Comparing Scoring Systems From Cluster Analysis and Discriminant Analysis Using Random Samples

William Wong and Chih-Chin Ho, Internal Revenue Service,

1111 Constitution Avenue, Washington, DC 20224, U.S.A.

KEY WORDS: Replication, Nonlinear Estimator, Variances, Detection, Profiling, Tax Compliance

1. Introduction

Currently, the Internal Revenue Service (IRS) calculates a scoring formula for each tax return and uses it as one criterion to determine which returns to audit. The IRS periodically updates this formula from a stratified random audit sample. In 1988, such an audit sample was selected. The sample was used to derive a new scoring formula. This score is one of the criteria used to determine whom to audit. In Wong and Ho (2002), we examined the effect of changing sample size on the scoring formula from discriminant analysis. We now extend that work by examining a method of deriving scoring functions using cluster analysis with a variety of distance functions and other options. Those results are compared, and the best results are then compared against those from discriminant analysis. For the evaluation, random subsamples of edited returns are selected, scoring functions developed and applied, and average performances and variances calculated.

Section 2 discusses the design of our analysis, our data, and our goals. Sections 3 and 4 describe our cluster analysis and discriminant analysis approaches. The results of our analysis are presented in Section 5, with the associated tables in the Appendix. Finally, we highlight our conclusions and future research.

2. Basic Analysis Framework

We studied one examination class with a sample of 4,356 audited returns. For our study purposes, we selected a fixed set of 100 original variables. For the cluster analysis procedures, we primarily used a fixed subset of 15 of the "best" variables. We also compared using the 15 "best" variables with using the full set of 100 variables in the cluster procedure. In the discriminant analysis procedures, for each random subsample, we used SAS Proc Stepdisc to determine a subset of the 100 variables to use to create our discriminant function. We used a cross-validation approach to evaluate the performances of the scoring formulas.

We start by selecting stratified random subsamples of 2,500 from our 4,356 sample returns using three strata. These subsamples of 2,500 returns serve as the **modeling data sets.** Thus, for each of these subsamples, we create the cluster analysis and/or discriminant analysis models we wish to compare. Our modeling goal is to maximize the likelihood of identifying returns that exceed a minimum **threshold** discrepancy between the reported and audited tax amounts. (Due to disclosure sensitivity, the threshold dollar amount is withheld.) We now apply the resulting models on the **test data sets** of the remaining 1,856 (= 4,356 - 2,500) returns to score each return. Here, a higher score means the model is predicting a higher probability of the return achieving the threshold. The test data set returns are sorted by descending scores and a **cutoff** percentage, *c*, of returns is selected for evaluation. The evaluation statistic, the "**hit rate**," is defined as the portion of the selected weighted returns achieving the threshold. Cutoff percentages of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, and 75 are analyzed. The cutoff percentage of 100 is also tabulated to provide the average hit rates over the entire test data sets. This procedure is repeated by reselecting 10 to 400 random subsamples, modeling, calculating hit rates for each cutoff percentage, averaging the hit rates over the subsamples, and calculating the variance of each average hit rate.

3. Cluster Analysis Framework

Motivation: Our approach is to identify returns that exceed the discrepancy threshold, find where they cluster, and score the returns based on their shortest distance to the cluster centroids.

Our cluster analysis proceeds as follows:

- Obtain modeling data set: Select a stratified random subsample of 2,500 of the 4,356 returns.
- Identify those returns that exceeded the threshold tax discrepancy. Typically, this would be around 10% of the subsample.
- Create clusters of these "threshold exceeders": Using those returns that exceed the threshold tax discrepancy, run SAS Proc Cluster to create clusters. To create these clusters, we use most of the distance functions available in SAS Proc Cluster: average, centroid, complete, EML, flexible, McQuitty, median, single, and Ward. Distance functions average, centroid, median, and Ward also have "nosquare" options where the distances are not squared.
- Find the centroids of each cluster: For each cluster, obtain the means and standard deviations for each variable.
- Develop raw predicted score functions: For each return exceeding the threshold, calculate its standardized distance to each cluster centroid. Thus,

for each variable, calculate the distance between the return value and the cluster mean and divide the result by the cluster standard deviation. Define the distance to each cluster centroid to be the square root of the sum of the squares of the distances across variables. The minimum of these distances across clusters is the raw predicted score. (When a cluster's average standard deviation is zero, the variable mean with a minimum of one is used.)

- Create cluster score adjustment factors: For each cluster, obtain both its average raw predicted score and its average real score, the tax discrepancy among its elements. The adjusted predicted score is then the raw score with a ratio adjustment to even out the cluster-to-cluster differences and prorate to the real score averages.

- Obtain the test data set: The test data set is the remaining 1,856 (= 4,356-2,500) returns.

- Score each test data set return: For each return, calculate raw scores using the same procedure as above and then apply the adjustment factors calculated above. Since a lower score currently means a higher likelihood of exceeding the threshold, the scores need to be inverted. Since the scores are only used in ranking returns, simply reverse the sort.

- Calculate hit rates for each cutoff percentage: After sorting the returns, apply the strata sampling weights to each return and calculate the weighted hit rates for each cutoff percentage.

- Select the next random subsample and repeat the procedure 10 or 400 times.

- Calculate average hit rates and standard deviations over the random subsamples.

4. Discriminant Analysis Framework

For our study purposes, we selected 100 original variables and use SAS Proc Stepdisc to determine which variables to use to create our discriminant function. Thus, the 100 variables are fixed, but the resulting subset of variables changes from sample to sample. The discrimination classification variable used is a zero-one indicator of whether a return exceeds the threshold tax discrepancy.

We start by selecting stratified subsamples of 2,500 from the 4,356 returns using three strata. The weighted samples are first processed through SAS Proc Stepdisc to determine which subset of variables will be used. This is done using two methods: stepwise with p=0.15 and forward discrimination with a maximum of 15 variables. The weighted subsamples are then processed through SAS Proc Discrim using only the variables identified by the Proc Stepdisc procedure. Only parametric discrimination is tested. These weighted subsamples serve as the discrimination modeling data set. The discrimination test data set is the remaining 1,856 (=4,356-2,500) returns. One output of Proc Discrim is the posterior probability of the test return exceeding the threshold. This posterior probability is used as the score. The test data set returns are sorted by descending scores and weighted, and hit rates are calculated for each cutoff percentage. This procedure is repeated over the 400 random subsamples, and average hit rates and their variances are calculated.

5. Results

For each of the methods, the mean hit rates across the 10 or 400 subsamples were calculated for each percentage cutoff. Along with each mean hit rate, the standard deviation of the mean was also calculated. (The standard deviations calculated were to determine whether the differences between the means are significant and are not sampling error estimates. Those estimates would require correction factors for the large subsampling fractions.)

As indicated above, the basic scoring function is an adjusted minimum distance between the return and the closest cluster centroid. Originally, the minimum cluster distances were not standardized. We found that standardized distances performed better. Various treatments of cluster variable means and variances when they were zero were tried. We settled on replacing the standard deviation with the variable mean with a minimum of 1 when the standard deviation was zero. (This is needed to standardize the distance.)

We tested minimum cluster sizes of 1, 2, 3, 4, 5, 6, 8, 10, and 16. High minimum sizes performed poorly and often did not yield any clusters. The results for minimum cluster sizes of 2 and 4 are given in Appendix Table A1. Since the main cutoffs of interest are 1% to 10%, we summarize the results by averaging the replicate Average Hit Rates (AHR) across these percentages and present them in Table 1. We see that a minimum cluster size of 2 performs better than 4. Furthermore, for distance functions: centroid nosquare, median nosquare, and singular, using a minimum cluster size of 4 did not yield clusters for every subsample.

Table 1 – Average Hit Rate (AHR) Means Across Cutoff Percentages 1% to 10%, by Min Cluster Size, Using 10 Replicates of 10 Clusters with 15 Variables

	Min Cluster Size		Best
	4	2	Size
Average	12.96	15.51	2
Average Nosquare	13.20	14.13	2
Centroid	11.25	14.52	2
Centroid Nosquare		11.88	2
Complete	13.21	16.50	2
EML	15.17	18.71	2
Flexible	16.13	18.89	2
McQuitty	13.08	15.61	2
Median	12.04	14.94	2
Median Nosquare		11.41	2
Single		10.44	2
Ward	15.58	18.66	2
Ward Nosquare	17.28	17.60	2

In parallel with deciding minimum cluster size, we needed to determine how many clusters we should form. We tested different numbers of clusters up to 20, but the higher values did not consistently yield clusters. Table 2 compares the results for forming 10, 8, 6, and 4 clusters using the thirteen distance measures. From the left-hand side of the table, we see that, if we average over the 1% to 10% cutoffs, the optimum number of clusters varies from 4 to 10. However, the 1% cutoff estimates are much larger than the rest. So, if the cutoffs of interest are likely to be in the 2% to 10% range, then the right-hand side of Table 2 shows that the optimum number of clusters is mainly 6 or 8. Most of the distance functions did reasonably well with 8 clusters; so, we pursued our analysis using 8 clusters.

Table 2 – Average Hit Rate (AHR) Means Across Cutoff Percentages to 10%, by Number of Clusters, Using 10 Replicates with Min Cluster Size of 2 and 15 Variables

	Mean of the AHR Over cutoffs 1% to 10% Number of Clusters:				Best	Mean of the AHR Over cutoffs 2% to 10% Number of Clusters:				Best
	10	8	6	t	t	10	8	6	4	t
Aver	15.51	15.96	15.25	13.94	8	14.97	15.34	14.21	13.03	8
AvNs	14.13	16.56	15.01	13.94	8	13.72	16.08	14.05	13.04	8
Cent	14.52	14.59	16.08	13.79	6	14.07	14.24	14.87	12.90	6
CntNs	11.88	13.40	14.80	13.61	6	11.39	13.06	14.30	12.48	6
Comp	16.50	17.79	17.92	15.28	6	16.42	17.31	17.05	14.63	8
EML	18.71	18.71	16.14	14.98	10	17.83	17.79	15.84	14.56	10
Flex	18.89	18.55	19.25	18.21	6	18.51	18.24	19.01	17.69	6
McQ	15.61	17.56	17.43	13.54	8	15.37	16.75	16.39	12.89	8
Med	14.94	16.64	16.42	12.63	8	14.60	15.78	15.42	12.16	8
MdNs	11.41	13.71	14.78	13.71	6	11.05	13.12	14.02	12.91	6
Single	10.44	11.31	11.03	11.67	4	10.35	11.12	10.84	10.85	8
Ward	18.66	19.18	16.50	15.00	8	17.76	18.14	16.23	14.58	8
WdNs	17.60	17.94	18.05	18.63	4	17.36	17.74	17.85	17.58	6

Now, would using 100 variables instead of 15 yield better results? The results in Table 3 show that using 100 variables was sharply poorer than using 15. Perhaps the distance formula needs sharper differential weights by variable when there are so many.

Table 3 – Average Hit Rate (AHR) Means Across Cutoffs Percentages of 1% to 10%, by Number of Variables, Using 10 Replicates of Forming 8 Clusters with Min Cluster Size of 2

	Using 15 vars	Using 100 vars	Best
Average	15.96	12.95	15
Average Nosquare	16.56	12.66	15
Centroid	14.59	11.92	15
Centroid Nosquare	13.40	11.85	15
Complete	17.79	12.12	15
EML	18.71	11.31	15
Flexible	18.55	10.71	15
McQuitty	17.56	12.91	15
Median	16.64	12.55	15
Median Nosquare	13.71	11.30	15
Single	11.31	8.10	15
Ward	19.18	12.53	15
Ward Nosquare	17.94	12.89	15

Just how stable are these average hits? Was using 10 replicates sufficient? Table 4 shows the mean Average Hit Rate and their ranks when using 10 replicates and 400 replicates. Although there is some difference in the means, their relative rankings only changed slightly. The top four distance functions: EML, flexible, Ward, and Ward nosquare, remained on top. The corresponding original tables and their standard deviations are given in Appendix Tables A2 and A3.

Table 4 – Average Hit Rate (AHR) Means Across Cutoffs of 1% to 10% and Their Ranks, by Number of Replicates, Using 8 Clusters with Min Cluster Size of 2 and 15 Variables

	Using 10 reps	Using 400 reps	Rank Using 10 reps	Rank Using 400 reps
Average	15.96	14.77	9	7
Average Nosquare	16.56	14.61	8	8
Centroid	14.59	14.29	10	10
Centroid Nosquare	13.40	13.30	12	11
Complete	17.79	15.99	5	5
EML	18.71	17.49	2	2
Flexible	18.55	17.46	3	4
McQuitty	17.56	15.25	6	6
Median	16.64	14.52	7	9
Median Nosquare	13.71	13.22	11	12
Single	11.31	10.71	13	13
Ward	19.18	17.47	1	3
Ward Nosquare	17.94	17.95	4	1

Finally, back to the original question of which is better, cluster analysis or discriminant analysis? Appendix Table A4 compares the best of the cluster analysis results with the discriminant analysis results. Discriminant analysis seems to do better, with forward discriminant doing the best. But, are we comparing the same things? Discriminant analysis used the package programs SAS Proc Stepdisc and Proc Discrim. Cluster analysis used the package program SAS Proc Cluster with a self-written scoring program. When writing the program, we noticed that the results were still rather sensitive to the parameters. These parameters need to be analyzed for improvement and robustness. Furthermore, we can interplay one method with the other and sharpen both results. We may also want to experiment with combining the methods with regression.

6. Conclusions

- High minimum cluster sizes, high numbers of clusters, and high numbers of variables perform poorly. High sizes and numbers of clusters may be difficult to create. Using 8 clusters with a minimum cluster size of 2 and 15 variables appeared to perform best for our data set. Using 100 variables overwhelmed the scoring algorithm.

- Among the cluster methods, EML, flexible, Ward, and Ward nosquare performed the best.

- Using standard discriminant analysis currently performs better than our cluster scoring procedure.

7. Future Research

In the future we would like to explore methods of enhancing our results, including:

- Combining the methods of cluster analysis, discriminant analysis, and regression for modeling.

- Studying alternative methods calculating and combining the distance functions between the test data set return and each cluster. One enhancement may be to tie the distance function to the function used in creating the clusters.

Finally, we need to test the different methods across years. Specifically, we wish to use one year's data to train the models and apply the results on a different year and then reverse roles. This will help determine the year-to-year deterioration of the models.

8. Footnote

Wong, William and Ho, Chih-Chin (2002), "Evaluating the Effect of Sample Size Changes on Scoring System Performance" 2002 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM], Alexandria, VA: American Statistical Association: 3777-3782.

Appendix

Table A1: Comparing Average % Hit Rates of 13 Clustering Methods by Minimum Cluster Sizes Using 10 Replicates of Forming 10 Clusters with 15 Variables

Cut-off Pct	Aver	Aver Nosq	Cent	Cent Nosq	Comp	EML	Flex	McQ	Med	Med Nosq	Sing	Ward	Ward Nosq
Using a minimum cluster size of 4:													
1	14.84	16.00	12.18	**	14.41	18.49	21.36	14.50	16.67	**	**	21.41	22.97
2	13.16	13.32	11.10	**	12.21	17.59	18.27	14.45	11.74	**	**	18.15	18.90
3	13.10	14.07	11.00	**	13.83	17.11	16.75	14.52	11.26	**	**	17.89	18.66
4	13.66	13.95	11.79	**	14.54	15.49	17.06	13.58	11.29	**	**	16.16	18.09
5	12.75	13.29	11.85	**	13.57	15.13	16.60	13.04	11.82	**	**	14.66	17.34
6	12.64	12.94	10.98	**	13.11	14.19	15.37	12.91	11.75	**	**	14.05	16.79
7	12.50	12.32	11.29	**	12.80	13.80	14.45	12.63	11.97	**	**	13.77	15.95
8	12.36	12.07	11.00	**	12.48	13.56	14.20	12.14	11.30	**	**	13.24	15.12
9	12.43	12.13	10.79	**	12.74	13.36	13.75	11.74	11.55	**	**	13.17	14.42
10	12.19	11.89	10.52	**	12.42	13.01	13.45	11.29	11.05	**	**	13.33	14.57
15	10.80	11.01	9.52	**	11.61	12.36	12.11	10.72	10.41	**	**	11.87	12.66
20	10.00	10.36	9.19	**	10.80	11.72	11.95	10.19	9.85	**	**	12.13	11.93
25	10.12	10.12	9.23	**	10.60	11.19	11.70	9.93	9.80	**	**	11.38	11.47
30	9.95	10.06	8.99	**	10.38	11.33	11.18	9.86	9.78	**	**	11.47	11.00
35	10.00	9.92	8.85	**	10.04	11.17	11.26	9.56	9.81	**	**	11.27	11.00
40	9.71	9.74	8.94	**	10.04	11.15	10.97	9.74	9.70	**	**	11.01	10.84
45	9.68	9.67	9.08	**	9.94	10.92	10.87	9.80	9.72	**	**	10.85	10.90
50	9.70	9.69	9.34	**	9.81	10.68	10.49	9.72	9.74	**	**	10.59	10.91
75	9.64	9.58	9.37	**	9.97	10.25	10.42	9.64	9.70	**	**	10.18	10.45
100	11.77	11.77	11.77	**	11.77	11.77	11.77	11.77	11.77	**	**	11.77	11.77
Using a minimum cluster size of 2:													
1	20.38	17.85	18.52	16.32	17.23	26.59	22.28	17.79	17.99	14.60	11.24	26.84	19.79
2	18.24	17.13	17.34	14.48	17.44	23.47	21.37	17.12	17.31	12.20	11.14	23.12	20.73
3	16.79	15.44	15.58	12.51	17.12	20.37	19.66	17.47	15.91	11.54	11.88	21.55	19.69
4	16.38	13.92	15.94	11.27	17.53	18.69	20.17	16.09	14.11	11.39	10.49	19.47	18.64
5	15.62	13.50	14.49	11.23	16.68	17.79	19.83	15.67	14.78	10.96	9.99	17.69	17.29
6	14.20	13.08	13.25	10.77	16.74	16.83	18.47	15.21	14.55	11.50	10.21	17.05	16.60
7	13.75	12.49	12.83	10.70	16.31	16.37	17.80	14.94	13.92	10.95	10.19	15.99	16.40
8	12.92	12.64	12.55	10.51	15.64	16.07	16.93	14.31	13.62	10.54	9.94	15.24	16.02
9	13.24	12.72	12.43	10.43	15.17	15.93	16.45	13.81	13.76	10.27	9.75	15.00	15.78
10	13.60	12.54	12.24	10.63	15.11	14.98	15.90	13.71	13.41	10.11	9.58	14.69	15.03
15	12.62	12.04	11.71	9.88	13.53	13.92	15.27	13.16	12.56	9.65	9.05	14.42	13.78
20	11.72	11.13	10.49	9.60	13.33	13.28	14.26	12.23	11.59	9.25	9.02	13.62	13.09
25	11.44	11.03	10.49	9.87	12.79	12.35	13.65	11.44	10.98	9.73	9.08	12.66	12.87
30	11.30	10.90	10.22	9.97	12.29	12.08	13.07	11.31	10.92	9.73	8.74	12.25	12.56
35	11.21	10.82	10.19	9.58	11.84	11.66	12.48	11.11	10.83	9.52	8.54	11.78	12.33
40	10.96	10.44	10.08	9.37	11.64	11.48	12.05	11.05	10.69	9.28	8.68	11.71	12.02
45	10.60	10.12	9.74	9.11	11.50	11.24	11.72	10.83	10.37	9.25	8.80	11.47	11.68
50	10.31	9.94	9.64	9.10	11.35	11.03	11.58	10.51	10.15	9.07	8.85	11.18	11.41
75	9.93	9.79	9.54	9.42	10.49	10.52	10.70	10.02	9.94	9.40	9.42	10.53	10.70
100	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77

Note: ** Ten clusters with cluster size >= 4 could not be formed for every replicate with this clustering method.

Table A2: Comparing Average % Hit Rates of 13 Clustering Methods by Number of Replicates When Forming 8 Clusters with 15 Variables and a Minimum Cluster Size of 2

Cut-off Pct	Aver	Aver Nosq	Cent	Cent Nosq	Comp	EML	Flex	McQ	Med	Med Nosq	Sing	Ward	Ward Nosq
Using 10 Replicates:													
1	21.54	20.85	17.77	16.54	22.16	27.02	21.35	24.88	24.32	19.07	13.07	28.47	19.80
2	20.98	20.50	17.88	15.64	19.08	23.24	20.47	22.02	21.04	14.65	13.07	25.14	20.88
3	17.66	18.08	16.72	13.95	19.09	20.28	18.24	17.92	17.59	14.87	12.06	21.26	20.22
4	16.44	17.89	15.25	13.69	18.52	19.51	18.87	17.52	16.77	14.40	11.24	19.20	19.91
5	15.05	16.18	13.63	13.34	17.15	17.67	17.60	16.29	16.06	13.21	11.25	17.86	18.33
6	14.12	15.50	13.52	12.76	17.27	16.68	18.00	15.77	14.66	12.90	10.72	17.09	17.03
7	14.12	14.78	13.50	12.40	16.80	16.24	18.41	16.00	14.81	12.12	10.29	16.41	16.47
8	13.52	14.27	12.76	12.13	16.29	15.66	18.07	15.68	14.04	12.28	10.58	15.91	16.11
9	13.17	14.06	12.43	11.70	15.96	15.62	17.30	15.14	13.60	12.13	10.60	15.52	15.61
10	12.99	13.45	12.49	11.89	15.60	15.20	17.18	14.42	13.47	11.46	10.26	14.92	15.06
15	12.55	12.89	12.00	11.03	14.18	14.23	15.42	13.30	12.36	11.12	9.35	14.37	14.41
20	12.20	12.11	11.50	10.44	13.51	13.12	14.44	12.92	11.88	10.56	9.70	13.67	13.68
25	11.56	11.67	11.19	10.35	13.14	12.37	13.43	12.31	11.57	10.46	9.65	12.79	13.21
30	11.39	11.65	11.17	10.46	12.74	12.07	13.04	11.74	11.22	10.55	9.37	12.29	13.04
35	11.24	11.33	10.98	10.24	12.39	11.87	12.66	11.58	11.14	10.35	9.20	12.09	12.52
40	10.94	11.12	10.77	10.21	11.99	11.54	12.45	11.46	11.04	10.10	8.90	11.83	12.09
45	10.65	10.83	10.43	9.97	11.76	11.25	12.39	11.16	10.77	9.72	9.00	11.48	11.94
50	10.48	10.41	10.33	9.73	11.44	11.09	12.06	10.86	10.65	9.66	9.05	11.34	11.45
75	10.06	10.09	9.99	9.69	10.52	10.41	10.84	10.24	10.29	9.53	9.47	10.27	10.70
100	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77	11.77
Using 400 Replicates:													
1	18.68	18.79	17.70	15.64	20.43	23.78	21.96	19.53	18.39	16.21	13.25	23.64	23.36
2	17.53	17.18	16.72	15.30	18.47	20.95	19.92	18.06	17.18	14.97	12.38	20.95	20.67
3	16.24	15.89	15.61	14.49	17.13	18.92	18.89	16.39	15.79	14.27	11.55	19.06	19.30
4	15.17	14.99	14.63	13.72	16.38	17.75	17.88	15.54	14.79	13.48	10.79	17.65	18.23
5	14.42	14.23	13.91	13.13	15.64	16.78	17.03	14.80	14.10	13.03	10.31	16.77	17.52
6	13.84	13.70	13.46	12.68	15.19	16.19	16.48	14.24	13.65	12.54	9.97	16.13	16.87
7	13.41	13.30	13.12	12.31	14.71	15.72	16.08	13.94	13.20	12.27	9.76	15.71	16.47
8	13.11	12.95	12.86	12.07	14.30	15.32	15.74	13.62	12.98	11.96	9.74	15.27	15.98
9	12.78	12.64	12.51	11.87	13.95	14.95	15.44	13.32	12.71	11.80	9.65	14.89	15.74
10	12.51	12.41	12.34	11.74	13.68	14.58	15.16	13.06	12.45	11.67	9.65	14.67	15.40
15	11.86	11.78	11.78	11.14	12.74	13.55	14.15	12.30	11.87	11.15	9.13	13.65	14.37
20	11.49	11.43	11.35	10.64	12.14	12.86	13.45	11.78	11.44	10.68	9.02	12.98	13.63
25	11.14	11.08	11.02	10.36	11.71	12.38	12.89	11.48	11.10	10.46	9.06	12.44	13.08
30	10.86	10.80	10.77	10.16	11.40	11.97	12.46	11.21	10.83	10.26	8.90	12.02	12.56
35	10.68	10.61	10.58	10.01	11.10	11.63	12.09	10.96	10.68	10.16	8.76	11.68	12.18
40	10.53	10.51	10.44	9.89	10.97	11.39	11.80	10.79	10.55	10.03	8.71	11.45	11.87
45	10.36	10.31	10.26	9.76	10.79	11.22	11.59	10.65	10.39	9.85	8.76	11.27	11.61
50	10.14	10.09	10.06	9.62	10.56	11.04	11.38	10.43	10.14	9.71	8.78	11.08	11.41
75	9.84	9.84	9.83	9.63	10.01	10.28	10.59	9.99	9.88	9.67	9.26	10.32	10.62
100	11.72	11.72	11.72	11.72	11.72	11.72	11.72	11.72	11.72	11.72	11.72	11.72	11.72

Table A3: Comparing Std Dev (Average % Hit Rates) of 13 Clustering Methods by Number of Replicates When Forming 8 Clusters with 15 Variables and a Minimum Cluster Size of 2

Cut-off Pct	Aver	Aver Nosq	Cent	Cent Nosq	Comp	EML	Flex	McQ	Med	Med Nosq	Sing	Ward	Ward Nosq
Using 10 Replicates:													
1	2.12	3.28	2.35	1.44	3.70	2.51	2.18	4.23	3.94	2.28	2.41	2.64	2.23
2	2.16	2.14	2.09	1.35	1.18	2.50	1.20	2.30	2.53	1.17	1.11	2.34	1.64
3	2.15	1.53	1.31	1.54	1.29	1.39	1.00	1.50	1.83	1.31	0.94	1.86	1.32
4	1.68	0.97	0.98	1.26	1.12	1.18	1.01	1.54	1.72	0.87	0.84	1.29	1.41
5	1.36	1.05	0.87	1.15	0.95	1.17	0.96	1.39	1.37	0.94	0.71	1.07	1.01
6	1.25	0.76	0.74	0.91	0.87	1.00	0.85	1.51	1.12	0.95	0.74	0.81	0.86
7	1.10	0.78	0.77	0.79	0.79	0.79	0.82	1.30	1.02	0.69	0.60	0.99	0.78
8	0.96	0.67	0.54	0.74	0.90	0.78	0.62	1.10	0.97	0.77	0.41	0.85	0.63
9	0.93	0.48	0.47	0.71	0.89	0.84	0.69	1.07	0.98	0.67	0.44	0.67	0.53
10	0.89	0.51	0.46	0.68	0.87	0.71	0.57	1.05	0.83	0.55	0.34	0.71	0.57
15	0.87	0.39	0.71	0.59	0.67	0.50	0.51	0.85	0.64	0.46	0.38	0.58	0.49
20	0.61	0.31	0.52	0.54	0.45	0.55	0.37	0.69	0.66	0.48	0.40	0.70	0.45
25	0.47	0.30	0.42	0.39	0.55	0.51	0.21	0.56	0.59	0.37	0.33	0.57	0.36
30	0.56	0.29	0.44	0.48	0.52	0.48	0.22	0.55	0.54	0.30	0.36	0.46	0.33
35	0.48	0.26	0.38	0.40	0.44	0.43	0.30	0.52	0.49	0.24	0.31	0.34	0.25
40	0.33	0.25	0.36	0.28	0.34	0.38	0.29	0.37	0.37	0.26	0.25	0.36	0.25
45	0.29	0.19	0.30	0.29	0.34	0.35	0.30	0.36	0.36	0.21	0.22	0.36	0.23
50	0.27	0.17	0.25	0.29	0.36	0.30	0.35	0.29	0.30	0.19	0.24	0.30	0.23
75	0.20	0.17	0.22	0.23	0.27	0.24	0.23	0.22	0.24	0.17	0.20	0.25	0.16
100	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16
Using 400 Replicates:													
1	0.41	0.40	0.41	0.39	0.43	0.43	0.46	0.45	0.41	0.41	0.34	0.42	0.51
2	0.28	0.27	0.29	0.27	0.30	0.31	0.31	0.30	0.29	0.27	0.23	0.32	0.35
3	0.23	0.22	0.22	0.22	0.23	0.26	0.26	0.23	0.23	0.23	0.18	0.25	0.27
4	0.20	0.18	0.18	0.19	0.20	0.22	0.22	0.20	0.19	0.19	0.15	0.21	0.24
5	0.18	0.17	0.17	0.17	0.17	0.19	0.20	0.18	0.17	0.16	0.13	0.18	0.21
6	0.16	0.16	0.15	0.16	0.17	0.17	0.18	0.17	0.16	0.15	0.12	0.17	0.19
7	0.15	0.15	0.15	0.14	0.16	0.16	0.17	0.15	0.15	0.14	0.11	0.15	0.17
8	0.14	0.14	0.13	0.13	0.15	0.15	0.16	0.14	0.14	0.13	0.10	0.14	0.15
9	0.13	0.12	0.13	0.12	0.14	0.14	0.15	0.14	0.13	0.12	0.09	0.13	0.15
10	0.12	0.12	0.12	0.12	0.13	0.13	0.14	0.13	0.12	0.11	0.09	0.12	0.14
15	0.10	0.09	0.10	0.09	0.11	0.10	0.11	0.10	0.10	0.09	0.07	0.10	0.10
20	0.08	0.08	0.08	0.08	0.09	0.09	0.10	0.09	0.08	0.08	0.06	0.09	0.09
25	0.07	0.07	0.07	0.07	0.08	0.08	0.09	0.08	0.07	0.07	0.05	0.08	0.08
30	0.06	0.06	0.06	0.06	0.07	0.07	0.07	0.07	0.07	0.06	0.05	0.07	0.07
35	0.05	0.05	0.05	0.06	0.06	0.06	0.07	0.06	0.06	0.05	0.05	0.06	0.07
40	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.05	0.05	0.05	0.04	0.05	0.06
45	0.04	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.05	0.05	0.04	0.05	0.05
50	0.04	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.05	0.04	0.04	0.05	0.05
75	0.03	0.03	0.03	0.03	0.03	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.04
100	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03

Table A4: Comparing Average % Hit Rates(AHR) & SD(AHR) Among Select Cluster & Discriminant Methods Using 400 Replicates where Clustering Is Done with 8 Clusters, 15 Variables, and a Minimum Cluster Size of 2

Cut-off Pct	Average % Hit Rate (AHR)						Standard Deviation (AHR)					
	Clustering				Discriminant		Clustering				Discriminant	
	EML	Flex	Ward	Ward Nosq	Step-wise	For-ward	EML	Flex	Ward	Ward Nosq	Step-wise	For-ward
1	23.78	21.96	23.64	23.36	27.03	27.65	0.43	0.46	0.42	0.51	0.49	0.49
2	20.95	19.92	20.95	20.67	27.47	28.85	0.31	0.31	0.32	0.35	0.33	0.34
3	18.92	18.89	19.06	19.30	27.29	28.42	0.26	0.26	0.25	0.27	0.28	0.28
4	17.75	17.88	17.65	18.23	26.70	27.44	0.22	0.22	0.21	0.24	0.24	0.23
5	16.78	17.03	16.77	17.52	26.06	26.56	0.19	0.20	0.18	0.21	0.21	0.21
6	16.19	16.48	16.13	16.87	25.38	25.79	0.17	0.18	0.17	0.19	0.19	0.18
7	15.72	16.08	15.71	16.47	24.85	25.17	0.16	0.17	0.15	0.17	0.17	0.17
8	15.32	15.74	15.27	15.98	24.23	24.63	0.15	0.16	0.14	0.15	0.16	0.16
9	14.95	15.44	14.89	15.74	23.76	24.02	0.14	0.15	0.13	0.15	0.15	0.14
10	14.58	15.16	14.67	15.40	23.29	23.49	0.13	0.14	0.12	0.14	0.14	0.13
15	13.55	14.15	13.65	14.37	21.29	21.38	0.10	0.11	0.10	0.10	0.11	0.11
20	12.86	13.45	12.98	13.63	19.68	19.86	0.09	0.10	0.09	0.09	0.09	0.09
25	12.38	12.89	12.44	13.08	18.69	18.71	0.08	0.09	0.08	0.08	0.08	0.08
30	11.97	12.46	12.02	12.56	17.80	17.79	0.07	0.07	0.07	0.07	0.07	0.07
35	11.63	12.09	11.68	12.18	17.09	17.05	0.06	0.07	0.06	0.07	0.06	0.06
40	11.39	11.80	11.45	11.87	16.45	16.42	0.06	0.06	0.05	0.06	0.06	0.05
45	11.22	11.59	11.27	11.61	15.89	15.90	0.05	0.06	0.05	0.05	0.05	0.05
50	11.04	11.38	11.08	11.41	15.40	15.42	0.05	0.05	0.05	0.05	0.05	0.05
75	10.28	10.59	10.32	10.62	13.34	13.41	0.04	0.04	0.03	0.04	0.04	0.04
100	11.72	11.72	11.72	11.72	11.72	11.72	0.03	0.03	0.03	0.03	0.03	0.03