Prev Sci (2015) 16:1007–1016
DOI 10.1007/s11121-015-0561-z

Clustering Methods with Qualitative Data: a Mixed-Methods Approach for Prevention Research with Small Samples

David Henry[1] · Allison B. Dymnicki[2] · Nathaniel Mohatt[3] · James Allen[4] · James G. Kelly[1]

Published online: 7 May 2015

Abstract Qualitative methods potentially add depth to prevention research but can produce large amounts of complex data even with small samples. Studies conducted with culturally distinct samples often produce voluminous qualitative data but may lack sufficient sample sizes for sophisticated quantitative analysis. Currently lacking in mixed-methods research are methods allowing for more fully integrating qualitative and quantitative analysis techniques. Cluster analysis can be applied to coded qualitative data to clarify the findings of prevention studies by aiding efforts to reveal such things as the motives of participants for their actions and the reasons behind counterintuitive findings. By clustering groups of participants with similar profiles of codes in a quantitative analysis, cluster analysis can serve as a key component in mixed-methods research. This article reports two studies. In the first study, we conduct simulations to test the accuracy of cluster assignment using three different clustering methods with binary data as produced when coding qualitative interviews. Results indicated that hierarchical clustering, K-means clustering, and latent class analysis produced similar levels of accuracy with binary data and that the accuracy of these methods did not decrease with samples as small as 50. Whereas the first study explores the feasibility of using common clustering methods with binary data, the second study provides a "real-world" example using data from a qualitative study of community leadership connected with a drug abuse prevention project. We discuss the implications of this approach for conducting prevention research, especially with small samples and culturally distinct communities.

Keywords Mixed methods · Cluster analysis · Simulation · Community leadership

Qualitative inquiry can be a valuable tool in prevention research, providing rich material on motivations, behaviors, thoughts, and feelings. Qualitative inquiry is a key tool in discovery-based research, in hypothesis generation, and in characterizing the mechanisms underlying quantitative findings (e.g., Farrell et al. 2007) and can provide guidance in the development and refinement of measures. The patterns discovered in qualitative data can advance our understanding of how people interact with prevention programs and the divergent contexts in which they are embedded. In the current paper, we describe two studies that evaluate the performance of three different quantitative approaches for clustering binary data, such as the type produced in coding of qualitative interviews. We then provide an example of the application of a cluster analytic approach as a mixed-methods component of a study of community leadership in drug abuse prevention.

For many years, researchers have advocated mixing qualitative and quantitative research methods, but it is often unclear what is meant by "mixed methods." In a seminal and highly influential early example of mixed methods, Campbell and Fiske (1959) used a multitrait-multimethod matrix to study the validity of psychological constructs. Following this thinking, Jick (1979) described the notion of triangulation that involves seeking convergent findings from multiple methods. He argued that although many scholars pushed others to think of quantitative and qualitative methods as complementary,

✉ David Henry
dhenry@uic.edu

[1] University of Illinois at Chicago, Chicago, IL, USA

[2] American Institutes for Research, Washington, D.C., USA

[3] University of Colorado, Nederland, CO, USA

[4] University of Minnesota, Minneapolis, MN, USA

few offered explicit descriptions of techniques that made such complementarity possible. This seems to remain the case to this day.

In their review of mixed-method studies in the health sciences, Ostlund et al. (2011) found that most mixed-method studies do not, in fact, directly mix methods. Of the 168 studies reviewed, only 22 actually combined qualitative and quantitative methods by converting one type of data for use with the other type of analysis. Instead, the most common approach was parallel analysis, where qualitative and quantitative analyses are conducted separately, and findings are not compared or combined until the interpretation stage. Sequential methods, where one method is used to inform the other, were also popular. For example, "Qual-Quan" studies begin with qualitative methods and then use quantitative methods to clarify or test findings from the qualitative analysis, whereas "Quan-Qual" studies analyze quantitative data using systematic qualitative methods. In such sequential approaches, one method is used either before or after the other, with the purpose of the one method informing the other as the predominant method.

Cluster analysis makes it possible to mix methods, by making use of a quantitative method to analyze data generated through qualitative research. However, it is unclear how well cluster analysis methods perform with the small samples that may be produced by qualitative prevention research with culturally distinct communities. Cluster analysis may be defined as "the classification of similar objects into groups, where the number of groups as well as their forms are unknown" (Aldenderfer and Blashfield 1984). Clustering methods attempt to define groups of cases by mapping the similarities or dissimilarities on multiple dimensions (Henry et al. 2005). Clustering techniques were developed and have been largely used with continuous data and are not recommended for binary data (Finch 2005). In contrast, latent class methods have excellent properties with binary data such as the data for the presence or absence of themes or statements. However, it is uncertain how the performance of latent class methods changes with smaller samples.

Given the importance of describing the processes underlying the effects of preventive interventions, cluster analysis has wide potential application in prevention research. However, a cursory review of articles in *Prevention Science* over the past 10 years turned up only five articles that employed clustering methods, and four of these used latent class analysis.

Three aims guided the current study. First, we sought to assess the accuracy of three clustering methods (latent class analysis, K-means, and hierarchical clustering) in assigning cases to their correct clusters. This test is important to establish the validity of the mixed-methods approach that we are proposing and to ascertain if one clustering technique is superior to others in this application. Our second aim was to evaluate the effects of sample size, number of clusters, and degree of certainty of cluster membership on the performance of each method. This is important to advancing understanding of how different factors influence the accuracy of different clustering methods and to informing researchers of potential limitations of each of the methods. Our third aim was to report a real-world example of the application of clustering to qualitative data from a study of community leadership for substance abuse prevention. The results increased the interpretability of complex qualitative data.

Study 1: Simulation Testing the Performance of Clustering Methods with Binary Data

Latent class analysis (LCA) seeks classes that result in local independence (i.e., the indicator variables are not correlated within the latent classes) and provides benefits over the other approaches because of its ability to produce statistics that permit determination of model fit to the data (McCutcheon 1987). LCA results provide latent class probabilities that indicate the proportion of the population expected in each class and conditional probabilities that describe the likelihood of each level of each indicator variable for members of each latent class. Although methods for the use of continuous data based on LCA exist (e.g., Vermunt and Magidson 1999), LCA itself is designed for binary or categorical variables.

Like LCA, K-means cluster analysis requires the researcher to specify the number of clusters. Beginning with randomly selected centers, cases are moved between clusters to maximize the between-cluster variance relative to within-cluster variance (MacQueen 1967). K-means is generally regarded to be inappropriate for binary data (IBM Support Portal 2012).

Hierarchical clustering is a method for mapping the distances among persons or variables according to a distance metric and linkage method chosen by the researcher. The results include a table of distances at which cases or clusters have been joined and a picture of the pattern of relations among individuals in a data set, referred to as a dendrogram or tree diagram. Researchers often use hierarchical clustering to determine an appropriate number of clusters and then follow this hierarchical analysis with another method for sorting cases into a specified number of clusters, such as K-means (Mandara and Murray 2002). The distances used in hierarchical clustering are assumed to be derived from continuous data (Anderberg 1973; Finch 2005).

Despite this general assumption that cluster analysis requires continuous data, there have been some attempts to evaluate the performance of clustering methods with binary data. Most of these evaluations have used large samples (e.g., Ordonez 2003; Finch 2005). Hands and Everitt (1987) conducted the simulation most relevant to the current study, testing the ability of hierarchical clustering techniques to recover known cluster structure in binary data. They generated data sets according to a latent class model, specifying differing conditional probabilities, numbers of clusters, and sample sizes (50, 100, 200) and compared the recovered conditional probabilities with the pre-specified conditional probabilities using a Euclidian distance measure. They found that larger samples tended to result in better recovery of known cluster structure but only when the mixing proportions (i.e., the proportions of cases in each cluster) were equal. As mixing proportions became less equal, the effect of sample size diminished. Hands and Everitt also found that larger numbers of variables and smaller numbers of clusters produced more accurate recovery of clustered structure.

Study 1 aims to assess the accuracy of hierarchical clustering and K-means, relative to LCA, for accurately assigning cases to known numbers of clusters. We follow Dimitriadou et al. (2002) in using LCA as a standard against which to evaluate the performance of clustering algorithms. We make no attempt to evaluate the ability of clustering algorithms to choose the correct number of clusters for two reasons. First, the optimal index for this purpose is not a settled issue either for K-means or hierarchical clustering (Dimitriadou et al. 2002). Second, we believe that the selection of the number of clusters is as much a substantive as a statistical issue (Henry et al. 2005).

Study 1 Methods

For this study, we used a latent class model cf. Hands and Everitt 1987) to create artificial data sets that varied the numbers of clusters, proportions of cases in each, and the conditional probabilities of the binary variables given cluster membership, while holding the proportion of the sample in each cluster constant. Each data set consisted of four binary variables and a cluster identifier. Four values of p (conditional probabilities or cluster means) were used: .9, .8, .7, and .6. For creating two clusters, the sample was divided .6 to .4 (meaning that 60 %of the sample was in cluster 1 and 40 % of the sample was in cluster 2). For three clusters, the division was .5, .3, and .2 and for four, .4, .3, .2, and .1.

The binary variables were created according to one of four patterns. The pattern for the first cluster across the four variables was (p, p, $1-p$, $1-p$), where p represents the conditional probability of a variable being a “1.” In other words, if p was equal to .9, a person in the first cluster would have a probability of .9 on the first and second variables and probability of .1 on the third and fourth variables. The pattern for the second cluster was ($1-p$, $1-p$, p, p), the third was ($1-p$, p, $1-p$, p), and the fourth was (p, $1-p$, p, $1-p$).

Creation of each data set began with random data from a uniform distribution. If the value of a data point exceeded the threshold (p), the value of the case on that variable was set at 1; otherwise, it was set at 0. A thousand simulated data sets for each cell of a 4 (sample size) × 3 (number of clusters) × 4 (conditional probabilities) design were created.

Each data set was analyzed using LCA, K-means clustering (MacQueen’s algorithm), and hierarchical clustering (Ward’s method with Euclidian distances), using the *e1071* and *stats* packages in *R* (R Core Team 2013). Cases were assigned to the known number of clusters, and the accuracy of each method for cluster assignment was evaluated by calculating the Cramer’s V value for the assigned clusters vs. the known clusters:

$$\text{Cramer's}V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

where N is the number of observations and k is the number of clusters. With two clusters, Cramer’s V is equal to the *phi* coefficient. Cramer’s V ranges from 0 to 1 and may be interpreted in approximately the same manner as a correlation coefficient.

Study 1 Results

Only hierarchical clustering consistently produced usable solutions with samples of 20. The complete simulation results for two to four clusters, sample sizes of 20 to 500, and conditional probabilities set to .6 to .9 for each of the three methods are available from the first author on request. The use of random data in constructing the data sets resulted in a normal distribution of inter-cluster distances, allowing the simulation to more closely approximate real studies, where inter-cluster distance would be a random variable. For interpretable analysis, we created data points by taking the mean of the 1000 repetitions of each data set and regressed these on dummy codes for method, number of clusters, number of observations, conditional probability, and all two-way interactions between method and the other predictors.

Table 1 reports the results of a linear model of the mean Cramer’s V values over 1000 simulations on each of the three cluster analytic methods, number of clusters, number of observations (sample size), conditional probability, and the two-way interactions between method and each of these other predictors. The overall difference in accuracy among clustering methods was not significant. Higher conditional probabilities linking the indicators to the clusters were associated with greater accuracy (B=2.04, standard error (SE)=0.08, t(131)=27.06, p<.01). For every increase of .10 in conditional probability, there was an associated increase of .20 in accuracy. The decrease in accuracy associated with increasing numbers of clusters was .14 for LCA (B=−.14, SE=0.01, t(131)= 10.93, p<.01) and was slightly less for hierarchical clustering (B=−.11). Higher conditional probabilities improved accuracy for LCA (B=.20, SE=0.01, t(131)=27.08, p<.01), with slightly less benefit for hierarchical clustering (B=.18).

Table 1 Linear model of performance of hierarchical clustering and K-means clustering methods relative to latent class analysis clustering by number of clusters, number of observations, and 144 combinations of conditional probability and by sample size

Parameter	*B*	*SE*	*t* value	*p* value	95 % CI	
					Lower	Upper
Intercept	0.66	0.05	14.22	0.00	0.57	0.75
Method						
Hierarchical clustering[a]	−0.04	0.03	−1.59	0.11	−0.10	0.01
K-means clustering[a]	−0.02	0.03	−0.67	0.50	−0.07	0.04
Number of clusters	−0.14	0.01	−10.93	0.00	−0.16	−0.11
Number of observations	0.00	0.00	−1.34	0.18	−0.01	0.00
Conditional probability	0.20	0.01	27.08	0.00	0.19	0.22
Hierarchical × number of clusters[a]	0.03	0.01	2.28	0.02	0.00	0.06
K-means × number of clusters[a]	0.00	0.02	0.13	0.90	−0.03	0.03
Hierarchical × number of observations[a]	0.00	0.00	−0.48	0.63	−0.01	0.01
K-means × number of observations[a]	0.00	0.00	−0.48	0.63	−0.01	0.01
Hierarchical × conditional probability[a]	−0.02	0.01	−1.98	0.05	−0.04	0.00
K-means × conditional probability[a]	0.00	0.01	0.01	1.00	−0.02	0.02
SD of arcsine-transformed accuracy	−0.24	0.38	−0.63	0.53	−1.00	0.51

Each data point is the mean of 1000 repetitions with the outcome being arcsine-transformed Cramer's *V*. Parameters for the number of observations and conditional probability are adjusted for interpretability; *N* refers to the number of simulations run. Simulations for sample size $n=20$ were not included in this analysis because these simulations did not produce interpretable results with two of the methods

[a] Comparison method is LCA

Study 1: Cluster Analysis Simulation Discussion

We found no overall difference in accuracy between LCA, hierarchical clustering, and K-means clustering for assigning cases to known numbers of clusters. It is important to point out that our simulations controlled for sample size, number of clusters, and conditional probabilities of cluster membership. We did find that some of the characteristics of the sample had effects on accuracy for hierarchical clustering only. However, overall, we found little difference in performance across all three methods.

The results of this study support the use of cluster analysis methods for binary data produced by coding qualitative data using grounded theory (Strauss and Corbin 1998) or any number of other qualitative approaches to coding. Clustering approaches can assist the researcher in identifying potential clusters by providing additional information regarding sets of co-occurring coding categories within subgroups of individual research participants in a sample.

Study 2: Application of Clustering Methods to a Study of Community Leadership

The following provides an example of the use of clustering methods with qualitative data. In it, we describe the process of preparing the data and conducting such an analysis. We apply all three clustering techniques to coded interviews collected through a research collaboration between the University of Illinois at Chicago and the Developing Communities Project of Greater Roseland (DCP), a church-related community organization. First, we present some background information about the original study and its initial findings. Second, we describe and present the results of a cluster analysis with one portion of the qualitative data collected that was aimed at furthering our understanding of the development of community leadership.

In 1990, DCP obtained funding from the Illinois Department of Alcohol and Substance Abuse (DASA) for substance abuse prevention services to the Greater Roseland community in Chicago. A key component of this community-based approach to prevention was to recruit community members who would provide community education and substance abuse prevention leadership. The funding to DCP also included an evaluation requirement. Because of the centrality of leadership development in DCP's approach to substance abuse prevention, the evaluators focused on obtaining a broad understanding of the involvement of 77 community leaders who had received training from the organization.

Typically, leadership research focuses on the personal qualities and attributes of individuals who are influential persons (Yukl 1998). This exclusive focus on personality ignores contextual influences on the development of leadership. In contrast, the aim of our collaborative research endeavor was instead to elucidate issues, processes, and motivating influences behind the emergence of ordinary citizens as leaders in substance abuse prevention.

The planning team sought to present the findings from these interviews in a manner that would be of maximum use to the community organization. Cluster analysis was used to reveal common themes that could be used by the organization to guide recruitment and training of future community leaders.

Study 2: Application of Clustering Methods to a Qualitative Study—Methods

Participants

Eighty African-American adults who participated in the DCP training for substance abuse prevention leadership were asked about their reasons for involvement in the training and their leadership activities. Due to missing data, the final sample size was 77. The sample was 54.5 % female, and 23 % were clergy (94 % male) serving community churches.

Procedure

Interviews

Interviewees were recruited through regular training held by the community organization. Following informed consent, the semi-structured interview lasted approximately 1.5 h. The interview included questions on four topics: (a) social support for the community leader, (b) skills learned and skills to be learned in future training, (c) communications with other community organizations, and (d) personal visions of the community leaders.

Initial Coding of Interviews

Data analysis consisted of data display, open coding, interpretation, and verification. This approach of generating codes from interview transcripts instead of using a priori codes, along with continuously reusing successfully applied codes, is similar to elements of a grounded theory approach (Strauss and Corbin 1990). In the first phase, the researchers examined interview transcripts and reached consensus on codes to classify sections of text. Fifty-six codes were generated and used by research staff to code the 77 interviews. Detailed description of the elaborate coding process along with results and interpretation is reported in Tandon et al. (1998).

When the research team moved to the step of displaying the data, it became apparent that the 56 codes could be grouped into five dimensions of community leadership representing social processes of community leadership, in contrast to personal qualities of any individual. These processes were (1) reasons for community involvement and activities, (2) the organization's impact on the leaders, (3) factors promoting continued and active involvement, (4) religious influences affecting leaders' commitment to community work, and (5) personal visions.

In its community dissemination of the research results, the research team used a graphic representation of five trees to organize the codes within each of these five dimensions. The citizen leader could then examine each of their five "trees" and then see how their responses compared more generally with the modal values of the entire sample of leaders.

Following dissemination, the Executive Director of the organization encouraged the research team to explore the interview data further to identify possible subgroups among its leaders. The original evaluators used cluster analysis as a tool to identify possible subgroups of leaders.

Analytic Approach

We present cluster analysis of coding within the first dimension of community leadership: reasons for community involvement. Based on hierarchical analysis, a three-cluster solution was chosen. Subsequent discussion regarding these subgroup results with the DCP leadership suggested ways that their leaders differed in their motivations for initial involvement. Those leaders who were grouped in the first cluster were leaders motivated by the desire to create community change. Leaders in the second cluster appeared to have more personal motives for community involvement, such as gaining personal knowledge and exchanging information with others. The third cluster was a group of people that had a specific agenda for community change, namely, systemic changes in the community via economic development.

In the end, the organization and its board felt that the results from the cluster analysis proved helpful information in guiding enhancements to the organization's recruiting and training activities. This positive outcome led us to select this example for the present paper. In our reanalysis of the cluster analysis used in the DCP data, we use each of the three methods tested in Study 1 to test using real-world data how each method would inform the presentation of the data and results.

Steps Taken to Cluster Analyze Qualitative Data in the DCP Study

Examine Descriptive Information Prior to clustering with binary data, we examined a table of means and standard deviations and a matrix of *phi* coefficients (Table 2), where the presence of one or more instances of each code in an individual record is scored "1" and the absence of the code is scored "0." Since the mean represents the proportion of 1's in the entire sample, variables with very low or high mean values may be excluded because they provide little information for differentiating among clusters of individuals. For example, the very low mean of reason 1 (giving back to the community) in

Table 2 Descriptive statistics and phi coefficients for reasons for involvement codes

Reason	Mean	SD	Correlations								
			R2	R3	R4	R5	R6	R7	R8	R9	R10
R1: Participates in order to give back to the community	0.06	0.25	0.1	0.04	−0.08	0.18	−0.12	0.19	−0.09	0.18	−0.24*
R2: Participation provides a sense of satisfaction	0.09	0.29		0	−0.1	0.11	−0.03	0.14	−0.11	0	−0.02
R3: Participates in order to encourage information sharing in the community	0.14	0.35			0.26*	0.13	−0.1	0.14	0.1	−0.18	0
R4: Participates to gain personal knowledge	0.09	0.29				−0.01	−0.03	0.14	0.04	−0.2	−0.02
R5: Participates in order to gain community influence	0.16	0.37					−0.11	−0.06	0.09	0.05	−0.1
R6: Participates in order to feel an increased sense of security	0.18	0.39						−0.01	−0.05	0.30*	0.18
R7: Participates because the individual was recruited	0.65	0.48							−0.02	−0.14	−0.04
R8: Participates in order to receive additional community resources	0.10	0.31								−0.12	0.03
R9: Participates in order to foster community change	0.29	0.45									0.12
R10: Participates to invest in neighborhood youth	0.45	0.50									

Means are for scores for if a code was or was not present for each individual participant, where "0" = the reason for involvement was not coded and 1 = the reason for involvement was coded; *phi* coefficients are interpretable similarly to the correlation coefficient

*$p<.05$

Table 2 suggests that it may be less likely to contribute to the cluster solution. If the cluster analysis yielded a difficult to interpret or poorly resolved solution, a more interpretable analysis may be facilitated through removal of this item. In this case, the solution proved highly interpretable when including all variables, so all were retained. We also chose a single variable to represent any two codes that were highly correlated, as including highly correlated variables would unduly influence the cluster solution. Table 2 shows that few variables were significantly correlated (only 3 out of 45 correlations were significant) and none that were highly correlated.

Select a Clustering Method

The DCP analysis used hierarchical clustering to decide on a number of clusters and K-means to classify observations. The present study provides information that might assist others in choosing a method. Though the results of Study 1 suggest that the three clustering methods produce essentially equivalent results, there may be features of each method that make them more or less desirable for a particular application. For example, Study 1 found that LCA had problems converging on a solution with samples smaller than 50. K-means also had difficulty producing interpretable solutions with samples of 20. By contrast, hierarchical clustering, which involves a relatively simple partitioning of a data set based on distances, produced interpretable solutions and should have no difficulty with very small samples. Other considerations for choosing and using the methods follow.

Latent Class Analysis

In an LCA, the number of latent classes to retain is chosen based on the fit of the model to the data. Fit may be evaluated by a chi-square goodness-of-fit test in which a significant result indicates a significant difference between the data as reproduced by the model and the actual data. One strategy is to select the most parsimonious solution with a non-significant goodness-of-fit test. If no solution has a non-significant test, comparing fit indices such as likelihood ratio tests and scree plots of the Bayesian information criterion (Kass and Raftery 1995) can be used to determine the point at which adding additional clusters does not result in an improved solution.

The solution will contain, in addition to the fit statistics noted above, two sets of probabilities. There will be as many latent class probabilities as there are classes in the solution. They represent the probability of an individual being assigned to a particular latent class. The conditional probabilities interpret the solution. There will be one probability for each variable and each class. These represent the probability of a case having a "1" on the variable if the case is truly part of the latent class.

Hierarchical Cluster Analysis No special preparation is required for conducting a hierarchical cluster analysis with binary data. Hierarchical clustering programs provide many choices of methods for joining cases and many possible distance metrics. Ward's minimum variance method is generally a good choice as a joining method, and, as can be seen in Study 1, Euclidian distances produce results that compare favorably to other methods. Interpreting a hierarchical analysis is best done with the assistance of a tree diagram or dendrogram, illustrated in Fig. 1. Most hierarchical clustering programs require the analyst to request such a diagram. As can be seen in Fig. 1, the height of the links between individuals and clusters represents the distance and can serve as a basis for choosing the number of clusters.

K-Means Analysis Several methods for determining the number of clusters have been proposed for use with K-means, but no single method has been widely adopted (Dimitriadou et al. 2002). As with hierarchical clustering, there is no special preparation required for using K-means with binary data, other than examining descriptive information and correlations to avoid inclusion of variables that would result in unequal weighting.

Interpretation of Clusters

The original DCP study used K-means clustering and worked with the community organization to reach an interpretable solution. The Executive Director and two members of the board of the sponsoring organization participated in multiple meetings with the research staff to interpret the cluster solutions. If they did not find the results interpretable for understanding leadership in their organization, the cluster analyses were redone. The results of a K-means analysis may be illustrated using a line plot or bar plot where the mean of each variable on each cluster is plotted, as in Fig. 2. Using such diagrams, we interpreted the different patterns of leadership involvement for each of the three proposed clusters. One cluster may be termed "gain personal benefits." Specifically, this cluster had the highest proportion who endorsed gaining personal knowledge, receiving additional community resources, and encouraging information sharing with others. A second cluster could be titled "moral obligation." This cluster had the highest values on giving back to the community and being a leader providing a sense of satisfaction. The final cluster included leaders who became involved to promote community change; specifically, this cluster had the highest values on wanting to create community change and being invested in the neighborhood youth.

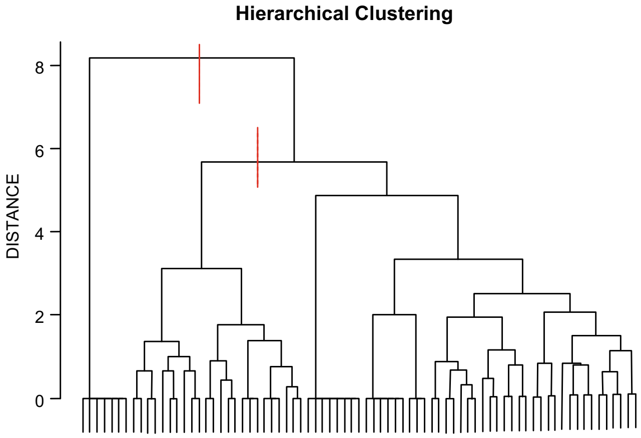

Fig. 1 Tree diagram or "dendrogram" from a hierarchical cluster analysis with cut lines illustrating division of the Developing Communities Project of Greater Roseland (DCP) reasons for involvement data into three clusters

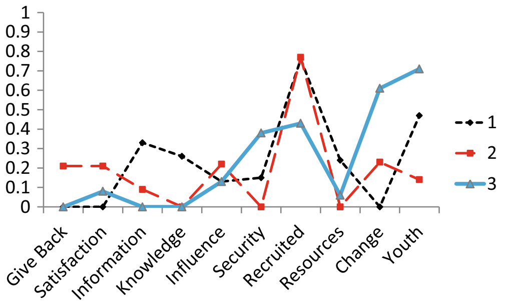

Fig. 2 K-Means analysis of the Developing Communities Project of Greater Roseland (DCP) reasons for involvement data

Study 2: Application of Clustering Methods to a Qualitative Study—Discussion

The cluster analysis of the DCP data provided new information about common patterns in the emergence of community leadership for substance abuse prevention. Specifically, the analyses conducted for DCP helped the organization revise its recruitment and training methods. This quantitative classification method served as a useful tool for making sense of a large amount of qualitative interview data. These results informed the work of the community organization staff, expanded our understanding of the multiple expressions of leadership, and informed our use of clustering methods within a mixed-methods research framework.

Despite the small sample size and other limitations of the data, cluster analysis provided a useful means for discovering natural groupings of community leaders in the complex data derived from these interviews. The three-cluster solution derived initially by the DCP investigators helped guide collaborative analysis informed through participation of the staff in the community organization. These analyses identified ways to better support and accommodate variation in motives and contextual circumstances of potential leaders for DCP when recruiting and maintaining active community leadership.

General Discussion

This article should provide encouragement to prevention researchers aiming to understand relationships within complex data produced by qualitative interviews. First, using simulation studies, we demonstrated that three different types of clustering methods accurately assigned individuals into known cluster solutions, controlling for number of clusters, conditional probabilities, and sample size, doing so with small samples using binary data. Second, we reported the results of an application of one approach to a prevention study that gathered complex qualitative interview data from community leaders who were recruited and trained as part of a substance abuse prevention program.

The Performance of Clustering Methods with Binary Data

The simulation produced evidence that hierarchical cluster analysis and K-means performed as well as LCA with binary data. Characteristics of the data sets (conditional probabilities, number of clusters, and sample size) differentially affected performance of hierarchical clustering to some extent. Performance was not hampered by sample sizes as small as N=50, contrary to the oft-repeated maxim that latent class methods require large samples (Haughton and Haughton 2011). Similarly, hierarchical clustering performed well with dichotomous data, contrary to admonitions that it should not be attempted with binary variables (IBM 2012). The one caveat arising from the simulations is that LCA and K-means may not perform well with binary data in samples smaller than N=50, which are common to many qualitative studies. In these cases, hierarchical clustering, which is essentially a simple descriptive method, appears to work well with binary data, even with the very small samples likely to be found in some qualitative studies.

Few comparisons of clustering methods exist in the literature. To our knowledge, this is only the second study to compare clustering and latent class methods on their performance with binary data and the first to do so with small samples. Dimitriadou et al. 2002 compared multiple indices for determining the correct number of clusters with binary data using K-means and another non-hierarchical method. Using simulated samples of 1000 cases with probabilities derived in a manner similar to the present study, they found that K-means accurately identified a four-cluster solution in approximately 50 % of the samples. The present study is the first study of which we are aware to vary sample size, conditional probability of cluster membership, and number of clusters to evaluate the performance of different methods in the same study. Nguyen and Rayward-Smith (2008) reviewed 40 metrics for measuring the quality of clustering solutions but did not arrive at a preferred measure. Eshghi et al. (2011) compared methods using the within- and between-variance of cluster solutions as a metric for comparison. Only the study by Hands and Everitt (1987) compared clustering methods on accuracy of cluster assignment with binary data. Such comparison would seem highly relevant for the purpose employed here, namely, reducing the complexity of coded qualitative data.

Application of Clustering Methods to a Qualitative Study of Community Leadership

During the course of the original DCP collaborative research study of community leadership for substance abuse prevention in Chicago, clustering provided several valuable insights. This led the community organization to a deeper understanding of the motives of its leadership development training participants and ultimately contributed to changes in the organization's recruiting and training practices.

Contributions of the Current Study

One contribution of the current study is to support the use of the older clustering methods, such as K-means and hierarchical clustering, with binary data. In the simulations, the performance of these methods was equal to that of latent class analysis. The second contribution is to identify hierarchical clustering as capable of producing valid solutions with samples as small as N=20. In the simulations, the performance of all three clustering methods was not affected as samples decreased in size, performing acceptably with samples as small as N=50, and hierarchical clustering in particular showed accuracy with samples as small as N=20. Third, researchers should exercise appropriate caution when using and interpreting cluster analysis results. One example of appropriate caution is choosing solutions of fewer clusters over solutions with many clusters, especially with smaller samples. In addition to the consideration of parsimony, the accuracy of assignment is higher with fewer clusters but can decline rapidly as the number of clusters increases. Given that prevention data are not likely to regularly provide a 90 % level of certainty of cluster membership, the results of these simulations emphasize how clustering is at core simply a tool for mapping multivariate distances. As such, meaningful application of clustering methods in mixed-methods research is heavily dependent upon strong grounding in theory, nuanced understandings of contexts, and close, deep understanding of the qualitative data set for valid interpretation.

The most important contribution of the current study is to provide a method for reducing the complexity of coded qualitative data in prevention studies with small samples. Qualitative research potentially produces volumes of coded and codeable data. The DCP study produced 56 codes in five different topic areas (Tandon et al. 1998). Although it is certainly possible to work with codes directly, as the DCP investigators did at first, the complexity of the analysis may exceed the resources of the researchers. Clustering produces groups of individuals, each with a specific pattern of responses. This can be extremely helpful in interpreting the results.

Limitations

There are some limitations that should be taken into account when considering this study. First, there exists only this single study providing precedent for testing the accuracy of cluster assignment with binary data. We varied parameters similar to those varied by Hands and Everitt (1987), but it would be possible also to vary other parameters. Since in actual practice, the underlying cluster structure is always unknown, we cannot determine the extent to which the conditional probabilities used in the simulations are similar to those likely to be encountered in prevention research. Despite these limitations, the results of these studies provide promising evidence supporting use of clustering methods as a mixed-method approach in the analysis of coded qualitative data.

Conclusions

Cluster analysis constitutes a promising tool that can guide theory and context-informed interpretation for qualitative researchers. The approach as used with the DCP interviews yielded new information regarding the qualitative coding and its interpretation. Its use here represents at true mixed method, where a quantitative method can assist in an analysis wherein the qualitative material is primary in importance for analytic yield. In addition, the approach opens new horizons in qualitative analysis by providing researchers a systematic method to guide subgroup analysis within a qualitative data set. Typically, qualitative findings are treated as reflective of the entire group of interviewees in a study, or alternatively, a particular finding may be reported as relevant to some but not all individuals or relevant to existing recognized social categories (e.g., gender). Clustering affords the qualitative researcher a means by which to identify possible subgroups within a sample that may differ in terms of the relevance of various codes. Perhaps more intriguing, though only minimally explored in the current analysis, it potentially allows more systematic exploration of the meaning of relational configurations of code structures over that of individual codes in isolation, as part of the interpretative work in qualitative analysis.

We emphasize that the clustering method that we have described requires combining multiple methods to confirm the accuracy of proposed cluster solutions. Cluster analysis only provides guidance in suggesting configurations and not firm decision rules. Use of cluster analysis and, in particular, the selection of meaningful clusters involve triangulation through use of interpretation processes embedded and require researcher grounding in theory and context, as well as full use of the rich immersion in narrative data that is only possible in qualitative analysis. This triangulation of the cluster analytic findings with convergences from theory, from a nuanced knowledge about context, and from deep understanding of rich narrative data moves us closer to the conceptualization of convergent findings from multiple methods of Jick (1979). In this way, the approach can provide an informative method of small-samples inquiry for difficult to access populations, including research with small and distinct populations that are often common to health disparity prevention research.

Acknowledgments The authors gratefully acknowledge the contributions of Mary Murray, M.A., who performed the cluster analyses for the original DCP study, Debra Strickland, Former Executive Director of the Developing Communities Project, and the community leaders who participated in the interviews for Study 2. Partial support for this study was provided by grant number R13 DA030834 (PI, Ching Fok, Ph.D.) for the conference, "Advancing Science with Culturally Distinct Communities."

Conflict of Interest The authors declare that they have no conflicts of interest.

References

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis: A Sage University paper*. Beverly Hills: Sage.

Anderberg, M. R. (1973). Cluster analysis for applications: DTIC document.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. Accessed 22 Feb 2014.

Dimitriadou, E., Dolnicar, S., & Weingessel, A. (2002). An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika, 67*, 137–160.

Eshghi, A., Haughton, D., Legrand, P., Skaletsky, M., & Woolford, S. (2011). Identifying groups: A comparison of methodologies. *Journal of Data Science, 9*, 271–291.

Farrell, A. D., Erwin, E. H., Allison, K., Meyer, A. L., Sullivan, T. N., Camou, S., Esposito, L. E. (2007). Problematic situations in the lives of urban African American middle school students: A qualitative study. Journal of Research on Adolescence, 17, 413-454.

Finch, H. (2005). Comparison of distance measures in cluster analysis with dichotomous data. *Journal of Data Science, 3*, 85–100.

Hands, S., & Everitt, B. (1987). A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques. *Multivariate Behavioral Research, 22*, 235–243.

Haughton, D., & Haughton, J. (2011). *Chapter 6: Grouping methods*. Springer Science+Business Media, LLC, Berlin.

Henry, D., Tolan, P. H., & Gorman-Smith, D. (2005). Cluster analysis in family psychology research. *Journal of Family Psychology, 19*, 121–132.

IBM Support Portal. (2012). Clustering binary data with K-means (should be avoided). *Technote* Retrieved March 4, 2013, from http://www-1.ibm.com/support/docview.wss?uid=swg21477401

Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly, 24*, 602–611.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of 5th Berkeley symposium on mathematical statistics* *and probability* (Vol. 1, pp. 281–297). Berkeley: University of California Press.

Mandara, J., & Murray, C. B. (2002). Development of an empirical typology of African American family functioning. *Journal of Family Psychology, 16*, 318.

McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park: Sage.

Nguyen, Q. H., & Rayward-Smith, V. J. (2008). Internal quality measures for clustering in metric spaces. *International Journal of Business Intelligence and Data Mining, 3*, 4–29. doi:10.1504/IJBIDM.2008.017973.

Ordonez, C. (2003). Clustering binary data streams with kmeans. Proceedings of the 8th ACM SIGMOD workshop on research issues in data mining and knowledge discovery, 12-19.

Ostlund, U., Kidd, L., Wengstrom, Y., & Rowa-Dewar, N. (2011). Combining qualitative and quantitative research within mixed method research designs: A methodological review. *International Journal of Nursing Studies, 48*, 369–383. doi:10.1016/j.ijnurstu.2010.10.005.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park: Sage.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Newbury Park: Sage.

Tandon, S. D., Azelton, L. S., Kelly, J. G., & Strickland, D. (1998). Constructing a tree for community leaders: Contexts and processes in collaborative inquiry. *American Journal of Community Psychology, 26*, 669–696.

Vermunt, J. K., & Magidson, J. (1999). Exploratory latent class cluster, factor, and regression analysis: the Latent GOLD approach. Paper presented at the Proceedings EMPS_99 conference, Lunenburg, Germany.

Yukl, G. (1998). *Leadership in organizations* (4th ed.). Englewood Cliffs: Prentice-Hall.