Cluster Example #3

There are lots of ways to use clustering to "sort out" kinds of folks, how they differ and what those differences portend!

A friend of mine runs a business that provided community-based treatment for adolescents with behavior disorders. Two of his major goals is to be able to anticipate who will and won't respond to the treatment and to anticipate who will and won't have problems at school. We've worked on several multiple regression and ldf models to do this over the years, with varied success. He became increasingly confident that it was important to assess *changes* in certain behaviors as the basis of prediction. We tried several different "behavior change indices" again with varied success. At one point we were working on this while I was teaching clustering and it occurred me to try using clustering to capture "behavior change profiles" to look for "kinds of folks". Remember the factor analysis suggesting that a pivotal variable in this population was extreme verbal abuse? This example shows the initial results from looking for groups of adolescents based on patterns of extreme verbal abuse over the first 6 weeks of treatment.

Here's the agglomeration schedule – Big jumps on steps 38-39, then the jumps get huge!

One approach is to start with 2 clusters and keep adding clusters until the clusters seem homogeneous (splitting clusters doesn't produce "meaningfully different" group). This approach also allows you to track "strays".

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	46	47	.000	0	0	2
2	13	46	.000	0	1	5
3	18	45	.000	0	0	11
4	39	43	.000	0	0	5
5	13	39	.000	2	4	7
6	33	38	.000	0	0	7
7	13	33	.000	5	6	9
8	28	30	.000	0	0	9
9	13	28	.000	7	8	12
10	16	32	.500	0	0	16
11	18	36	1.167	3	0	16
12	13	40	2.067	9	0	25
13	9	29	3.067	0	0	15
14	12	14	4.067	0	0	15
15	9	12	6.067	13	14	18
16	16	18	8.900	10	11	20
17	34	35	12.400	0	0	24
18	2	9	16.800	0	15	22
19	10	15	22.800	0	0	28
20	1	16	29.967	0	16	22
21	41	42	38.467	0	0	28
22	1	2	47.082	20	18	25
23	8	19	56.082	0	0	29
24	34	44	65.915	17	0	33
25	1	13	76.357	22	12	33
26	3	7	91.857	0	0	30
27	26	27	107.857	0	0	35
28	10	41	124.607	19	21	37
29	8	37	145.607	23	0	31
30	3	31	168.774	26	0	35
31	8	25	198.774	29	0	37
32	6	20	230.774	0	0	38
33	1	34	268.458	25	24	39
34	21	24	314.458	0	0	36
35	3	26	377.792	30	27	40
36	5	21	443.125	0	34	38
37	8	10	519.250	31	28	39
38	5	6	656.317	36	32	40
39	1	8	830.775	33	37	42
40	3	5	1071.675	35	38	41
41	3	23	1478.375	40	0	42
42	1	3	3586.512	39	41	0

2-cluster Solution

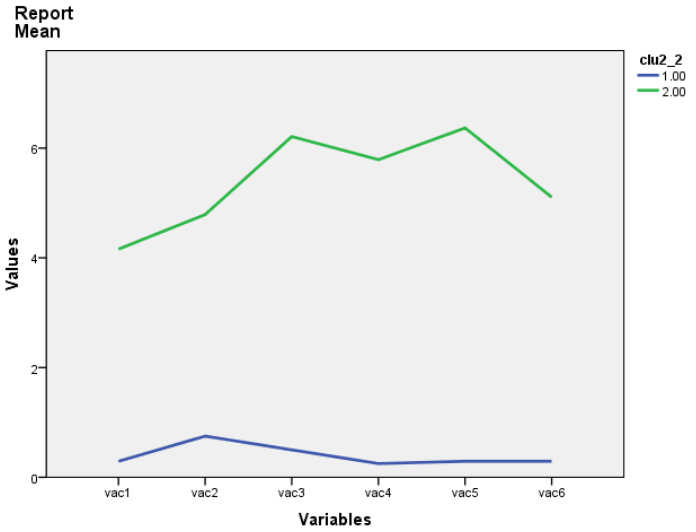
Ward Method

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	32	68.1	74.4	74.4
	2	11	23.4	25.6	100.0
	Total	43	91.5	100.0	
Missing	System	4	8.5		
Total		47	100.0		

One of the least interesting interesting cluster solutions is to find groups that have only level differences, like these…

The only "pattern" is that the elevated group seems to show an increase in events/week across the 6 weeks.

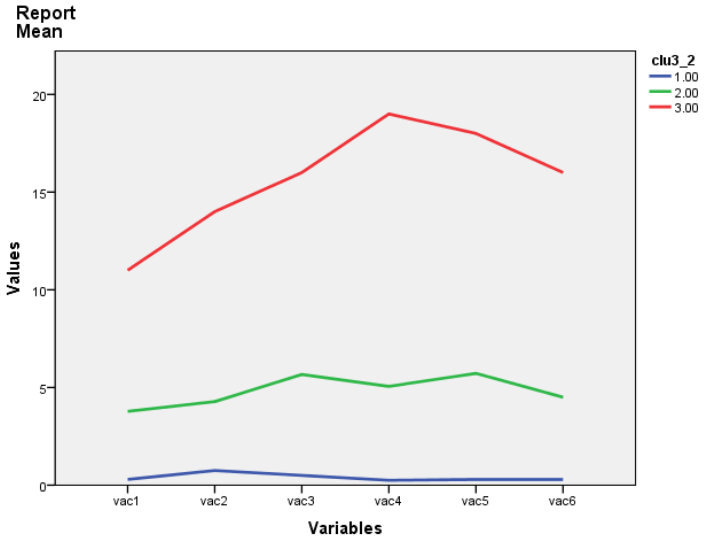
When that happens, you often find that the original quantitative variable provides better association with other characters or behaviors than the binary grouping variable (much like quantitave variable related better to other variables than does the "median split" version of the same variable).



3-cluster Solution

Ward Method

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	32	68.1	74.4	74.4
	2	10	21.3	23.3	97.7
	3	1	2.1	2.3	100.0
	Total	43	91.5	100.0	
Missing	System	4	8.5		
Total		47	100.0		



Not much more interesting than the 2-cluster solution. Removing the "stray" led to a lower estimate of verbal abuse by the second group (outlier effect) and shows that the pattern of incresing verbal abuse across the weeks was also an artifact of including this one individual.

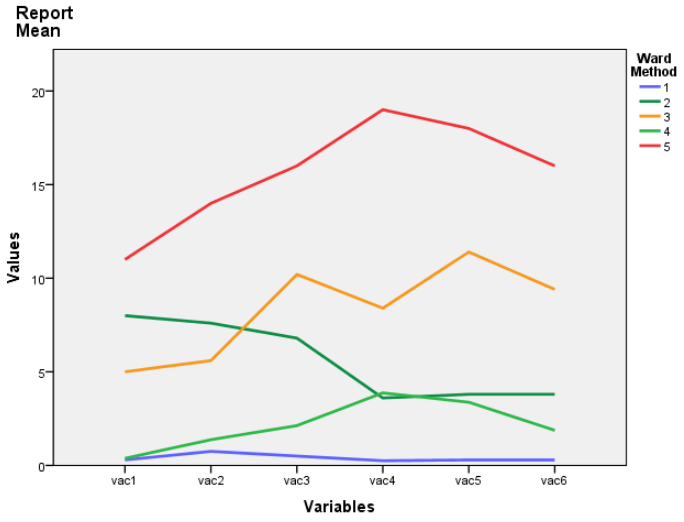
These are the types of results that sometimes lead folks to become disenchanted with clustering methods.

Sometimes, the key is to remember that this is an exploratory process, and give the patterns a chance to emerge…

5-cluster Solution

Ward Method

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	24	51.1	55.8	55.8
	2	5	10.6	11.6	67.4
	3	5	10.6	11.6	79.1
	4	8	17.0	18.6	97.7
	5	1	2.1	2.3	100.0
	Total	43	91.5	100.0	
Missing	System	4	8.5		
Total		47	100.0		



Now it gets interesting!

Still have a dominant (56%) group that produces little verbal abuse and the one that has 1-2 events per day.

The other three groups may be interesting, especially if group membership is related to treatment outcome or school behavior variables.

The cluster profiles also lead to the question, What part of the cluster differences is related to other behaviors? Which carries more information, initial behavior rate (say, weeks 1-2) or later behavior rate (say, weeks 4-6). Groups 2 & 3 both show relatively elevated initial behavior rates. Group 3 shows increased behavior rate over time, however the behavioral rate of group 2 drops over time to the level of groups 4 & 1. (With larger samples ANOVA & pairwise comparisons would be used as the basis for these statements.)

What about the small group memberships? In terms of absolute numbers, it is always nice to have larger samples, and even nicer to have replications (that are really from the same population). In terms of relative sample sizes, keep in mis that groups 2 & 3 are each about 11% of the sample and group 4 is 17% -- lots of clinically relevant populations make up smaller proportions of the "general population"!!

What do the groups tell us?

In one analysis we looked at who was removed from treatment by the presiding judge during the eight month treatment and probation period. The results were…

removed from treatment by judge * Ward Method Crosstabulation

Count

		Ward Method					Total
		1	2	3	4	5	
removed from treatment by judge	kept	22	5	1	8	0	36
	gone	2	0	4	0	1	7
Total		24	5	5	8	1	43

Not surprisingly, the single member of group 5 didn't last the course! The only group to have substantial proportions removed by the judge was group 3, who has shown high initial levels of verbal abuse that escalated over the following weeks.

In another follow-up analysis, we looked at group differences in number of in-school and from-school suspensions during the 6 months following the 6-week intensive treatment program (the same six weeks that the verbal abuse data were collected). The individual who was isolated into the fifth group was not included in the analysis. The results were…

Descriptives

		N	Mean	Std. Deviation
number of in-school disciplinary actions	1.00	23	.8261	1.55657
	2.00	5	8.0000	5.78792
	3.00	5	7.6000	3.20936
	4.00	8	3.8750	6.31184
	Total	41	2.5122	4.29606
number of suspensions from school	1.00	23	.8696	1.28997
	2.00	5	4.6000	4.27785
	3.00	5	5.2000	2.94958
	4.00	8	2.8750	4.08613
	Total	41	1.8780	2.87398

As expected, the majority group (who gave less verbal abuse) had the lowest number of both types of suspensions.

Notice that groups 2 and 3 both had more of each type of suspensions than group 4.

This pattern suggests that it is a high rate of verbal abuse behaviors during the first two weeks of treatment that predicts who will have troubles in school, not whether that behavior increases or decreases.

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
number of in-school disciplinary actions	Between Groups	230.865	3	76.955	5.612	.003
	Within Groups	507.379	37	13.713		
	Total	738.244	40			
number of suspensions from school	Between Groups	68.907	3	22.969	3.250	.033
	Within Groups	261.484	37	7.067		
	Total	330.390	40			

The differential prognosis of the groups depending upon "outcome" is assessed by school suspensions or removal from treatment is interesting, and suggests the importance of predicting specific behaviors, rather than identifying "problem individuals"!!