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. One of his major goals is to be able to anticipate who will and won't respond to the treatment. 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". This example uses just to variables measured during each of the 6 months of treatment (pd = property damage, VA = extreme verbal abuse of a therapist or supervisor) and a small sample. The findings hold with a larger set of variables and different datasets!

Agglomeration Schedule

	Cluster Combined			Stage Cluster First Appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	46	47	.000	0	0	2
2	13	46	.000	0	1	5
3	39	43	.000	0	0	10
23	1	34	15.333	21	19	26
24	10	41	23.500	22	0	26
25	8	19	25.000	0	0	27
26	1	10	35.478	23	24	28
27	8	37	39.500	25	0	30
28	1	44	47.769	26	0	31
29	3	31	55.000	0	0	33
30	8	25	56.667	27	0	32
31	1	42	64.111	28	0	32
32	1	8	75.089	31	30	41
33	3	7	75.500	29	0	35
34	26	27	96.000	0	0	35
35	3	26	105.333	33	34	38
36	5	24	128.000	0	0	39
37	20	21	158.000	0	0	38
38	3	20	197.800	35	37	40
39	5	6	214.000	36	0	40
40	3	5	238.000	38	39	41
41	1	3	320.856	32	40	42
42	1	23	1079.071	41	0	0

There were 12 items in the profile -- 6 months each of property damage & extreme verbal abuse-- and 47 cases.

Here's the agglomeration schedule -- pretty messy!

I kept the cluster memberships for several solutions…

Average Linkage (Between Groups)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	32	68.1	74.4	74.4
	2	7	14.9	16.3	90.7
	3	2	4.3	4.7	95.3
	4	1	2.1	2.3	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		

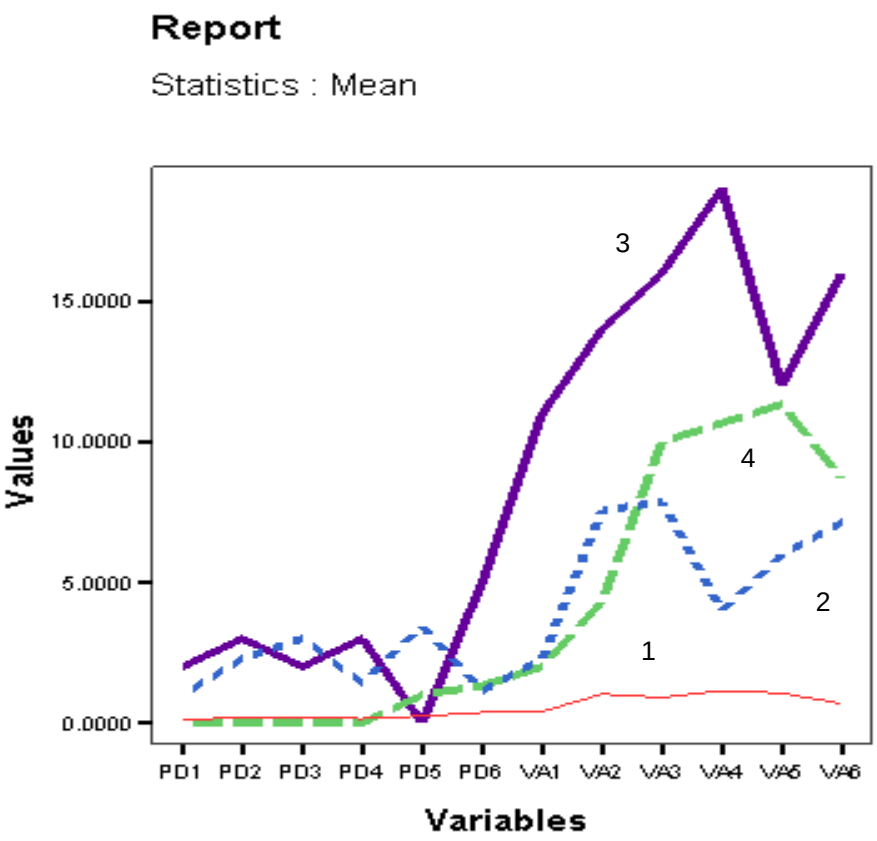
Average Linkage (Between Groups)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	32	68.1	74.4	74.4
	2	7	14.9	16.3	90.7
	3	3	6.4	7.0	97.7
	4	1	2.1	2.3	100.0
	Total	43	91.5	100.0	
Missing	System	4	8.5		
Total		47	100.0		

Average Linkage (Between Groups)

		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		

We liked the 4-cluster solution… and obtained the following graph for it.



Remember that the intent of the analysis is to find a way to anticipate who will have “troubles after treatment” so we looked for group differences on several measure ofs “trouble”.

ANOVA

		F	Sig.
remanded to penal system by judge	Between Groups	1.801	.178
number of in-school	Between Groups	3.362	.045
number of suspensions	Between Groups	1.867	.168
Teacher's rating of	Between Groups	5.015	.012
Teacher's rating of	Between Groups	4.983	.012
Parent's rating of	Between Groups	14.336	.000
Parent's rating of	Between Groups	4.455	.018

Pairwise follow-ups showed that (for the measures with significant differences) Group 1 was different from Groups 2-4, though these groups were seldom different from each other.

We’re working to identify similar models that work using data form earlier months -- to provide “faster” prediction.