Use of Clustering Methods to Understand More about the Case

Allison B. Dymnicki\textsuperscript{a} and David B. Henry\textsuperscript{a}

\textsuperscript{a}Institute of Health Research and Policy, University of Illinois at Chicago

Abstract

During the past seventy years, the field of cluster analysis has emerged, accompanied by a plethora of methods, algorithms, concepts, and terminology that are used in cluster-related research. We refer to cluster analysis (CA) as a general approach composed of several multivariate methods for delineating natural groups or clusters in data sets. In this paper, we describe the ability of CA to provide rich information about the individual case and highlight potential underlying social processes. First, we discuss the theory behind CA as well as differentiate between more and less familiar clustering approaches. Second, we illustrate the value of less familiar clustering techniques by comparing the results of a four wave growth mixture model of family variables versus clustering the same data with a more familiar two-step approach. The growth mixture modelling approach suggested a one-class cluster solution where all families shared similar growth trajectories in parenting practices and family relationship characteristics. However the two-step clustering approach suggested a four-class solution. Finally, we describe ways that CA allows researchers to model processes whose outcomes are the results of a combination of multiple factors and additional benefits of less familiar clustering methods.

Keywords: Cluster analysis, case-based approaches, families, growth mixture modelling

Introduction

Recently, a number of authors have criticized variable-based approaches which typically employ the use of regression analyses and assume causal homogeneity across cases (Abbott, 2001; Ragin, 2006; Cooper, 2005). In response to such criticisms, researchers have developed a variety of newer case-based approaches that have the ability to explore how certain predictor variables and groups of predictor variables are sufficient and/or necessary conditions for a specific outcome (Byrne & Ragin, 2009). The purpose of this article is threefold: (a) to discuss the theoretical underpinnings of cluster analysis (CA) and distinguish between more and less familiar clustering methods, (b) illustrate the benefits of the less familiar techniques with an empirical example where we applied two cluster approaches to understand the social processes of families in urban neighbourhoods, and (c) describe how CA can allow researchers to model processes whose outcomes are the results of a combination of several factors and some additional benefits of less familiar clustering methods.

The Theory of Cluster Analysis

When monographs and books emerged as early as 1939 (Tryon, 1939) establishing CA as a new field, researchers distinguished this approach from variable-based approaches by its aim to identify and describe
groups of individuals within a sample who share similar constellations of attributes (Aldenderfer & Blashfield, 1984). We use the term ‘cluster analysis’ (CA) to refer to a general approach composed of several multivariate methods for delineating groupings in the data that occur at greater than chance frequency (Seber, 1984). CA identifies and describes groups of cases defined by similarities or dissimilarities on multiple dimensions.

Such modelling, however, employs a different logic than that used when studying average effects of causal or predictor variables on a dependent variable, as with correlation or regression. Fundamentally, clustering involves sorting cases or variables according to their similarity on one or more dimensions and focuses on the interdependence among variables. CA creates groups that maximize within-group similarity and minimize between-group similarity (Henry et al., 2005).

From a theoretical perspective, we can define a cluster as an approximation to uniqueness (Dymnicki & Henry, in press). Just as a sample mean produces an unbiased approximation of a population mean, a cluster approximates the unique configurations of characteristics found in individuals. As sample sizes increase, estimates of population means become more accurate and we draw closer to sampling an entire population. Clusters become more accurate for identifying individuals as the number of identifying characteristics and the number of clusters increases. Researchers, however, must strive for a balance in determining the final cluster solution. Choosing the solution with too many groups reduces CA’s ability to aid in reducing complexity, but choosing a solution with too few groups reduces CA’s accuracy in classifying individual cases (Macrae & Bodenhausen, 2000).

Different Clustering Approaches

We are making a distinction in the current paper between older clustering approaches offered in statistical packages such as SAS or SPSS, and newer methods that may be less familiar because of their relative absence from such packages. Focusing on these newer approaches we highlight methods that have advanced the field including: (a) fuzzy clustering which allows researchers to place a case into multiple clusters instead of somewhat arbitrarily assigning a case to only one cluster when it may be close to several (de Oliveira & Pedrycz, 2007, Cooper & Glaesser, 2011); (b) model-based clustering which allows researchers to compare multiple solutions based on different models of cluster shapes (e.g., spherical or ellipsoidal) instead of assuming a spherical shape (Fraley & Rafferty, 1998); and (c) latent class analysis/latent profile analysis which provides fit indices for different cluster solutions found with continuous data (Mitchell & Plunkett, 2000; Vermunt & Magidson, 2000).

The Decision to Use Latent Class Analysis

In the current paper, we apply a type of latent class analysis, termed growth mixture modelling, that greatly expands the ability of researchers to understand patterns of growth that characterize individual cases. We applied growth mixture modelling to data from a study by Gorman-Smith et al. (1998) in order to compare the results to their analysis of family variables, conducted before methods were available to model clusters of variables over time. Other longitudinal approaches that could have been used to understand and group differences in developmental trajectories over time include (a) latent transition analysis which is primarily used for categorical or binary data (Lanza et al., 2008) or (b) a semi-parametric mixture model that is appropriate for data with skewed distributions (Jones et al., 2001). The growth mixture modelling approach we used was the most appropriate approach given the type of data we employed and the fact that our variables were approximately normally distributed.

A Brief Description of Growth Mixture Modelling

The basic idea behind growth mixture modelling is that individual growth patterns can be clustered much in the same way that individual observations are clustered in a cluster analysis. A clustering method based on latent class analysis is applied to the intercepts and slopes of one or more variables and groups of subjects that are similar in their growth parameters (Muthén & Muthén, 2000; Muthén & Khoo, 1998). Classes of growth patterns could reveal three types of differences in the growth of family characteristics: (a) youth could differ
in their initial levels of any or all of the family variables (explored through including random intercepts), (b) youth could differ in the linear growth of family variables (explored through including random linear slopes), and (c) youth could differ in the acceleration or deceleration of family variables (explored through including random quadratic slopes). This approach makes it possible to determine whether a single growth pattern is adequate to describe the sample (as is borne out in our analyses) or multiple growth patterns are indicated. Growth mixture modelling both estimates mean growth curves for each class and captures individual variation around these growth curves by the estimation of variances in the growth parameters within each class (Muthén & Muthén, 2000).

Consider a family variable score $y_{it}$ for individual $i$ at time point $t$. In this analysis, $t$ indicates one of four different waves of measurement. The growth model for each family is specified as

$$y_{it} = \beta_{0i} + \beta_{1i}t + \beta_{2i}t^{2} + e_{it}$$

where $\beta_{0i}$, $\beta_{1i}$, and $\beta_{2i}$ are family-specific parameters describing initial level of the family variable ($\beta_{0i}$), and rate of linear growth ($\beta_{1i}$) and acceleration ($\beta_{2i}$). Thus, each family will have three parameters that describe the growth curve of the variable of interest for that family. Growth mixture modelling describes the distribution of those growth curves in two ways. First, it determines whether solutions with multiple groups of growth curves improve the fit of the model to the data. Second, using methods related to latent class analysis, growth mixture modelling estimates the variability in growth parameters within groups of families. Growth mixture modelling can incorporate multiple variables into clusters of growth patterns. Thus, a cluster might be described as stable on one variable, increasing linearly on another, and decelerating on a third.

**Using Cluster Analysis to Understand Growth and Change in Families**

Extensive research has identified family functioning as among the strongest predictors of risk for delinquent behaviour (Loeber & Dishion, 1983; McCord, 1991, Patterson & Stouthamer-Loeber, 1984). Much of this work focuses on parenting practices and specifically a parent’s ability to effectively monitor and discipline a child. Regardless of ethnicity or socioeconomic status, research suggests the importance of parents being involved with their children, monitoring their child’s whereabouts, and implementing consistent and effective disciplinary practices (Gorman-Smith et al., 1999; Patterson et al., 1992). Evidence also points to other aspects of family functioning that explain additional variance in an individual’s risk of delinquency (Farrington, 1994). These include low levels of emotional warmth and cohesion, poor communication styles, lack of organizational structure, and low beliefs about the importance of family (Gorman-Smith et al., 2000b; Tolan et al., 1997). Our exemplar study, conducted by Gorman-Smith and colleagues (Gorman-Smith et al., 1998), built upon this previous work to classify families based on measures of parenting practices (discipline and monitoring; Tolan et al., 1997) and family functioning (cohesion, beliefs about family, and family structure; Gorman-Smith et al., 1996).

As described below, Gorman-Smith and colleagues (1998) used a two-step clustering approach (i.e., hierarchical CA followed by K-means non-hierarchical splitting) to cluster four waves of family variables, producing a four-fold typology of families. The family types were then related to a youth’s risk of engagement in different delinquency pathways. In the current paper, we used latent class analysis to cluster four waves of the same measures, producing evidence that a single class of growth patterns was sufficient to describe families. We describe a youth’s likelihood of involvement in different delinquency pathways for the one-class solution versus the previously found four-class solution. In addition, we illustrate the benefit of CA methods to identify multivariate outliers by exploring ten cases identified as a second cluster in a separate part of the latent class analysis. These ten cases have different growth trajectories in parenting practices and family functioning variables from the rest of the sample and upon further investigation, had interesting differences in ratings of neighbourhood safety.
Study Design and Sample

The current study used data from four waves (collected 1 year apart) of the Chicago Youth Development Study (CYDS), a longitudinal study of the development of serious delinquent behavior among inner-city young adolescent males. Boys were initially recruited from fifth- and seventh-grade classrooms in 17 Chicago public schools. After obtaining parental permission, teachers completed the Teacher Rating Form of the Child Behavior Checklist (Achenbach, 1991) for 1,105 boys (92% of the population of fifth and seventh grade boys in the schools). The sample was selected so that 50% were considered ‘high risk’ for development of serious aggression based on teacher ratings (i.e., boys received a rating at the 98th percentile for aggressive behavior on the Teacher Rating Form) and the remaining 50% were randomly selected from the remaining sample. Seventy-five percent (341) of the eligible participants completed interviews during the first wave of interviews.

Participants who participated in at least three waves of data collection over the first four waves were included in the current analyses (N = 263). Comparison of the non-interviewed and the interviewed of the targeted sample yielded no significant differences using initial teacher ratings of aggression \( F(1, 524) = .57, p = .45 \), or official arrest records (ever arrested) as of Wave 4 \( X^2(1, N = 298) = .37, p = .54 \). Comparisons were also made, for each subsequent wave, between those continuing to participate and those not continuing on all predictors and behaviors. Based on over twenty comparisons, we found only a single significant difference between groups that accounted for less than 2% of the variance, indicating little bias due to attrition.

At the first wave of interviews, participants were between ages 11 and 14 and at the fourth wave of interviews, between ages 15 and 18. Adolescents were primarily African-American or Latino/Hispanic, living in 36 inner-city communities. Sixty-two percent lived in single-parent homes, 74% of their families had incomes below $20,000 per year, and 48% of the families had a total income below $10,000 per year. More information about sample selection and information regarding attrition analyses have been described in detail elsewhere (Gorman-Smith et al., 1996; 1998; 2000a).

Independent and Dependent Variables

Parenting Practices.

We measured parenting practices using the same questions included in the Pittsburgh Youth Study (Loeber et al., 1991) which involves four scales of mother and child ratings: (a) discipline effectiveness, (b) discipline avoidance, (c) positive parenting, and (d) extent of monitoring/involvement in the child’s life. Internal consistency reliabilities, as measured by alpha, ranged from 0.68 to 0.81 for these constructs. Using confirmatory factor analysis, two higher order factors were identified that represent Discipline and Monitoring (Gorman-Smith et al., 1996; Patterson et al., 1992). Discipline refers to the mother’s report of how effective their disciplinary practices are in controlling their son’s behaviour and how often they avoid disciplining their son for fear of the son’s behaviour escalating. Monitoring refers to the extent of parent involvement in the son’s daily activity and routines, awareness of the son’s whereabouts, and use of encouragement of appropriate behaviour.

Family Relationship Characteristics.

Caregivers and youth completed a 92-item measure assessing family relationship characteristics (Tolan et al., 1997). The Family Relationship Scale, originally developed by Gorman-Smith et al. (1996), taps three aspects of family relationship characteristics that distinguish risk for serious antisocial behavior: Beliefs, Cohesion, and Structures. Beliefs represented expectations regarding the importance of the family, purpose of the family, and expectations about child development, as well as tolerance for deviant beliefs. Cohesion represented the extent to which emotional closeness, dependability, support, and clear communication within families existed. Structure represented the family’s organization and the support provided to family members. Besides fitting the data extremely well according to fit indicators, these three constructs represent theoretically important aspects of family functioning from a family systems perspective (Nichols & Schwartz, 1991).
**Engagement in Different Delinquency Pathways.**

Non-hierarchical k-means cluster analysis identified four delinquency pathways based on frequency, seriousness, violence, and consistency in involvement in antisocial behaviour over four waves (Gorman-Smith et al., 1998; Loeber et al., 1993). The ten variables used in the cluster analysis included parent and youth reports of Authority Conflict, Covert Behaviour, and Overt Behaviour taken from the Maternal Child Behaviour Checklist, Youth Self-Report of Delinquency Scale, and Youth Self Report of the Child Behaviour Checklist (YSR; Achenbach, 1991). Based on the pathways identified through cluster analysis, each youth was classified into one of four groups: (1) nonoffenders (i.e., those with some aggression and fighting, but no delinquent behaviours, 26%); (2) chronic minor offenders (i.e., those consistently involved in only minor offences over each of the four waves, 34%); (3) escalators (i.e., those starting delinquent involvement at a later wave and quickly escalating to more seriousness and violent offending, 12%); and (4) serious, chronic, and violent offenders (i.e., those involved in serious and violent offending at every wave, 28%). The four groups and percentages of involvement in each are consistent with pathways reported by others (Loeber et al., 1991; LeBlanc & Kaspy, 1998).

**Neighbourhood Safety.**

Mothers and their sons rated to what extent they felt safe in their neighbourhood by completing the Fear of Crime Scale and the Extent of Neighbourhood Problems Scale. The first scale asks respondents about the extent of their safety concerns and how the fear of crime has impacted on various aspects of people's lives (e.g., has the fear of crime limited people to where they can shop, work, or go by themselves). The second scale asked mothers and sons about a range of problems (e.g., graffiti, drugs, and gangs) in their neighbourhoods and the seriousness of these problems, with responses on a Likert-type scale from 1 (a little) to 5 (serious problem). Internal consistencies were 0.76 and 0.84 for the mothers' and sons report on the Fear of Crime Scale; and 0.84 and 0.83 for mothers and sons' reports on the Extent of Neighbourhood Problems Scale. Confirmatory factor analysis indicated these two scales should be combined into one latent construct representing Concerns for Safety.

**Results of Family Types Identified by Gorman-Smith et al. (1998)**

Using four waves of the family variables, Gorman-Smith et al. (1998) first used hierarchical cluster analysis through SYSTAT (Wilkinson et al., 1996) to identify the optimal number of clusters. Deciding on a four-cluster solution, they extracted the clusters using a K-means non-hierarchical splitting method to find solutions that provided maximum differentiation of cases on the variables they entered and the greatest interpretability.

Table 1 presents the means and standard errors of the four cluster solution found based on the five family variables. The first type, referred to as ‘Exceptionally Functioning’ (n = 72, 27%) had high levels of parenting practices and structures over time, high levels of cohesion, and strong beliefs about the importance of family. The second group, referred to as ‘Task-Orientated’ (n = 67, 25%), had high levels of parenting practices and structure but low levels of emotional warmth and low beliefs about the family. The third group, referred to as ‘Struggling Families’ (n = 52, 20%) had consistently low levels of discipline, monitoring, structure, cohesion, and beliefs about the family. The fourth group, referred to as ‘Moderately Functioning’ (n = 72, 27%) had adequate but not high levels of discipline and monitoring over time and high levels of beliefs.
Table 1: Means and Standard Errors of Family Cluster Characteristics at Four Waves

| Cluster          | Wave | Beliefs | Cohesion | Structure | Discipline | Monitoring |
|------------------|------|---------|----------|-----------|------------|------------|
| 1. Exceptionally functioning families | 1    | -0.28 (0.06) | -0.13 (0.03) | -0.59 (0.03) | 0.52 (0.06) | 0.21 (0.05) |
|                  | 2    | 0.40 (0.07)  | 0.33 (0.04)  | 0.18 (0.05)  | -0.06 (0.06) | 0.23 (0.04) |
|                  | 3    | 0.22 (0.07)  | 0.24 (0.04)  | -0.07 (0.04) | 0.09 (0.06)  | 0.24 (0.06) |
|                  | 4    | 0.31 (0.07)  | -0.23 (0.04) | 0.16 (0.04)  | 0.01 (0.06)  | 0.24 (0.06) |
| n = 72           |      |          |          |           |            |            |
| 2. Task-oriented families               | 1    | -0.54 (0.05) | -0.25 (0.03) | -0.67 (0.03) | 0.44 (0.06) | 0.15 (0.04) |
|                  | 2    | -0.33 (0.06) | -0.00 (0.04) | 0.09 (0.05)  | 0.06 (0.06)  | 0.20 (0.04) |
|                  | 3    | -0.57 (0.07) | -0.06 (0.04) | -0.01 (0.04) | 0.00 (0.06)  | 0.07 (0.05) |
|                  | 4    | -0.47 (0.06) | -0.52 (0.04) | 0.02 (0.04)  | -0.05 (0.06) | 0.01 (0.05) |
| n = 67           |      |          |          |           |            |            |
| 3. Struggling families                  | 1    | -0.62 (0.07) | -0.35 (0.03) | -0.84 (0.04) | 0.16 (0.07) | -0.33 (0.05) |
|                  | 2    | -0.33 (0.08) | -0.24 (0.05) | -0.16 (0.06) | -0.12 (0.07) | -0.34 (0.05) |
|                  | 3    | -0.28 (0.08) | -0.19 (0.05) | -0.01 (0.05) | -0.32 (0.07) | -0.36 (0.06) |
|                  | 4    | -0.37 (0.08) | -0.79 (0.04) | -0.17 (0.05) | -0.41 (0.07) | -0.54 (0.07) |
| n = 52           |      |          |          |           |            |            |
| 4. Moderately functioning families     | 1    | -0.28 (0.06) | -0.24 (0.03) | -0.72 (0.03) | 0.27 (0.06) | -0.02 (0.04) |
|                  | 2    | 0.07 (0.07)  | -0.06 (0.04) | -0.07 (0.05) | 0.00 (0.06)  | -0.05 (0.04) |
|                  | 3    | 0.35 (0.07)  | -0.05 (0.05) | -0.05 (0.04) | -0.12 (0.06) | -0.04 (0.05) |
|                  | 4    | 0.36 (0.06)  | -0.56 (0.04) | -0.02 (0.04) | -0.16 (0.06) | -0.14 (0.06) |
| n = 72           |      |          |          |           |            |            |

Note. From ‘A developmental-ecological model of the relation of family functioning to patterns of delinquency.’ By D. Gorman-Smith, P. H. Tolan, & D. B. Henry, 2000. Journal of Quantitative Criminology, 16, p. 184. Copyright 2000 by Springer. Reprinted with permission.

These four family types were significantly related to adolescents’ delinquency patterns. Youth from struggling families were more likely to have a pattern of chronic minor delinquent involvement (odds ratio = 20.09, p < .001) and escalating delinquent involvement (odds-ratio = 68.7, p < .001) than those from moderately functioning families, whereas youth in task-oriented families were more likely to be involved in serious, chronic, and violent delinquency (odds-ratio = 4.80, p < .001) than youth in moderately functioning families. Youth in exceptional families were less likely to engage in any delinquent involvement than youth in moderately functioning families (odds-ratios = 0.22, 0.01, 0.03, respectively). Thus, the four family types demonstrated their value in their ability to predict future adolescent delinquent behaviour.
Table 2: Fit Indices for One through Five Class Cluster Solutions of Family Characteristics

| No. of Free Parameters | BIC       | LMR LRT            | Entropy |
|------------------------|-----------|--------------------|---------|
|                         | With Quadratic Term                      |                     |         |
| One-class               | 157       | 3496.732           | NA      | 1.00    |
| Two-class               | 171       | 3545.734           | 79.596, $p = 0.2521$ | 0.989 |
| Three-class             | 187       | 3571.446           | 62.739, $p = 0.4761$ | 0.851 |
| Four-class              | 203       | 3638.379           | 22.976, $p = 0.7098$ | 0.931 |
| Five-class              | 219       | 3680.642           | 9.771, $p = 0.7465$  | 0.916 |
|                         | Without Quadratic Term                    |                     |         |
| One class               | 93        | 3664.201           | NA      | 1.00    |
| Two-class               | 107       | 3686.512           | 69.622, $p = 0.3661$ | 0.992 |
| Three-class             | 117       | 3710.629           | 54.021, $p = 0.2365$ | 0.903 |
| Four-class              | 118       | 3478.174           | 33.336, $p = 0.3341$ | 0.945 |
| Five-class              | 129       | 4146.228           | 36.395, $p = 0.3718$ | 0.877 |

Note. BIC stands for Bayesian Information Criteria and LMR LRT stands for Lo-Mendell-Rubin likelihood ratio test comparing the fit of a solution to a solution with one fewer cluster.

Family Clusters Identified Using Growth Mixture Modelling.

Although Gorman-Smith and colleagues included four waves of measures of the family and parenting variables, clustering methods in use at the time did not allow appropriate modelling of change over time, like growth mixture modelling does. For this paper, we fit growth mixture models of one to five classes. Table 2 presents the fit indices from each solution including only a linear term or including linear and quadratic terms. We used three indices to select the best fitting and most parsimonious model. Smaller values of the Bayesian Information Criterion (BIC; Schwartz, 1978) indicate better model fit, with a difference of 10 considered to favour one model over another. We also used the Lo-Mendell-Rubin likelihood ratio test (LMR LRT) to quantify the likelihood that the data can be described better by a model with one fewer class. On the LMR LRT a $p$ value less than .05 indicates a significantly better fit of the model with fewer classes (Muthén, 2003). The entropy statistic provides a summary measure of the accuracy with which cases can be classified into the latent classes (Muthén et al. 2002). Values close to one indicate clear classifications. In addition to these formal criteria, we considered (and suggest others do the same) class sizes and interpretability of the solution in selecting the final model. We recommend that all clusters contain at least ten cases to ensure that the cluster describes a substantial number of cases and not a few outlying cases (Loureiro et al., 2004). If clusters are identified containing fewer than ten cases or if one cluster is less than 10% of the size of the next smallest cluster, we suggest conducting further analyses to understand more about those cases, as well as removing those cases from analyses and repeating the CA to see if a similar solution is produced.
Table 3: Estimates for Intercepts and Slopes of the Five Family Variables Using Four Waves of Data

|                      | Estimate | Standard Error | t-value | Two-tailed p value |
|----------------------|----------|----------------|---------|-------------------|
| **Intercept**        |          |                |         |                   |
| Family Beliefs       | -0.458   | 0.028          | -15.274 | 0.000             |
| Family Cohesion      | -0.226   | 0.012          | -18.075 | 0.000             |
| Family Discipline    | 0.344    | 0.028          | 12.462  | 0.000             |
| Family Monitoring    | 0.018    | 0.024          | 0.750   | 0.453             |
| Family Structure     | -0.699   | 0.015          | -45.268 | 0.000             |
| **Linear Slope**     |          |                |         |                   |
| Family Beliefs       | 0.548    | 0.049          | 11.145  | 0.000             |
| Family Cohesion      | 0.461    | 0.022          | 20.655  | 0.000             |
| Family Discipline    | -0.371   | 0.037          | -10.039 | 0.000             |
| Family Monitoring    | 0.024    | 0.026          | 0.918   | 0.359             |
| Family Structure     | 0.757    | 0.032          | 23.517  | 0.000             |
| **Quadratic Slope**  |          |                |         |                   |
| Family Beliefs       | -0.133   | 0.015          | -8.613  | 0.000             |
| Family Cohesion      | -0.183   | 0.007          | -25.353 | 0.000             |
| Family Discipline    | 0.077    | 0.012          | 6.608   | 0.000             |
| Family Monitoring    | -0.017   | 0.008          | -1.977  | 0.048             |
| Family Structure     | -0.181   | 0.011          | -16.570 | 0.000             |

Given the four-class solution found by Gorman-Smith and colleagues (1998), we were surprised by fit indices suggesting a single-class quadratic solution for growth in the family variables. Consistent with the Gorman-Smith and colleagues’ 1998 findings, the BIC was lowest for a four-class linear solution. However, other indices were not consistent with a four-class solution. The numbers of youth in Classes 2 and 3 were lower than 10 (n = 3, n = 7, respectively) and the LMR LRT did not indicate that a four-class model fit better than a three-class model. In fact, the LMR LRT did not indicate that any model with more than a single class fit better than the single class model. Table 3 presents estimates and associated standard errors for the intercept, linear slope, and quadratic slope of the five family variables for the single-class solution. All variables besides family monitoring (see Figure 1d) had significant estimates for intercept, linear slope, and quadratic slopes. Figure 1a and b presents the growth pattern for Family Beliefs and Family Cohesion which are similar. Levels of Family Discipline started off high and decelerated over time whereas Family Structure started off low and initially increased (see Figure 1c and 1e).
Figure 1. Mean growth curve trajectories for the one-class solution on (a) family beliefs, (b) family cohesion, (c) family discipline, (d) family monitoring, and (e) family structure. Parameters include an intercept, linear, and quadratic effect. The vertical axis is in standard deviation units at Wave 1.

(a) Growth of Family Beliefs

(b) Growth of Family Cohesion
(c) Growth of Family Discipline

(d) Growth of Family Monitoring
There were small differences between the one and two class solutions that suggested the presence of outlying cases rather than clustered structure of the growth patterns. In the two-class solution, the two-class GMM returned a second class of only ten youth and a non-significant LMR LRT test. The 10 families in Class 2 had substantially different growth patterns. First, there was no linear or quadratic change in the slope of Family Beliefs or Discipline for youth in Class 2 but there was a significant quadratic slope for Family Monitoring. Figure 2a-e presents growth patterns for the same family characteristics as shown in Figure 1a-e. As compared to the families in Class 1, the ten families in Class 2 had consistently lower Family Beliefs over the four waves, similar initial levels of Discipline that increased at Wave 4, and similar initial levels of Structure which then declined at Wave 4.
Figure 2. Mean growth curve trajectories for the two-class solution on (a) family beliefs, (b) family cohesion, (c) family discipline, (d) family monitoring, and (e) family structure. Parameters include an intercept, linear, and quadratic effect. The vertical axis is in standard deviation units at Wave 1.

(a) Growth of Family Beliefs

(b) Growth of Family Cohesion
(c) Growth of Family Discipline

(d) Growth of Family Monitoring
Further examination led us to identify other differences in the community characteristics of these ten families, especially with regards to ratings of Neighbourhood Safety. The mean for neighbourhood safety greatly increased between Wave 1 and Wave 2 for the ten youth in Class 2 (-0.34 at Wave 1 to 0.57 at Wave 2) but remained consistent for the remainder of the sample (-0.20 in Wave 1 to -0.12 at Wave 2). Further investigation revealed that half of the families in Class 2 moved between Wave 1 and Wave 2, and cited ‘better living conditions’ as the reason for moving. The transition of these families to safer communities may have contributed to a number of factors distinguishing these cases from the rest of the sample. Moving to a safer community could be associated with a host of other attitudinal/motivational variables that may have led to, or resulted from, community transition. For example, research suggests that a safe and supportive environment fulfils youths’ basic psychological needs for belonging, autonomy, competence, and security. As these needs are met, youth tend to report increased psychological sense of community (Brodsky et al., 1999) and commitment to the community’s norms, rules, and values (Learning First Alliance, 2001). Increased commitment to a safer community could be associated with increased involvement with prosocial individuals/organizations and decreased involvement with antisocial behaviour (Aspy et al., 2004; Catalano & Hawkins, 1996). Youth might also place less importance on family given their increased attachment to community. On the other hand, we saw that family monitoring and discipline at Wave 4 increased in these families while it did not increase for the general sample, perhaps suggesting unique characteristics of the parents of these youth. Perhaps these parents were particularly concerned about relocating to a safer community for their children’s future and feel strongly about remaining involved in their children’s life even during adolescence. However, we do not have measures of these attitudinal/motivational variables and acknowledge a much deeper investigation is needed to adequately capture the story of development for these ten youth.

The Gorman-Smith et al. (1998) study found clustered structure in its examination of four waves of family and parenting data. This led us to expect, reasonably we believe, that we would find distinct groups of families with similar patterns of growth in those family characteristics. However, we found no such evidence after applying a clustering method for growth patterns that allowed us to test solutions with greater numbers of clusters against those with fewer clusters. There was, however, evidence that, although a single growth pattern fit most of the cases in the sample, a few cases had anomalous growth patterns. This points to two valuable functions of clustering methods – first to determine when a single growth pattern fits an entire
sample, and second, to detect multivariate outliers; observations that do not appear to belong to the same population as the rest of the sample (Henry et al., 2005; Dymnicki & Henry, in press).

CA’s Ability to Model Processes whose Outcomes Result from the Combined Effect of Several Factors

By focusing on the outcomes associated with the combination of several factors, more and less familiar clustering methods can overcome some of the limitations of regression-based approaches discussed at the beginning and are especially useful for analyzing survey data. We believe benefits of this approach can be classified into four categories that we describe briefly below and also describe three additional benefits of less commonly used clustering methods.

First, modelling time and longitudinal development for cases may result in more parsimonious cluster solutions. While regression can model time, it does not take into account differences in development for different cases within the sample. We found when incorporating time into our models that a one-class solution was sufficient to describe most of the cases in the sample (i.e., multiple family characteristics had significant linear and quadratic change over time but all cases had similar patterns of growth), providing insight on the developmental dimension of the cases. Clustering approaches, particularly those that employ mixture models, allow multiple ways of modelling development (e.g., with linear and quadratic terms, orthogonal polynomials free growth parameters, and survival analysis). These approaches can shed light on what we consider to be ‘normative development’ and cases that present interesting variations from this pattern. For example, Mitchell and Plunkett (2000) modelled the lifetime substance use of American Indian youth and found a four class solution: (a) abstainers, (b) predominantly alcohol users, (c), predominantly alcohol and marijuana users, and (d) plural substance users. They then demonstrated the usefulness of the solution by differentiating between levels of social cognitive variables such as attitudes towards alcohol, community mindedness, and peer values with this four-class typology.

It is important to point out that incorporating time into clustering does not require multiple waves of measurement. Time may be incorporated by including the occurrence of specific behaviours or situations surrounding the initiation, maintenance, or cessation of risk behaviours. For example, classifying patterns of situations occurring before or after relapse might assist in understanding cases that relapse. Shiffman (1986) clustered relapsed situations for smokers into four categories: social, work, upset, and relaxation, using k-means clustering. Time could also be incorporated in clustering methods by looking at changes in neighbourhood characteristics over different periods. For example, a ‘geodemographic’ database using CA assesses characteristics of neighbourhoods over ten-year periods based on demographic characteristics, urbanization, lifestyle, and consumer behaviour patterns (Heitgerd & Lee, 2003).

Second, all clustering methods have the ability to situate the case in multivariate space (Uprichard, 2009). CA can identify patterns of contexts revealing types of families, schools, and communities and their possible influence on the case. One study identified different types of urban communities in Chicago and found that these community types differentially influenced delinquency patterns (Gorman-Smith et al., 2000a). Defining inner-city communities as those with greater than 40% of the population below the poverty level, these investigators found that the social processes in communities differentiated levels of youth crime. Youth living in inner-city communities without functioning social processes (i.e., where youth had a low sense of neighbourhood belonging, support, and involvement as well as high safety concerns) were significantly more likely to engage in serious chronic, escalating, or chronic minor delinquency than youth living in inner-city communities with functioning social processes or youth living in other urban poor communities. It is important to acknowledge that CA cannot be used to infer causality in these dynamic relationships. Thus, we cannot rule out that delinquent youth and their families are more likely to gravitate to communities without functioning social processes versus the characteristics of communities without functioning social processes prompting youth to engage in delinquent behaviour.

Third, a common criticism of CA involves the possibility that CA creates artificial clusters that do not exist in the data. As seen in our exemplar study, newer, less familiar methods provide ways to determine when multiple clusters do not fit the data well (e.g., the LMR LRT comparing the fit of a two-class and one-class
solutions). There are also a number of cluster validity measurement techniques that have been developed to provide internal, external, and relative criteria for determining the correct cluster solution (Halkidi et al., 2001). The first type of validity implies that researchers evaluate the results of a pre-specified clustering algorithm that is imposed on a data set and reflects their intuition about the clustering structure of the data set. The second type of validity usually involves calculating a Cophenetic Correlation Coefficient to test the presence of hierarchic structure in one's data (Rohlf & Fisher, 1968) or determining the degree of agreement between a given clustering scheme C, consisting of nc (i.e., a given number of clusters), and the proximity matrix P for non-hierarchical data. The third type of validity involves evaluating the clustering structure by comparing it to other clustering schemes, resulting from the same algorithm but with different parameter values. While these methods have reduced concerns about creating artificial clusters, they have not eliminated the possibility. On the other hand, it can be equally problematic to assume that samples are not clustered as it is to retain an inappropriate number of clusters. Researchers should, where possible, test the assumption that a sample is homogeneous and lacks sub-samples.

Fourth, case-based research aims to describe cases with a high degree of faithfulness to the characteristics of the case. Like solutions produced by other multivariate methods, cluster solutions are influenced by outlying observations (Everitt, 2003). This is particularly problematic when parsimonious solutions with few clusters are desired, or when samples are small. Several clustering methods can be used to detect outlying observations. In hierarchical solutions, look for one or more observations connected to the remainder of the distribution by large joining distances. Non-hierarchical solutions will also produce small clusters that may suggest outlying observations upon examination (cf., Gorman-Smith et al., 1998). Clusters smaller than 10% of the size of the next smallest cluster should be explored as possible outliers.

In the current study, exploring the family characteristics of the ten youth in Class 2 suggested some interesting differences in terms of the growth patterns of these ten individuals. Specifically, a considerable number of these youth moved to a safer community after Wave 1 which might be associated with different family beliefs (e.g., youth might place less emphasis on the importance of family as they become more committed to their safer neighbourhood) or different characteristics of the youth’s parents (e.g., parents of these youth might have been more concerned about their children’s safety and future than other parents in the sample and thus relocated to a safer area as well as continued to monitor and discipline their children throughout adolescence). Finding differences in concerns for safety and relocation in the 10 outlying cases in this study leads us to recommend that investigators in case-based research use CA to identify elusive outliers and unusual cases that may provide richer information about the cases in the data.

In addition to these four types of benefits, we believe less familiar cluster approaches can overcome three limitations frequently associated with clustering methods: lack of fit indices to determine best-fitting cluster solution, representing non-spherical clusters, and indeterminate cluster membership. Some less commonly used clustering approaches provide fit indices such as those in Table 2, making identification of the appropriate solution potentially more data-driven. Furthermore, while this did not occur in our analyses of family variables, in poorly fitting models, the analyses will not converge (i.e., the programme will produce error messages and will not provide fit statistics). In addition, even when models do converge, we advise researchers to continue trying to find better fitting models if entropy values do not exceed .70. Being guided by fit indices reduces some of the subjectivity in determining the best fitting solution.

Second, clusters in multivariate space may not be consistent in shape. This is particularly to be expected in growth mixture models, where cluster shapes are determined by the particular parameters used to describe growth. These parameters can vary among clusters, so that the most parsimonious and best-fitting model may involve linear growth for one cluster, no growth for another, and quadratic or cubic growth for a third. When time is not included in clustering, the relation between two variables may be strong and positive within one cluster, but strong and negative in another, resulting in clusters that resemble bread sticks tilted in different directions. The weakness of CA when clusters are not spherical in shape is shared by more and less familiar clustering methods (Dymnicki & Henry, in press). However, this weakness can be addressed by using a model-based approach to determine the optimal combination of number, shape, and variance of clusters. Model-based clustering (Fraley & Raftery, 1998) tests multiple possible models of cluster shape and variance,
comparing up to 8 different models using the BIC as a criterion of model fit. This methodology also opens up the possibility of more accurately reflecting similarities or differences that exist among cases.

Third, the problem of indeterminate cluster membership in data that have clustered structure can be addressed to some extent by making use of the probabilities of cluster assignment that are available in latent class clustering methods (e.g., Vermunt & Magidson, 2002; Muthén et al., 2002). Using these probabilities the researcher can evaluate the degree of certainty of cluster assignment, create overlapping or ‘fuzzy’ clusters (where an individual can be assigned to more than one cluster if there is uncertainty about assignment to a single cluster), or regard those with relatively high probabilities of membership in more than one cluster as a separate group. Cooper and Glaesser (2011) compared the results of analyses done with fuzzy CA and fuzzy Qualitative Comparative Analysis, another promising approach for case-based research, and found converging findings. They point out however, that other data sets are likely to generate greater differences when the two approaches are compared.

Conclusion

In this paper we have argued for the use of clustering methods to understand features of a case over regression analyses that assume each independent variable acts ‘independently.’ As we highlight with our study of family social processes, different combinations of family relationship characteristics and parenting practices work together to influence children’s delinquent behaviour. Clustering approaches have the ability to place the case in time and in space by modelling the developmental nature of the case, the role of specific events, and the influence of different contexts. Though it is still possible for CA to find clusters in random data, tests comparing the fit of different solutions as well as different cluster validity techniques can help determine if there is any inherent clustering in the data. Multivariate outliers and unusual cases that may illuminate different patterns of social processes can also be detected.

Furthermore, in some less familiar CA techniques, fit indices can be used to determine when an adequate number of clusters have been identified and non-spherical clusters can be modelled more accurately. Indeterminate cluster membership can be overcome by the use of overlapping or fuzzy clustering methods. We temper our recommendations for this approach by acknowledging some limitations such as producing multiple or conflicting solutions that can leave researchers unsure when choosing the final solution (Blashfield & Aldenderfer, 1988). Incorporating theory in forming a basis for the clustered structure of the cases as well as using fit indices, multiple visual representations, and cluster validity techniques should guide researchers in determining useful and interpretable final solutions. The clustering methods described here can advance our understanding of cases by shedding light on grouping in samples that vary in underlying social processes.

Notes

1 This research was supported, in part, by the National Center for Injury Prevention and Control, Centers for Disease Control and Prevention and the National Institute on Drug Abuse.
2 If they are in the same model, \( t \) and \( t^2 \) may be collinear, making it difficult to interpret the growth parameters. If it is desired to interpret linear growth and acceleration independently, time can be centered at the mean or coefficients of orthogonal polynomials can be used in the place of actual measures of time or wave numbers. Both will result in uncorrelated linear and quadratic growth measures.

References

Abbott, A. (2001) Time Matters: On Theory and Change. Chicago: University of Chicago Press.

Achenbach, T. M. (1991) Integrative Guide for the 1991 CBCLS4-18, YSR, and TRF profiles. Burlington: University of Vermont, Department of Psychiatry.
Aldenderfer, M. S., and Blashfield, R. K. (1984) *Cluster Analysis: A Sage University Paper*. Beverly Hills, CA: Sage Publications.

Aspy, C. B., Oman, R. F., Vesely, S. K., McLeroy, K., Rodine, S., and Marshall, L. (2004) Adolescent Violence: The Protective Effects of Youth Assets. *Journal of Counseling and Development*, 82, 269-277.

Blashfield, R. K., and Aldenderfer, M. S. (1988) The Methods and Problems of Cluster Analysis. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of Multivariate Experimental Psychology*, 2nd ed., pp. 447–473. New York: Plenum Press.

Brodsky, A., O’Campo, P., and Aronson, R. (1999) PSOC in Community Context: Multi-Level Correlates of a Measure of Sense of Community in Low-Income, Urban Neighborhoods. *Journal of Community Psychology*, 27, 659-679.

Byrne, D. and Regin, C.C. (Eds.) (2009) *The Sage Handbook of Case-Based Methods*. London: Sage Publications.

Catalano, R., and Hawkins, J. D. (1996) *The Social Development Model: A Theory of Antisocial Behavior*. Cambridge: Cambridge University Press.

Cooper, B. (2005) Applying Regin’s Crisp and Fuzzy Set QCA to Large Datasets: Social Class and Educational Achievement in the NCDS. *Sociological Research Online*, 10(2).

Cooper, B., and Glaesser, J. (2011). Using Case-Based Approaches to Analyse Large Datasets: A Comparison of Regin’s fsQCA and Fuzzy Cluster Analysis. *International Journal of Social Research Methodology*, 14, 31–48.

de Oliveira, J.V., and Pedrycz, W. (Eds.). (2007). *Advances in Fuzzy Clustering and Its Applications*. New York: Wiley.

Dymnicki, A. B. and Henry, D. B. (in press). ‘Grappling with Grouping: Clustering and Its Applications in Community Research’. In L. A. Jason and D. S. Glenwick (Eds.), *Innovative Methodological Approaches to Community-Based Research: Theory and Application*. Washington, DC: American Psychological Association.

Everitt, B.S. (1993). *Cluster Analysis*. New York: Wiley.

Farrington, D. P. (1994). ‘Childhood, Adolescent, and Adult Features of Violent Males.’ In L. R. Huesmann (Ed.), *Aggressive Behavior: Current Perspectives*, pp. 215–240. New York: Plenum Press.

Fraley. C. and Raftery, A. (1998). ‘How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis.’ *The Computer Journal*, 41, 578–588.

Gorman-Smith, D., Tolan, P. H., and Henry, D. (1999) ‘The Relation of Community and Family to Risk Among Urban Poor Adolescents.’ In Cohen, P., Robins, L., and Slomkowski, C. (Eds.). *Where and When: Influence of Historical Time and Place on Aspects of Psychopathology*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gorman-Smith, D., Tolan, P. H., and Henry, D. B. (2000a) ‘A Developmental-Ecological Model of the Relation of Family Functioning to Patterns of Delinquency.’ *Journal of Quantitative Criminology*, 16, 169–198.
Gorman-Smith, D., Tolan, P. H., Henry, D. B., and Florsheim, P. (2000b) ‘Patterns of Family Functioning and Adolescent Outcomes among Urban African American and Mexican American Families.’ *Journal of Family Psychology*, 14, 436-457.

Gorman-Smith, D., Tolan, P. H., Loeber, R., and Henry, D. B. (1998) ‘Relation of Family Problems to Patterns of Delinquent Involvement among Urban Youth.’ *Journal of Abnormal Child Psychology*, 26, 319–333.

Gorman-Smith, D., Tolan, P. H., Zelli, A., and Huesmann, L. R. (1996) ‘The Relation of Family Functioning to Violence among Inner-City Minority Youths.’ *Journal Family Psychology*, 10, 115-129.

Halkidi, M., Batistakis, Y., Vazirgiannis, M. (2001) ‘On Clustering Validation Techniques.’ *Journal of Intelligent Information Systems*, 17, 107–145.

Heitgerd, J. L., and Lee, C. V. (2003) ‘A New Look at Neighborhoods Near National Priorities List Sites. *Social Science and Medicine*, 57, 1117-1126.

Henry, D. B., Tolan, P. H., and Gorman-Smith, D. (2005) ‘Cluster Analysis in Family Psychology Research.’ *Journal of Family Psychology*, 19, 121–132.

Jones, B. L., Nagin, D. S., and Roeder, K. (2001) ‘A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories.’ *Sociological Methods and Research*, 29, 374-393.

Lanza, S. T., Lemmon, D. R., Schafer, J. L. and Collins, L. M. (2008) *PROC LCA and PROC LTA User's Guide*. State College, PA: The Methodology Center, Penn State University.

Learning First Alliance. (2001) *Every Child Learning: Safe and Supportive Schools*. Washington, DC: Association for Supervision and Curriculum Development.

LeBlanc, M., and Kaspy, N. (1998) ‘Trajectories of Delinquency and Problem Behavior: Comparison of Social and Personal Control Characteristics of Adjudicated Boys on Synchronous and Non-Synchronous Paths.’ *Journal of Quantitative Criminology*, 14, 181–214.

Loeber, R., and Dishion, T. (1983) ‘Early Predictors of Male Delinquency: A Review.’ *Psychological Bulletin*, 94, 68–99.

Loeber, R., Stouthamer-Loeber, M., Van Kammen, W. B., and Farrington, D. P. (1991) ‘Initiation, Escalation and Desistance in Juvenile Offending and their Correlates.’ *Journal of Criminal Law, and Criminology*, 82, 36–82.

Loeber, R., Wung, P., Keenan, K., Giroux, B., Stouthamer-Loeber, M., Van Kammen, W. B., and Maughan, B. (1993) ‘Developmental Pathways in Disruptive Child Behavior.’ *Development and Psychopathology*, 5, 101–133.

Loureiro, A., Torgo, L., and Soares, C. (2004) *Outlier Detection Using Clustering Methods: A Data Cleaning Application*. In proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector. Bonn, Germany.

Macrae, C. N., and Bodenhausen, G. V. (2000) ‘Social Cognition: Thinking Categorically about Others.’ *Annual Review of Psychology*, 51, 93–120.

McCord, J. (1991) ‘Family Relationships, Juvenile Delinquency, and Adult Criminality.’ *Criminology*, 29, 297–417.
Mitchell, C. M., and Plunkett, M. (2000) ‘The Latent Structure of Substance Use Among American Indian Adolescents: An Example Using Categorical Variables.’ *American Journal of Community Psychology, 28*, 105-125.

Muthén, B. (2003) ‘Statistical and Substantive Checking in Growth Mixture Modelling: Comment on Bauer and Curran.’ *Psychological Methods, 8*, 369-377.

Muthén, B., Brown, C. H., Masyn, K., Jo, B., Khoo, S. T., Yang, C. C., Wang, C. P., Kellam, S. G., Carlin, J. B., and Liao, J. (2002) ‘General Growth Mixture Modelling for Randomized Preventive Interventions.’ *Biostatistics, 3*, 459–475.

Muthén, B., and Khoo, S. T. (1998) ‘Longitudinal Studies of Achievement Growth Using Latent Variable Modelling.’ *Learning and Individual Differences, 10*, 73–101.

Muthén, B., and Muthén, L. K. (2000) ‘Integrating Person-Centered and Variable-Centered Analyses: Growth Mixture Modelling with Latent Trajectory Classes.’ *Alcoholism: Clinical and Experimental Research, 24*, 882-891.

Nichols, M. P., & Schwartz, R. C. (199 I). Family therapy: Concepts and methods. Boston: Allyn & Bacon.

Patterson, G. R., and Stouthamer-Loeber, M. (1984) ‘The Correlation of Family Management Practices and Delinquency.’ *Childhood Development, 55*, 1299–1307.

Patterson, G. R., Reid, J. B., and Dishion, T. J. (1992) *Antisocial Boys: A Social Interactional Approach, Vol. 4*. Eugene, OR: Castalia Publishing Company.

Ragin, C. C. (2006) ‘The Limitations of Net Effects Thinking.’ In B. Rihoux and H. Grimm (Eds.) *Innovative Comparative Methods for Political Analysis*, pp. 13-41. New York: Springer.

Rohlf, F. J., and Fisher. D.L. (1968) ‘Test for hierarchical structure in random data sets’, *Systematic Zool.*, 17:407-412.

Schwartz, G. (1978) ‘Estimating the Dimension of a Model.’ *The Annals of Statistics, 5*, 461-464.

Seber, G. (1984) *Multivariate Observations*. New York: John Wiley and Sons.

Shiffman, S. (1986) ‘A Cluster-Analytic Classification of Smoking Relapse Episodes.’ *Addictive Behaviors, 11*, 295-307.

Tolan, P. H., Gorman-Smith, D., Huesmann, L. R., and Zelli, A. (1997) ‘Assessing Family Processes to Explain Risk for Antisocial Behaviour and Depression among Urban Youth.’ *Psychological Assessment, 9*, 212-223.

Tryon, R. C. (1939) *Cluster Analysis*. Oxford: Edwards Bros.

Uprichard, E. (2009) ‘Introducing Cluster Analysis: What Can It Teach Us About the Case?’ In D. Bryne and C. C. Ragin (Eds). *The Sage Handbook of Case Based Methods*, pp. 132-147. London: Sage Publications.

Vermunt, J. K., and Magidson, J. (2002) ‘Latent Class Cluster Analysis,’ In J. A. Hagenaars and A. L. McCutcheon (Eds.). *Applied Latent Class Analysis*, pp. 89-106. Cambridge: Cambridge University Press.
Wilkinson, L., Blank, G., and Gruber, C. (1996) *Desktop Data Analysis with SYSTAT*. Upper Saddle River, NJ: Prentice Hall.

**Biographies:**

Allison B. Dymnicki is a Postdoctoral Fellow at the Institute of Health Research and Policy at the University of Illinois at Chicago. Her research interests include applications of cluster analysis and longitudinal modeling to understand how contexts influence individual behavior. Applications include prevention and promotion, normative development and risk-taking behaviors of children, adolescents, and their families.

David B. Henry is a Professor of Health Policy and Administration at the Institute of Health Research & Policy, University of Illinois at Chicago. He studies contextual processes that influence individual behavior, child and adolescent development and psychopathology, and prevention.