven cluster vs. four cluster solutions (step 1)

Cluster	Frequency	Nearest Cluster		Cluster	Frequency	Nearest Cluster
1	51	6		1	20	2
2	58	7		2*	91	3
3	15	4		3	44	2
4	11	3		4	155	2
5	43	7				
6	17	2				
7*	115	5				

*target institution

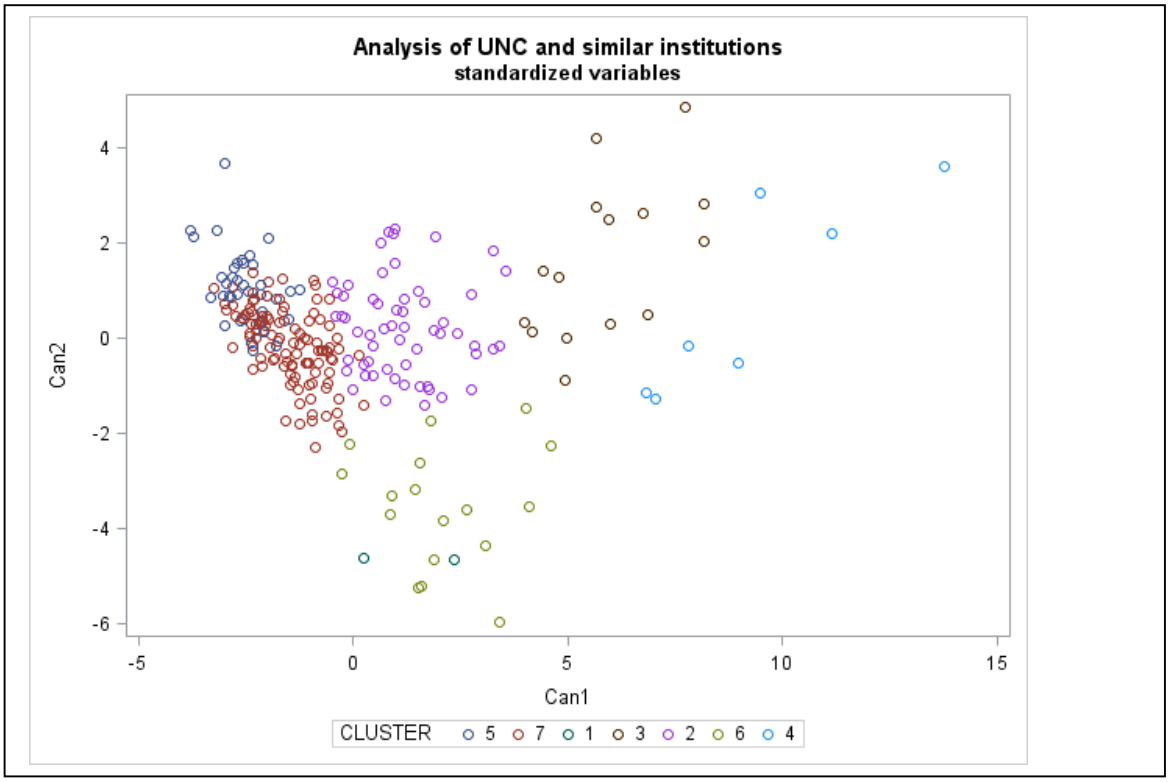

Figure 1. Scatter plot of seven clusters (step 1)

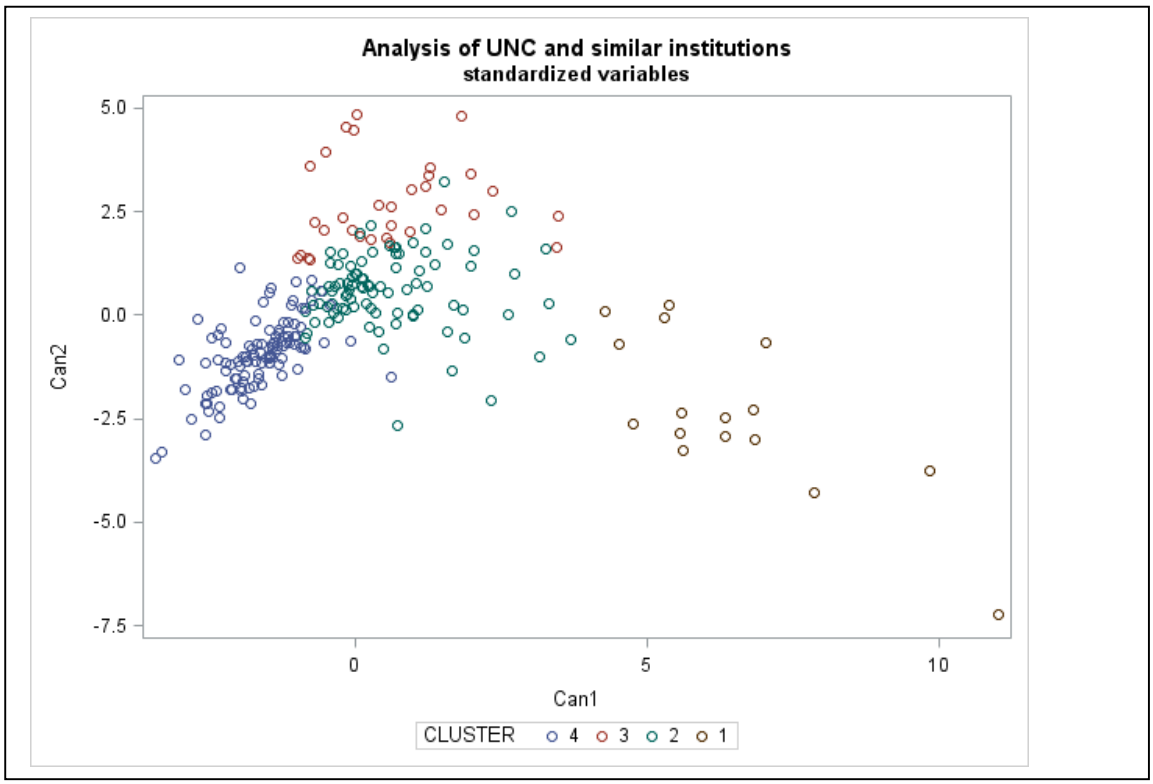

Figure 2. Scatter plot of four clusters (step 1)

Step 2

The second step included using a subset of the original data to examine eigenvalues, pseudo F, and cubic clustering criterion with results from PROC CLUSTER. A seven cluster solution was selected from the first step. Observations from cluster seven (n=115), the target institution cluster, were analyzed with PROC CLUSTER. Syntax for PROC CLUSTER and the table of eigenvalues is shown below.

```
proc cluster data=rawsub7 std method=average ccc pseudo;
  var TotalPriceInstate_OnCampus--InstrExp_Percent;
  id unitid;
run;
```

Table 3. Eigenvalues (step 2)

	Eigenvalue	Difference	Proportion	Cumulative
1	3.0294	0.7132	0.2754	0.2754
2	2.3162	0.6140	0.2106	0.4860
3	1.7022	0.4111	0.1547	0.6407
4	1.3011	0.3338	0.1183	0.7590
5	0.9673	0.3132	0.0879	0.8469
6	0.6540	0.2417	0.0595	0.9064
7	0.4123	0.1163	0.0375	0.9438
8	0.2960	0.0381	0.0269	0.9708
9	0.2579	0.2234	0.0234	0.9942
10	0.0345	0.0051	0.0031	0.9973
11	0.0294		0.0027	1.0000

The code below standardizes data with a mean of 0 and a standard deviation of 1 (PROC STANDARD), uses PROC FASTCLUS to extract 7 clusters, uses PROC CANDISC to create variables to produce a scatter plot with PROC SGPLOT. Similar code can be used to extract 4 factors, using maxclusters=4.

```
proc standard data=rawsub7 mean=0 std=1 out=stan6;
  var TotalPriceInstate_OnCampus--InstrExp_Percent;
proc fastclus data=stan6 out=clus6 maxclusters=6 summary;
  var TotalPriceInstate_OnCampus--InstrExp_Percent;
title2'standardized variables';
proc candisc data=clus6 out=can6 noprint;
  var TotalPriceInstate_OnCampus--InstrExp_Percent;
  class cluster;
proc sgplot data=can6;
  scatter y=can2 x=can1 / group=cluster;
run;
```

Results from a six cluster solution found an observed over-all r-squared equal to 0.4167, cubic clustering criterion equal to 13.796, and pseudo F equal to 15.58. Table 4 shows frequencies and nearest clusters. Figure 3 illustrates a scatterplot of the six clusters.

Table 4. Six cluster solution (step 2)

Cluster	Frequency	Nearest Cluster	Distance Between Cluster Centroids
1	22	2	2.2675
2	28	1	2.2675
3	12	4	2.8527
4	20	2	2.8527
5	21	1	2.9269
6*	12	4	2.8824

*target institution

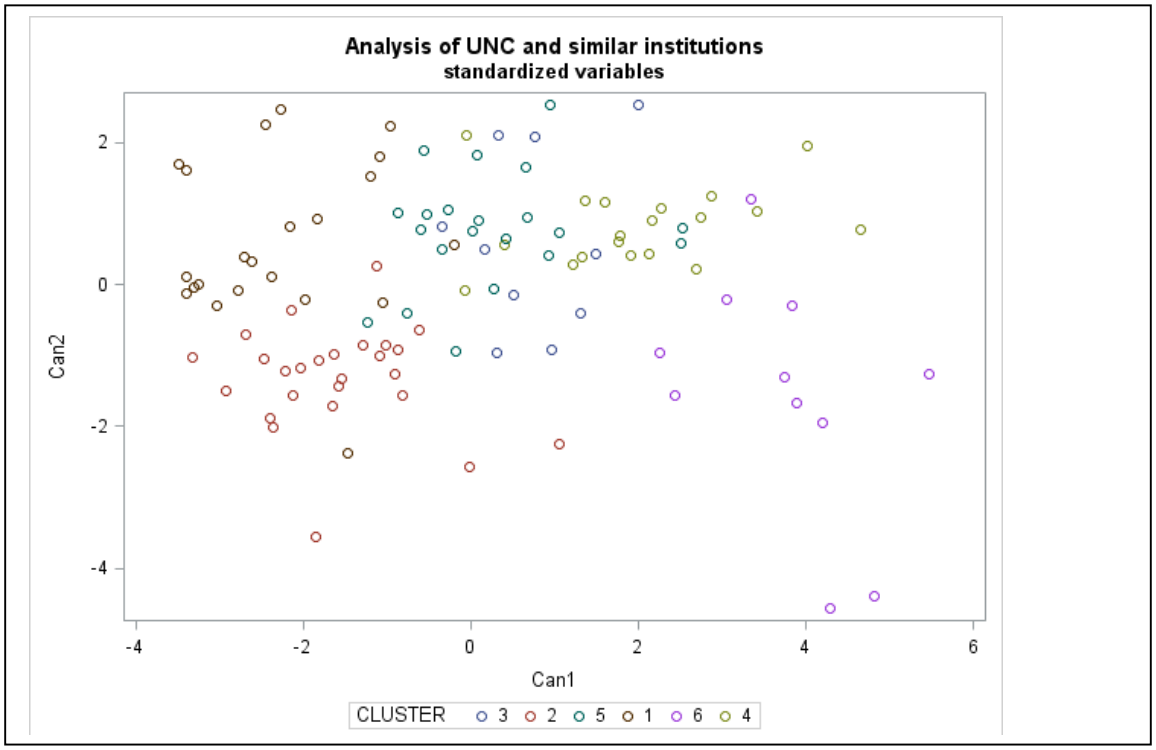

Figure 3. Scatter plot of six clusters (step 2)

Final Cluster

The final cluster or peer institution group selected with the specified variables included twelve institutions (including the target institution). For this simulation, Table 5 shows the institutions included in the peer group. Descriptive statistics, means and frequencies, indicated similarities between institutions. IPEDS data feedback report comparison group categories for the twelve institutions include categories 130 through 140. Eight of the twelve institutions are in the same geographic region. Table 6 illustrates mean, standard deviation, minimum and maximum values for the peer group.

Table 5. Peer Institutions

	Institution
1	University of Colorado Colorado Springs
2	University of Northern Colorado*
3	Central Michigan University
4	Eastern Michigan University
5	Ferris State University
6	University of Michigan – Dearborn
7	Oakland University
8	Plymouth State University
9	University of Akron Main Campus
10	Cleveland State University
11	University of Toledo
12	Portland State University

*target institution

Table 6. Descriptive statistics

Variable	Mean	Standard deviation	Minimum	Maximum
Undergraduate enrollment	14,558	6,362.75	4,453	22,966
Graduate enrollment	3,694	1,908.08	1,203	6,496
Graduation Rate	43.7	7.67	30	59
Core revenues total dollars	277,828,414	142,092,410	76,104,138	537,893,860
Tuition fees percent of core revenues	50.08	6.42	38	61
State appropriations percent of core revenues	16.92	8.07	0	23
Revenues from tuition fees per FTE	8,669.92	576.46	7739	9688
Revenues from state appropriations per FTE	3070.83	1546.83	0	4702
Instruction expenses percent of total core expenses	50.50	4.70	44	59

Conclusions

Cluster analysis is an exploratory data analysis technique for classifying and organizing data into meaningful clusters, groups, or taxonomies by maximizing the similarity between observations in each cluster. Members of the groups (clusters) share properties in common with a strong relationship between members of the same cluster and weak relationship between members of different clusters. Cluster analysis creates groups without any preconceived notion.

Cluster analysis methods will always produce a grouping. The groupings produced by cluster analysis may or may not prove useful for classifying objects. Clusters are dependent on the methods or approach, on the variables used to differentiate clusters and the goal or purpose for determining clusters. PROC CLUSTER and PROC FASTCLUS provide methods to perform cluster analysis.

References

Bailey, Ken (1994). "Numerical Taxonomy and Cluster Analysis". *Typologies and Taxonomies*. p. 34. ISBN9780803952591.

Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. Journal of Abnormal and Social Psychology, 38, 476-506.

Cluster analysis. (2006). In Collins Dictionary of Sociology. Retrieved from http://0-www.credoreference.com.source.unco.edu/entry/collinssoc/cluster_analysis.

Estivill-Castro, Vladimir (June 2002 2002). "Why so many clustering algorithms —A PositionPaper" (http://cs.nju.edu.cn/zhouzh/zhouzh.files/course/dm/reading/reading05/estivill- castro_sigkddexp02.pdf). *ACM SIGKDD Explorations Newsletter* 4 (1): 65–75. doi:10.1145/568574.568575 (http://dx.doi.org/10.1145%2F568574.568575).

Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek (2011). "Density-based Clustering" (http://wires.wiley.com/WileyCDA/WiresArticle/wisId-WIDM30.html).*WIREs Data Mining and Knowledge Discovery* 1 (3): 231–240. doi:10.1002/widm.30 (http://dx.doi.org/10.1002%2Fwidm.30).

Lloyd, S. (1982). "Least squares quantization in PCM". *IEEE Transactions on Information Theory*28 (2): 129–137. doi:10.1109/TIT.1982.1056489.edit

Retrieved from support.sas.com
Cluster_PROCFASTCLUSOverview_support.sas.com_documentation_cdl_en_statug_63347_HTML_d.
ClusterAnalysisOverview_support.sas.com_documentation_cdl_en_statug_63347_HTML_d.

Contact Information

Diana Suhr, PhD
University of Northern Colorado
Greeley CO 80634
diana.suhr@unco.edu