A Method for Improving List Building: Cluster Profiling

Will Cecere[1], Denise Abreu[1], Jaki McCarthy[1], Thomas Jacob[1]

[1]National Agricultural Statistics Service Research and Development Division, U.S. Department of Agriculture, 3251 Old Lee Highway, Fairfax, VA 22030

Abstract

The USDA National Agricultural Statistics Service (NASS) conducts the quinquennial Census of Agriculture in years ending in 2 and 7. Also, NASS conducts an annual area frame based survey, the June Area Survey (JAS). The census has a dual frame: an independent list frame and the area frame from the JAS. The JAS is used to identify farming operations missed on the list frame. In 2007, a full census questionnaire was sent to all JAS records that were not found on the census mail list. Multiple clustering techniques were used to characterize farming operations missed during the census mail list building. Hierarchical methods (average linkage, centroid, and Ward's method) and non-hierarchical k-means clustering were used to identify groupings. Through cluster profiling, potential improvements to future list building efforts are discussed.

Key Words: hierarchical clustering, k-means, cluster profiling, dual frame

1. Introduction

The USDA National Agricultural Statistics Service (NASS) conducts the quinquennial Census of Agriculture in years ending in 2 and 7. The Census of Agriculture is a complete count of United States (U.S.) farms and ranches as well as the people who operate them. A farm is defined as a place from which $1,000 or more of agricultural products were produced and sold, or normally would have been sold during the census year, including agriculturally related government payments. The census collects data on land use, ownership, operator characteristics, production practices, income and expenditures, and many other characteristics. The outcome, when compared to earlier censuses, helps to measure trends and new developments in the agricultural sector of the national economy. The information is used only for statistical purposes and data are published only in tabulated totals. The census provides the only source of uniform, comprehensive agricultural data for every county in the nation.

NASS maintains a list of farmers and ranchers from which the Census Mail List (CML) is compiled. Census forms are sent to the CML of all known and potential agricultural operations in the U.S. The goal is to build as complete a CML as possible of agricultural places that meet the NASS farm definition. NASS builds and improves the list on an ongoing basis. To achieve this NASS obtains special commodity lists to address specific list deficiencies.

Despite the agency's best efforts in building as complete a list as possible, however, there will ultimately be some level of incompleteness in covering the farm population in the resulting CML and NASS uses its area frame based June Area Survey (JAS) to measure this incompleteness. For the 2007 JAS, a supplemental sample was selected which targeted farming demographics that typically had lower coverage rates on the list. Farming operations from the 2007 JAS (and its supplemental sample) that did not match those on the CML were determined to be Not-on-the-Mail List (NML). These operations were mailed a census report form to collect information about them. Data from the NML operations provided a measure of the undercoverage of the CML as well as information on their size, commodities produced, operator demographics and other descriptive information.

1.1 The Census of Agriculture and Mail List Development

The goal with the CML is to build as complete a list as possible of agricultural places that meet the NASS farm definition. The CML compilation begins with the list used to define sampling populations for NASS surveys conducted for its annual agricultural estimates program. NASS builds and improves the list on an ongoing basis by obtaining outside source lists. Sources include State and federal government lists, producer association lists, seed grower lists, pesticide applicator lists, veterinarian lists, marketing association lists, and a variety of other agriculture related lists. NASS also obtains special commodity lists to address specific list deficiencies. These outside source lists are matched to the NASS list using record linkage programs. Most names on newly acquired lists are already on the NASS list. Records not on the NASS list are treated as potential farms until NASS can confirm their existence as a qualifying farm.

List building activities for developing the 2007 CML started in 2004. Between 2004 and 2007, NASS conducted a series of Agricultural Identification Surveys (AIS) to screen approximately 1.7 million records for agriculture activity, which included nonrespondents from the 2002 Census of Agriculture and newly added records from outside list sources. The AIS report form collected information that was used to determine farm/non-farm status. Reports identified as farms were added to the NASS list and subsequently to the CML. The official CML was finalized on September 1, 2007 and contained 3,194,373 records. There were 2,198,410 records that were thought to meet the NASS farm definition and 995,963 potential farm records.

To account for farming operations not on the CML, NASS used its area frame. The NASS area frame covers all land in the U.S. and includes all farms. The land in the U.S. is stratified by characteristics of the land. Segments of approximately equal size are delineated within each stratum and designated on aerial photographs (See red outlined boundary in Figure 1). A probability sample of segments is drawn within each stratum for the NASS annual area frame-based JAS.

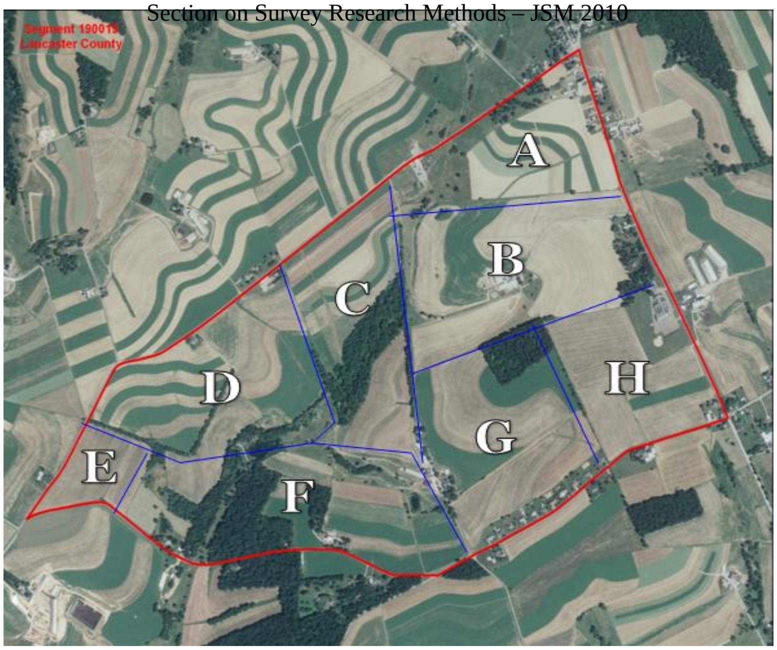

Figure 1: JAS segment with tract boundaries

The JAS sample of segments is allocated to strata to provide accurate measures of acres planted to widely grown crops and inventories of hogs and cattle. Sampled segments in the JAS are personally enumerated. Each operation identified within a segment boundary is known as a tract (See blue outlined areas labeled A through H in Figure 1). The 2007 JAS consisted of 10,912 regular sampled segments and it was supplemented with 3,692 Agricultural Coverage Evaluation Survey (ACES) segments. ACES segments were selected to provide measures of small and minority owned farms. These additional ACES segments targeted farming demographics that typically had lower coverage rates on the list. The information from each tract (operation) within a segment is matched against operations on the NASS list to determine the amount of undercoverage that exists for a wide range of farming sectors and farmer demographics.

Data from the NML operations provided a measure of the undercoverage of the CML operations. In general, NML farms tended to be small in acreage, production, and sales of agricultural products. Farm operations were missed for various reasons, including the possibility that the operation started after the mail list was developed, the operation was so small that it did not appear in any agriculture related source lists, or the operation was erroneously classified as a nonfarm prior to mailout.

The objective in this research was to find ways to improve our list building through a better understanding of our NML population. It was thought that knowing more about the NML would help NASS find farm operations from outside sources more easily. In order to achieve this, a way to partition or group operations that are similar must be employed to identify areas that list building efforts may be targeted.

2. Methods

In order to achieve the goal of characterizing the NML operations, we must look at techniques that allow for the partitioning of the operations based on a set of variables. One insightful way of looking at this problem is through the use of a multivariate technique called cluster analysis. Cluster analysis seeks to find optimal groupings or clusters which minimize differences within a cluster while maximizing differences across clusters.

The intended use of cluster analysis in the context of this research is similar to that of businesses using a form of cluster analysis called customer segmentation. Here clustering is performed to segment a customer base in order to get useful results; in this context useful typically means that the results will aid in a marketing process. The usual goals in this process are to build customer segments in order to understand how to best market a product or set of products to each customer group. These techniques gained popularity due to the fact that businesses could avoid mass marketing and thus save on costs by having their marketing plan customized to specific marketing groups (Collica 2007). This concept is related to the objectives of this research in that the NML population represents a portion of our customer base. It is important to better understand the NML operations with the use of clustering in order to better target common groupings of operations to optimize list building efforts.

One important aspect of cluster analysis is the use of similarity or proximity measures. To accurately depict the degree of closeness from one observation to another, a quantitative measure must be selected for all variables used in the analysis. Common measures of similarity for categorical data often involve calculating a similarity coefficient for whether two observations have the same values. For continuous data there are more options for measures of distance, ranging from a simple Euclidean distance to correlation measures such as Pearson’s. A common situation is to have mixed mode data, continuous and categorical, in which case a similarity matrix is often used as a measure of proximity.

2.1 Clustering techniques

There are numerous techniques available for cluster analysis due to the wide range of application it has. A popular approach to clustering is to employ hierarchical methods, all of which use a series of partitions to arrive at the final number of clusters. There are two categories of hierarchical clustering, agglomerative, and divisive. In an agglomerative method, we start out with *n* clusters and end with a single cluster containing all observations. In a divisive method, a single cluster with all observations is broken up until there are *n* clusters. Criteria are examined in either case to determine which set of clusters most appropriately distinguishes the data.

For this analysis, three agglomerative hierarchical methods were evaluated: average, centroid, and Ward’s method. In the average linkage method, the distance between two clusters A and B is the average of the distances between all observations in A and all observations in B. The centroid method examines the Euclidean distance between the mean vectors of two clusters to determine distance. Ward’s method seeks to minimize the total within-cluster error sum of squares. Consequently, Ward’s method selects the minimum between-cluster distances before merging them.

Another common approach to clustering is to use optimization techniques. These techniques involve maximizing or minimizing a set of numerical criteria in order to produce a preselected number of clusters. One such popular method examined is called the k-means method. Once the number of clusters k is preselected, various algorithms depending on the software package are performed so that the sum of squares within each cluster is minimized.

When working with larger data files, often it is easier to use a two-stage clustering approach. Under this method, a pre-cluster stage is performed in order to reduce a large data file into cluster seeds. From the cluster seeds, typically a hierarchical method is used to determine a final number of clusters. One major advantage of the two-stage clustering approach is that it offers a Euclidean distance measure for continuous variables as well as a likelihood function for categorical variables, making it convenient for mixed mode data. One critical assumption for using a two-stage clustering approach is that all continuous variables follow the normal distribution.

An important aspect of cluster analysis is that there is no "correct" solution. Results may vary greatly depending on what method is employed and how the data are used. The goal of the researcher in using cluster techniques should be to come out with practical results. If the clusters that result from using any method cannot be linked to some form of useful interpretation with respect to the subject matter, then the results are of no use. A quote by Dr. George Box accurately describes our approach. He stated about statistical models in general "All models are wrong, some are useful". Therefore we must be discriminating with results so that we may get some use out of them.

2.2 Data and software preparation

The data file used for this project consisted of 4,810 tracts from the 2007 June Area Survey. These tracts represent all of the NML operations qualifying as farms, and they expanded to a total of 361,687 farming operations. The data analyzed came from 2007 Census of Agriculture questionnaires that were sent to these operations.

Starting with a data file with over 400 variables, criteria were established in order to trim the number of variables to a more appropriate list from which useful interpretation could be drawn. If a variable had a large number of missing observations or valid zeros, we removed it from the analysis. For several specialty commodity variables that didn't contain enough observations (i.e. fruits, nuts, and livestock), indicator variables were created to account for them. If a variable displayed an unusually high correlation with another variable, it was also removed. Highly correlated variables have a tendency to skew cluster formations in their direction, which in turn conceals other variables that may be more significant in the cluster formation. Additional subject matter knowledge and expertise were used to remove further variables not eliminated previously.

A final list of 70 variables was arrived at for our analysis. A representation of the kind of variables used is shown in Table 1.

Table 1: Types of variables used in cluster analysis

Operator expenditures	Commodities raised
Farm Type	Value of sales
Operator Demographics	Cropland

The SAS software package JMP was initially used to examine one-stage methods. The hierarchical methods as well as k-means clustering were tested using JMP's procedures. It was very difficult to arrive at any form of interpretable results from the one-stage clustering methods. The software struggled with the mixed mode data as well as the quantity of variables used as inputs. Graphical outputs such as dendrograms were of no use given the quantity and type of data used.

SAS Enterprise Miner data mining software package was used to examine two-stage cluster methods. For the Enterprise Miner two-stage cluster procedure, the first stage utilizes an optimization method and the final stage uses a hierarchical method. The k-means method was used for all analysis to make the cluster seeds and then the three hierarchical methods discussed (average, centroid, and Ward's method) were performed separately in the second stage.

Since the variables in the study are not all measured in the same units (i.e. acres, dollars, etc), they were standardized by dividing by their respective standard deviations. This assures that no additional weight is given to variables with a larger scale. Log transforms were used in order for the positively skewed continuous variables to meet the normality assumptions.

The cluster procedure in Enterprise Miner used a k-means algorithm to select the cluster seeds, and then selected in the second stage the smallest number of clusters such that two constraints were met. The first was that at least two clusters and no more than the maximum number of clusters requested were produced. The second was that the cubic clustering criteria (testing the hypothesis that all data are from the same uniform distribution) had to be greater than the preset cutoff. After the clusters were formed, they could be further analyzed by using segment profiling in order to gain a greater understanding of the variable values in each cluster.

3. Results

The clustering was performed using the three hierarchical methods in the second stages. Both the centroid and the average linkage yielded a five cluster solution while Ward's method gave a three cluster result. A closer look at the solution given by Ward's method showed that it was difficult to distinguish the defining variable values. For each cluster the values for the variables most important to that cluster were not distinctly separate from those of the other cluster. This made characterizing the clusters difficult so the solution from Ward's method was not chosen.

The two separate five cluster solutions were practically identical so either one could have been used for interpretation. The sizes of the clusters in terms of the number of tracts and expanded farms in each cluster are displayed in Table 2.

Table 2: Cluster sizes

Cluster	1	2	3	4	5	Total
Tracts	1,800	1,783	588	323	316	4810
Expanded number of farms	158,687	141,053	19,458	18,566	23,922	361,687

Cluster 1 is the largest group and represents almost 160,000 farm operations. It is characterized by a high quantity of point farms. A point farm is defined as an operation that didn't report enough agricultural sales but had enough agriculture inventory to qualify as a farming operation. When compared to the overall NML population, this point farm cluster has a much higher proportion of cattle, equine, and other livestock.

One aspect that the segment profiling examined in SAS Enterprise Miner is the logworth statistic, which measures how well a variable partitions observations into a cluster. For each cluster, the defining variables of the cluster are listed in order of their logworth value. Some defining variables with a high logworth value for cluster one include Total Value of Production (TVP) and Farm Type. Figure 2 shows the overall distribution of TVP as compared to the operations in the point farms cluster. The inner circle displays the overall population distribution while the outer circle shows the cluster distribution.

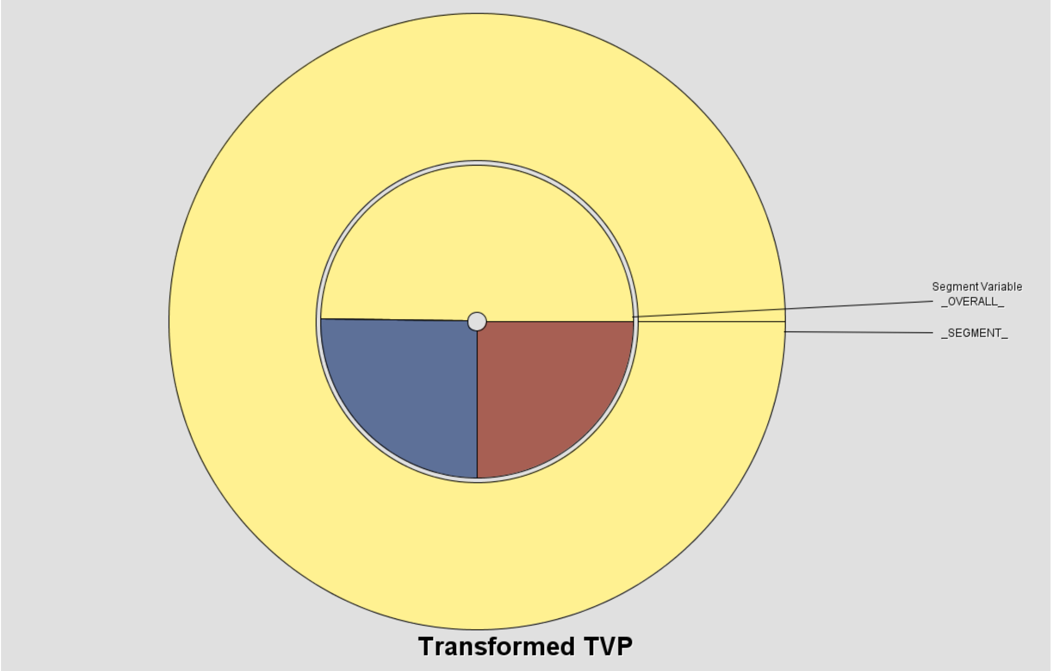

Figure 2: Segment profile of TVP for cluster 1

Here the yellow indicates a Total Value of Production ranging from $0 to $900. The blue indicates values from $900 - $8500 and the red represents values above $8500. It is clear from this chart that cluster 1 in the outer ring or the point farm cluster has observations with a low TVP relative to the overall NML population.

Cluster 2 can be described as a group of operations that represent the overall NML population closely. All variables examined for cluster 2 showed that they were reflective of the overall NML population. Defining variables for this cluster include Total Sales and Cropland Harvested.

Cluster 3 can be described as the high value of sales cluster. The majority of the operations in this cluster have a high sales value and the defining variables are primarily sales variables such as TVP and total sales. This group is much smaller than the previous two clusters with 588 tracts representing over 19,000 operations. It contains mainly full time operators (primarily males) who have been in operation for more than 20 years.

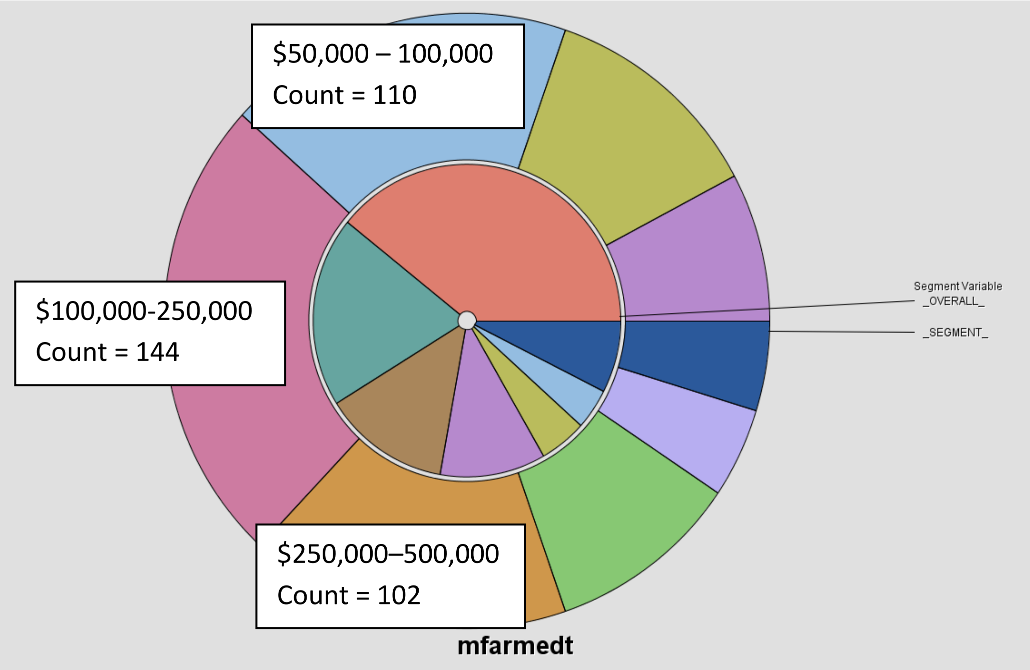

Figure 3: Census final farm value of sales for cluster 3

The discrepancy in value of sales between cluster 3 and the overall NML population is shown in Figure 3. In the inner circle representing the NML population, the highest sales class displayed ranges from $50,000 to $100,000 and is shown by the light blue. The majority of the outer circle, representing the distribution of value of sales for cluster 3, shows that the majority of operations have a sales class of greater than $50,000 with several over $1,000,000.

Operations renting land was an important characteristic of the fourth cluster. These are mostly part-time operations that have not been in operation until more recently. Its defining variables include Land Rented from Others and low Dollar Value of Owned Land.

Finally, the fifth and smallest cluster contains mostly operations that have idle cropland. Many operations in this cluster have hay or idle cropland.

A common practice once the clusters are formed is to examine variables of interest across the clusters. This can provide insight as to additional characteristics that each cluster may possess and ultimately will aid in targeting that group. A total of 17 variables of interest were examined across the clusters ranging from operator characteristics to geographic variables.

Table 3: Part Time operator status across clusters

Frequency	**Cluster Number**					
Code	1	2	3	4	5	Total
Full Time	41780	43049	12595	7714	6347	**111488**
Part Time	116907	98003	6862	10851	17574	**250198**
Total	**158687**	**141053**	**19458**	**18566**	**23921**	**361687**

Table 3 shows a binary variable called Part Time that tells whether an operation is a full time or part time operation. It can be seen that the majority of the operations across the NML tracts are part time, 250,198 out of 361,687. However, the number of full time operations within cluster 3 (the high sales cluster) is almost double that of part time operators.

4. Discussion

Analyzing variables across clusters gives the ability to target multiple characteristics that are specific to subgroups. For instance in the example of the Part Time variable, adding more knowledge of the high sales cluster can potentially make it easier for operators with those characteristics to be found on an outside source list and added to the CML.

After presenting results of this research to the NASS List Frame Section, it was recommended that analysis be done to compare the clusters formed from the NML records to the same cluster definitions on the CML. This showed which areas of the CML we are missing most in proportion to the NML. A simple examination of the clusters when applied to the CML showed that while the high sales cluster on the NML looks concerning, over 40 times the number of operations are assigned to this cluster for the CML. This indicates that the high sales operations are well represented on the CML. In cluster 1 or the point farms cluster, there are roughly double the number of operations in the NML than in the CML. This may be a sign that point farms are under-represented on the CML.

Figure 4 shows a comparison of the CML vs. the NML across cluster 3 for a variable called Start Year. The years on the bottom indicate the decade in which an operation started, i.e. 30 means that an operation started in the 1930s and 0 means an operation began in the 2000s. From the data, it is clear that a much larger percentage of the NML population in cluster 3 began operating in the 2000s. This makes sense given that newer operations would be more difficult to capture on the NASS CML. However, information such as this also provides a valuable comparison of the NML cluster to the CML.

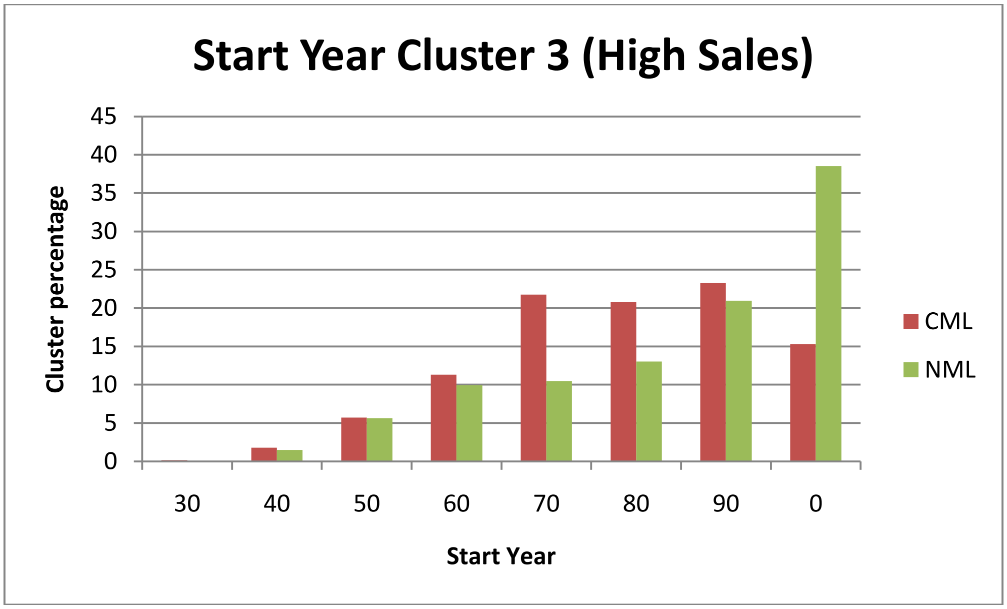

Figure 4: NML vs. CML comparison of Start Year in cluster 3

The efforts of the cluster analysis have yielded a combination of results, some of which were known and some that provided new insights about NML operations. The use of this exploratory technique allowed for the ability to use a wide variety of variables in order to gain insight as to which operations on the NML are most similar and why. It was clear from our results that all NML operations are not alike. It is useful to know the characteristics of clusters within the NML and the relative size of the clusters. Future efforts will focus on trying to incorporate this information into efforts to obtain outside source lists to add to the CML. Through this effort we hope to make improvements to the CML for the 2012 Census of Agriculture.

References

Everitt, B.S., Landau, S, Leese, M. (2001), *Cluster Analysis (Fourth Edition),* Arnold

Rencher, A.C. (2002), *Methods of Multivariate Analysis (Second edition)*, Wiley Series in Probability and Statistics

Collica, R.S. (2007), CRM Segmentation and Clustering *Using SAS Enterprise Miner*, SAS Press Series