United States Department of Agriculture

National Agricultural Statistics Service

Research and
Development Division
Washington DC 20250

RDD Research Report
Number RDD-11-03

September 2011

A Method for Improving List Building: Cluster Profiling

Will Cecere, Denise A. Abreu, Jaki McCarthy, and Thomas Jacob

This report has been prepared for limited distribution to the research community outside the United States Department of Agriculture. The views expressed herein are not necessarily those of the National Agricultural Statistics Service or the United States Department of Agriculture.

EXECUTIVE SUMMARY

The 2007 June Area Survey (JAS) was used to identify farming operations that were not found on the Census Mail List (CML). These Not-on-the-Mail-List (NML) operations were used as a measure of undercoverage for the 2007 Census of Agriculture. The operations were mailed a census report form to collect information about them. The NML farms consisted of 4,810 JAS tracts, representing an expanded number of 361,687 operations.

Given that all NML farms are not alike, an examination of their differences was proposed. Using a selected number of variables obtained from the census questionnaire and the JAS, a variety of clustering techniques were performed on the data. The objective was to partition or group observations such that differences were minimized within each cluster while maximizing differences across clusters. After several cluster methods were performed, a solution was chosen that used 5 clusters to describe the data. Segment profiling was applied to characterize each cluster in terms of the variables that best defined them. Results showed that the clusters were able to distinguish operations that had such characteristics as: a low Total Value of Production (TVP) and a lot of point farms, a high TVP, part-time status and renting land, and idle cropland.

Variables of interest were examined across the clusters. This analysis pointed out such things as, the majority of the operations in the cluster with many point farms were of part-time status, as well as that almost two thirds of the operations in the cluster with rented land had an operator who became the primary operator after the year 2000. An additional approach was taken to compare the same cluster definitions when applied to the CML. This gave perspective as to how the NML compared to the CML. The results showed that the NML contained more operations in the land rented cluster and roughly double the number of operations in the point farm cluster as the CML. The clusters were also used to compare the CML to the NML across matching variables of interest. The results of this cluster analysis could be used to target operations for future building of the list frame.

RECOMMENDATIONS

1. Use comparisons between the CML and NML by variable as a gauge for what is missing on the CML and needs to be targeted for list building.
2. Examine needed areas of CML list building using the results of the NML clusters across variables examined and utilize this information to match to outside list sources.

A Method for Improving List Building: Cluster Profiling

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

Abstract

The 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 employs a dual frame: an independent list frame and the area frame component 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] Will Cecere, Denise A. Abreu and Jaki McCarthy are statisticians with the National Agricultural Statistics Service (NASS) Research and Development Division located at Room 305, 3251 Old Lee Highway, Fairfax, VA 22030. Thomas Jacob is a NASS statistician with the Agency's Information Services Section, located at 1400 Independence Avenue SW, Washington, DC 20250.

1. INTRODUCTION

The 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 enumeration 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 using the CML to all known and potential agricultural operations in the U.S. The goal is to build as complete a CML as possible of all agricultural places that meet the NASS farm definition. NASS builds and improves the list on an ongoing basis. To achieve this, NASS obtains information from outside sources as well as special commodity lists.

Despite the agency's best efforts in building as complete a list as possible, there will ultimately be some level of incompleteness in covering the farm population in the resulting CML. To measure this incompleteness, NASS uses its area frame based June Area Survey (JAS). For the 2007 JAS, prior to the census, an additional supplemental area sample was selected which targeted farming demographics that typically have lower coverage rates on the list frame, the foundation for the CML. Any farming operations found on the 2007 JAS or the 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 information from outside sources. These sources include lists from state and federal government agencies, producer associations, seed growers, pesticide applicators, veterinarians, marketing associations, and a variety of other agriculture related areas. 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. Within this, 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 thus 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.

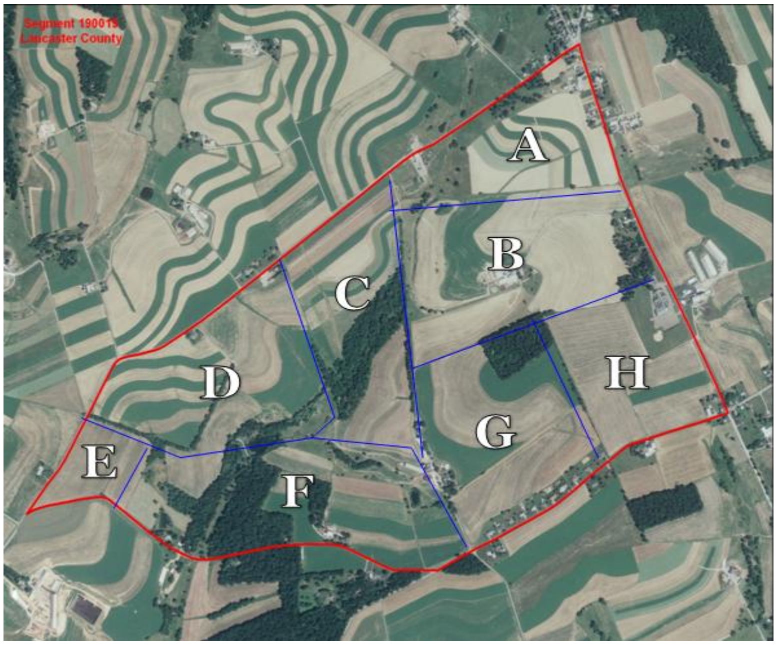


Figure 1. JAS segment with tract boundaries

The JAS sample of segments is allocated to strata to provide accurate measures of acres planted to widely grown crops and not-on-the-list cattle inventory. 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 a supplemental sample of 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 CML to determine the NML operations to which a census report form was mailed.

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 (Eldridge, 2007). However, it is important to keep in mind that NML operations are not all alike. 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 non-farm prior to mailout.

The objective of this research was to find ways to improve our list through a better understanding of our NML population. It was thought that knowing more about distinct subgroups within the NML would help NASS find farm operations more easily from outside sources. In order to achieve this, a way to partition or group similar operations was needed to identify areas where list building efforts could 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 an application 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 the NASS user or 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 of distance 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.

Variables used in cluster analysis in most cases are not measured using the same units. For example, continuous variables Total Value of Production (TVP) and Cropland Harvested are recorded using a different scale. Therefore, it would not make sense to treat measures of distance the same for variables using different units. A common solution to this problem is to standardize each continuous variable using the standard deviations calculated from the dataset.

2.1 Clustering techniques

There are numerous techniques available for cluster analysis due to its wide range of applications. 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 purposes of this research, three agglomerative hierarchical methods were evaluated: average linkage, 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 approach to clustering is to use optimization techniques. These techniques involve maximizing or minimizing a set of numerical criteria in order to produce a pre-selected number of clusters. One such popular method examined is called the k-means method. Once the number of clusters k is pre-selected, 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 produce 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

Of the 14,604 tracts in the 2007 JAS, there were 4,810 tracts utilized for this project. These tracts represented 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 which were sparse with data but for which we thought could be deterministic (e.g. 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 broad representation of the kind of variables used is shown in Table 1. For a complete list of variables along with their descriptions, see Appendix A.

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.

The 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 utilized an optimization method and the final stage used 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 linkage, 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 assured that no additional weight was given to variables with a larger scale. Log transforms were used in order for the positively skewed continuous variables to meet the normality assumptions.

As previously stated, the cluster procedure in Enterprise Miner used a k-means algorithm to select the cluster seeds in the first stage. In the second stage, the smallest number of clusters was selected 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 (which tested the hypothesis that all data are from the same uniform distribution) had to be greater than the pre-set 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 clusters. This made characterizing the clusters difficult, so the solution from Ward's method was not chosen.

The two separate five cluster solutions from the centroid and average linkage methods were practically identical so either one could have been used for interpretation. This report will show results from the centroid method. The sizes of the clusters in terms of the number of tracts and expanded farms in each cluster are displayed in Table 2. Further clustering results can be found in Appendix B.

Table 2. Cluster sizes for Centroid Method

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 an expanded number of almost 160,000 farm operations. It is characterized by a high quantity of point farms. A point farm is defined as an operation that reports less than $1,000 of sales, but has 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 1 included 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 NML population distribution, while the outer circle shows the cluster distribution.

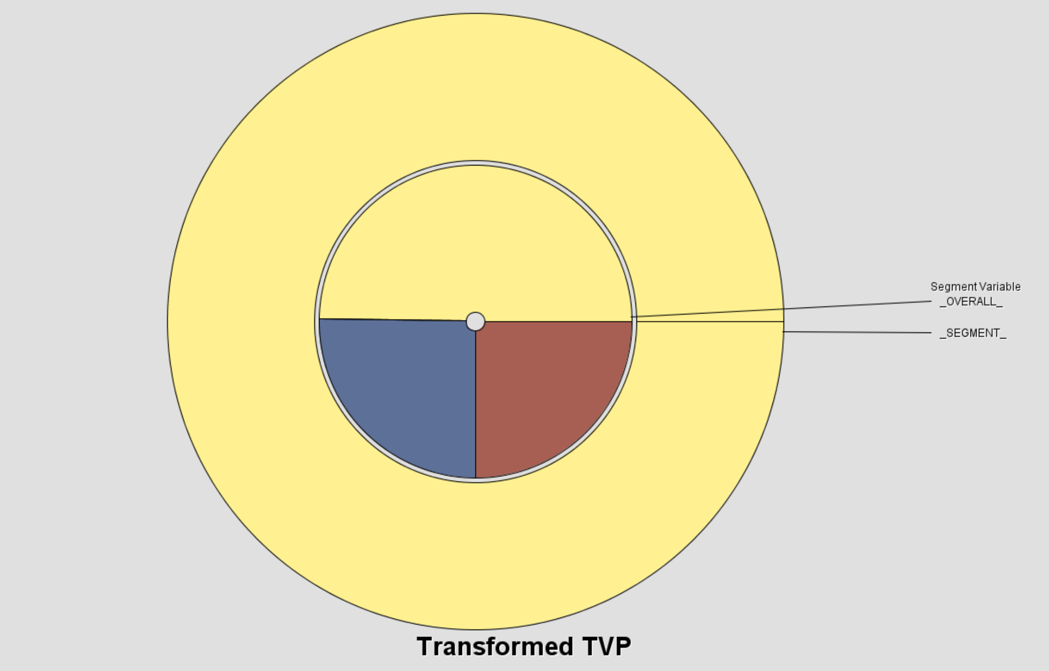


Figure 2. Segment profile of TVP for Cluster 1

Here the yellow indicates a Total Value of Production ranging from $0 to $900. The blue indicates values from $900 - $8,500 and the red represents values above $8,500. 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 is slightly smaller than Cluster 1 and 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. This cluster is of relatively little use since it does not distinguish any unique features of the NML.

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 involved in the operation for more than 20 years.

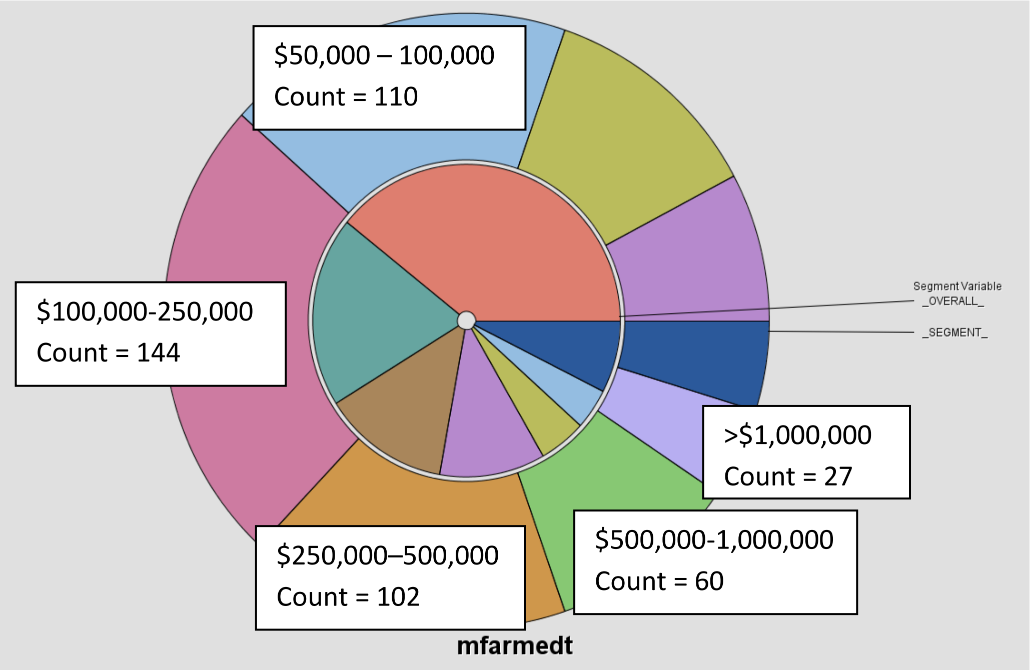


Figure 3. Census final farm value of sales for Cluster 3

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 most of the operations have a sales class of greater than $50,000 with several over $1,000,000.

Cluster 4 is characterized by operations that rented land. These are mostly part-time operations that have not been in operation for a long time. Its defining variables include Land Rented from Others and low Dollar Value of Owned Land.

Lastly, Cluster 5, 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 all variables of interest across the clusters. These variables of interest are not limited to ones used in the clustering procedures. This can provide insight as to additional characteristics that each cluster may possess and ultimately will aid in targeting a specific subgroup. A total of 8 variables of interest were examined across the clusters ranging from operator characteristics to geographic variables.

Table 3. Part-Time operator status across clusters

Cluster Number	1	2	3	4	5	
Description	Point Farms	Typical NML	High Sales	Rented Land	Idle Cropland	Total
Full-time	41,780	43,049	12,595	7,714	6,347	**111,488**
Part-time	116,907	98,003	6,862	10,851	17,574	**250,198**
Total	**158,687**	**141,053**	**19,458**	**18,566**	**23,921**	**361,687**

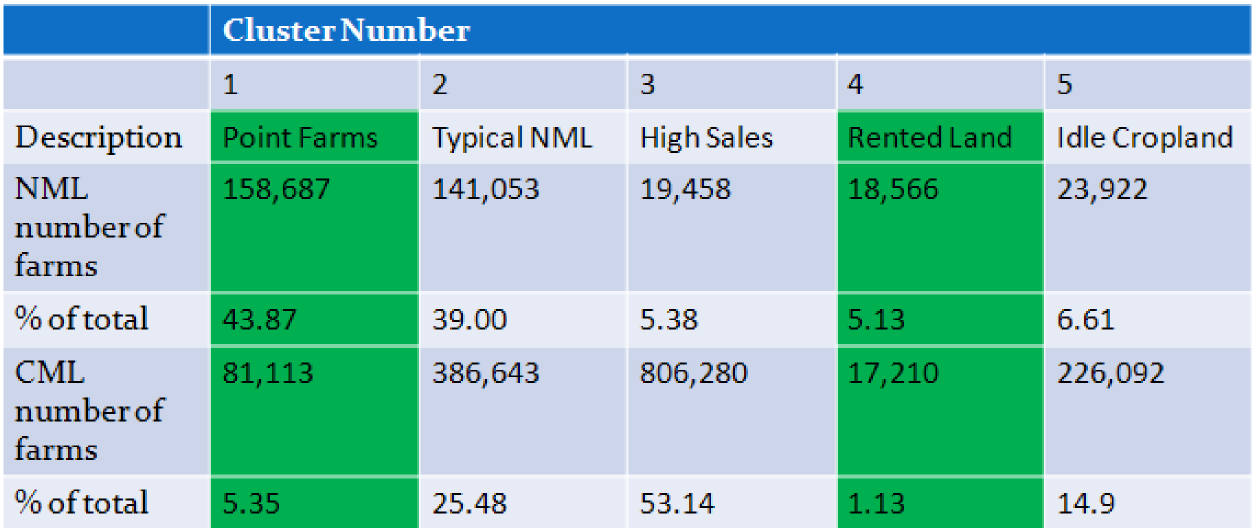
Table 3 shows the expanded number of farms for a binary variable called Part-Time that tells whether an operator is full-time or part-time. Included with the cluster number is a description of the cluster in terms of what best defines it from the NML segment profiling. Although the description does not give the entire picture of the cluster, it is a way to attach a name to the cluster that characterizes it. It can be seen that the majority of the operators across the NML tracts are part-time, 250,198 out of 361,687. However, the number of full-time operators within Cluster 3 (the high sales cluster) is almost double that of part-time operators. This variable illustrates a good example of how differences can be identified by examining variables across the clusters. Results for all variables examined across clusters may be found in Appendix C.

3.1 CML vs. NML comparisons

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 above, to clusters on the CML, formed using the same definitions. Research conducted by Eldridge (2007), compared CML vs. NML for the 2002 Census for a number of characteristics. His research identified the characteristics of records on the NML and whether or not they were properly covered on the CML. This section of the report intends to supplement the 2002 results with information from the clustering to fine tune the CML categories into other areas not previously explored.

Due to the amount of information needed to score cluster definitions to the CML, only respondents from the 2007 Census were used. There were 1,517,338 of these records used from the CML in this analysis. As opposed to using the nonresponse weighted total of CML records, the unweighted records were utilized to avoid any potential effects of nonresponse adjustment bias.

Table 4. NML vs. CML by cluster



	Cluster Number				
	1	2	3	4	5
Description	Point Farms	Typical NML	High Sales	Rented Land	Idle Cropland
NML number of farms	158,687	141,053	19,458	18,566	23,922
% of total	43.87	39.00	5.38	5.13	6.61
CML number of farms	81,113	386,643	806,280	17,210	226,092
% of total	5.35	25.48	53.14	1.13	14.9

Table 4 displays the cluster definitions fitted to the 2007 CML along with the original NML number of farms used to make the clusters. This analysis highlighted 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 in the NML looks concerning, over 40 times the number of operations are assigned to this cluster for the CML. This indicates that high sales operations are well represented, making up over 53 percent of the CML. Highlighted in green in Table 4, the point farms cluster shows that there are roughly double the number of operations in the NML than in the CML within this cluster. This cluster makes up over 43 percent of the total NML while it accounts for only about 5 percent of the CML. Also highlighted is the rented land cluster. While the number of farms in the NML is not much greater for this cluster, it accounts for over 5 percent of the NML compared to over 1 percent of the CML.

Just as specific variables were compared across clusters for the NML population, 8 variables were compared in the same way across the CML. The results are shown in Appendix C.

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 that the 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 CML. However, information such as this also provides a valuable comparison of the NML cluster to the CML. Histograms comparing the CML and NML results for the remaining variables are shown in Appendix D.

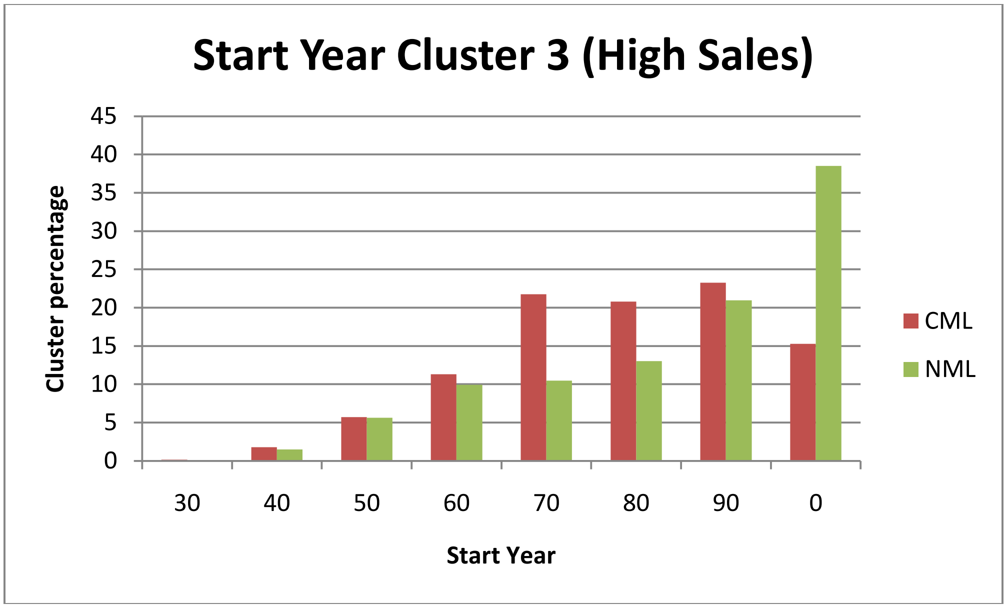


Figure 4. NML vs. CML comparison of Start Year in Cluster 3

4. CONCLUSIONS

Analyzing variables across clusters gives an 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 thus, added to the CML.

The efforts of the cluster analysis have yielded a combination of results, some of which were known anecdotally, 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. Through efforts of examining the details shown in this research, we hope to make improvements to the CML for the 2012 Census of Agriculture.

5. RECOMMENDATIONS

1. Use comparisons between the CML and NML by variable as a gauge for what is missing on the CML and needs to be targeted for list building.
2. Examine needed areas of CML list building using the results of the NML clusters across variables examined and utilize this information to match to outside list sources.

6. REFERENCES

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

Eldridge, H.H. (2007). NASS Census Not on Mail List (NML) Farms. Research and Development Division. RDD Research Report: RDD-07-02.

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

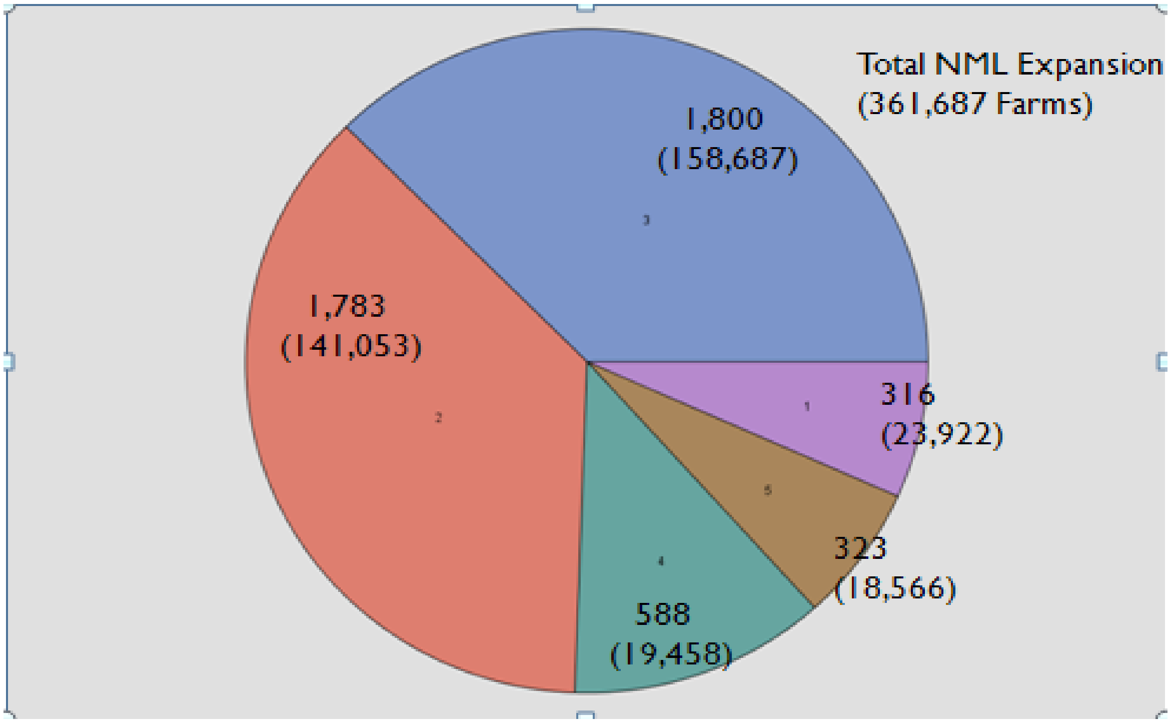
Appendix A. Variables included in the cluster procedures

VARIABLE	DESCRIPTION
EXP_K930	Principal Operator--year Began Operation
FARMTYPE	
K1080	Possible Duplicate -- Y/N?
K1086	Any Other Farm - Y/N?
K1153	Any woodland crops, Y/N
K1157	Any woodland crops, Y/N
K1237	Any Other Livestock?--Y/N
K1671	Type of Organization
K55	Principal County
K924	Principal Operator retired, Y/N
K925	Principal Operator - age
K926	Principal Operator - sex
K927	Principal Operator – Spanish Origin
K928	Principal Operator – Principal Occupation
K9903	Reporting mode code
LOG_CALCPTS	Calculated census points
LOG_CLANDNTR	Non-ag tract acres
LOG_CTRACTAC	Ag tract acres
LOG_FARM_WT	Tract to farm weight
LOG_K1021	Acres from Which All Hay & Forage was Harvested
LOG_K103	Alfalfa Hay Harvested, Acres
LOG_K106	Small Grain Hay Harvested, Acres
LOG_K1062	Cropland Idle or Used for Cover Crops, Acres
LOG_K1229	Layers- table egg types Inventory
LOG_K1347	Total Sales--NUPC (Not Under Production Contract)
LOG_K1501	Operator's (+LL*) Expenditure for Commercial Fertilizer
LOG_K1503	Operator's (+LL) Expenditure for Seeds, Bulbs, Etc
LOG_K1506	Operator's (+LL) Expenditure for Feed
LOG_K1507	Operator's (+LL) Expenditure Dollars for Fuels and Oils
LOG_K1509	Operator's (+LL) Expenditure for Supplies, Repairs, and
LOG_K1513	Operator's (+LL) Cash Rent Paid for Land & Buildings
LOG_K1517	Operator's (+LL) Property Taxes Paid
LOG_K1518	Operator's (+LL) All Other Production Expenses
LOG_K1520	Operator's Depreciation Expenses
LOG_K1540	Operators's (+LL) Total Production Expenses
LOG_K43	Land Owned, Acres
LOG_K44	Land Rented from Others, Acres
LOG_K45	Land Rented to Others, Acres
LOG_K46	Total Acres of Land in This Place

* LL = Landlord

LOG_K685	Government Payments Received from CRP and WRP
LOG_K787	Cropland Harvested, Acres
LOG_K788	Cropland Used for Pasture, Acres
LOG_K790	Cropland on Which All Crops Failed, Acres
LOG_K791	Cropland in Summer Fallow, Acres
LOG_K794	Woodland Pastured, Acres
LOG_K796	Permanent Pasture and Rangeland (Name Change Only from Other Pasture), Acres
LOG_K797	All Other Land, Acres
LOG_K803	Total Cattle and Calf--inventory
LOG_K805	Milk Cow--inventory
LOG_K904	Duck Inventory
LOG_K996	Dollar Value of Owned Land
LOG_TVP	Total value of production
LOG_TVPG	Total value of production minus government payments
LOG_gfarmpnt	Area farm points
MOLNOLAC	Overlap indicator
STRATUM	Area stratum
gfarmdef	Area farm value of sales
gfarmedt	Area edited farm value of sales = edited to include point farms
gfarmtyp	Area type of farm
gqstrsps	Area response code (helps identify refusals and inaccessible)
mdemhisp	Census Hispanic status
mdemoage	Census age of operator
mdemosex	Census gender
mdemrace	Census race indicator
mfarmedt	Census final farm value of sales
yqstrsps	Census response code
Livestock	A binary variable indicating the presence of any one of (K830, K892, K898, K1221, K916, K1225, K910, K908, K820, K852, or K825)
Fruits_nuts	A binary variable indicating the presence of any one of (K1045, K121, K137, K299, K125, K368)

Appendix B. Centroid Cluster Results

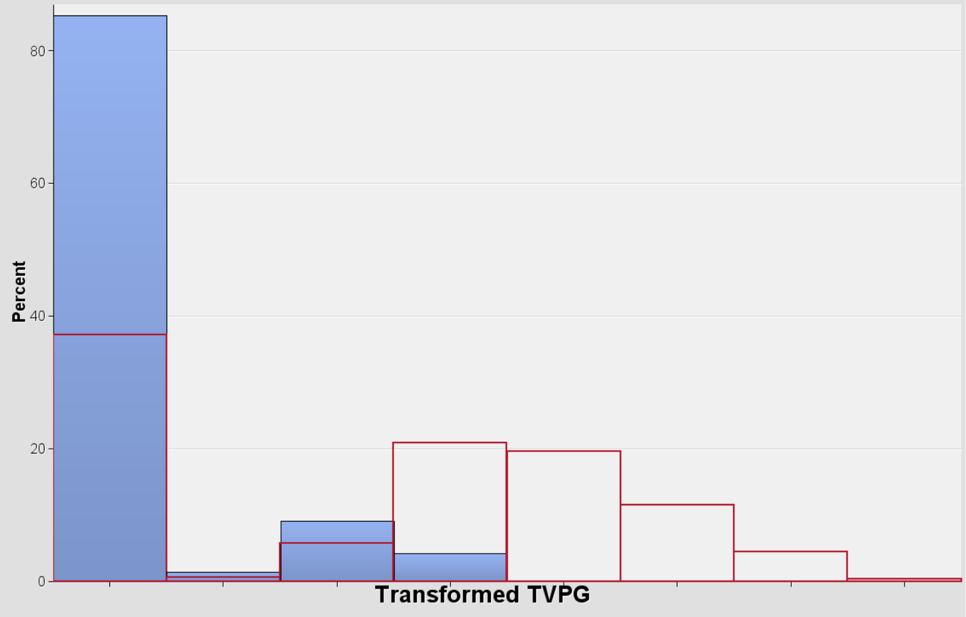


Variable Importance		
Name	**Label**	**Importance**
Log_K1347	Transformed Total Sales	1.0000
Log_TVPG	Transformed TVPG	0.9061
Log_TVP	Transformed TVP	0.8377
Log_1501	Transformed Fertilizer Expenditures	0.8336
Log_K787	Transformed Cropland Harvested	0.8333
Log_CALCPTS	Transformed Calculated Points	0.832

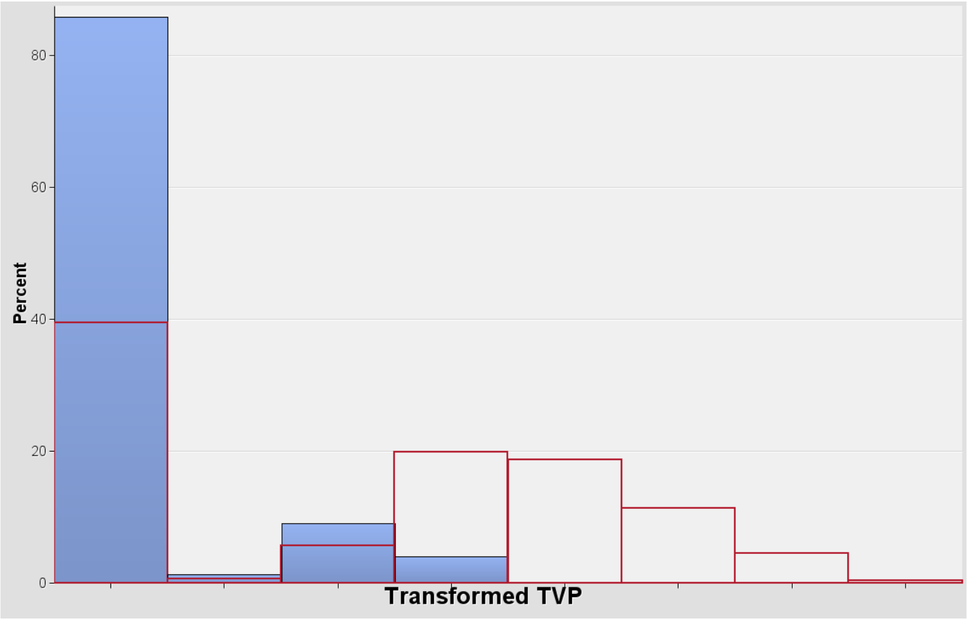
Our five cluster solution is shown in the chart above with the number of tracts as the top number and the expanded number of farms in parenthesis. In the table above are the variables that defined the cluster breaks in order of their importance.

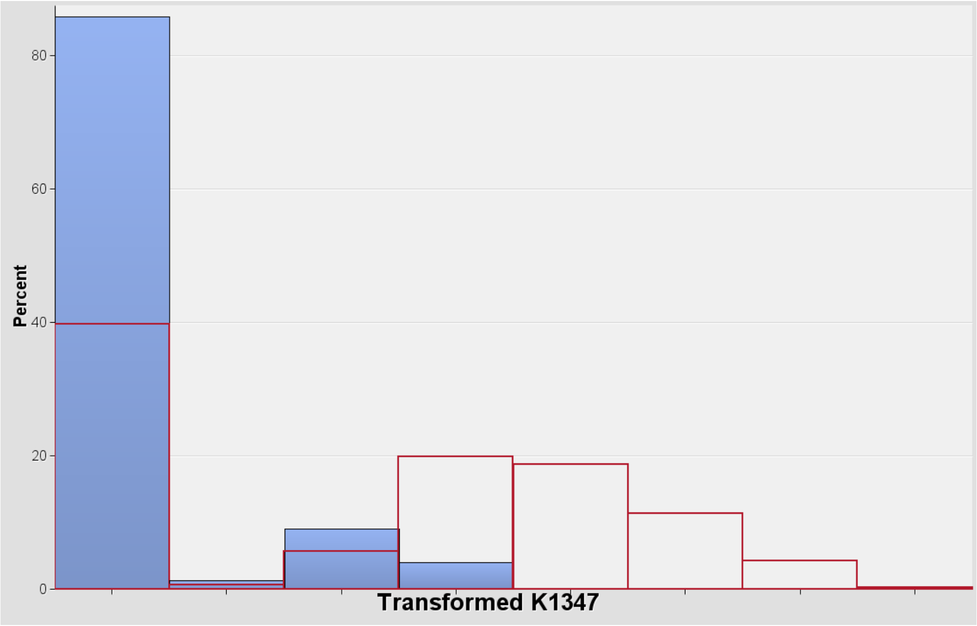
Below are tables of the variables that define each cluster in order of their log worth. Following each table are graphic depictions of the distribution of each variable for that cluster.

Cluster 1 Variable Importance Profile			
Variable	**Label**	**Log Worth**	**Rank**
Log_TVPG	Transformed TVPG	0.331	1
Log_TVP	Transformed TVP	0.304	2
Log_K1347	Transformed Total Sales	0.300	3
Log_CALCPTS	Transformed Calculated Points	0.287	4
FARMTYPE	Farm Type	0.135	5

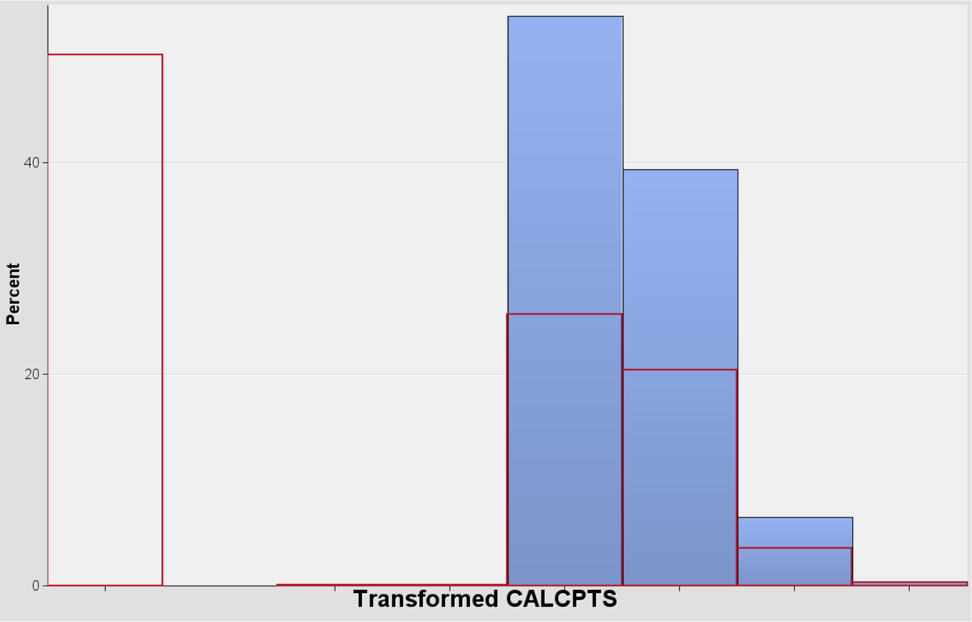


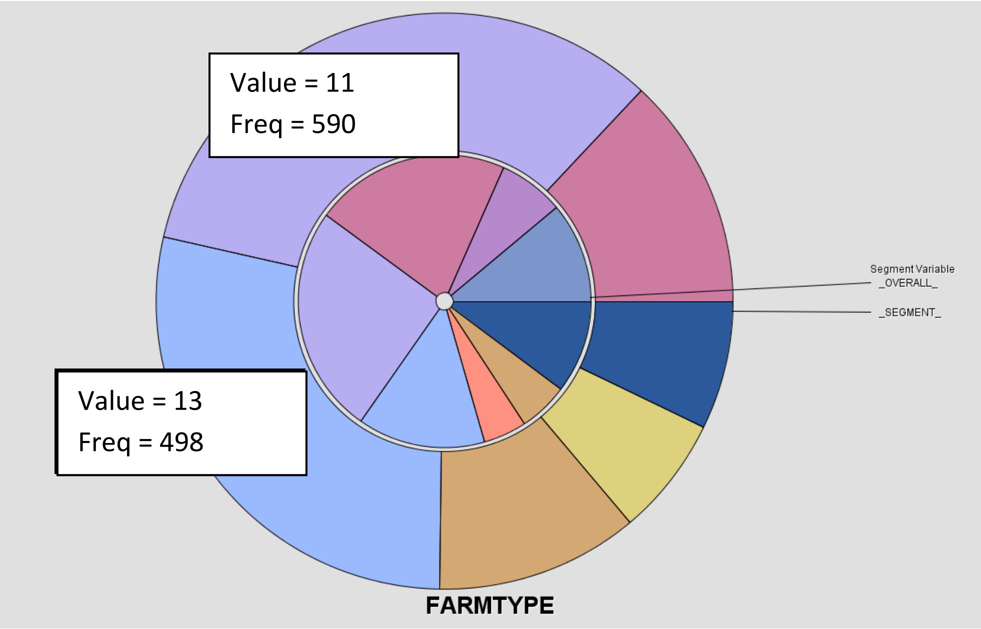
The red outlines show the total NML distribution across transformed TVPG whereas the blue represents the cluster. The large bar on the left represents missing values.





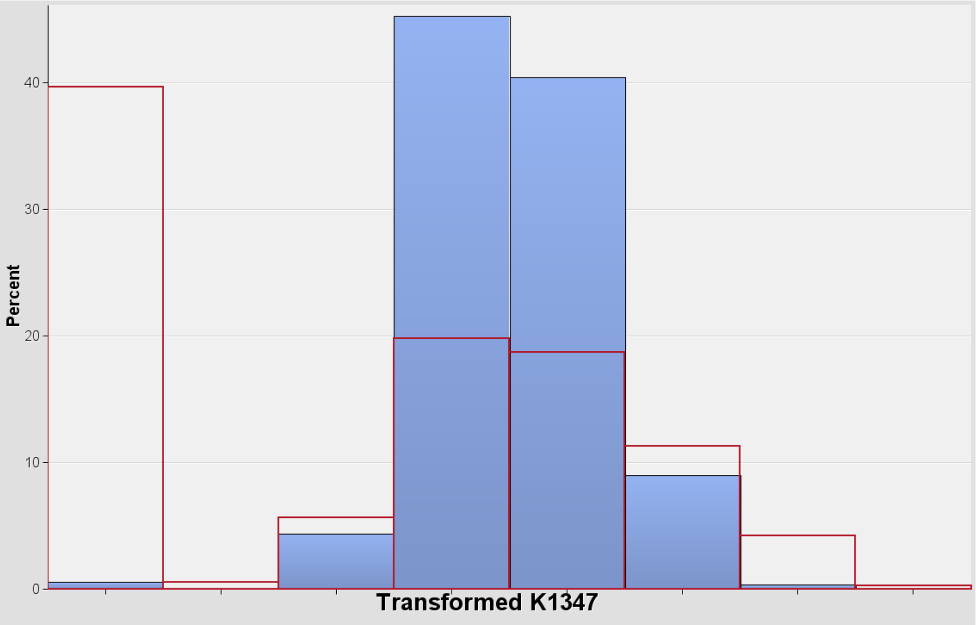
Total Sales - NUPC

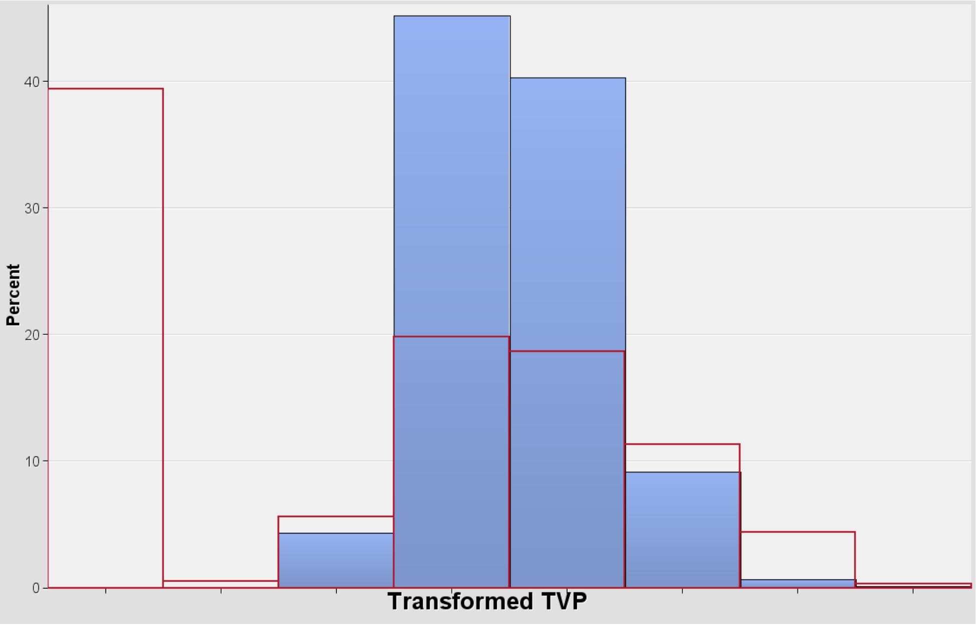


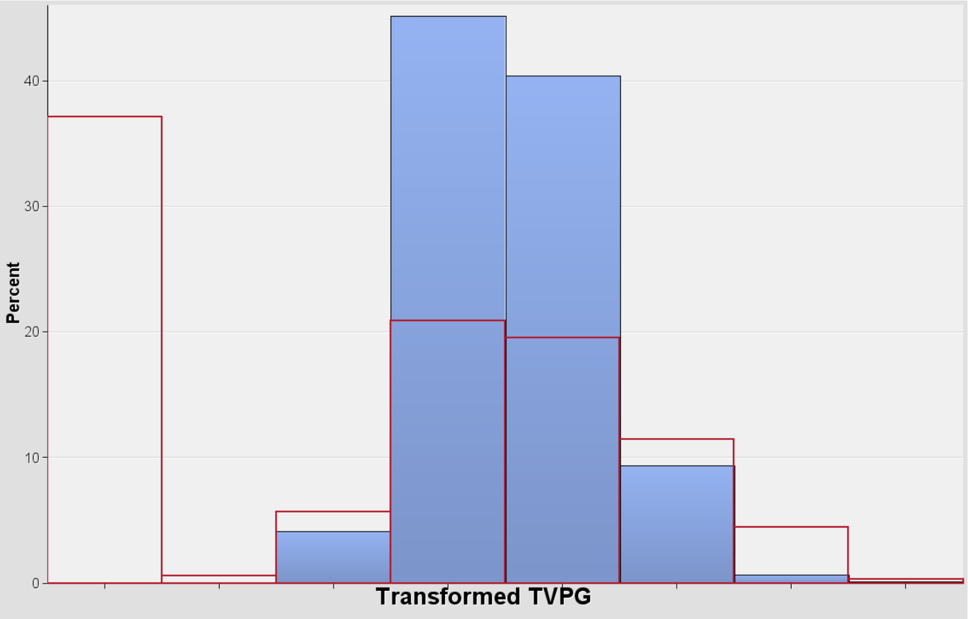


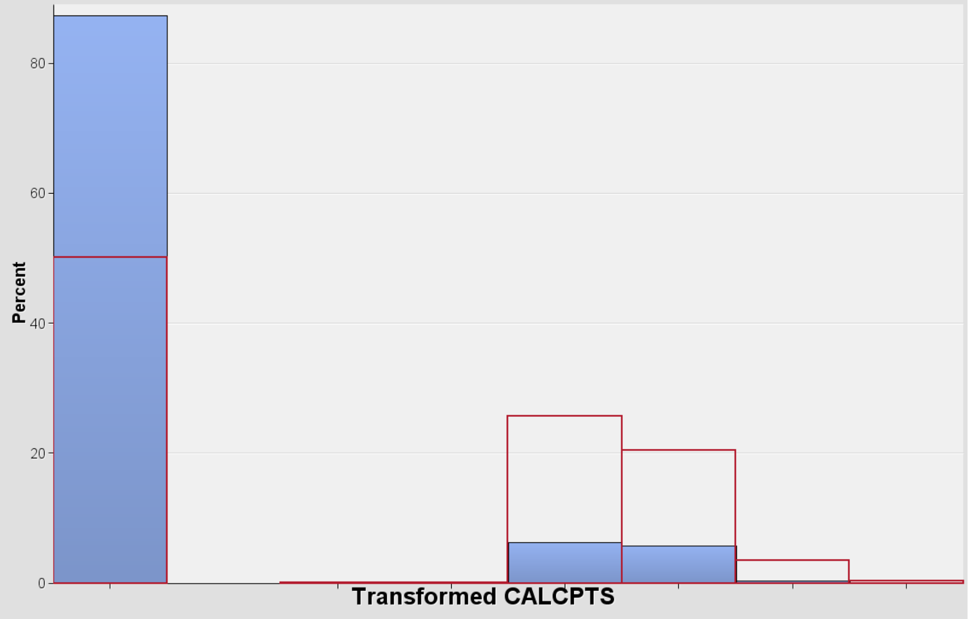
Farmtype 11 is Cattle and Calves
Farmtype 13 is Horse, Ponies, Mules, Burros and Donkeys

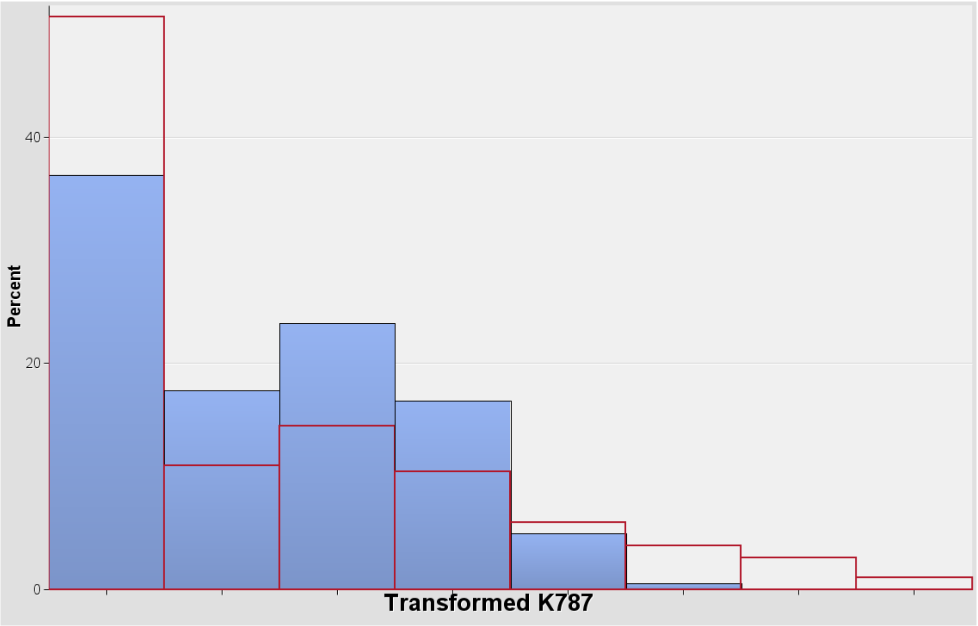
Cluster 2 Variable Importance Profile			
Variable	**Label**	**Log Worth**	**Rank**
Log_1347	Transformed Total Sales	0.286	1
Log_TVP	Transformed TVP	0.286	2
Log_TVPG	Transformed TVPG	0.261	3
Log_CALCPTS	Transformed Calculated Points	0.153	4
Log_787	Transformed Cropland Harvested	0.086	5



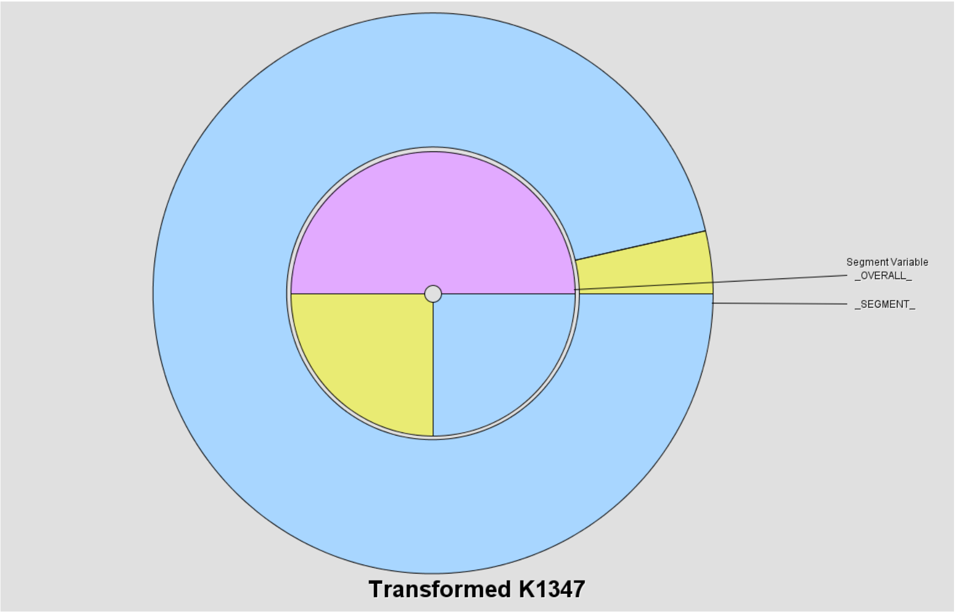



Percent
0
10
20
30
40
Transformed TVPG

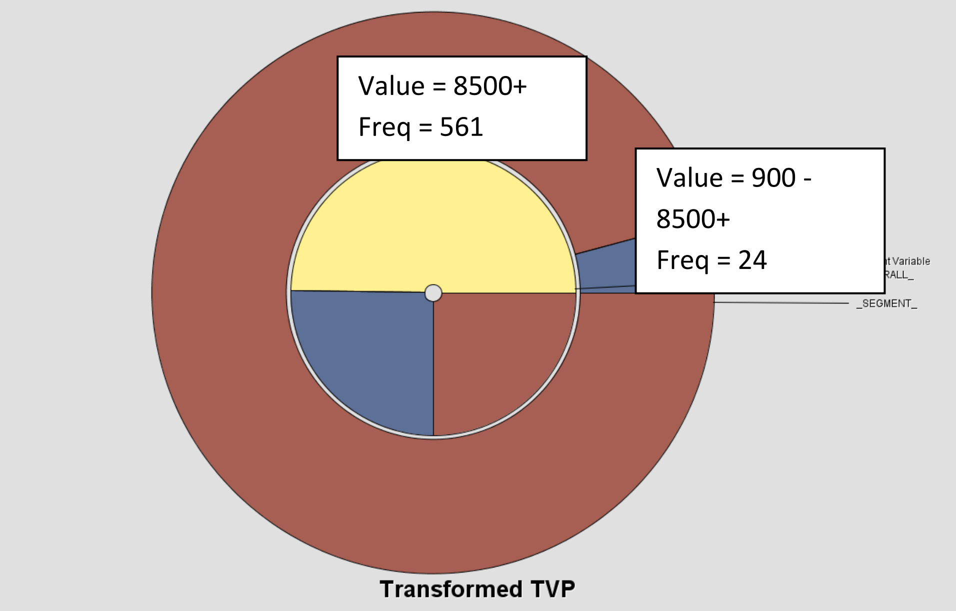

Percent
0
20
40
60
80
Transformed CALCPTS



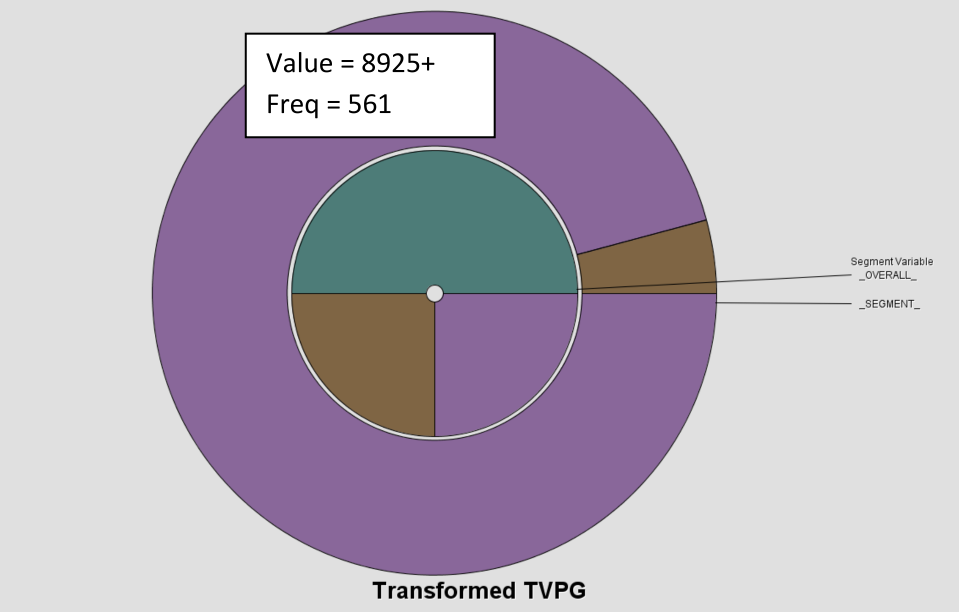
Cluster 3 Variable Importance Profile			
Variable	**Label**	**Log Worth**	**Rank**
Log_K1347	Transformed Total Sales	0.131	1
Log_TVP	Transformed TVP	0.129	2
Log_TVPG	Transformed TVPG	0.129	3
Log_K787	Transformed Cropland Harvested	0.125	4
Mfarmedt	Census Value of Sales	0.124	5



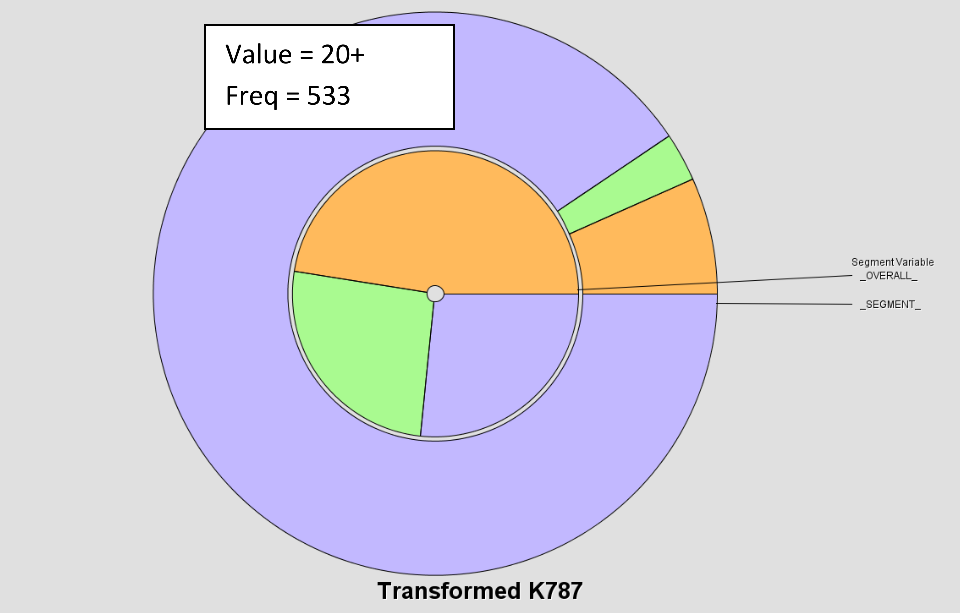
Total Sales - NUPC



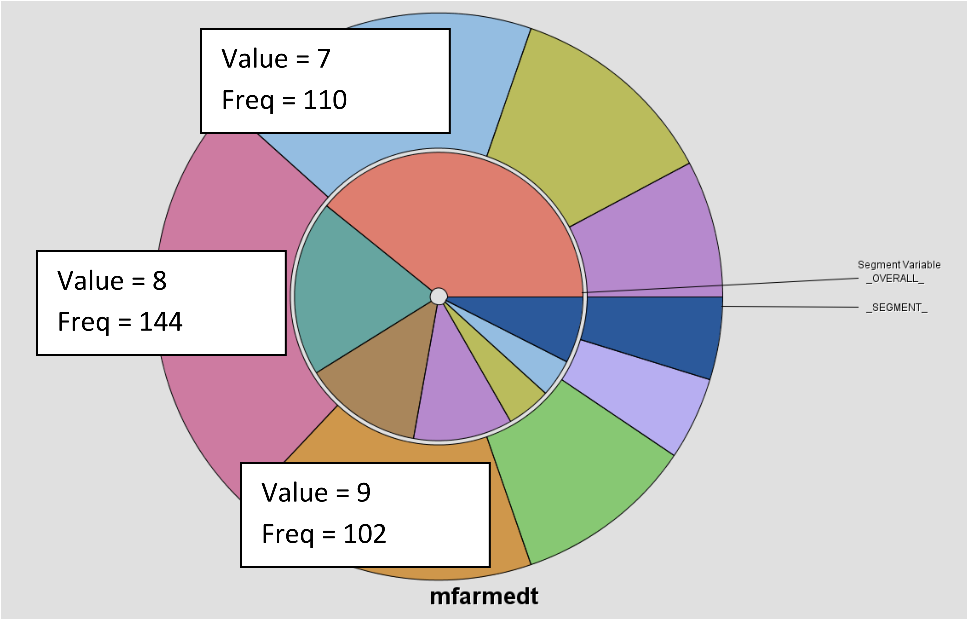
Transformed TVP



Transformed TVPG



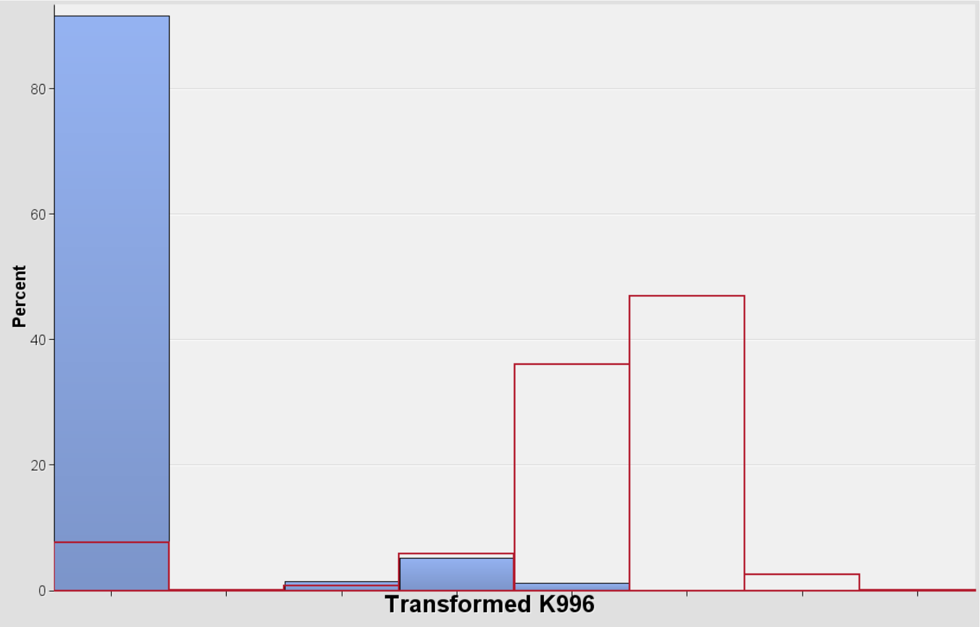
“Cropland Harvested, acres”



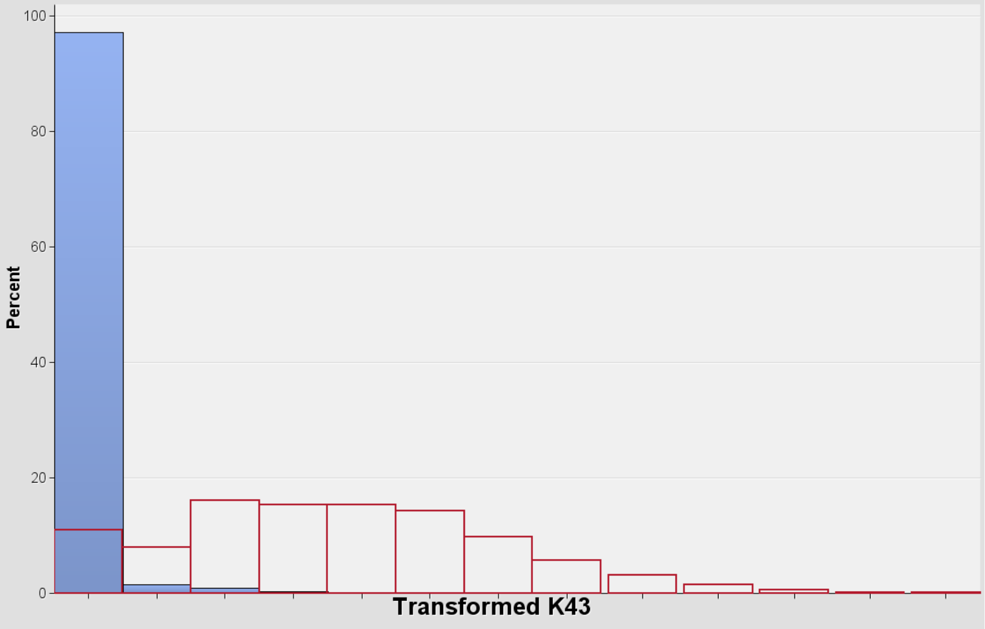
“Census Final Farm Value of Sales”

More than half of the cluster is above the $50,000 category

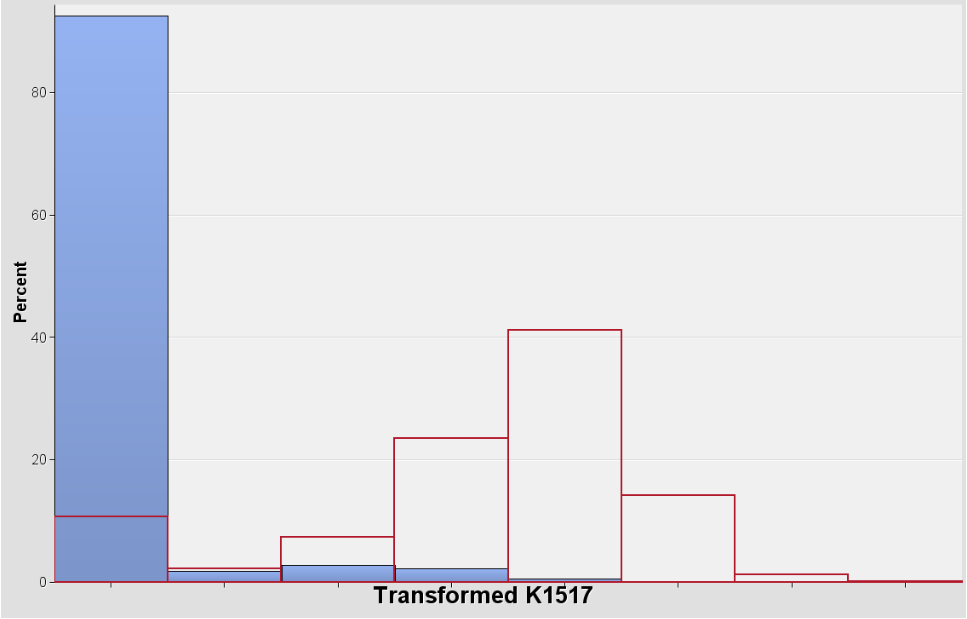
Cluster 4 Variable Importance Profile			
Variable	**Label**	**Log Worth**	**Rank**
Log_K996	Transformed Land Owned Value	0.0927	1
Log_K43	Transformed Land Owned	0.0909	2
Log_K1517	Transformed Property Taxes Paid	0.0656	3
Log_K44	Transformed Land Rented From	0.0347	4
Log_K1513	Transformed Cash Rent Paid	0.0146	5



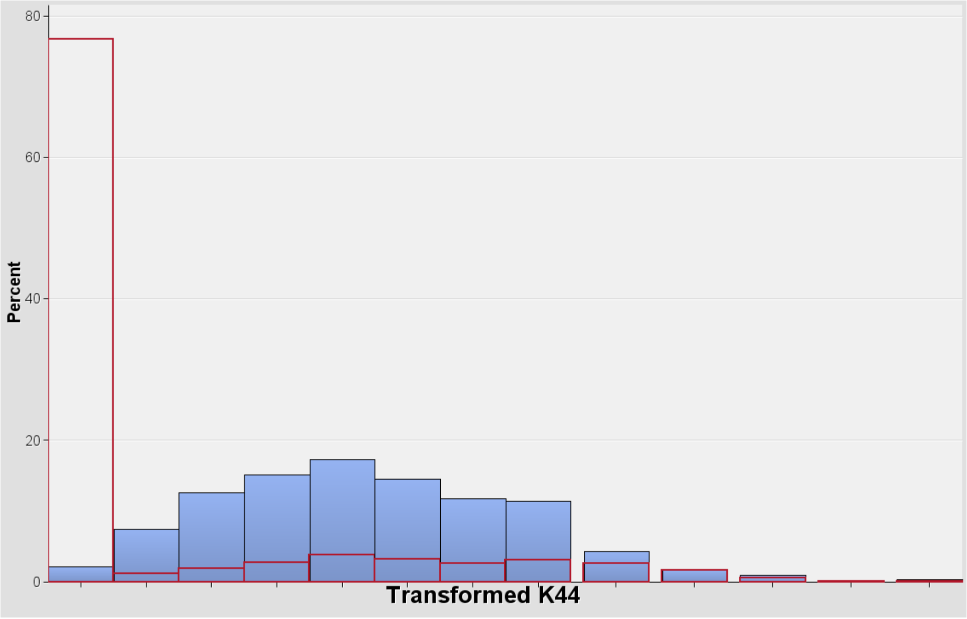
"Dollar Value of Owned Land"



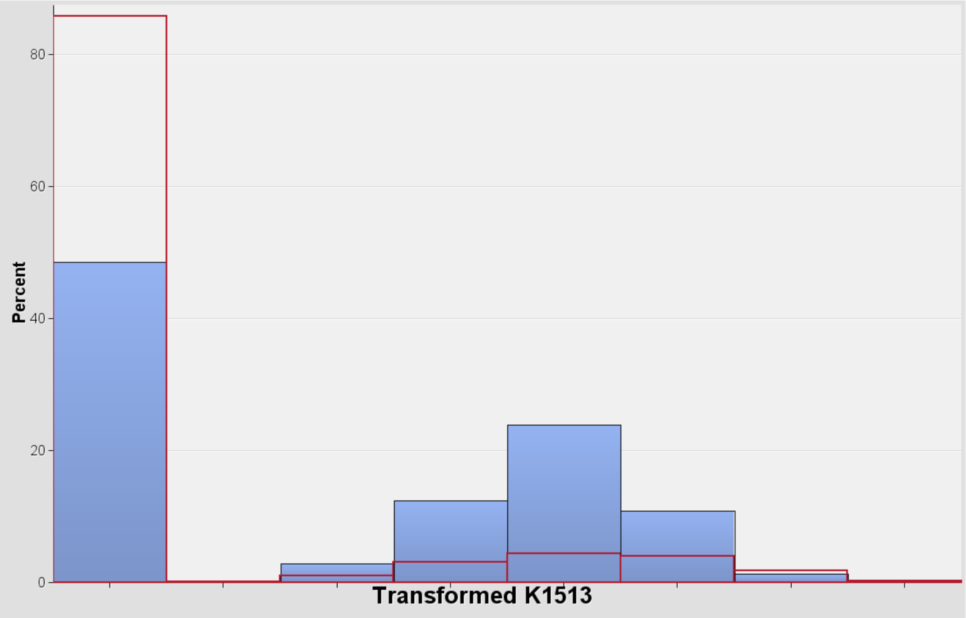
"Land Owned, Acres"



“Operator's (+LL) Property Taxes Paid”

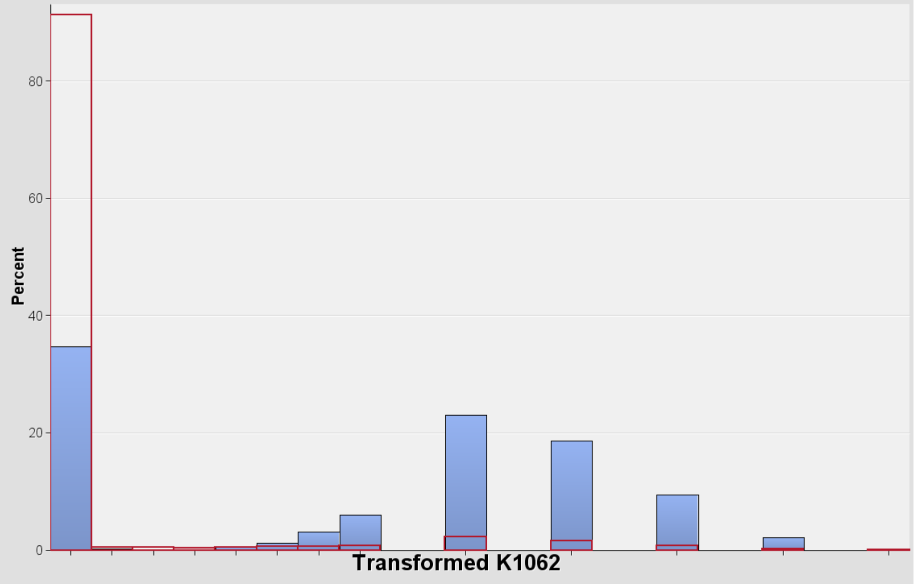


“Land Rented from Others, Acres”

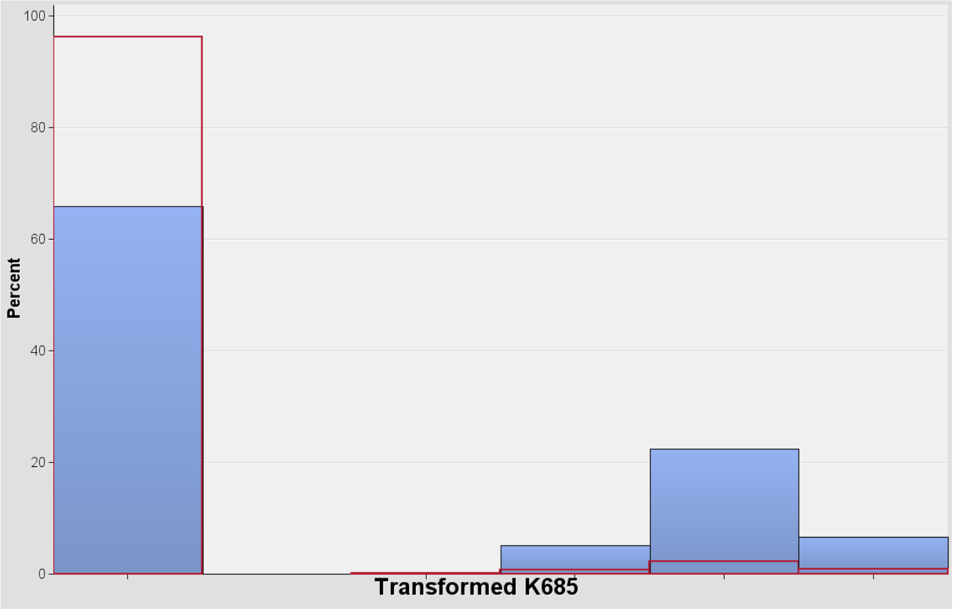


“Operator's (+LL) Cash Rent Paid for Land & Buildings”

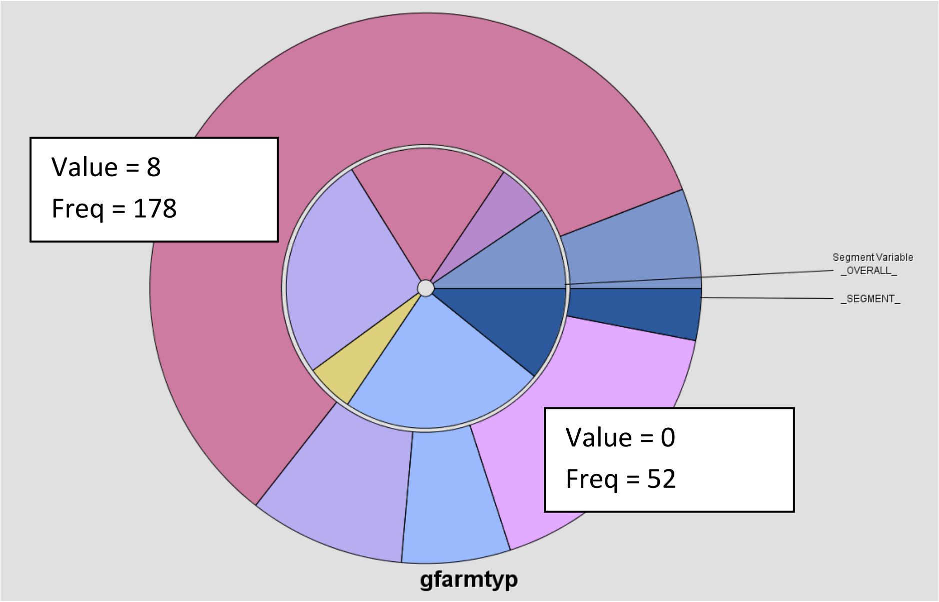
Cluster 5 Variable Importance Profile			
Variable	**Label**	**Log Worth**	**Rank**
Log_K1062	Transformed Cropland Idle	0.0505	1
Log_K685	Transformed Government Payments Received	0.0245	2
Gfarmtyp	JAS farmtype	0.0156	3
FARMTYPE	Farm Type	0.0156	4
Log_K43	Transformed Land Owned	0.0106	5



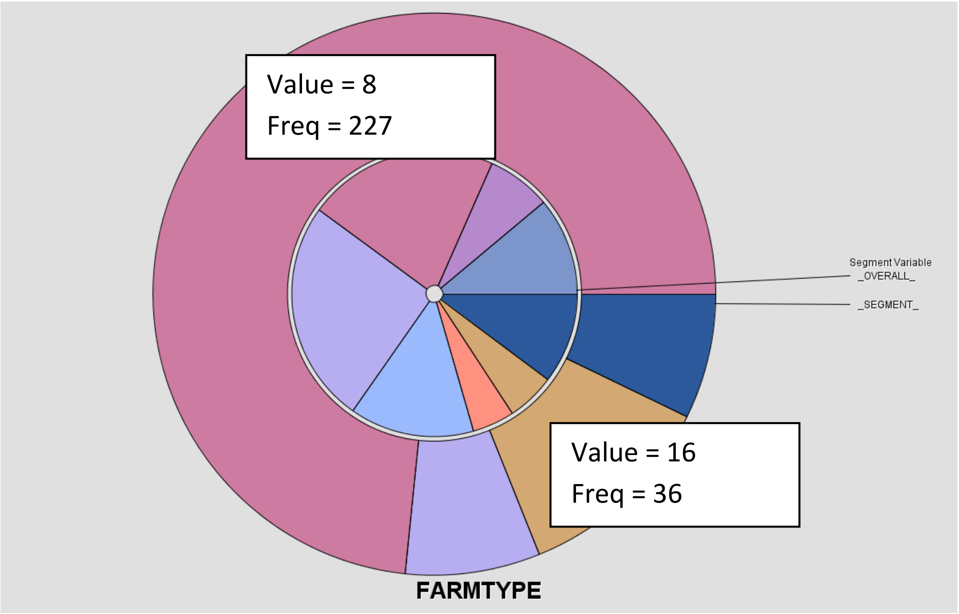
“Idle Cropland”



“Government Payments Received from CRP and WRP”



“June Area Farm Type”



“Census Farm Type”

Appendix C. Cluster frequency tables by variable (Expanded number of farms)

NML Farm Type						
FarmTypes	**Cluster Number**					
	1	2	3	4	5	**Total**
Grain(1)	1293	10927	7766	3423	698	**24107**
Tobacco(2)	0	796	146	0	0	**942**
Cotton(3)	0	159	229	212	0	**600**
Vegetable(4)	69	6859	269	919	0	**8116**
Fruit(5)	1020	13338	994	913	325	**16590**
Nursery(6)	194	2786	185	295	0	**3461**
Christmas Tree(7)	1195	309	8	63	96	**1672**
Other Crop(8)	21820	39552	1548	2897	17983	**83800**
Hog(9)	3470	2907	463	504	31	**7375**
Dairy(10)	0	2723	1216	45	0	**3984**
Cattle(11)	49774	31427	6208	4114	1409	**92932**
Sheep(12)	10097	3796	0	343	187	**14423**
Horse(13)	49454	10328	213	3222	59	**63276**
Poultry(14)	5605	13561	125	767	817	**20875**
Aquaculture(15)	0	33	8	0	34	**75**
Other Animal(16)	14699	1552	80	849	2281	**19460**
Total	**158687**	**141053**	**19458**	**18566**	**23921**	**361687**

NML Gender						
Gender	**Cluster Number**					
	1	2	3	4	5	**Total**
Male	123174	114810	17563	13873	16233	**285653**
Female	35514	26243	1895	4693	7689	**76034**
Total	**158687**	**141053**	**19458**	**18566**	**23921**	**361687**

NML Part-time Operator Status						
Status	**Cluster Number**					
	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**

NML Retired Status						
Status	**Cluster Number**					
	1	2	3	4	5	**Total**
Retired	40677	38466	2954	3241	9670	**95008**
Not Retired	118011	102587	16504	15325	14252	**266678**
Total	**158687**	**141053**	**19458**	**18566**	**23921**	**361687**

NML Spanish Origin						
Status	**Cluster Number**					
	1	2	3	4	5	**Total**
Hispanic	12595	7278	355	1202	261	**21691**
Non-Hispanic	146092	133775	19104	17364	23661	**339995**
Total	**158687**	**141053**	**19458**	**18566**	**23921**	**361687**

NML Start Year						
Decade	**Cluster Number**					
	1	2	3	4	5	**Total**
1930s	0	136	0	0	0	**136**
1940s	922	707	286	0	639	**2555**
1950s	2325	3007	1091	258	1219	**7899**
1960s	5494	5931	1934	475	2389	**16223**
1970s	13569	13810	2039	1442	2995	**33855**
1980s	22090	19311	2531	1778	2327	**48037**
1990s	48268	35173	4082	3555	6950	**98028**
2000+	66019	62989	7495	11058	7403	**154954**
Total	**158687**	**141053**	**19458**	**18566**	**23921**	**361687**

NML State						
State	**Cluster Number**					
	1	2	3	4	5	**Total**
AL(1)	3564	3332	178	113	598	**7785**
AZ(4)	2352	1426	30	206	0	**4015**
AR(5)	2417	2463	384	157	69	**5490**
CA(6)	7403	7264	746	834	427	**16674**
CO(8)	4516	1574	403	177	504	**7175**

CT(9)	295	453	0	0	141	**889**
DE(10)	116	144	15	266	34	**574**
FL(12)	6224	3042	289	293	150	**9997**
GA(13)	4293	3292	579	799	347	**9309**
HI(15)	251	1595	2	609	4	**2461**
ID(16)	2512	1602	123	259	405	**4901**
IL(17)	3541	3998	983	681	1139	**10343**
IN(18)	4040	3684	816	680	569	**9790**
IA(19)	1328	3128	1602	421	757	**7236**
KS(20)	4212	3296	1040	346	524	**9419**
KY(21)	4516	5752	364	665	1578	**12875**
LA(22)	3268	2334	262	112	644	**6620**
ME(23)	777	1301	6	0	42	**2125**
MD(24)	773	1482	44	69	124	**2491**
MA(25)	809	908	11	118	0	**1847**
MI(26)	3073	4200	448	454	666	**8841**
MN(27)	2765	3699	395	256	776	**7890**
MS(28)	2463	1897	310	316	556	**5542**
MO(29)	4679	5252	579	385	1485	**12381**
MT(30)	1238	878	129	242	0	**2487**
NE(31)	781	825	657	288	731	**3283**
NV(32)	292	326	0	0	0	**617**
NH(33)	449	807	43	42	118	**1460**
NJ(34)	826	485	36	80	54	**1480**
NM(35)	1253	2121	48	1368	0	**4790**
NY(36)	2316	2766	705	0	487	**6274**
NC(37)	5202	4102	167	751	248	**10470**
ND(38)	590	888	332	424	1504	**3737**
OH(39)	3283	3966	610	465	130	**8455**
OK(40)	7823	5904	1367	868	1537	**17499**
OR(41)	3008	2784	501	147	214	**6654**
PA(42)	6578	5492	741	742	1801	**15353**
RI(44)	0	403	0	0	0	**403**
SC(45)	1977	1317	477	106	585	**4462**
SD(46)	360	205	790	79	722	**2156**
TN(47)	5216	2788	207	251	313	**8775**
TX(48)	24009	14024	1292	1400	1511	**42235**
UT(49)	628	1708	33	239	127	**2734**
VT(50)	731	786	100	57	144	**1815**
VA(51)	5897	3597	588	537	412	**11032**
WA(53)	6083	6217	174	795	98	**13367**
WV(54)	2987	2650	70	231	451	**6389**
WI(55)	5218	7976	707	786	1171	**15857**
WY(56)	1758	918	78	454	23	**3232**
Total	**158687**	**141053**	**19458**	**18566**	**23921**	**361687**

NML Operator Status						
K1671	**Cluster Number**					
	1	2	3	4	5	**Total**
Family or Individual 1	138354	130147	15548	16404	17527	**317980**
Partnerships 2	9642	7114	1932	695	3020	**22403**
Incorporated 3	8251	2528	1592	896	885	**14152**
Other 4	2440	1265	386	571	2489	**7152**
Total	**158687**	**141053**	**19458**	**18566**	**23921**	**361687**

CML Farm Type						
FarmType	**Cluster Number**					
	1	2	3	4	5	**Total**
Grain(1)	121	17086	224472	1568	8715	**251962**
Tobacco(2)	0	1464	5152	101	129	**6846**
Cotton(3)	1	187	7507	30	130	**7855**
Vegetable(4)	166	14457	11437	704	1249	**28013**
Fruit(5)	1219	33714	28331	405	3505	**67174**
Nursery(6)	38	14622	14682	571	535	**30448**
Christmas Tree(7)	305	5252	2363	86	1956	**9962**
Other Crop(8)	6047	83750	88871	2039	169276	**349983**
Hog(9)	969	5857	12970	263	504	**20563**
Dairy(10)	26	2095	43000	43	66	**45230**
Cattle(11)	27272	128052	301610	7579	20398	**484911**
Sheep(12)	8602	21119	6514	893	2328	**39456**
Horse(13)	28736	31235	20291	1893	3520	**85685**
Poultry(14)	1183	16590	22842	260	1715	**42590**
Aquaculture(15)	3	1357	2283	15	128	**3786**
Other Animal(16)	6425	9806	13955	760	11938	**42884**
Total	**81113**	**386643**	**806280**	**17210**	**226092**	**1517338**

CML Gender						
Gender	**Cluster Number**					
	1	2	3	4	5	**Total**
Male	62051	323245	749127	14414	180009	**1328846**
Female	19062	63398	57153	2796	46083	**188492**
Total	**81113**	**386643**	**806280**	**17210**	**226092**	**1517338**

CML Part-time Status						
Status	**Cluster Number**					
	1	2	3	4	5	**Total**
Full-time	20526	113370	535303	4783	49566	**723548**
Part-time	60587	273273	270977	12427	176526	**793790**
Total	**81113**	**386643**	**806280**	**17210**	**226092**	**1517338**

CML Retired Status						
Status	**Cluster Number**					
	1	2	3	4	5	**Total**
Retired	19475	126891	174181	2102	102059	**424708**
Not Retired	61638	259752	632099	15108	124033	**1092630**
Total	**81113**	**386643**	**806280**	**17210**	**226092**	**1517338**

CML Spanish Origin						
Status	**Cluster Number**					
	1	2	3	4	5	**Total**
Hispanic	2889	9879	10779	475	2110	**26132**
Non-Hispanic	78224	376764	795501	16735	223982	**1491206**
Total	**81113**	**386643**	**806280**	**17210**	**226092**	**1517338**

CML Start Year						
Decade	**Cluster Number**					
	1	2	3	4	5	**Total**
1930s	31	267	1123	3	821	**2245**
1940s	330	2779	14500	30	6735	**24374**
1950s	1035	8581	45972	126	14258	**69972**
1960s	2872	20767	91171	328	21241	**136379**
1970s	7904	48712	175476	1046	34741	**267879**
1980s	13561	70644	167527	2156	39405	**293293**
1990s	26778	121315	187422	5240	60183	**400938**
2000+	28594	113556	123064	8280	48685	**322179**
Total	**81113**	**386643**	**806280**	**17210**	**226092**	**1517338**

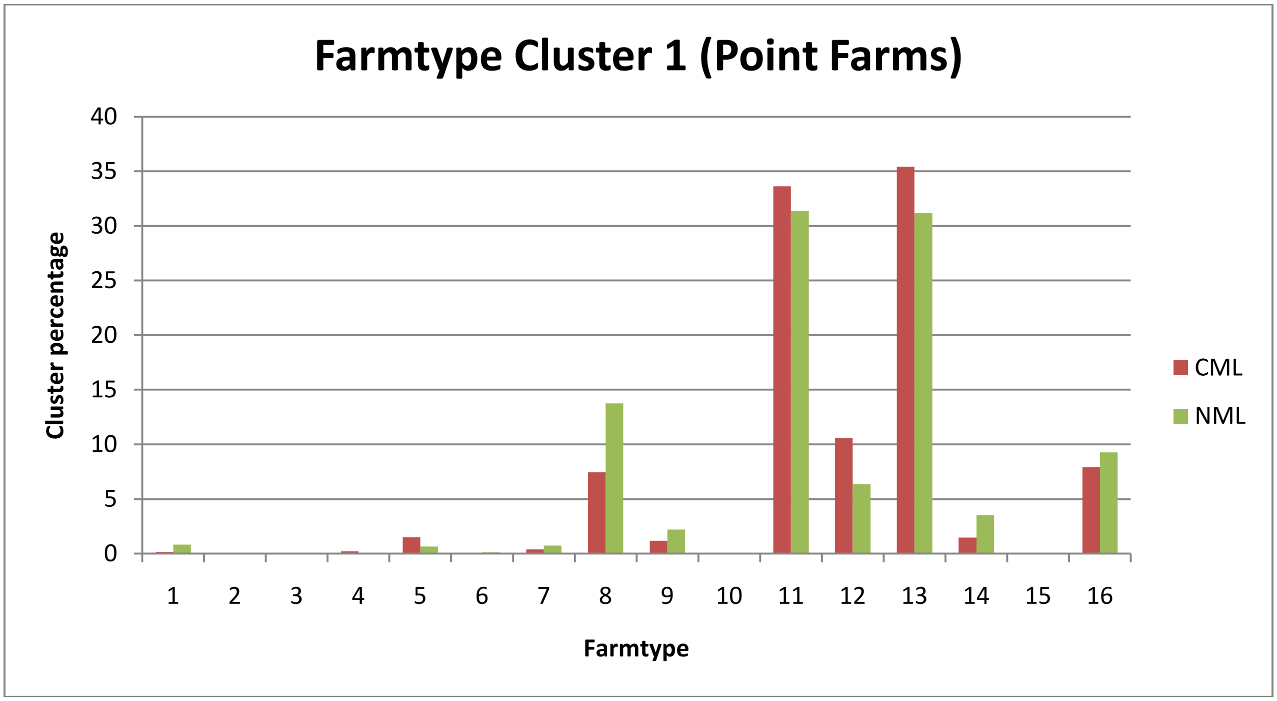
CML State						
State	**Cluster Number**					
	1	2	3	4	5	**Total**
AL(1)	2204	9791	15539	622	5367	**33523**
AZ(4)	2654	5155	2514	134	305	**10762**
AR(5)	2308	9551	20894	423	3693	**36869**
CA(6)	2470	19706	27674	596	1554	**52000**
CO(8)	2121	6358	13231	337	3899	**25946**
CT(9)	121	1446	1150	47	81	**2845**
DE(10)	70	491	1025	22	149	**1757**
FL(12)	3151	13795	11633	219	1717	**30515**
GA(13)	1819	7948	14833	266	5275	**30141**
HI(15)	992	5068	8892	191	1983	**17126**
ID(16)	1049	7641	33190	502	13169	**55551**
IL(17)	1824	10396	24090	347	6901	**43558**
IN(18)	991	7236	46222	523	17126	**72098**
IA(19)	992	5577	29621	411	9521	**46122**
KS(20)	3338	17921	28161	682	8451	**58553**
KY(21)	1344	5522	8894	452	3018	**19230**
LA(22)	189	1857	1974	63	610	**4693**
ME(23)	368	2462	4676	140	1306	**8952**
MD(24)	288	2170	1782	85	120	**4445**
MA(25)	1674	9856	20126	251	7825	**39732**
MI(26)	1145	7126	34918	337	15413	**58939**
MN(27)	1487	5971	12872	404	7966	**28700**
MS(28)	3632	16481	43942	699	13151	**77905**
MO(29)	1526	3804	12909	269	3833	**22341**
MT(30)	785	3401	26684	436	5344	**36650**
NE(31)	190	616	1233	26	105	**2170**

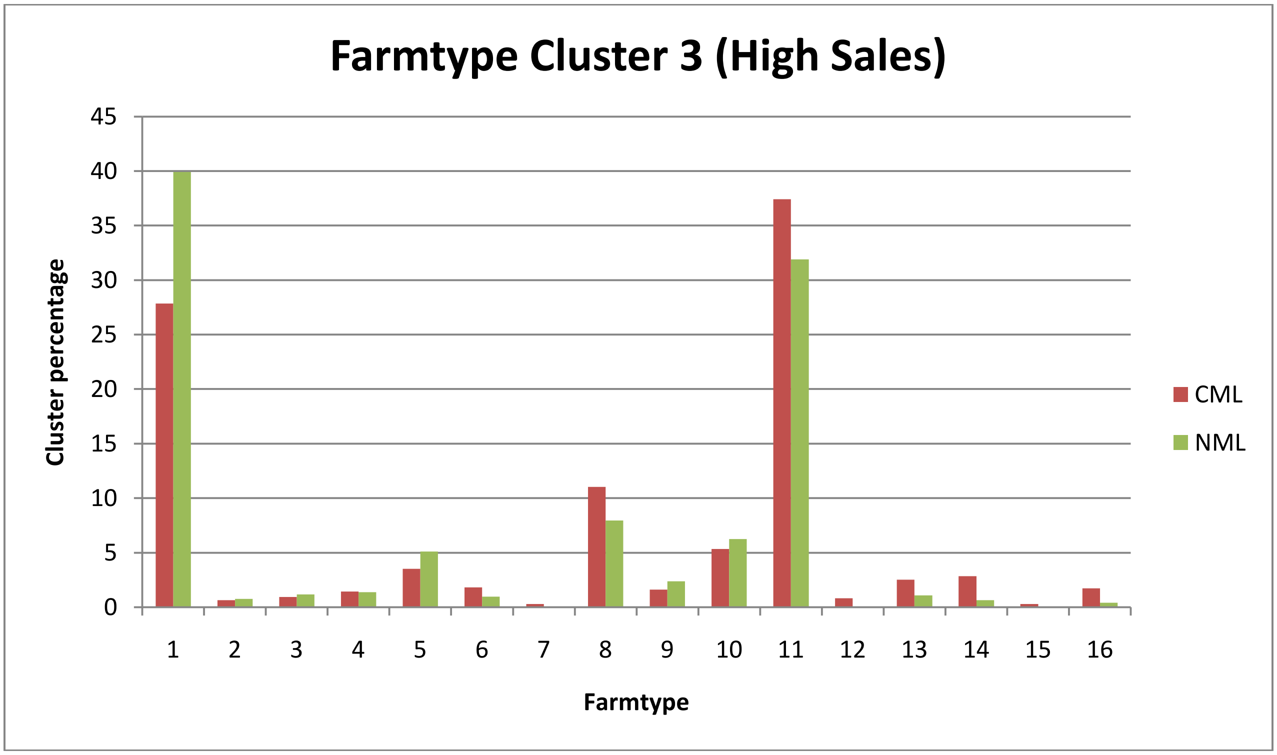
NV(32)	152	1162	920	44	123	**2401**
NH(33)	333	4365	2522	108	237	**7565**
NJ(34)	1716	4310	5353	284	1440	**13103**
NM(35)	1111	5673	14821	199	2354	**24158**
NY(36)	1723	11714	17498	569	3743	**35247**
NC(37)	186	687	14659	206	7150	**22888**
ND(38)	2112	15209	29866	539	7623	**55349**
OH(39)	3154	12415	32909	886	5499	**54863**
OK(40)	1874	12179	10478	375	1592	**26498**
OR(41)	1701	10710	21494	382	6389	**40676**
PA(42)	42	437	311	17	16	**823**
RI(44)	1233	5618	6656	223	3748	**17478**
SC(45)	495	1812	16727	230	3980	**23244**
SD(46)	4091	22357	23945	547	7278	**58218**
TN(47)	15111	50301	82893	2657	17504	**168466**
TX(48)	1100	4296	5222	224	795	**11637**
UT(49)	154	1366	2234	53	234	**4041**
VT(50)	1683	9286	15812	334	2710	**29825**
VA(51)	1697	9189	10564	251	2490	**24191**
WA(53)	883	6287	6459	191	1190	**15010**
WV(54)	1365	8642	31022	318	11367	**52714**
WI(55)	465	1282	5236	89	748	**7820**
Total	**81113**	**386643**	**806280**	**17210**	**226092**	**1517338**

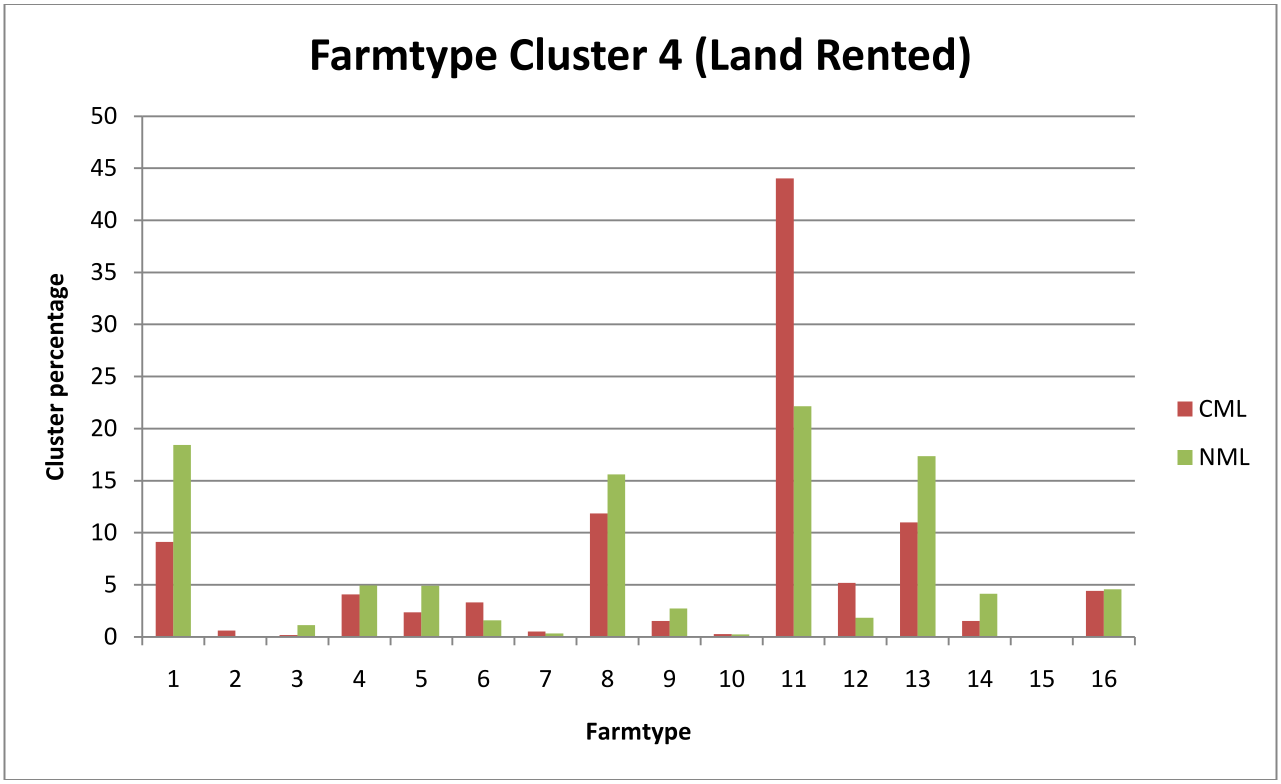
CML Operator Status

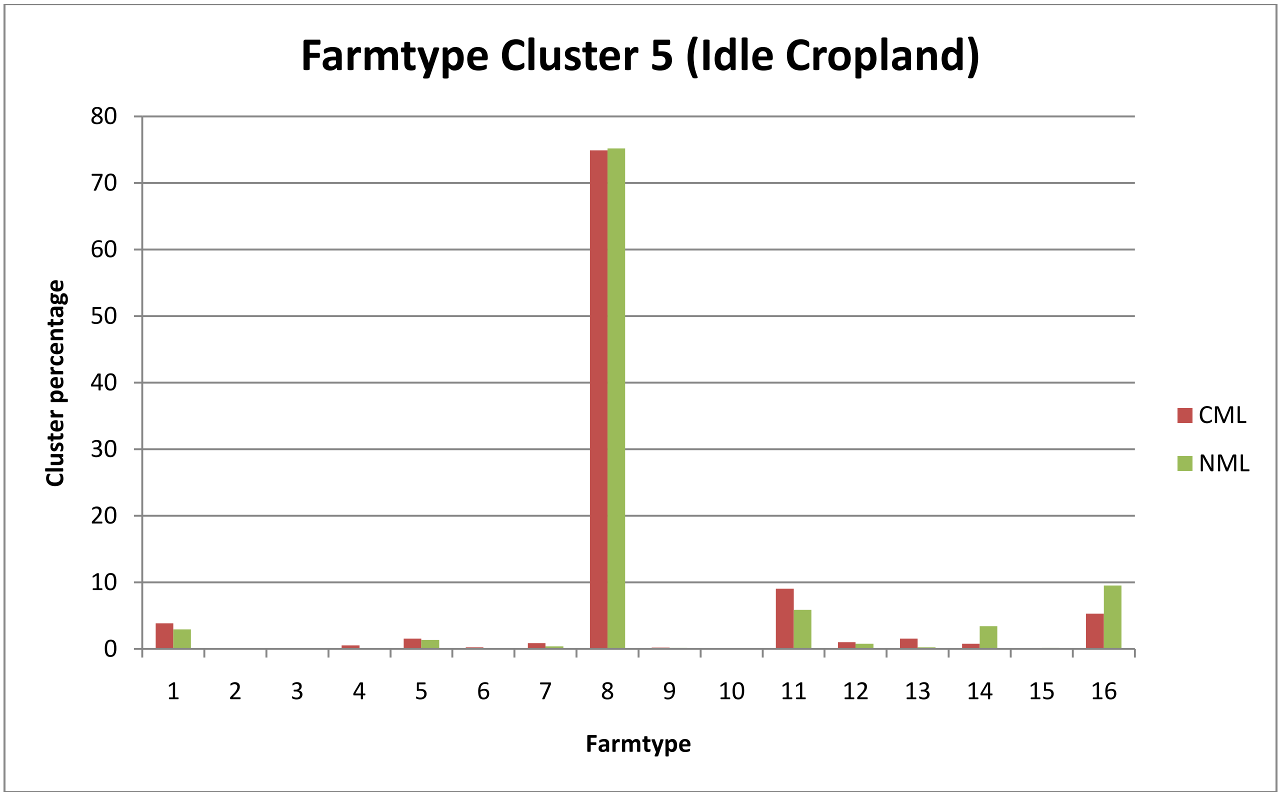
K1671	**Cluster Number**					
	1	2	3	4	5	**Total**
Family or Individual 1	75477	361587	651374	15492	190242	**1294172**
Partnerships 2	3616	16474	86961	988	20109	**128148**
Incorporated 3	1506	6049	59560	473	6970	**74558**
Other 4	514	2533	8385	257	8771	**20460**
Total	**81113**	**386643**	**806280**	**17210**	**226092**	**1517338**

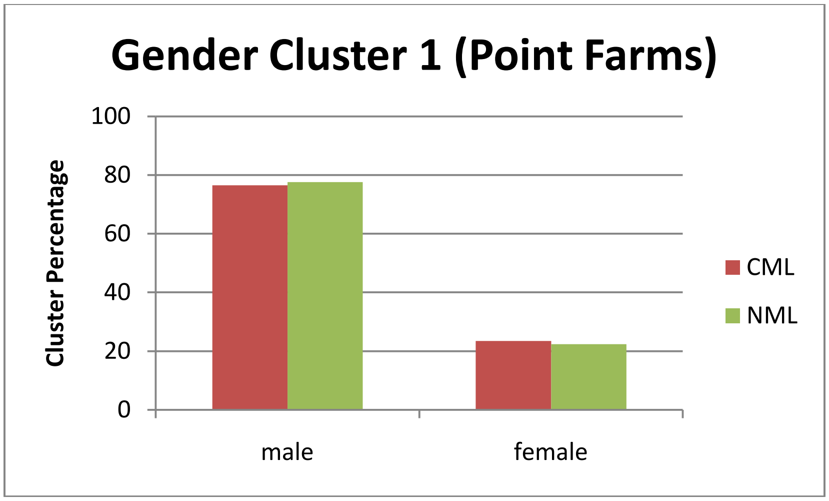
Appendix D. CML-NML comparison graphs

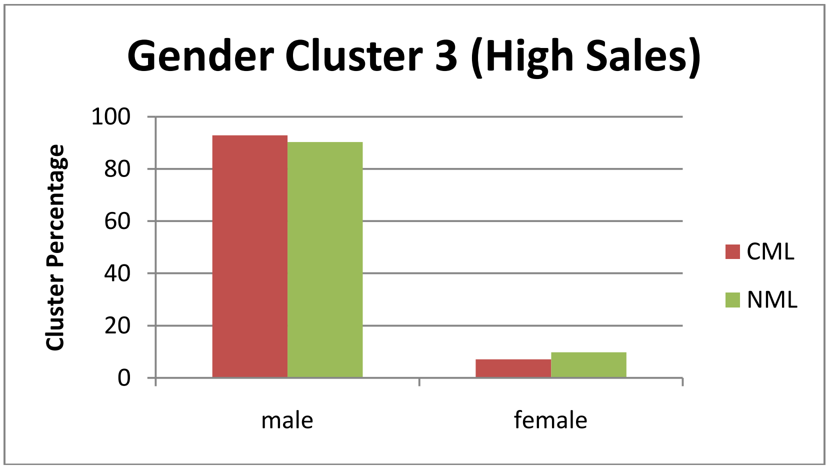


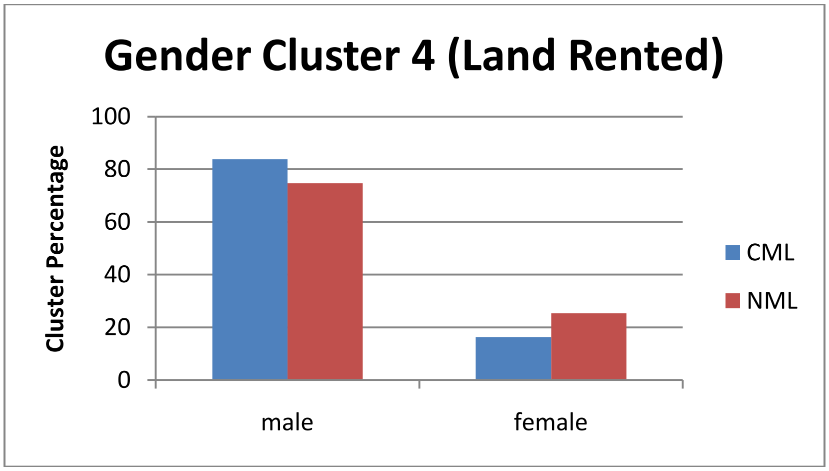


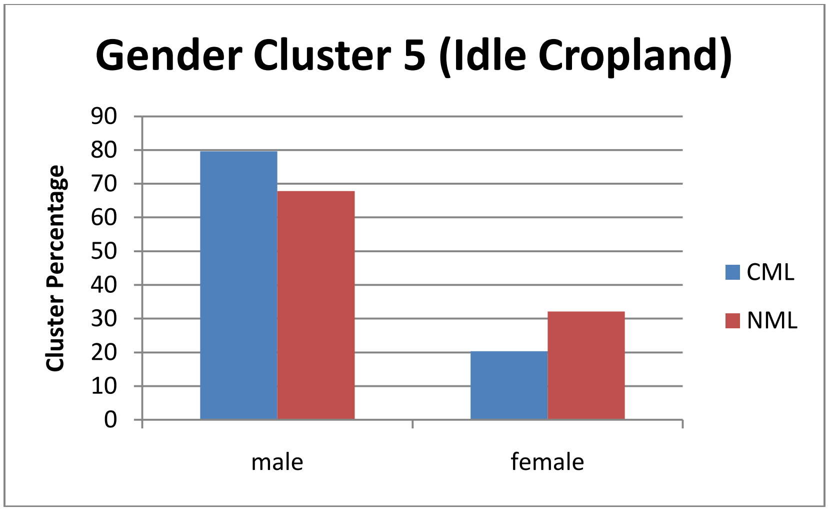

Farmtype Cluster 4 (Land Rented)
Cluster percentage
50
45
40
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Farmtype
🟥 CML
🟩 NML

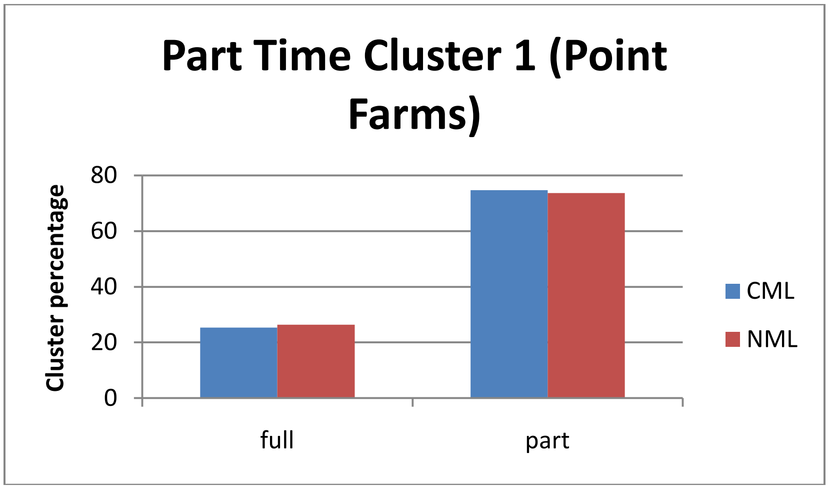

Farmtype Cluster 5 (Idle Cropland)
Cluster percentage
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Farmtype
🟥 CML
🟩 NML

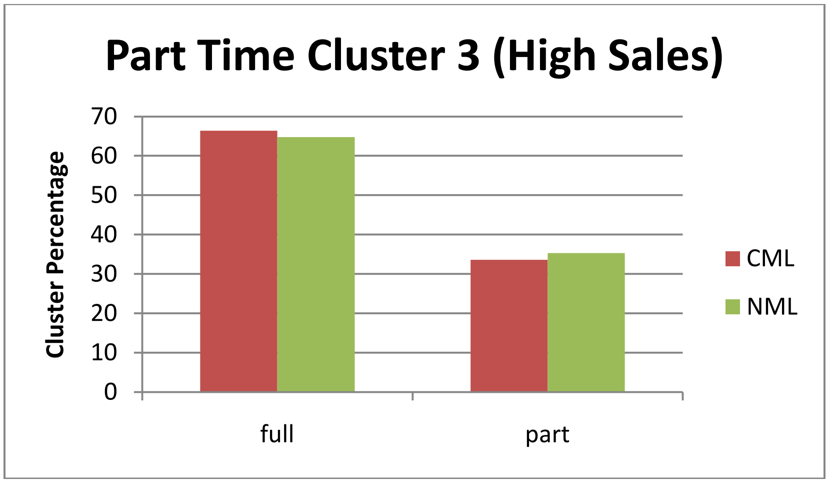

Gender Cluster 1 (Point Farms)
Cluster Percentage
100
80
60
40
20
0
male
female
CML
NML

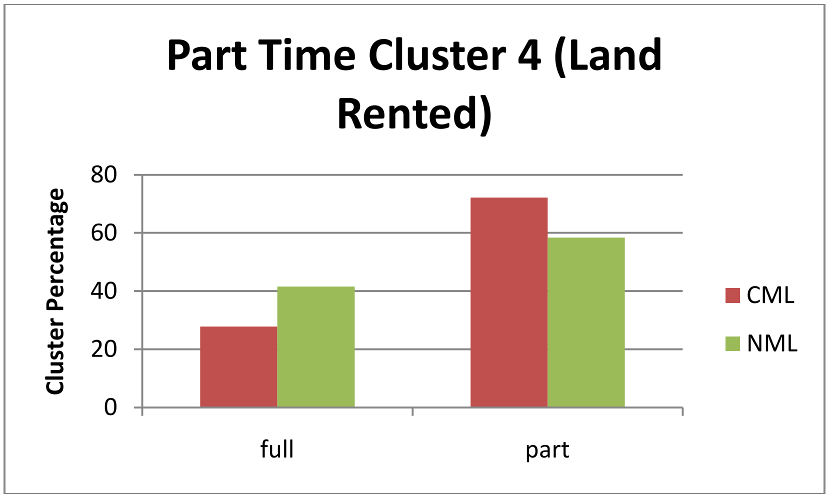

Gender Cluster 3 (High Sales)
Cluster Percentage
100
80
60
40
20
0
male
female
CML
NML

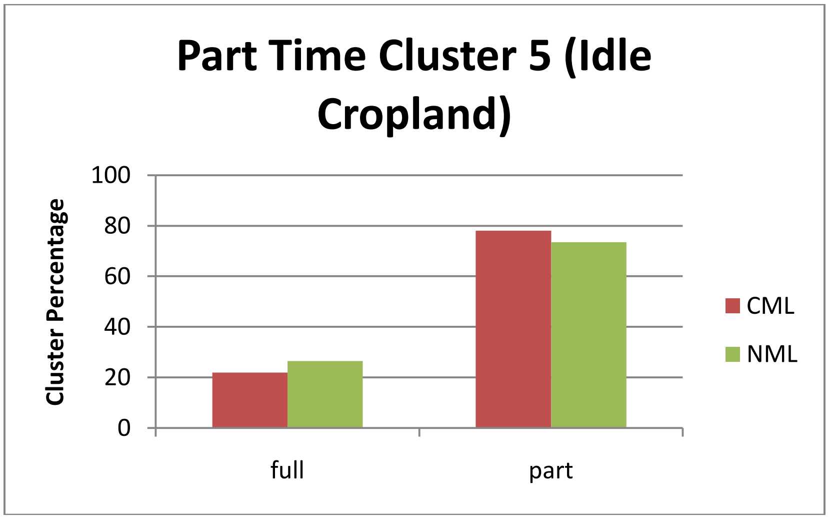

Gender Cluster 4 (Land Rented)
Cluster Percentage
100
80
60
40
20
0
male
female
CML
NML

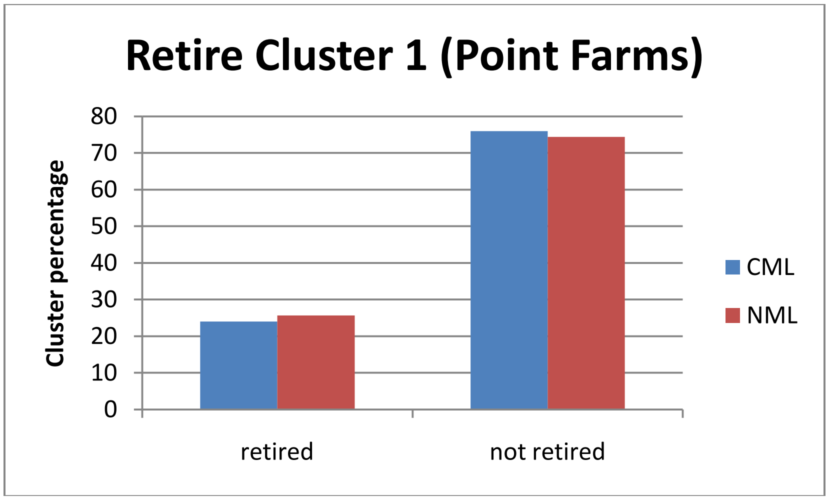

Gender Cluster 5 (Idle Cropland)
Cluster Percentage
90
80
70
60
50
40
30
20
10
0
male
female
CML
NML

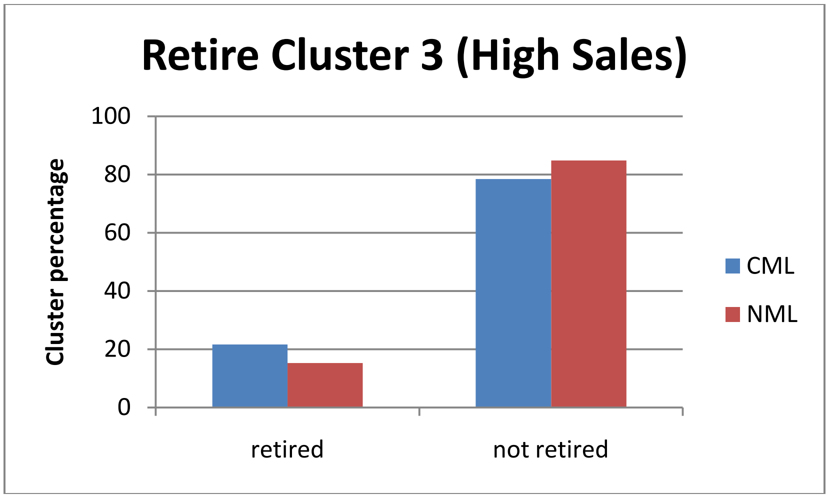

Part Time Cluster 1 (Point Farms)
Cluster percentage
80
60
40
20
0
full
part
CML
NML

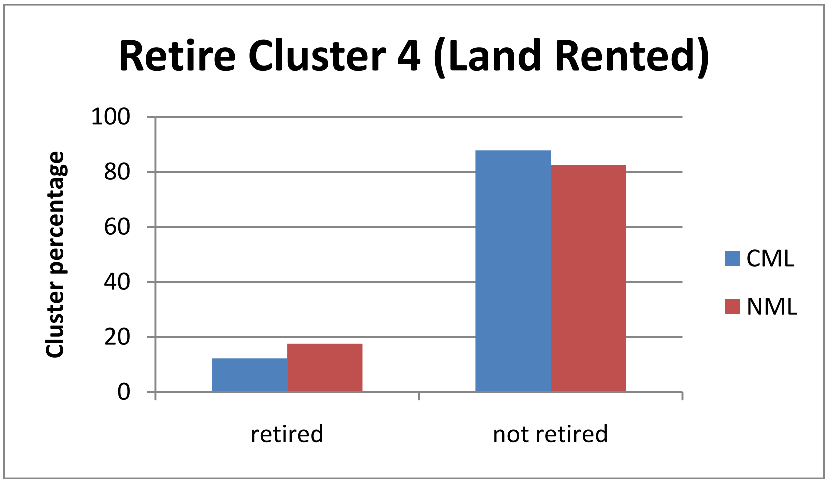

Part Time Cluster 3 (High Sales)
Cluster Percentage
70
60
50
40
30
20
10
0
full
part
CML
NML

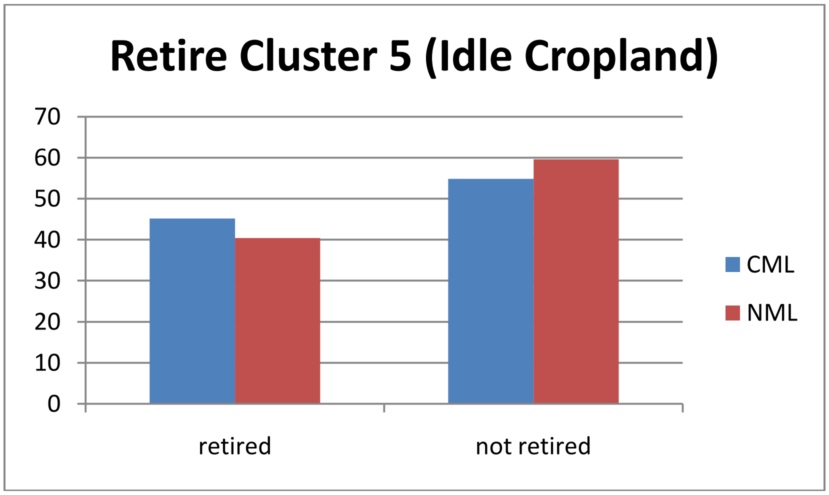

Part Time Cluster 4 (Land Rented)
Cluster Percentage
80
60
40
20
0
full
part
CML
NML

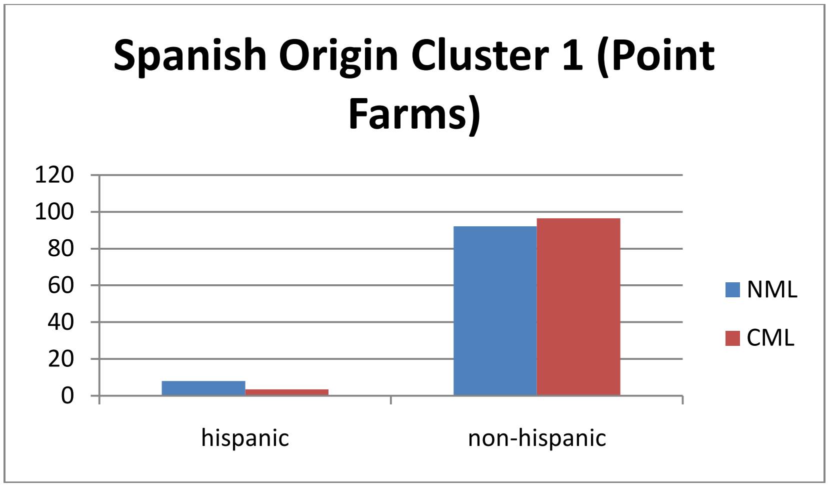

Part Time Cluster 5 (Idle Cropland)
Cluster Percentage
100
80
60
40
20
0
full
part
CML
NML

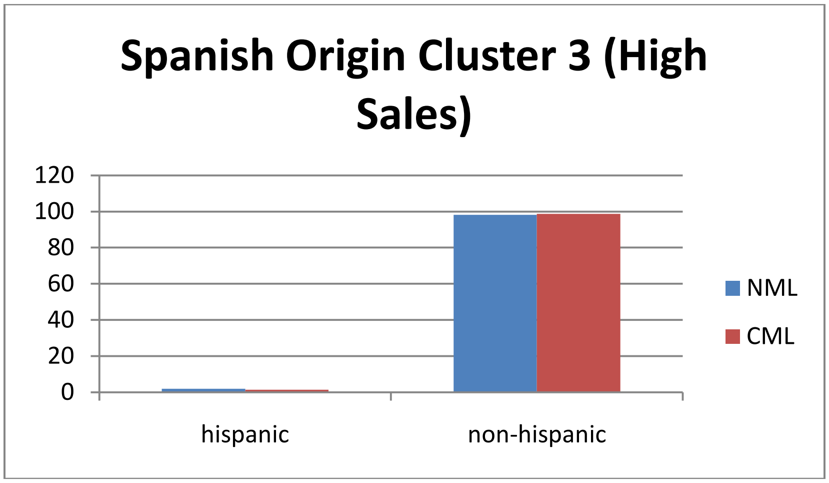

Retire Cluster 1 (Point Farms)
Cluster percentage
80
70
60
50
40
30
20
10
0
retired
not retired
CML
NML

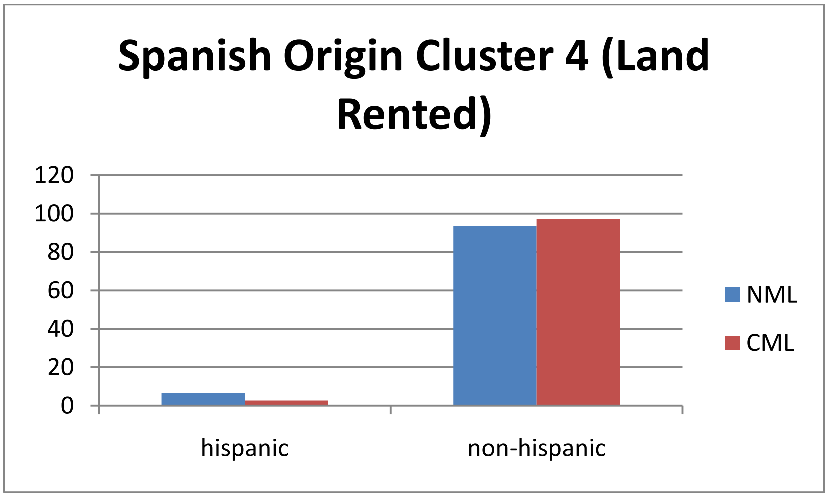

Retire Cluster 3 (High Sales)
Cluster percentage
100
80
60
40
20
0
retired
not retired
CML
NML

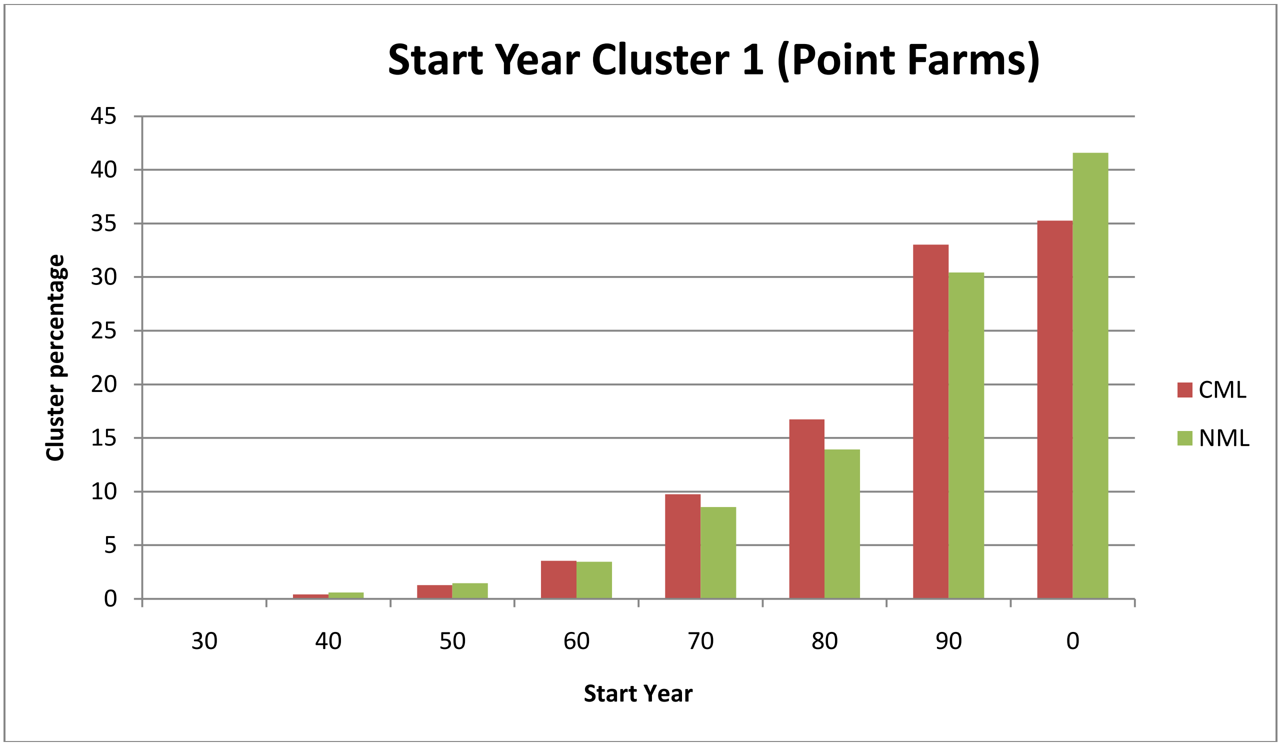

Retire Cluster 4 (Land Rented)
Cluster percentage
100
80
60
40
20
0
retired
not retired
CML
NML

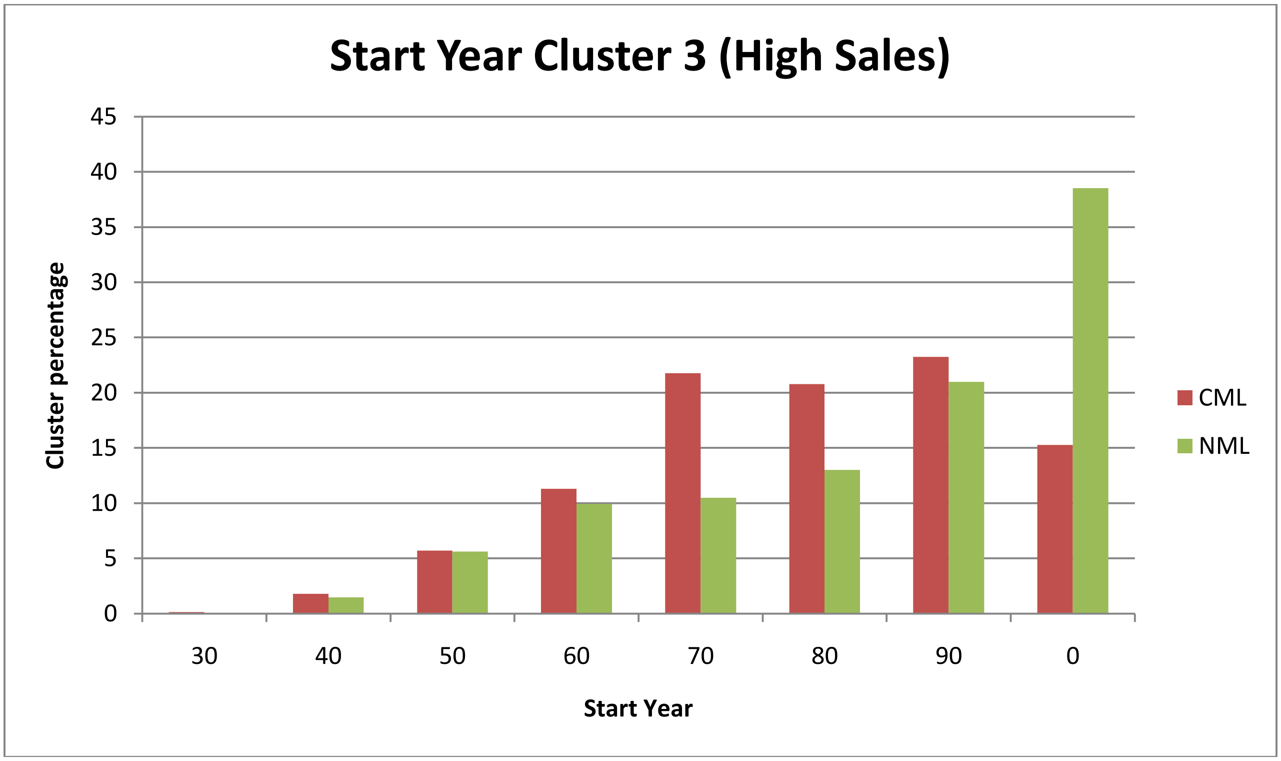

Retire Cluster 5 (Idle Cropland)
70
60
50
40
30
20
10
0
retired
not retired
CML
NML

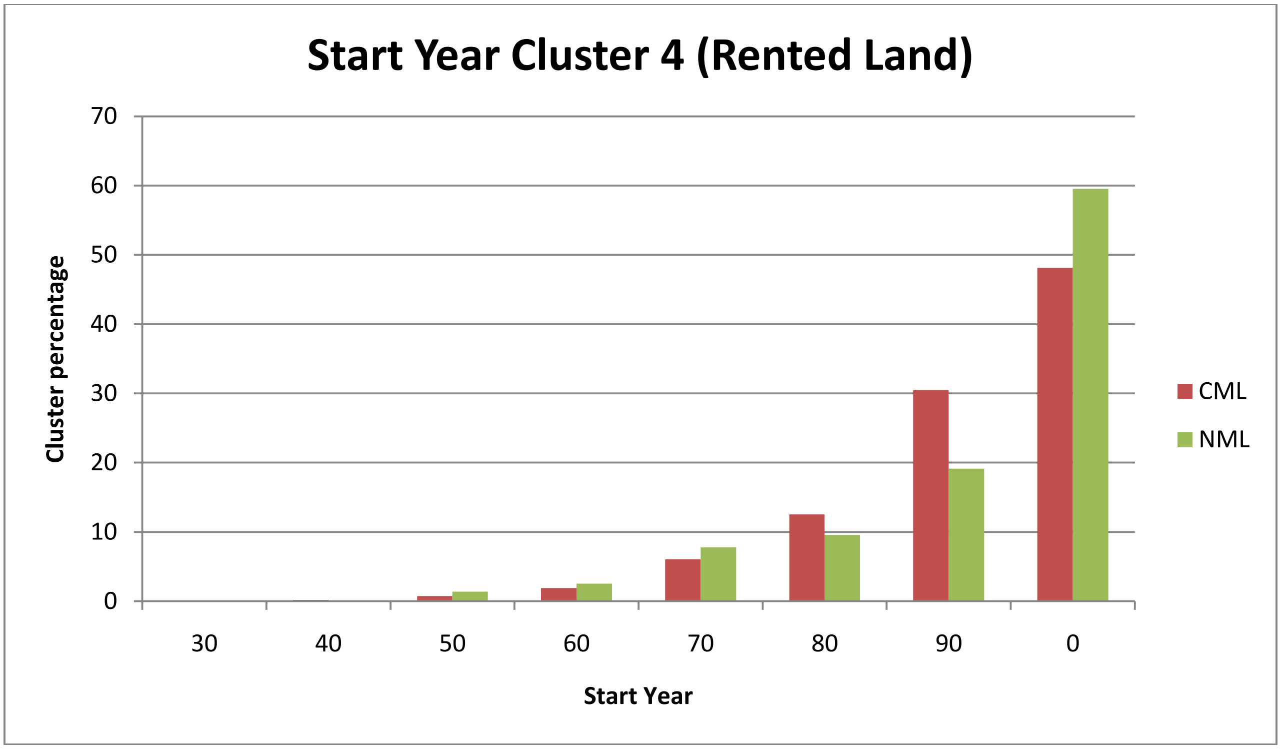

Spanish Origin Cluster 1 (Point Farms)
120
100
80
60
40
20
0
hispanic
non-hispanic
🟦 NML
🟥 CML

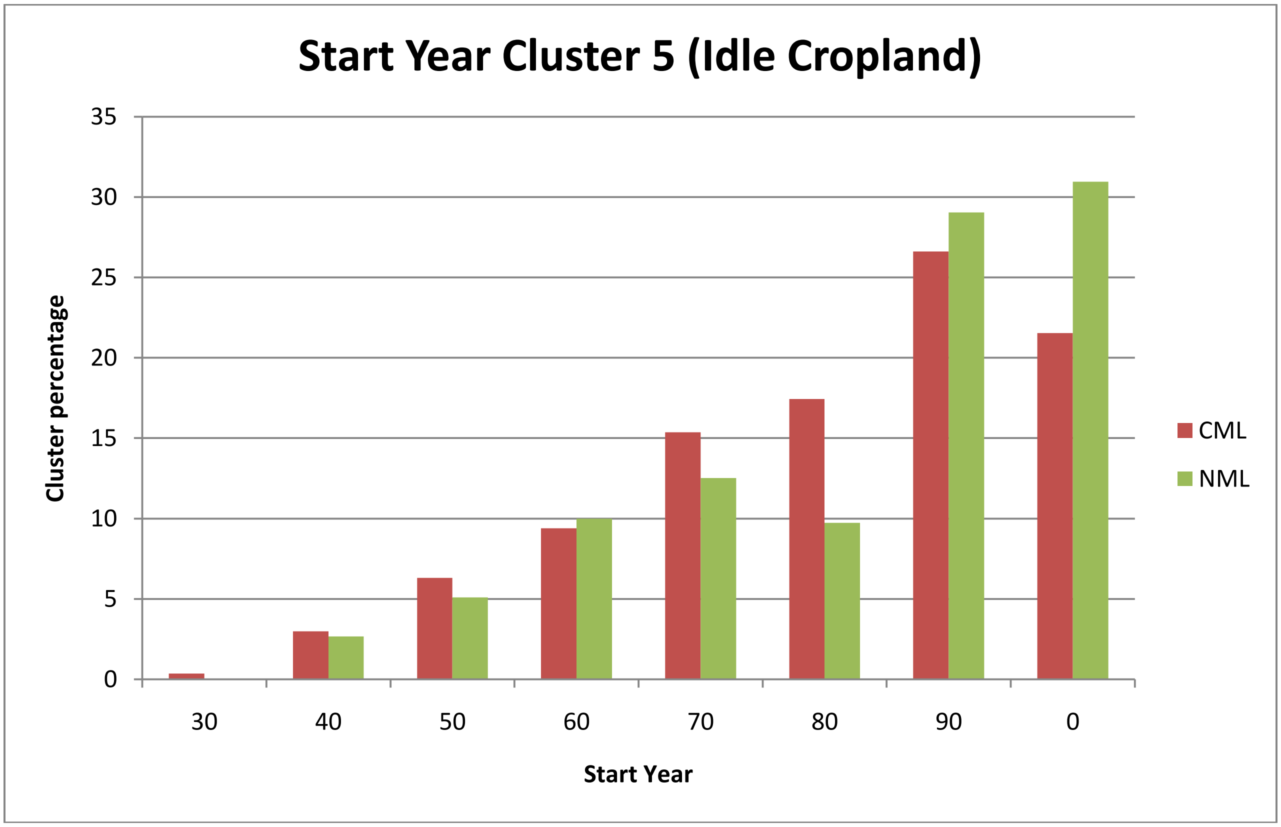

Spanish Origin Cluster 3 (High Sales)
120
100
80
60
40
20
0
hispanic
non-hispanic
🟦 NML
🟥 CML


Spanish Origin Cluster 4 (Land Rented)
120
100
80
60
40
20
0
hispanic
non-hispanic
🟦 NML
🟥 CML


Start Year Cluster 1 (Point Farms)
Cluster percentage
45
40
35
30
25
20
15
10
5
0
30
40
50
60
70
80
90
0
Start Year
🟥 CML
🟩 NML


Start Year Cluster 3 (High Sales)
Cluster percentage
45
40
35
30
25
20
15
10
5
0
30
40
50
60
70
80
90
0
Start Year
🟥 CML
🟩 NML


Start Year Cluster 4 (Rented Land)
Cluster percentage
70
60
50
40
30
20
10
0
30
40
50
60
70
80
90
0
Start Year
🟥 CML
🟩 NML


Start Year Cluster 5 (Idle Cropland)
Cluster percentage
35
30
25
20
15
10
5
0
30
40
50
60
70
80
90
0
Start Year
🟥 CML
🟩 NML