HBAT Cluster Analysis (August 2019)

Jason Huggy, Student, Lewis University

I. INTRODUCTION

The purpose of this study was to utilize cluster analysis to analyze the HBAT dataset found in *Multivariate Data Analysis* [1]. The data set includes 99 lines of unlabeled data, and 23 variables. Cluster analysis is a group of methods used to group together observations based on their features. The end goal of cluster analysis is to group similar items very close to each other in characteristics (homogeneity) while having dissimilar objects remain very far from each other (heterogeneity) [1]. Cluster analysis is most used for data reduction and hypothesis testing, due to its ability to classify groups. In this study, SAS Enterprise Guide (EG) is used to predict all outcomes and to analyze the relationship between variables.

II. METHODOLOGY

To conduct this study, SAS Enterprise Guide is used to produce all data tables. To begin, the HBAT data is imported into EG and verified to be free of missing values or errors. A quick examination of the data shows that the first five columns of data, as well as the last, are nonmetric features. The rest of the features are all metric. For this study only 5 variables are used as the analysis variables, and then another four are used as dependent variables in order to validate the results.

A. Building the Hierarchical Cluster Model

In the first part of this study, hierarchical cluster analysis is used. This means the process uses a top bottom-up approach in a tree like manner. All observations are initially placed in their own cluster, then they are combined until the greatest amount of heterogeneity exists between each cluster, and the greatest homogeneity exists within each cluster. The first model is built using the *Cluster Analysis* method in EG. I set the options to use the Wards minimum variance method to create the hierarchical model. I set the option to return a CCC plot to show how the model chose how many clusters to use. Variables 6, 8, 12, 15, and 18 are used as the analysis variables.

B. Building the Nonhierarchical Cluster Model

Next, the nonhierarchical cluster model is built. In this model the number of clusters is specified before beginning. The number of clusters returned from the previous step is the same number of clusters used in this step. This model is again built using the *Cluster Analysis* method in EG. I set the options to use the k-means method to create the nonhierarchical model. As will be seen in the results section, I set the number of clusters to two. I set the option to return k-means clustered data in the data output. This output data is used in the next step. Again, variables 6, 8, 12, 15, and 18 are used as the analysis variables.

C. Validating the Results

In the validation stage of this study, the output data from the previous step is used to create a one-way ANOVA statistic. To create the ANOVA, variables 19, 20, 21, and 22 are used as dependent variables, then the clusters returned from the previous stage are used as the independent variable.

III. RESULTS AND DISCUSSION

A. Hierarchical Cluster Model

The end results for the hierarchical model suggest that two clusters be used. As can be seen in figure 1, the CCC increases dramatically between using five clusters and using two clusters. The CCC being highest at 2 clusters shows the greatest agreement between the variables. Looking at the increase in the Pseudo F statistic and Pseudo T-Squared, you can also see how the ratio of inner-cluster homogeneity to cluster difference increases as the number of clusters decreases. With this information, two clusters will be used in the nonhierarchical process to conduct the next stage of the study. Figure 2 shows how the hierarchical process goes through the process of slowly combining clusters until there are only two preferred clusters left.

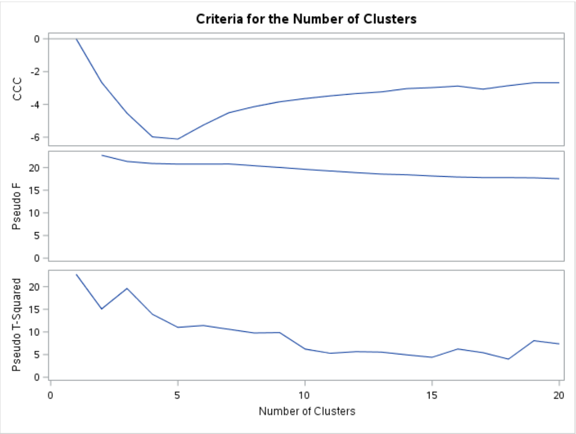


Fig. 1. Results from the hierarchical clustering model

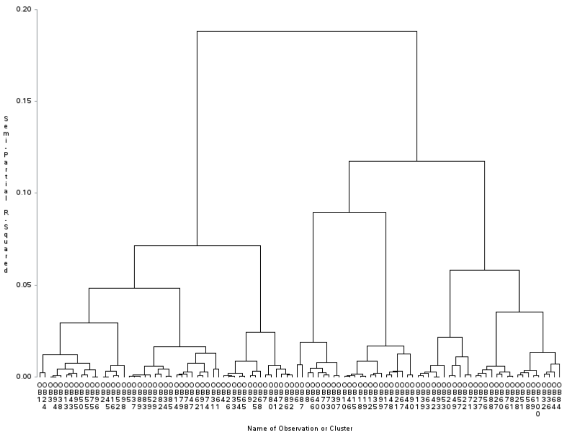
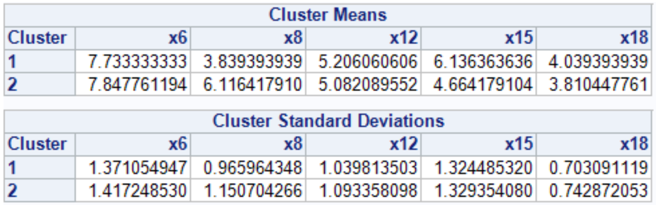


Fig. 2. Tree results from the hierarchical clustering model. The reader can see that the model ends with two clusters at the top of the tree.

B. Nonhierarchical Cluster Model

Next, the nonhierarchical cluster model is evaluated. In this model the number of clusters is set to two based on the previous stage. No plots are available to be plotted in EG using this method, so table 1 is returned to show the results. Table 1 shows the mean and standard deviation for each variable within each cluster. Looking at the means, variables 8 and 15 shows good separation in means. Ideally, the greater the separation in means between the two cluster, the more we are able to rely on the results. However, there is a decent amount of variability within each cluster as see by the standard deviation. Therefore, only variable 8, even with its variability in observations, shows a significant separation that can be used to classify the observations. Variable 8 represents technical support in the HBAT data, which may show the formation of two different customers based on how they rated the company's technical support.



Cluster Means

Cluster	x6	x8	x12	x15	x18
1	7.733333333	3.839393939	5.206060606	6.136363636	4.039393939
2	7.847761194	6.116417910	5.082089552	4.664179104	3.810447761

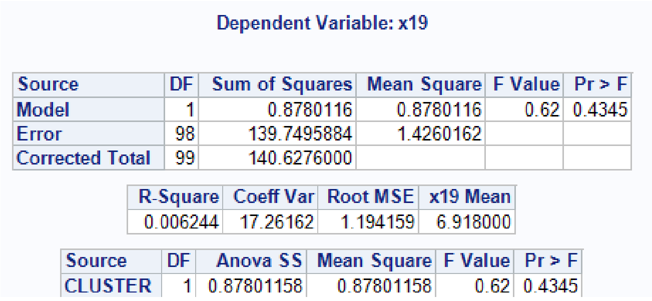
Cluster Standard Deviations

Cluster	x6	x8	x12	x15	x18
1	1.371054947	0.965964348	1.039813503	1.324485320	0.703091119
2	1.417248530	1.150704266	1.093358098	1.329354080	0.742872053

Table. 2. Results from the nonhierarchical clustering model

C. Validation Results

In the final stage of the study a one-way ANOVA metric is used to validate the results from the previous step. Table 2 shows the ANOVA results for each dependent variable. The F value for each variable shows how significant the difference is in means between both clusters. There is not a significant difference for most of the dependent variables, except for variable 22. Variable 22 has an F value of 1.12 and significance at 0.29. This is not great, because we would prefer the significance to be below 0.05. This means that the cluster model created in the previous steps would not do a very good job at predicting the cluster using variables 19 through 22.

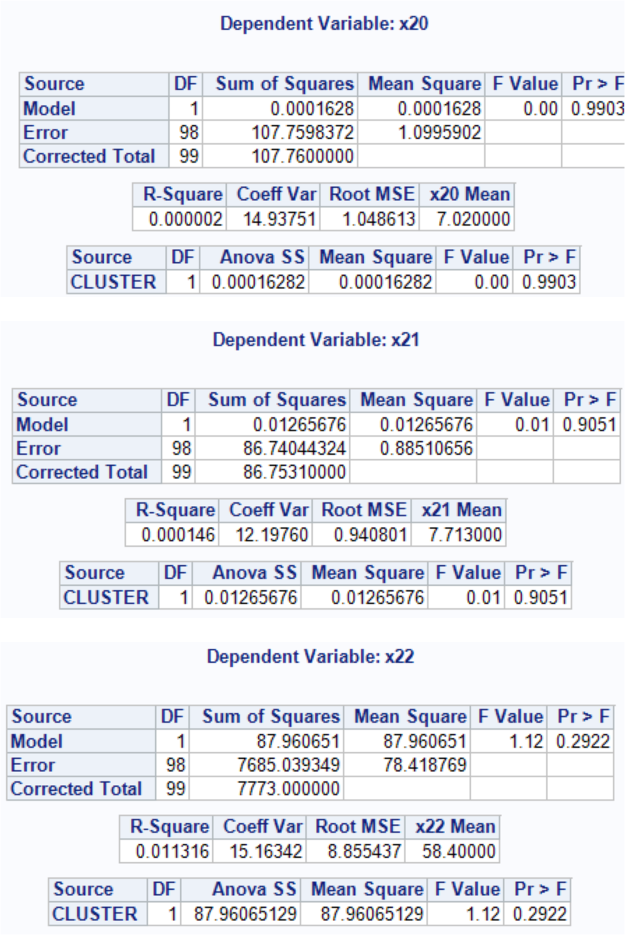


Dependent Variable: x19

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.8780116	0.8780116	0.62	0.4345
Error	98	139.7495884	1.4260162		
Corrected Total	99	140.6276000			

R-Square	Coeff Var	Root MSE	x19 Mean
0.006244	17.26162	1.194159	6.918000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
CLUSTER	1	0.87801158	0.87801158	0.62	0.4345



Dependent Variable: x20

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.0001628	0.0001628	0.00	0.9903
Error	98	107.7598372	1.0995902		
Corrected Total	99	107.7600000			

R-Square	Coeff Var	Root MSE	x20 Mean
0.000002	14.93751	1.048613	7.020000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
CLUSTER	1	0.00016282	0.00016282	0.00	0.9903

Dependent Variable: x21

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.01265676	0.01265676	0.01	0.9051
Error	98	86.74044324	0.88510656		
Corrected Total	99	86.75310000			

R-Square	Coeff Var	Root MSE	x21 Mean
0.000146	12.19760	0.940801	7.713000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
CLUSTER	1	0.01265676	0.01265676	0.01	0.9051

Dependent Variable: x22

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	87.960651	87.960651	1.12	0.2922
Error	98	7685.039349	78.418769		
Corrected Total	99	7773.000000			

R-Square	Coeff Var	Root MSE	x22 Mean
0.011316	15.16342	8.855437	58.40000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
CLUSTER	1	87.96065129	87.96065129	1.12	0.2922

Table. 2. Results of one-way ANOVA to validate cluster results.

IV. CONCLUSION

While the hierarchical model suggested two clusters, the ANOVA results show that the model does not do a very good job at predicting the values of other dependent values. In the results, it seemed like variable 8 was the only variable that showed significant difference between each cluster. This makes the model less reliable. If new data was used, there is probably no guarantee that variable 8 would see the same results. It would be best to attempt this study again, but with multiple samples to see if there are any consistent results amongst each sample.

REFERENCES

[1] J. Hair, W. Black, B. Babin, R. Anderson, *Multivariate Data Analysis: Pearson New International Edition*. 2013.