5

This chapter provides an introduction to the family of partitional cluster analytical methods, with specific attention to research on community college students. Key decision points and common approaches in the use of cluster analysis are described.

The Use of Cluster Analysis in Typological Research on Community College Students

Peter Riley Bahr, Rob Bielby, Emily House

In Chapter Three in this volume, Peter Riley Bahr makes the case for the need to differentiate and identify types of community college students, described there as "the varied answers to who is enrolling in a given community college, how they are using the community college, and to what end they are using it." The chapter demonstrates through an application that understanding who students are with respect to their use of the community college has substantial value to community college stakeholders as they seek to address a range of policy- and practice-relevant questions. In other words, there is much to be gained by community college stakeholders from a comprehensive system of classification (a typology) of student behavioral profiles (Bahr, 2010, 2011).

One useful and increasingly popular method of classifying students is known commonly as cluster analysis (Ammon, Bowman, and Mourad, 2008; Bahr, 2010; Boughan, 2000; Hagedorn and Prather, 2005; VanDerLinden, 2002). The variety of techniques that comprise the cluster analytic family are intended to sort observations (for example, students) within a data set into subsets (clusters) that share similar characteristics and differ in meaningful ways from other subsets (Borden, 2005; Jain and Dubes, 1988; Punj and Stewart, 1983). In the case of community college students, for example, clusters may be formed on the basis of student aspirations (VanDerLinden, 2002), student course-taking and enrollment behaviors (Bahr, 2010), student demographic characteristics (Ammon, Bowman, and Mourad, 2008), or any combination of these or other measures (Boughan, 2000). These clusters assist in the identification of

Published online in Wiley Online Library (wileyonlinelibrary.com) • DOI: 10.1002/ir.417

patterns of behaviors or characteristics in data sets that contain an otherwise incomprehensible amount of student information.

The goal of cluster analysis, broadly stated, is to find the arrangement of observations and clusters that maximizes both within-group homogeneity and between-group heterogeneity (Borden, 2005; Everitt, Landau, Leese, and Stahl, 2011). Within-group homogeneity refers to the extent to which observations that are assigned to a given cluster share similar attributes on the variables included in the cluster analysis. Between-group heterogeneity refers to the extent to which each cluster is dissimilar in the aggregate from other clusters with respect to the variables included in the analysis. As Cormack (1971) described it, the goal of cluster analysis is to arrive at groups of observations that have "internal cohesion and external isolation" (p. 329). Thus, the optimal cluster solution places together into clusters those students who are most alike on the variables of interest and simultaneously creates clusters that on average are most different from one another on the variables of interest.

In this chapter, we provide an introduction to the cluster analytic method as it pertains to research on community college students. The execution of cluster analysis requires a number of carefully considered methodological decisions. The first set of decisions concerns the selection of the variables that will be included in the analysis, the scaling of these variables, and the structure of the data. The second involves the selection of a proximity metric by which within-group homogeneity and between-group heterogeneity will be calculated. Then, the cluster technique itself must be selected. Finally, one must decide how to make sense of the identified clusters.

In the sections that follow, we describe each stage of the decision-making process and detail several common approaches at each stage. Although the intent of this chapter is not to develop a detailed manual for the use of cluster analysis, we cite many comprehensive sources throughout to aid researchers in locating additional informational resources.

Data Preparation

Thanks in part to the increasing focus on data-driven decision making in community colleges (Morest and Jenkins, 2007) and the growth of statewide data systems (Ewell and Boeke, 2007; Ewell and Jenkins, 2008), the number and size of data resources that are available to community college researchers have grown substantially in recent years. However, the increase in data resources has been accompanied by an increase in the complexity of the data as well, making cluster analysis all the more useful as an analytical tool. Still, cluster analysis has its own complexities (Jain, Murty, and Flynn, 1999; Punj and Stewart, 1983; Rapkin and Luke, 1993), beginning with the preparation of the data. In this section, we describe data preparation for cluster analysis: selecting variables, transforming the scales of variables, and structuring the data.

Variable Selection. As with most other analytical procedures, variable selection is a critical aspect of cluster analysis (Borden, 2005). Perhaps more than most other analytical procedures, though, the results of cluster analysis are highly sensitive to variable selection (Fowlkes, Gnanadesikan, and Kettenring, 1988). For example, demographic characteristics such as race, sex, age, and citizenship frequently are available in data sets that pertain to community college students, and such variables are used in analyses nearly as a matter of course. However, as Bahr (2010) explained, "The inclusion of race, sex, and age presumes that these variables are not just important predictors of cluster membership, but, rather, important variables in their own right for defining a 'type' of student" (pp. 727–728), which may or may not agree with objectives that underlie a researcher's use of cluster analysis

In this respect, researchers must be wary of using every variable that is available in a given data set simply because it is available and instead focus on selecting variables that are pertinent to the research questions of interest (Punj and Stewart, 1983; Rapkin and Luke, 1993). For example, if a researcher is seeking to understand students' enrollment or course-taking patterns, demographic variables are not dimensions on which the clusters should be projected. Instead, one might consider course credit load, number of enrolled semesters, course success rate, number of courses attempted in math or English, and the like. Furthermore, simulation models have shown that the inclusion of variables that are unrelated to the true clusters in a data set reduces the capability of cluster analytic algorithms to return an optimal solution (Milligan, 1980). Therefore, whenever possible, it is important to consider carefully the variables that are available and select only those that are relevant for the research questions that underlie the analysis.

Of note, one aspect of variable selection with which researchers need not be concerned when using cluster analysis is the distinction between independent and dependent variables (Punj and Stewart, 1983). As Bahr (2010) explained, "Cluster analytic techniques do not presuppose the segregation of *type* and *outcome* variables" (p. 745, italics in original). Thus, variables that typically would be considered outcomes, such as degree completion, may be employed together in a cluster analysis with variables that typically would be considered predictors, such as course credit load, if the research questions justify this joint use.

While selecting variables based on specific research questions is greatly preferred, a number of other variable selection techniques exist for instances in which the appropriate set of variables is not clear (Steinley, 2006). Similar to stepwise techniques in multiple regression (Beale, 1970), these methods compare the statistical properties of cluster solutions based on varying sets of variables and select the set of variables that optimizes the chosen metric. Although we do not discuss these alternatives here, we refer readers to Steinley (2006) and Steinley and Brusco (2008) for further information.

Transforming the Scales of Variables. If the variables selected for the cluster analysis are continuous (as opposed to dichotomous, nominal, or ordinal), it is common to transform the scales of the variables in order to reduce the effect on the cluster analytic process of differences in ranges, magnitudes, and units of measurement (Gnanadesikan, Kettenring, and Tsao, 1995; Hunt and Jorgensen, 2011; Jain, Murty, and Flynn, 1999; Milligan and Cooper, 1988; Rapkin and Luke, 1993). In many cases, this transformation is accomplished by what is known commonly as standardization or autoscaling. The mean of a given variable is subtracted from each value of that variable, and this difference is divided by the standard deviation of the variable, resulting in a transformed variable that has a mean of zero and a standard deviation of one. The autoscaling equation is presented below:

$$k_i^* = \frac{\left(k_i - \bar{k}\right)}{\sigma_k}$$

In this equation, k_i is the unstandardized value of variable k for observation i (the value of a variable for a given student), $\bar{k}$ is the mean of variable k for all observations in the data set, σ_k is the standard deviation of variable k for all observations in the data set, and k_i^* is the autoscaled value of variable k for observation i.

While autoscaling is a convenient method to ensure that continuous variables in a cluster analysis have the same mean and variance, it is somewhat arbitrary and applies an implicit weighting scheme to the data (Everitt, Landau, Leese, and Stahl, 2011; Steinley, 2006). In effect, each variable is weighted by the inverse of its standard deviation, thereby decreasing the weight of variables that have greater variance and, conversely, increasing the weight of variables that have lower variance. This weighting scheme may have significant consequences for the solution that is returned by the cluster analytic procedure because a variable in the data that, as a practical matter, varies little is given equal standing with a variable that varies greatly. As Milligan and Cooper (1988) argued, “There is no compelling reason to practice democracy while performing all cluster analyses” (p. 183).

As an alternative, Milligan and Cooper (1988) suggested that each value of a given variable be divided by the range of that variable rather than the standard deviation. They offered two different formulas to accomplish this end:

$$k_i^* = \frac{k_i}{k_{max} - k_{min}}$$

$$k_i^* = \frac{k_i - k_{min}}{k_{max} - k_{min}}$$

Here, $k_{\min}$ and $k_{\max}$ refer to the minimum and maximum values, respectively, of variable k for all observations in the data set. The benefit of the second option over the first is that it produces scaled values that are bounded between zero and one, so long as the unscaled variable does not have negative values.

Milligan and Cooper (1988) argued that these range-based scaling techniques produce scaled variables with desirable qualities for cluster analysis, but their conclusions are not universally accepted (Gnanadesikan, Kettenring, and Tsao, 1995). In addition, the range-based transformation limits the interpretability of the variables, particularly when compared with traditional autoscaling. With autoscaling, a one-unit change in a standardized variable represents a change of one standard deviation in the unstandardized variable. The same does not hold for range-based scaling.

We should note here that scale transformation does not necessarily entirely solve the problem of extreme values and outliers on a given variable. Some of the most common cluster analytic techniques are based on the means of variables and therefore may be influenced greatly by extreme values even when the variables are standardized (Rapkin and Luke, 1993). Therefore, it is important for researchers to examine the distribution of each variable to be included in the cluster analysis in order to identify extreme values, explore why these values exist in the data, and consider appropriate solutions if an intervention appears to be warranted. Some of the potential solutions include deleting extreme values (which generally is not preferred), adjusting the values of extreme observations to a certain percentile of the distribution of that variable (for example, adjusting downward to equal the 99.9th percentile those values on a given variable that exceed the 99.9th percentile), and accommodating extreme values through the choice of cluster analytic technique. Although a detailed discussion of extreme values and outliers is outside the scope of this chapter, we refer readers to Fox (2008) for a discussion of outlier identification and to Punj and Stewart (1983) and Steinley (2006) for discussions of outliers as they pertain specifically to cluster analysis.

Data Structure. Whether or not a researcher elects to transform the scales of variables, the next step to consider is the manner in which the data will be structured. Prior to the cluster analysis, the data should be assembled in one of two forms (Jain and Dubes, 1988; Jain, Murty, and Flynn, 1999; Everitt, Landau, Leese, and Stahl, 2011). The first possible form is a typical N (observations) by p (variables) matrix, with each row representing one observation (for example, a student) and each column representing one variable. The second possible form is an $N \times N$ proximity matrix in which each cell represents the distance between two observations on the selected proximity metric, which we discuss later.

Computationally, most statistical software will transform a standard $N \times p$ data matrix into an $N \times N$ proximity matrix in the process of calculating a cluster solution. As data generally are arranged in the $N \times p$

format, leaving one's data in this form removes a step from the process of executing a cluster analysis and reduces the possibility of user error.

However, there are some advantages to converting an $N \times p$ data matrix into an $N \times N$ proximity matrix prior to running the cluster analysis (Everitt, Landau, Leese, and Stahl, 2011). First, some proximity metrics may not be available in a particular statistical program and therefore must be calculated by hand. Second, the $N \times N$ matrix allows one to explore the data visually before executing the cluster analysis (Borden, 2005). This visual exploration may aid in the selection of an appropriate cluster technique and help the researcher develop expectations concerning the findings.

Such visual exploration, however, is constrained by the size of the data set. A data set that contains only 500 students would produce an $N \times N$ dissimilarity matrix containing 250,000 cells. Although the matrix is symmetric (Jain and Dubes, 1988), meaning that the diagonal contains zeros and the upper and lower halves of the matrix are identical, one still would be faced with an unwieldy amount of information. Thus, the choice of data format likely will be governed as much by the need for efficiency as by the desire for control over the analysis.

Proximity Metrics

The next step in the cluster analytic process is selecting a proximity metric. A proximity metric is used to measure the distance between a given observation and another such observation, as well as the distance between an observation and a cluster of other such observations, with respect to the variables selected for the cluster analysis. In other words, in order to evaluate a particular set of assignments of observations to clusters, one must decide on a method of measuring the proximity of observations to one another, keeping in mind that the goal of cluster analysis is to produce the set of clusters that minimizes the distance between observations that share a cluster and maximizes the distance between clusters.

In this section, we first define important terminology and then describe several proximity metrics that are used frequently in cluster analysis. We organize the proximity metrics by the level of measurement of variables to which the metrics apply (for example, continuous, dichotomous, nominal, and ordinal). The set of metrics detailed here is in no way exhaustive, but the citations we provide will guide readers to more extensive lists and descriptions.

Terminology. Three terms occur especially frequently in the literature on proximity metrics: *dissimilarity*, *distance*, and *similarity* (Gower, 1985). *Dissimilarity* typically describes proximity metrics that are used for continuous variables (Hunt and Jorgensen, 2011), and these measure the degree to which two observations differ on the variables included in the cluster analysis. Dissimilarity metrics may be referred to as distance metrics when they satisfy the triangular inequality (Everitt, Landau, Leese, and Stahl, 2011; Gower and Legendre, 1986). In contrast, *similarity* metrics typically are used with dichotomous, nominal, and ordinal variables, that is, categorical variables (Hunt and Jorgensen, 2011). Similarity metrics, when used with categorical variables, may be understood as the degree to which two observations share the same values on the variables included in the cluster analysis (StataCorp, 2007).

As a general rule, researchers are constrained to select a proximity metric that applies to either continuous variables or categorical variables (Borden, 1995, 2005), although there are a few exceptions to this rule. In other words, in most cases, all of the variables to be used in the cluster analysis share a level of measurement (for example, all continuous, all dichotomous). Note that nominal and ordinal variables may be treated as such, though this constrains the researcher to a subset of all proximity metrics. Alternatively, nominal and ordinal variables may be converted into sets of dichotomous dummy variables (Everitt, Landau, Leese, and Stahl, 2011; Hunt and Jorgensen, 2011). However, this conversion raises some analytical concerns, which we discuss later.

Dissimilarity Metrics for Continuous Variables. A large number of dissimilarity metrics are available for use with continuous variables, and Euclidean distance is among the most commonly used (Jain and Dubes, 1988). Euclidean distance represents an intuitive understanding of the measurement of the distance between two observations as the linear distance between points in space (Everitt, Landau, Leese, and Stahl, 2011; Jain, Murty, and Flynn, 1999; Rapkin and Luke, 1993). It is calculated as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

Here, d_{ij} is the dissimilarity of observations i and j. The terms x_{ik} and x_{jk} represent the values of continuous variable k for observations i and j, respectively. The portion within the parentheses represents the difference between observations i and j on variable k. Much like the calculation of a standard deviation, the difference between the two observations on this variable is squared to eliminate negative values, the squared differences of the comparisons on all of the p variables (that is, all of the variables included in the analysis) are summed, and then the square root of the sum is calculated.

Although conveniently intuitive and used frequently in cluster analysis, Euclidean distance is sensitive to the influence of outliers due to the squaring of the parenthetical term (Cormack, 1971). In addition, it is highly sensitive to differentials in the scales of the variables, generally requiring that the scales be transformed prior to the analysis (Jain and Dubes, 1988).

Another common distance metric for continuous data is the absolute value distance (Everitt, Landau, Leese, and Stahl, 2011; Jain and Dubes, 1988). Absolute value distance is calculated as follows:

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

Similar to the benefits of using the least absolute deviations estimator in a regression context (Angrist and Pischke, 2009), the use of absolute value distance reduces the impact of extreme values and outliers on the cluster analytic process because the differences between observations are not squared in the calculation. However, this metric lacks the intuitive interpretation offered by Euclidean distance. In fact, it has been referred to as the taxicab metric, the city block metric, and the Manhattan metric because proximity is measured not linearly but instead with joints comparable to corners on a street map (Borden, 2005; Jain and Dubes, 1988; Rapkin and Luke, 1993). Moreover, absolute value distance is less computationally efficient than is Euclidean distance. Regardless, the central concern in selecting between these and other dissimilarity metrics should be the degree to which they fit the characteristics of the data and the degree of clarity that they provide to the proximity of observations (Everitt, Landau, Leese, and Stahl, 2011).

Similarity Metrics for Dichotomous variables. In circumstances in which the variables to be used in the cluster analysis are dichotomous (variables that have only two values, typically zero and one), one would select from among a wide range of similarity metrics designed for this level of measurement (Borden, 2005; Gower and Legendre, 1986; Hubálek, 1982; StataCorp, 2007). As a general rule, these similarity metrics draw on the idea that on each dichotomous variable that is included in the cluster analysis, a given two observations (for example, two students) will match by sharing a value of one (a 1–1 match), match by sharing a value of zero (a 0–0 match), or not match because one observation has a value of one on the variable while the other observation has a value of zero. Each of the similarity metrics that is available for such data combines the information on all comparisons for two observations into a particular expression of the ratio of matches to comparisons. The various similarity metrics differ from each other primarily with respect to the weight given to 1–1 matches versus 0–0 matches and the weight given to matches versus mismatches.

The most common of the similarity metrics for dichotomous variables is a simple matching coefficient, which represents the proportion of the variables included in the analysis on which two observations match (Jain and Dubes, 1988). Many of the more complex expressions of similarity metrics arise from concerns about how to deal with 0–0 matches, where two observations match because neither has the characteristic indicated by a given variable, as opposed to 1–1 matches, where two observations match because they share the characteristic indicated by that variable (Cormack, 1971). For example, noting that two students enrolled in remedial mathematics (a 1–1 match on a hypothetical variable *remedial math*) may be more informative for the purpose of a given cluster analysis than noting that two students match because neither enrolled in remedial mathematics (a 0–0 math). Not matching on this variable could mean that one student enrolled in college-level math, while the other did not enroll in math at all. In other words, the presence of a characteristic may differ in importance from the absence of this characteristic (Gower and Legendre, 1986). In such cases, when a 1–1 match is more informative than is a 0–0 match, a recommended similarity metric is the Jaccard coefficient (Jain and Dubes, 1988), which is the ratio of 1–1 matches to the sum of 1–1 matches and all mismatches.

Similarity Metrics for Nominal and Ordinal Variables. When the cluster analysis to be performed draws on variables that are nominal or ordinal, one method of handling the variables is to recode them into a series of dichotomous dummy variables, where each dummy variable represents one value or level of a given nominal or ordinal variable. However, such manipulations result in large numbers of 0–0 matches and, consequently, even greater concern about the selection of an optimal similarity metric (Cormack, 1971). Instead, Everitt, Landau, Leese, and Stahl (2011) suggest the use of the following metric:

$$s_{ij} = \frac{1}{p}\sum_{k=1}^{p} s_{ijk}$$

Here, s_{ij} denotes the similarity between observations i and j across p variables. The match or mismatch of observations i and j on variable k is denoted by s_{ijk}, which is assigned a value of one if the two observations match on k and a value of zero if the two observations do not match on k. The sum of the matches and mismatches is calculated and then divided by the total number of variables (p), resulting in a similarity coefficient that is equivalent to the simple matching coefficient discussed earlier for dichotomous variables. Unfortunately, this metric is not a standard option for cluster analysis in many statistical programs (StataCorp, 2007). Therefore, its use requires the researcher to calculate the $N \times N$ proximity matrix.

Gower: A Versatile Similarity Metric. Gower (1971) proposed a similarity metric that may be applied to data sets that contain variables of any level of measurement (Everitt, Landau, Leese, and Stahl, 2011; StataCorp, 2007). It is calculated as follows:

$$s_{ij} = \sum_{k=1}^{p} s_{ijk}\delta_{ijk} \Bigg/ \sum_{k=1}^{p} \delta_{ijk}$$

Here again, s_{ij} is the similarity of two observations i and j across p variables. The term δ_{ijk} is a quantity that is equal to one when a valid comparison may be made between the two observations on a given variable k and zero otherwise. Cases in which valid comparisons cannot be made typically occur when one or both of the observations have a missing value on a given variable. However, listwise deletion of observations with missing values on any variable included in the analysis often is a default setting of cluster analytic procedures in statistical software (such as StataCorp, 2007). Therefore, the denominator simply will be a count of the total number of variables on which the two observations were compared.

The innovative aspect of the Gower coefficient is the handling of s_{ijk}, which is the similarity of two observations i and j on a given variable k. When the variable of interest is dichotomous, nominal, or ordinal, a simple matching metric (detailed earlier) is used, where matches are assigned a value of one and mismatches are assigned a value of zero. When k is continuous, s_{ijk} is calculated as follows:

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{k_{max} - k_{min}}$$

Here, x_{ik} and x_{jk} represent the values for observations i and j, respectively, on variable k. The denominator is the range of variable k. The quotient of the absolute difference between x_{ik} and x_{jk} and the range of k is the dissimilarity of observation i and j on variable k, and it has a potential range of zero to one, assuming that the variable k does not have negative values. Subtracting this quotient from one produces a similarity that then may be combined with the similarities calculated for dichotomous, nominal, and ordinal variables.

Returning to the general Gower equation presented earlier in this section, the sum of the similarities (s_{ijk}) across all of the variables is divided by the number of variables on which comparisons were made (δ_{ijk}), resulting in a coefficient (s_{ij}) that represents the degree of similarity between observations i and j across all of the variables. Hence, the Gower coefficient is quite versatile in that it may be used with variables that have any level of measurement. Still, it is important to note that, with respect to categorical variables, the Gower coefficient suffers the limitation discussed earlier of treating 1–1 matches and 0–0 matches as equal in importance. This limitation should be contemplated carefully in light of the research questions to be addressed and weighed against the benefits of the versatility.

Cluster Techniques

Once the proximity metric has been selected, one may proceed to the cluster analysis itself. Here we discuss two common and closely related methods of cluster analysis that are classified as partition methods. In addition to the partition methods, a number of other classes of cluster analytic techniques are available, but partition methods typically are preferred when one has a large data set (Hunt and Jorgensen, 2011; Jain and Dubes, 1988; Jain, Murty, and Flynn, 1999), as often is the case with data sets composed of student records.

***k*-Means.** The cluster analysis technique known as *k*-means is an iterative algorithm that attempts to generate the most appropriate fit of observations to clusters, given the number of clusters (*k*) selected by the researcher prior to the execution of the cluster analysis (Everitt, Landau, Leese, and Stahl, 2011; Hunt and Jorgensen, 2011). In other words, the researcher selects the number of clusters in advance, which is the *k* in "*k*-means" (Rapkin and Luke, 1993). The algorithm generates the assignment of students to clusters that minimizes differences between observations within a cluster and maximizes the differences between clusters.

The algorithm begins by selecting starting sets of observations equal to the number of clusters that the researcher selects. These starting sets of observations constitute the initial centroids on which the clusters are built. In *k*-means cluster analysis, a *centroid* may be understood as the multivariate mean of each cluster—the "center" of each cluster (Steinley, 2006). Interestingly, the starting sets need not be equal in size. However, no observation ever is assigned to more than one cluster at a given time (Hautamäki and others, 2005). Therefore, using one very large starting set to form one of the initial centroids would require that the remaining starting sets for the other initial centroids be correspondingly small.

The selection of the starting sets of observations may be accomplished in a number of ways (Punj and Stewart, 1983; Steinley, 2006). One common method is to select randomly *k* observations such that each randomly selected observation becomes the initial centroid of a cluster. Another common method is to partition the data set randomly into *k* groups, allocating randomly each observation in the data to one starting set. Both methods depend on the random number generator of the statistical software, which necessitates setting the starting number so that the results may be replicated. As a third alternative, a researcher may have an idea about which observations should be assigned to the starting sets to form the initial centroids, perhaps based on some criterion in the data, and these observations may be identified and used in this manner. In fact, these "rational starts" (Steinley, 2006) have been recommended by some to achieve optimal cluster solutions when using *k*-means (Milligan, 1980).

Once the initial centroids are constructed from the starting sets, the algorithm proceeds by assigning each observation to the cluster to which it is closest in terms of its multivariate mean (Steinley, 2006). Then the multivariate mean of each centroid is recomputed, and the proximity of each observation to its assigned cluster and the centroids of the other clusters is reevaluated. Observations are moved if they are closer in terms of multivariate mean to the centroid of another cluster than to the centroid of the cluster to which they currently are assigned, and the centroids again are recomputed. The process continues in this iterative fashion until no observations are moved after the centroids are recomputed, until a predetermined number of iterations have been completed, or until some other user-determined stopping rule is achieved.

Here *k*-means sometimes suffers from problems related to local optima, meaning that the algorithm may produce a cluster solution that is optimal with respect to the starting sets of observations but not optimal with respect to the data set as a whole (Falkenauer and Marchand, 2001; Hunt and Jorgensen, 2011; Jain and Dubes, 1988; Jain, Murty, and Flynn, 1999; Milligan, 1980; Steinley, 2003, 2006). Therefore, if one uses a random number generator to build starting sets, it may be prudent to explore several different starting numbers in succession and then compare the resulting cluster solutions to ascertain that the clusters are similar regardless of the starting number.

***k*-Medians.** As with most other statistical procedures that depend on the mean, one of the major challenges of *k*-means cluster analysis is its sensitivity to the influence of outliers and extreme values (Hautamäki and others, 2005). Similar to the calculation of a univariate mean, the multivariate mean of a centroid tends to be drawn disproportionately toward extreme values. As the centroids shift, the measured proximity of individual observations changes, possibly resulting in observations being reassigned to the "wrong" clusters (assuming that there is a "correct" underlying cluster structure to the data). The impact on cluster solutions of this sensitivity to extreme values requires determined efforts by the researcher to identify and resolve outlying observations prior to performing *k*-means cluster analysis.

The technique known as *k*-medians cluster analysis seeks to remedy this problem by relying on the multivariate median as the center of each cluster rather than a mean-based centroid. Under some circumstances, this may result in decreased influence of outliers on the cluster solution because the calculation of a median is dependent only on the rank order of values rather than the distance between values (Steinley, 2006). Still, *k*-means remains a highly popular cluster analytic method, in part for its intuitive nature and in part because it produces cluster solutions that are competitive with those produced by other partitioning procedures (Milligan, 1980), particularly when it is implemented thoughtfully, with attention to variable selection and scale, extreme values, selection of starting sets, and so forth.

Examination of the Results

No discussion of cluster analysis would be complete without some consideration of how to examine and make sense of the identified clusters. The possibilities in this respect are numerous (Punj and Stewart, 1983) but ultimately should be guided by the nature of the data, the particular analysis that was executed, and the focal research questions of the study. Still, we recommend that, at a minimum, the researcher conduct three preliminary steps to begin making sense of the identified clusters.

First, calculate measures of central tendency, such as the mean and the median, for each variable in each cluster and for the data set as a whole, and determine the size of each cluster relative to the data set as a whole. These statistics will provide the researcher with a sense of where the center of mass lies for each cluster, relative to the other clusters and relative to all students in the data set, answering the question, "Who are the students in each cluster?"

Second, calculate measures of dispersion, such as the standard deviation and the range, for each variable in each cluster and for the data set as a whole. These statistics will allow the researcher to determine how much variation exists on each variable within each cluster. This is an important step because the center of mass for a particular variable in a particular cluster may be deceiving in that the cluster may contain students who, in the aggregate, exhibit a high level of variability on that variable. To aid in this examination, we recommend that the researcher calculate the 10th and 90th percentiles of any continuous variables, thereby trimming the extreme values to garner a sense of the amount of variation that exists among the middle 80 percent of the students in a cluster (Bahr, 2010). Taken together, these statistics allow the researcher to answer the question, "How consistent are the characteristics that I observe in each cluster?"

Finally, if the researcher has executed several different cluster analyses, using the same data set but differing numbers of presumed clusters (differing values of k), different starting sets, or different proximity metrics, we recommend the execution of simple cross-tabulations of the several cluster solutions (Bahr, 2010; Rapkin and Luke, 1993). These cross-tabulations can be enormously informative concerning how the cluster algorithm perceives the similarities or differences of the clusters in that one will be able to observe, for example, how clusters of students in one cluster solution are collapsed to produce a solution that entails fewer presumed clusters or subdivided to produce a solution that entails a greater number of presumed clusters.

Conclusion

With respect to the diversity of the student body and the variability of students' means and ends, the community college is a complex postsecondary environment. As a result, the community college researcher is faced with a complex and challenging task in answering questions about who is enrolling in the college, how they are using the college, and to what end they are using it. Cluster analysis is one promising technique for answering such questions, but its use requires careful consideration of a number of methodological decisions. In this chapter, we provided an introduction to some of the critical decision points and common approaches in executing a cluster analysis of data that address community college students.

References

Ammon, B. V., Bowman, J., and Mourad, R. "Who Are Our Students? Cluster Analysis as a Tool for Understanding Community College Student Populations." *Journal of Applied Research in the Community College*, 2008, *16*, 32–44.

Angrist, J. D., and Pischke, J. S. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, N.J.: Princeton University Press, 2009.

Bahr, P. R. "The Bird's Eye View of Community Colleges: A Behavioral Typology of First-Time Students Based on Cluster Analytic Classification." *Research in Higher Education*, 2010, *51*, 724–749.

Bahr, P. R. *Classifying Community Colleges Based on Students' Patterns of Usage*. Ann Arbor: Center for the Study of Higher and Postsecondary Education, University of Michigan, 2011.

Beale, E.M.L. "Note on Procedures for Variable Selection in Multiple Regression." *Technometrics*, 1970, *12*, 909–914.

Borden, V.M.H. "Segmenting Student Markets with a Student Satisfaction and Priorities Survey." *Research in Higher Education*, 1995, *36*, 73–88.

Borden, V.M.H. "Identifying and Analyzing Group Differences." In M. A. Coughlin (ed.), *Intermediate/Advanced Statistics in Institutional Research* (pp. 132–168). Tallahassee, Fla.: Association for Institutional Research, 2005.

Boughan, K. "The Role of Academic Process in Student Achievement: An Application of Structural Equations Modeling and Cluster Analysis to Community College Longitudinal Data." *AIR Professional File*, 2000, *74*, 1–18.

Cormack, R. M. "A Review of Classification." *Journal of the Royal Statistical Society, Series A,* 1971, *134*, 321–367.

Everitt, B., Landau, S., Leese, M., and Stahl, D. *Cluster Analysis* (5th ed.). Hoboken, N.J.: Wiley, 2011.

Ewell, P., and Boeke, M. *Critical Connections: Linking States' Unit Record Systems to Track Student Progress*. Indianapolis, Ind.: Lumina Foundation for Education, 2007. Retrieved Feb. 1, 2011, from http://www.luminafoundation.org/publications/Critical_Connections_Web.pdf.

Ewell, P., and Jenkins, D. "Using State Student Unit Record Data to Increase Community College Student Success." In T. H. Bers (ed.), *Student Tracking in the Community College*. New Directions for Community Colleges, no. 143. San Francisco: Jossey-Bass, 2008.

Falkenauer, E., and Marchand, A. "Using K-Means? Consider ArrayMiner." Paper presented at the 2001 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, Las Vegas, 2001.

Fowlkes, E. B., Gnanadesikan, R., and Kettenring, J. R. "Variable Selection in Clustering." *Journal of Classification*, 1988, *5*, 205–228.

Fox, J. *Applied Regression Analysis and Generalized Linear Models*. Thousand Oaks, Calif.: Sage, 2008.

Gnanadesikan, R., Kettenring, J. R., and Tsao, S. L. "Weighting and Selection of Variables for Cluster Analysis." *Journal of Classification,* 1995, *12*, 113–136.

Gower, J. C. "A General Coefficient of Similarity and Some of its Properties." *Biometrics*, 1971, *27*, 857–872.

Gower, J. C. "Measures of Similarity, Dissimilarity, and Distance." In S. Kotz, N. L. Johnson, and C. B. Read (eds.), *Encyclopedia of Statistical Sciences* (Vol. 5, pp. 397–405). Hoboken, N.J.: Wiley, 1985.

Gower, J. C., and Legendre, P. "Metric and Euclidean Properties of Dissimilarity Coefficients." *Journal of Classification*, 1986, *5*, 5–48.

Hagedorn, L. S., and Prather, G. "The Community College Solar System: If University Students Are from Venus Community College Students Must Be from Mars." Paper presented at the 2005 annual forum of the Association for Institutional Research, San Diego, 2005.

Hautamäki, V., and others. "Improving *K*-Means by Outlier Removal." *Image Analysis*, 2005, *3540*, 219–227.

Hubálek, Z. "Coefficients of Association and Similarity, Based on Binary (Presence-Absence) Data: An Evaluation." *Biological Reviews*, 1982, *57*, 669–689.

Hunt, L., and Jorgensen, M. "Clustering Mixed Data." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2011, *1*, 352–361.

Jain, A. K., and Dubes, R. C. *Algorithms for Clustering Data*. Upper Saddle River, N.J.: Prentice Hall, 1988.

Jain, A. K., Murty, M. N., and Flynn, P. J. "Data Clustering: A Review." *ACM Computing Surveys*, 1999, *31*, 264–323.

Milligan, G. W. "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms." *Psychometrika*, 1980, *45*, 325–342.

Milligan, G. W., and Cooper, M. C. "A Study of Standardization of Variables in Cluster Analysis." *Journal of Classification*, 1988, *5*, 181–204.

Morest, V. S., and Jenkins, D. *Institutional Research and the Culture of Evidence at Community Colleges*. New York: Community College Research Center, Teachers College, Columbia University, 2007. Retrieved Feb. 1, 2011, from http://www.achievingthedream.com/publications/research/institutionalresearchccrc.pdf.

Punj, G., and Stewart, D. W. "Cluster Analysis in Marketing Research: Review and Suggestions for Application." *Journal of Marketing Research*, 1983, *20*, 134–148.

Rapkin, B. D., and Luke, D. A. "Cluster Analysis in Community Research: Epistemology and Practice." *American Journal of Community Psychology*, 1993, *21*, 247–277.

StataCorp. *Stata Multivariate Statistics Reference Manual, Release 10*. College Station, Tex.: StataCorp LP, 2007.

Steinley, D. "Local Optima in *K*-Means Clustering: What You Don't Know May Hurt You." *Psychological Methods,* 2003, *8*, 294–304.

Steinley, D. "*K*-Means Clustering: A Half-Century Synthesis." *British Journal of Mathematical and Statistical Psychology*, 2006, *59*, 1–34.

Steinley, D., and Brusco, M. J. "Selection of Variables in Cluster Analysis: An Empirical Comparison of Eight Procedures." *Psychometrika*, 2008, *73*, 125–144.

VanDerLinden, K. *Credit Student Analysis: 1999 and 2000*. Annapolis Junction, Md.: Community College Press, American Association of Community Colleges, 2002.

PETER RILEY BAHR is an assistant professor in the Center for the Study of Higher and Postsecondary Education at the University of Michigan's School of Education.

ROB BIELBY is a doctoral student in the Center for the Study of Higher and Postsecondary Education at the University of Michigan's School of Education.

EMILY HOUSE is a doctoral student in Educational Studies at the University of Michigan's School of Education.