

LI Analysis Training Series

Cluster Analysis

(Last Revised: 3/21/2006)

Rodney Funk, Melissa Ives & Michael Dennis

Chestnut Health Systems

Bloomington IL 61701

309-827-6026

www.chestnut.org

Acknowledgement: *This document was developed under grant No. T111320 from the Center for Substance Abuse Treatment (CSAT) of the Substance Abuse and Mental Health Services Administration (SAMHSA). The opinions expressed here are solely those of the authors and do not reflect official positions of the U.S. Government.*

Purpose: Cluster analysis is an exploratory technique used to develop a typology for describing the major subgroups in a population that is heterogeneous on more than one dimension of variables, settings, or over time. The aim of this memo is to explain the method the authors typically use to obtain a cluster analysis and its subsequent validation using SPSS (Version 7.5 to 14.0).

Background: We have conducted cluster analyses to help develop typologies for several clinically heterogeneous behaviors including: HIV risk behaviors (Dennis, Wechsberg et al., 2004; Wechsberg et al., 1998), psychopathology (Rief et al., 2001), post treatment relapse patterns (Godley et al., 2004), changes in crime trajectories (Chan et al., 2006) and changes in psychiatric symptoms (Hawke et al., 2006). Using Rapkin & Luke's (1993) list of key decisions for cluster analysis, below is a summary of our general approach.

- **Identifying cases for the analysis:** We have generally done cluster analyses on clinical populations and have examined issues where we are interested in identifying small high-risk groups, thus we typically include everyone (vs. excluding outliers).
- **Selecting, reducing and scaling variables:** The interpretability of a cluster-based typology grows exponentially complex as different types of variables and dimensions are added- thus we tend to focus on a single substantive area of behavior. We also include any time varying covariates that are likely to have a strong influence on the trajectories over time (e.g., days in a controlled environment, amount of treatment received). As in factor analysis, we standardize all variables into z-scores by variable (within time as well if for a trajectory).
- **Deriving proximity measures among cases:** We use the Squared Euclidean Distance between cases/cluster centers. This places greater weights on cases that are further apart and serves to isolate high-risk groups faster.
- **Choosing a clustering algorithm:** We use Ward's (1963) minimum distance method. This is a hierarchical method that groups cases to maximize between group differences and minimize within group differences (i.e., optimizes an F-

- Statistic). It keeps grouping the most similar pair of cases/clusters until there is just one cluster. This can graphically be depicted as a dendrogram tree.
- Determining the number of clusters: We generally save and evaluate the solutions with 2 to 12 clusters and select the final solution based on multiple criteria including: (a) cluster eigenvalue greater than one; (b) smallest group of at least 30 people (or 0.5% of total if larger); (c) a solution that explained 70% or more of each of the measures being clustered; (d) a solution that explained 99% or more of the variance in the joint distribution (based on 1-Wilks' Lambda) of the measures; and (e) the interpretability of the solution when profiled (i.e., the source variables plotted by cluster subgroup).
 - Choosing an appropriate statistical package: We have used SPSS (Version 7.5 to 14.0).
 - Interpreting cluster profiles: We generally identify the 2-3 solutions that are statistically acceptable, profile them based on the variables used for cluster and demographics, and review these materials with a team of researchers and clinicians to identify which group is the most interpretable.
 - Determining cluster stability: We generally test the cluster solution on split half or independent samples to make sure that they come out approximately the same. If there has been extensive replacement of missing data we also compare solutions based on list wise deletion vs. imputing missing data to see if the solutions there is a point at which their solutions diverge (which might be due to missing data imputation and hence less reliable).
 - Determining cluster validity: As part of the selection process we check whether the clusters predict the variables from which they were derived. We will also use the Fisher discriminant function coefficients to classify a split half or additional sample of clients to see it predicts their scores as well.
 - Presenting cluster results: We generally present the overall pattern of outcomes, a pie chart showing the size of the subgroup groups, the differences at intake, the differences over time (profiling one group at time), then differences in their correlates (e.g., HIV status, diagnosis etc by cluster) and/or composition (e.g., % in each cluster by site, gender, or other subgroup).

Below we have walked through the steps of our approach and provided a detailed example of syntax and results using a cluster analysis of criminal activity trajectories (Chan et al., 2006).

Data Requirements: The data may be quantitative, dichotomous or count data. It is also important that the data be of the same scale. If you have 0/1 variables and a variable that goes from 0 to 90, a value of 90 will be weighted as 90 times more important than a 1 in the 0/1 variable. So if the variables you are clustering on are of different scales, we recommend standardizing them, such as transforming them to z-scores. Cluster analysis also expects there to be data for every variable used in the analysis. If you are missing just one variable for a record, no clusters will be calculated for that record. Since listwise deletion is considered the most biased possible approach (Little & Rubin, 1987; Figueredo et al., 2000), we recommend using some method for replacing missing data before performing the cluster analysis (e.g., replacing within individual using a mean

scale value or regression, hot deck imputation, median/mean replacement). When looking at trajectories over time, the data should be set up in a horizontal data file with one record per case and one variable per measurement (i.e. intake, 3, 6, 9, 12 and 30-month data). Because you are looking for trends, rather than taking the mean of a variable over all times it would be better to interpolate between the two most recent times or estimate with regression from other time points.

Procedure: In this example, we will be looking at scales measuring different types of crime, days of illegal activity, days of illegal activity for money and days in a controlled environment taken from the Global Appraisal of Individual Need (GAIN) measured at intake and 3, 6, 9, 12 and 30-months post intake. Since the crime scales are symptom counts going ranging from 0 to 7 and the days variables range from 0 to 90, all the variables were transformed to z-scores to put them all on the same scale prior to clustering. Below is SPSS syntax used for this cluster analysis:

```
CLUSTER  Zpcs_0 Zrpcs_3 Zrpcs_6 Zrpcs_9 Zrpcs_12 Zpcs_30 Zics_0 Zrics_3
Zrics_6
  Zrics_9 Zrics_12 Zics_30 Zmdcs_0 Zrdcs_3 Zrdcs_6 Zrdcs_9 Zrdcs_12
Zrdcs_30 Zl3d_0
  Zr13d_3 Zr13d_6 Zr13d_9 Zr13d_12 Zl3d_30 Zl3e_0 Zr13e_3 Zr13e_6
Zr13e_9 Zr13e_12
  Zl3e_30 Zmaxce_0 Zrmaxce_3 Zrmaxce_6 Zrmaxce_9 Zrmaxce_12 Zmaxce_30
/METHOD WARD
/MEASURE= SEUCLID
/PRINT SCHEDULE
/PLOTS Dendrogram
/SAVE CLUSTER(2,12) .
```

PCS refer to the Property Crime Scale, ICS to the Interpersonal (violent) Crime Scale, DCS to the Drug Crime Scale, L3d to the days of illegal activity, L3e to the days of illegal activity for money and MAXCE refers to days in a controlled environment. The “Z” means the variable was transformed into a z-score with a mean of 0 and standard deviation of 1.0. The “r” on waves 3 to 12 means that the small amount of missing data per wave was imputed with “regression (missing data values were not extrapolated to intake or 30 months). The method used is Ward’s minimum distance with the measure of Squared Euclidean Distance. The dendrogram plot is a horizontal plot showing the linkage of the cases into clusters with the distance being on the horizontal axis. The final line asks for the cluster solutions for 2 through 12 groups to be saved.

The next step is deciding on the correct number of clusters. We use the following stopping rules in deciding on the proper number of clusters: a) initial groups that comprised 5% or more of the total; b) a solution that explained 70% or more of each of the measures being clustered; c) a solution that explained 99% or more of the variance in the joint distribution (based on 1-Wilks’ Lambda) of the measures; and d) the interpretability of the solution when profiled (i.e., the source variables plotted by cluster subgroup).

First, we run frequencies on the cluster solutions to check on when small groups break out. These cluster solution variables are saved at the end of the data file with the names

CLU2_1 (2 groups), CLU3_1 (3 groups), and on up to CLU12_1. (If another cluster command is run, additional variables are created with an increase in the last number (CLU2_2, CLU_3, etc.)) The next step is to run GLM multivariate analysis on the variables in the cluster analysis by the cluster groups; starting with the 2 group solution on up to a solution with a group size is smaller than 5% of the total sample. When running the GLM's you want to be sure you print out means and the Eta-Squares. You can get this in the GLM Multivariate dialogue box by clicking on Options and under Display, click on descriptive and estimates of effect size. Here is an example of how it would look in syntax:

```
GLM
  Zpci_0 Zrpci_3 Zrpci_6 Zrpci_9 Zrpci_12 Zpci_30 Zici_0 Zrici_3 Zrici_6
  Zrici_9 Zrici_12 ZSco01 Zmdci_0 Zrdci_3 Zrdci_6 Zrdci_9 Zrdci_12
ZSco02 Zl3v_0
  Zr13d_3 Zr13d_6 Zr13d_9 Zr13d_12 Zl3d_30 Zl3w_0 Zr13e_3 Zr13e_6
Zr13e_9 Zr13e_
  12
  Zl3e_30 Zmaxce_0 Zrmaxce_3 Zrmaxce_6 Zrmaxce_9 Zrmaxce_12 Zmaxce_30 BY
CLU2_2
/METHOD = SSTYPE(3)
/INTERCEPT = INCLUDE
/PRINT = DESCRIPTIVE ETASQ OPOWER HOMOGENEITY
/CRITERIA = ALPHA(.05)
/DESIGN = CLU2_2 .
```

In the /print subcommand, DESCRIPTIVE will give you the mean and standard deviation of the cluster variables by the cluster groups, ETASQ will give the Eta-squares for the variance explained in the cluster variables by the cluster groups. From the multivariate tests table in the output we get Wilks' Lambda and Roy's Largest Root. See example below:

Multivariate Tests(c)

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.421	15.387(b)	36.000	763.000	.000	.421
	Wilks' Lambda	.579	15.387(b)	36.000	763.000	.000	.421
	Hotelling's Trace	.726	15.387(b)	36.000	763.000	.000	.421
	Roy's Largest Root	.726	15.387(b)	36.000	763.000	.000	.421
CLU2_2	Pillai's Trace	.684	45.842(b)	36.000	763.000	.000	.684
	Wilks' Lambda	.316	45.842(b)	36.000	763.000	.000	.684
	Hotelling's Trace	2.163	45.842(b)	36.000	763.000	.000	.684
	Roy's Largest Root	2.163	45.842(b)	36.000	763.000	.000	.684

a Computed using alpha = .05

b Exact statistic

c Design: Intercept+CLU2_2

Another important piece of output is the tests of between-subjects effects from which we get the Eta-squares. A partial table is given below for how well the 3-group solution predicts each of the variables it was derived from. Rapkin & Luke (1993) refer to this as “reverse validation”.

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	Zpci_0 Zscore: pci_0: Property Crime Index	164.587(b)	2	82.293	102.048	.000	.204
	Zrpci_3 Zscore(rpci_3)	150.815(c)	2	75.408	88.900	.000	.182
	Zrpci_6 Zscore(rpci_6)	222.848(d)	2	111.424	148.571	.000	.272
	Zrpci_9 Zscore(rpci_9)	104.004(e)	2	52.002	62.926	.000	.136
	Zrpci_12 Zscore(rpci_12)	166.819(f)	2	83.409	119.377	.000	.231
	Zpci_30 Zscore: FU Property Crime Index	79.882(g)	2	39.941	43.703	.000	.099

For each cluster solution that a GLM multivariate analysis was performed, we paste the “Dependent Variable” and “Partial Eta-squared” columns from the above table into Excel. The Eta-squares are formatted as percents (Note that the highlighted percents below come from the Eta-Squared table above, and the highlighted 1-Wilks’ Lambda and Roy’s Largest Root come from the earlier Multivariate Test table.) We also calculate 1-Wilks’ Lambda, to estimate the variance of the multivariate distribution, and take Roy’s largest root, to see if on cluster group is very different from the others from the first table above. Together we make a table in Excel like the example below to help evaluate the cluster solutions.

Tests of Between-Subjects Effects

	# of Clusters (n=800)			
	2	3	4	5
Zpci_0 Zscore: pci_0: Property Crime Index	9%	20%	22%	23%
Zpci_3 Zscore: pci_3: Property Crime Index	18%	18%	21%	22%
Zpci_6 Zscore: pci_6: Property Crime Index	26%	27%	28%	29%
Zpci_9 Zscore: pci_9: Property Crime Index	11%	14%	20%	20%
Zpci_12 Zscore: pci_12: Property Crime Index	23%	23%	26%	38%
Zpci_30 Zscore: FU Property Crime Index	7%	10%	10%	14%
Zici_0 Zscore: ici_0: Interpersonal Crime Index	8%	20%	24%	24%
Zici_3 Zscore: ici_3: Interpersonal Crime Index	17%	18%	26%	27%
Zici_6 Zscore: ici_6: Interpersonal Crime Index	19%	20%	27%	27%
Zici_9 Zscore: ici_9: Interpersonal Crime Index	12%	14%	27%	27%
Zici_12 Zscore: ici_12: Interpersonal Crime Index	17%	17%	24%	31%
Zici_30 Zscore: FU Interpersonal Crime Index	6%	11%	12%	12%
...(one row per variable)				
Zmaxce_30 Zscore: Maximum days in a controlled environment	2%	3%	4%	5%
1 - Wilks’ Lambda	0.684	0.833	0.936	0.972
Roy’s Largest Root	2.16	2.25	3.74	3.94
smallest group n	169	169	23	23

The bold percents indicate where there is a jump in the increase of variance explained in a certain variable due to the break out of a new cluster group. We need to look at one last thing to evaluate the cluster solutions: the interpretability of the solutions. We do this by graphing the means of the variables in the cluster analysis by the groups from the cluster solution. First is to go to the Descriptive Statistics table in the SPSS output. Double click on the table, then go to the Pivot menu and select Pivoting Trays so the screen looks like below.

The screenshot shows the SPSS Output Viewer window with the 'Descriptive Statistics' table selected. The table displays mean and standard deviation for various variables across two clusters (CLU2_2 Ward). A 'Pivoting Trays' dialog box is open, showing a diagram with 'Layers', 'Columns', and 'Rows' sections. The 'Rows' section is highlighted, indicating that the cluster groups are being moved to the columns of the pivot table.

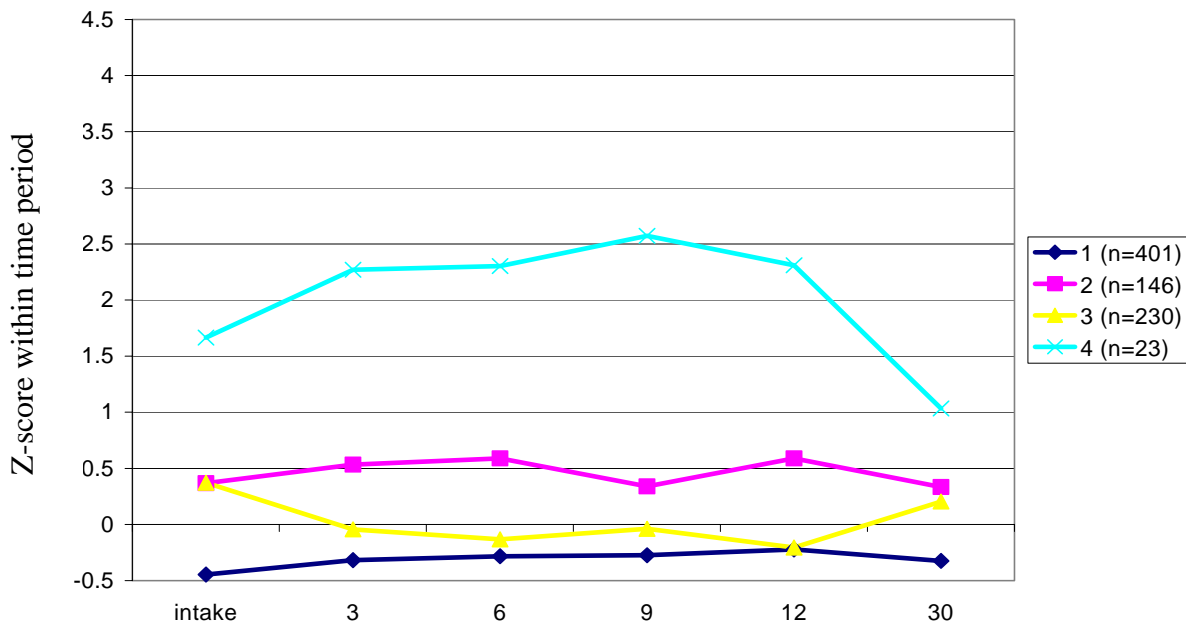
	CLU2_2 Ward	Mean	Std. Deviation	N
Zpci_0 Zscore: pci_0: Property Crime Index	1	-.1538473	.92945859	631
	2	.6031223	1.05409214	169
	Total	.0060625	1.00518050	800
Zrpci_3 Zscore(rpci_3)	1	-.1850672	.77273488	631
	2	.8619159	1.34640316	169
	Total	.0361080	1.01728140	800
Zrpci_6 Zscore(rpci_6)	1	-.2553193	.59885068	631
	2	1.011593	1.50532655	169
	Total	.0123159	1.01341115	800
Zrpci_9 Zscore(rpci_9)	1	-.1800771	.68899567	631
	2	.6303033	1.49608137	169
	Total	-.0088842	.97698006	800
Zrpci_12 Zscore(rpci_12)	1	-.2409687	.46262921	631
	2	.8724596	1.58789713	169
	Total	-.0057569	.95170475	800
Zpci_30 Zscore: FU Property Crime Index	1	-.1404322	.73862497	631
	2	.5203066	1.55529368	169
	Total	-.0008511	1.00579284	800
Zici_0 Zscore: ici_0: Interpersonal Crime Index	1	-.1482354	.89291970	631
	2	.5449495	1.21757151	169
	Total	-.0018001	1.01021688	800
Zrici_3 Zscore(rici_3)	1	-.2172283	.60251743	631
	2	.7708619	1.57105043	169

In the pivoting tray, click on the inside symbol of the rows at the bottom of the Pivoting Tray dialogue box (holding the cursor over the symbol will show it is CLU2_2 Ward Method) and drag the cluster groups to the columns over to the right. Now click and drag the Statistics symbol from the columns to the rows so it is on the very outside (furthest left). Your table should now look like this:

		1	2	Total
Mean	Zpoi_0 Zscore: poi_0: Property Crime Index	-.1538473	.6031223	.0060625
	Zpoi_3 Zscore(rpoi_3)	-.1850672	.8619159	.0361080
	Zpoi_6 Zscore(rpoi_6)	-.2553193	1.011593	.0123159
	Zpoi_9 Zscore(rpoi_9)	-.1800771	.6303033	-.0088842
	Zpoi_12 Zscore(rpoi_12)	-.2409687	.8724596	-.0057569
	Zpoi_30 Zscore: FU Property Crime Index	-.1404322	.5203066	-.0008511
	Zici_0 Zscore: ici_0: Interpersonal Crime Index	-.1482354	.5449495	-.0018001
	Zici_3 Zscore(rici_3)	-.2172283	.7708619	-.0084943
	Zici_6 Zscore(rici_6)	-.2281903	.8220565	-.0063257
	Zici_9 Zscore(rici_9)	-.1878594	.6431200	-.0123150
	Zici_12 Zscore(rici_12)	-.2159879	.8231302	.0035258
	Zsco01 Zscore(ici_30) FU Interpersonal Crime Index	-.1315948	.4284498	-.0132853
	Zmdoi_0 Zscore: mdoi_0: Drug Crime Index-Imputed by regression for LA	-.1479710	.5406065	-.0025090
	Zrdci_3 Zscore(rdci_3)	-.2313495	.9151387	.0108462
	Zrdci_6 Zscore(rdci_6)	-.2677373	1.021866	.0046914
	Zrdci_9 Zscore(rdci_9)	-.2380568	.8292088	-.0125970
	Zrdci_12 Zscore(rdci_12)	-.2574862	.9236404	-.0079732
	Zsco02 Zscore(dci_30) FU Drug Crime Index	-.1592730	.5061794	-.0186962

Next we copy this revised table from the SPSS output to a blank worksheet in the Excel workbook with our Eta-square table. In Excel, we then create line graphs for the variables where each line represents a cluster group. Since in this example we are clustering variables over time, we graph a single variable over time by the cluster groups. This could also be a different style graph, e.g. a scatter plot of two continuous variables where the points differ by cluster.

Cluster 2 Interpersonal Crime

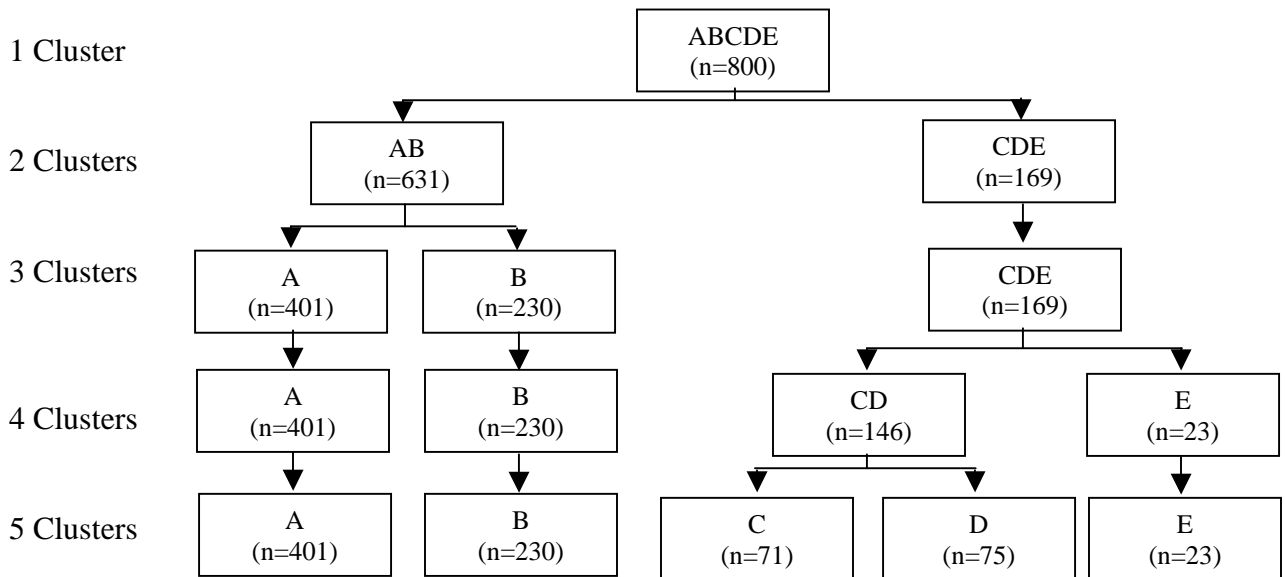


The n sizes are added to the cluster group labels to help show how the clusters break out as you go up in cluster groups. Note that the cluster numbers (e.g., 1, 2, 3) are derived

with each iteration and are generally in order by sample size. This means that if you compare different solutions the clusters may not be in the same order. It is therefore useful to help track the groups manually by tracking how they collapse from the largest number of groups you would consider (e.g., individual groups A-E) to just one group (ABCDE) as shown in the simple dendrogram below. Again, this has to be done manually by seeing which groups stayed intact vs. which groups were combined at each stage.

It is usually relatively obvious that one to two of the cluster solutions are optimal statistically. Deciding which one to use is often a subjective decision made by the authors meeting to discuss which one fits best with the literature, practice, and would be the most straight forward story to tell. In the absence of a clear preference, we usually opt for the simpler (i.e. fewer groups) solution.

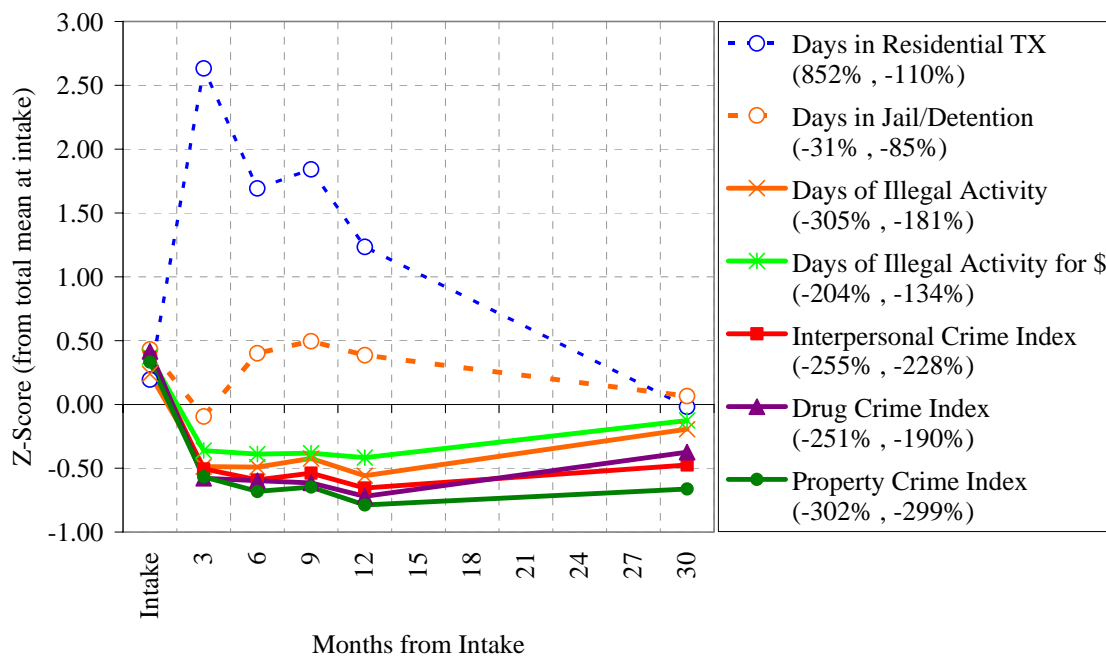
Dendrogram of How Cluster Subgroups Break Out



Once you have a cluster solution and rough storyline for the interpretation, it is often useful to review the distribution of groups and variables to decide how to best present them. Typically, we would sort groups so that they go from lowest to highest severity on the total score or primary factor of variation. We generally use variations of line types or colors to group together variables that sit together substantively (e.g., types of treatment vs. types of symptoms) or methodologically (days vs. symptoms) and sort them so that those that typically are higher in the graph (or which are going to be talked about first) are at the top of the legend. We would then generate a profile for each group on the variables used in the cluster as shown below.

Note that depending on the kind of variables used, you may have to make another transformation in order to plot a profile across variables. If the variables are all on the same metric (e.g. days of use) you could use the raw metric. But if they are on different scales (e.g., days vs. symptoms vs. scales) or have very different distributions, it is often useful to norm them based on the total baseline score (i.e., subtracting the total mean at baseline from each variable x time mean and dividing by the total Standard Deviation at baseline). This gives one metric for comparing groups at baseline, change within and across groups over time. Though they look similar, this is not the same as Z-scores calculated earlier – which were within variable by time and would produce a residual plot with time being taken out (i.e., relative score within time at each time).

Group 3 (n=230): Moderate Crime With More Days in Controlled Environment
(Percent change intake to average for months 3 to 12 and intake to month 30)



Describing These Procedures:

Below is an example of how to describe the above methods in an article for this example.

A cluster analysis was conducted with SPSS Version 14.0.1 (2006) using squared Euclidean distance on 6 measures (property crime index, interpersonal crime index, drug crime index, days of illegal activity, days of illegal activity for money and days in a controlled environment) at each of 6 time points (intake plus 3, 6, 9, 12, and 30 months post intake). To avoid the bias associated with listwise deletion (see Little & Rubin, 1987; Figueredo et al., 2000) missing data were replaced within scale or by regression within person at a given time point, if the case was missing a single time point, it was interpolated from the wave before and after. Each of the 36 measure x time variables was standardized and Ward's (1963) minimum distance was used to create groups by optimizing the ratio of between-group variance to within-group variance (i.e., groups that were maximally different from each other and maximally homogenous within group). The number of groups for the cluster solution was selected based on multiple criteria including: (a) cluster eigenvalues greater than one; (b) smallest group of no less than 30; (c) a solution that explained 70% or more of each of the measures being clustered; (d) a solution that explained 99% or more of the variance in the joint distribution (based on 1-Wilks' Lambda) of the measures; and (e) the interpretability of the solution when profiled (i.e., the source variables plotted by cluster subgroup). Clusters of four to five subgroups met criteria a-d, but the fifth group was sensitive to missing data replacement so we used the simpler and more stable four-group solution.

Any and all parts of the documents can be used at will with or without attribution.

References

- Chan, Y.F. (2006). Correlates with trajectory of criminal behavior for adolescent substance Users during treatment and thirty-month follow-up. Presentation at the 2006 Joint Meeting on Adolescent Substance Abuse Treatment Effectiveness (JMATE). Rockville, MD, March 27-29, 2006. <http://www.mayatech.com/cti/jmate/>
- Dennis, M. L., Wechsberg, W. M., McDermeit (Ives), M., Campbell, R. S., & Rasch, R. R. (2001). The correlates and predictive validity of HIV risk groups among drug users in a community-based sample: Methodological findings from a multi-site cluster analysis. *Evaluation and Program Planning*, 24, 187-206.
- Figueredo, A. J., McKnight, P. E., McKnight, K. M., & Sidani, S. (2000). Multivariate modeling of missing data within and across assessment waves. *Addiction*, 95, S361-S380.
- Godley, S. H., Dennis, M. L., Godley, M. D., & Funk, R. R. (2004). Thirty-month relapse trajectory cluster groups among adolescents discharged from outpatient treatment. *Addiction*, 99, 129-139.
- Hawke, J. (2006). Conceptualizing the Trajectories of Change in Internalizing and Externalizing Disorders on Alcohol and Drug Treatment Outcomes. Presentation at the 2006 Joint Meeting on Adolescent Substance Abuse Treatment Effectiveness (JMATE). Rockville, MD, March 27-29, 2006. <http://www.mayatech.com/cti/jmate/>
- Little, R. A., & Rubin, D. A. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.
- Rapkin, B. D., & Luke, D. A. (1993). Cluster analysis in community research: Epistemology and practice. *American Journal of Community Psychology*, 21, 247-277.
- Reif, S., Wechsberg, W. M., & Dennis, M. L. (2001). Reduction of co-occurring distress among women substance abusers. *Journal of Prevention and Intervention in the Community*, 22, 61-80.
- Wechsberg, W. M., Dennis, M. L., & Stevens, S. J. (1998). Cluster analysis of HIV intervention outcomes among substance abusing women. *American Journal of Drug and Alcohol Abuse*, 24, 239-257.