AUGUST 1971

A Use of Cluster Analysis in Outdoor Recreation Research

AGRICULTURAL EXPERIMENT STATION / AUBURN UNIVERSITY

E. V. Smith, *Director* / Auburn, Alabama

SUMMARY AND CONCLUSIONS

Researchers in the Department of Agricultural Economics and Rural Sociology at Auburn University Agricultural Experiment Station developed a cluster program for grouping appropriate data. Entirely mechanical in operation, the procedure delineates groups based entirely on joint variances among the input variables. The program will accept any number of observations with any number of variables and will group the data into all possible clusters from one through the number of observations. The operator specifies the number of clusters desired.

The program described in this report was designed for ease of use. The user has only to identify the amount of data included, the number of variables for each data item, the number of cluster groupings required, and the type of options desired. Included in the program are options for weighting, standardizing, listing, and normalizing the input data. By use of these options, unlike variables can be grouped and significant factors can be used to force contiguity.

The cluster procedure was used to identify various types of outdoor recreation regions based on certain variables. Examples presented in the report indicate that outdoor recreation regions can be derived. Before actual outdoor regions are isolated by a cluster procedure, however, additional analysis of the factors influencing outdoor recreation is necessary.

Delineation of outdoor recreation regions represents only one use of a cluster procedure. In the field of rural sociology, cluster analysis can be used to identify socioeconomic types by various quantified measurements on individuals. A cluster procedure can be used in marketing to identify types of market areas and, in combination with transportation analysis, to identify optimum locations for processing or storage facilities. In production economics, cluster analysis permits identification of types of farming areas according to soil type and socioeconomic data.

Data in this report were weighted by use of a two-variable location factor, which tended to cause contiguous groupings. Data can be weighted by other means. For example, if ease of access and mobility are considered significant the data can be weighted according to its location with respect to transportation routes.

Cluster analysis is also a useful tool in sampling, to identify strata. The sample can then be collected from the various strata to ensure additional validity.

Cluster analysis is useful in economic research, particularly for grouping data items that have many variables. The procedure is not restricted to outdoor recreation or to research, but has wide applicability for any identification problem.

ACKNOWLEDGMENTS

Research on which this report is based was carried out under Project Ala.-299 supported by Hatch and State Research funds. Special appreciation is expressed to the Alabama Department of Conservation for use of data gathered for the development of the Alabama Statewide Comprehensive Outdoor Recreation Plan. Appreciation is also expressed to J. H. Blackstone, Department of Agricultural Economics and Rural Sociology, for providing valuable guidance in preparing the study. The assistance of Mrs. B. T. Kenney in all phases of the study also is acknowledged.

CONTENTS

FIRST PRINTING 2M, AUGUST 1971

A Use of Cluster Analysis in Outdoor Recreation Research

E. W. McCOY and W. C. WILLIS, III[1]

OUTDOOR RECREATION as an area of study received major impetus from the reports by the Outdoor Recreation Research Review Committee in 1959 *(1)*. The ORRRC was appointed by the President to determine the status of outdoor recreation in the United States and to make recommendations for improving the availability of outdoor recreation opportunities for all people in the Nation.

The Committee published 27 volumes of reports and recommended the creation of a Federal department to coordinate the development of outdoor recreation. Based on this recommendation, Congress authorized creation of the Bureau of Outdoor Recreation (BOR) within the Department of the Interior. This bureau was given the responsibility of creating the Nationwide Outdoor Recreation Plan, coordinating the Statewide Outdoor Recreation Plans, and determining distribution of the Land and Water Conservation Fund.

A primary responsibility of BOR was fostering and supporting research in all areas of outdoor recreation. This responsibility committed BOR to long range planning for meeting outdoor recreation needs. The planning objective included the concept of research and development of recreational districts. Creation of recreational districts had no firm theoretical foundation.

The Alabama Legislature authorized creation of planning and development districts in compliance with BOR requirements. Twelve districts have officially been designated. Planning agencies within each district are responsible for planning aspects

[1] Assistant Professor, Department of Agricultural Economics and Rural Sociology; and former Programmer Assistant, Department of Agricultural Economics and Rural Sociology, now Systems Analyst, The Bell Company Labs, Berkeley, California.

of growth and development. These planners require data on relationships between outdoor recreation demand, supply, and natural resources before they can formulate programs to satisfy recreational needs.

The arsenal that agricultural economists have available for attacking problem areas can be used in outdoor recreation research. Among the weapons that have proven fruitful are input-output analysis for recreational impact studies *(2)*, regression and correlation for participation and user preference studies *(3)*, and many types of descriptive and non-parametric analyses where the data do not warrant parametric procedures *(4)*. The lack of sufficient quantitative data limits application of many research tools. Theoretically, any procedure applicable to other researchable areas could be used in outdoor recreation research. Linear programming and simulation have been applied to recreational firms when demand curves for services could be derived *(5)*. Demand curves can be derived if actual or proxy prices have been developed.

Cluster analysis is one of the methodological procedures with a potential for use in outdoor recreation research. It is a method whereby a mass of data is compacted into a number of homogeneous groupings. At least two functions are served by grouping data. First, since the amount of data a human being can comprehend is limited, grouping the data allows analysis of manageable portions of the universe. Second, analysis of the data within groups may reveal factors that cause or are associated with agglomeration. Cluster analysis is especially useful when neither the significant factors involved in grouping nor the number of significant groups is known.

BASIS FOR CLUSTER ANALYSIS

The ideal grouping of any data set should meet certain minimum standards. Included are these standards: (1) that each data item belongs uniquely to its own group, i.e., there would be no overlap between groups; (2) each group should be unique from every other group, i.e., the group differences would meet the statistical criteria for valid differences; and finally, (3) each item arrangement should be optimum for all groups, i.e., the movement of any item from one group to another would reduce the total fit for all groups.

A decision model to meet these criteria was developed for

clustering data items into groups. For any "n" data items there is only one optimum grouping. Additionally, there is only one optimum arrangement of the "n" items in X groups. With the aid of electronic computers it is possible to arrange the data items in every possible grouping and every possible arrangement for any possible grouping. Thus, for 10 data items there are a maximum of 10 possible groups. In addition, there are 1,023 possible arrangements in deriving the 1 through 10 groups.

The optimum number of groups can be derived by subdividing into two parts the total difference of each value from every other value. This procedure is synonomous with apportioning the variation between and among all groups.

For each individual value:

$D_T = D_W + D_B$

where

D_T = total difference between the value and all other values

D_W = total difference between the value and all other values within the same group

D_B = total difference between the value and all values in other groups.

The optimum grouping is obtained within any specified number of groups when any further movement of data from one group to another decreased the size of D_B. The optimum number of groups is obtained when the decrease in size in D_W is less than the degree of freedom lost in creating an additional group.

Number of clusters and amount of data within each cluster can be decided on statistical grounds. The total difference measure has a constant relationship to the variance.

$S^2 = D_T^2/2N$

where

S^2 is the variance

D_T^2 is the sum of squares of each difference between values

N is the number of data items

Using summation signs

$\epsilon D_T^2 = \epsilon D_W^2 + \epsilon D_B^2$ since the cross-product term is equal to zero. An indication of the correct number of clusters can be obtained by comparing the amount of total difference allocated to

the within-cluster groupings. Selected clusters can then be verified or rejected by discriminant analysis or other statistical procedures.

THE CLUSTER PROGRAM

Agricultural economists at Auburn University Agricultural Experiment Station have developed a cluster procedure to divide any given set of data into a number of homogeneous groups *(6)*. The procedure is particularly useful for screening a relatively large number of variables to determine their effect on cluster groupings. As with experiments in the physical sciences, a control grouping can be made with one variable. Modifications from the control grouping can be examined by inserting additional variables into the cluster procedure.

A primary use of this cluster procedure includes deriving the group arrangement of a data set for testing by discriminate analysis. Given a set of "n" data items there are $2^n - 1$ possible permutations of these data. As the size of "n" increases, the possibility of selecting optimum groupings or of determining if grouping exists becomes extremely difficult. A set of 20 observations with one variable per observation has approximately 1 million possible groupings. Increasing the number of variables to two does not increase the number of possible groupings, but makes it much more difficult to find the correct grouping among all possible arrangements. Clearly, it is inconceivable that the optimum grouping would be selected from many data items by observation unless some exogenous factor forced the data into specified groups. An example of an exogenous factor is state boundaries. Quantified county data arranged into groups by the cluster program would not necessarily follow state boundaries. The program developed for this report is designed to manipulate quantifiable variables. If institutional variables can be quantified they can be included in the clustering decision.

Outdoor Recreation Examples of Cluster Program Use

The cluster program was applied to three types of outdoor recreation data: resources, supply, and participation. Cluster analysis had previously been used to delineate economic regions *(7)*. Thus, it was hypothesized that outdoor recreation regions

could be delineated based on type of resources available, population, and certain measurements of outdoor recreation activity.

A complete inventory of all State outdoor recreation resources and facilities was available in the Alabama Statewide Comprehensive Outdoor Recreation Plan *(8)*. Volume 6 of the report was devoted to recreational resource needs, Volumes 3 and 4 to the supply of facilities, and Volume 2 to the demand for outdoor recreation. Data from these volumes were used as input for the cluster analysis.

The resource, supply, and demand regions, as well as any other data grouped by the cluster program, were based entirely on mathematical multivariate relationships of the input data. Analysis and interpretation of the clusters were dependent on additional statistical tests, knowledge of the data, and plans for ultimate use of the results. Recreational regions presented in this report were intended as an example of use of the cluster procedure and not as outdoor recreation regional analysis per se.

Since, in this instance, the number of regions was not known in advance, the cluster program was allowed to group the data in all possible clusters through 20. To indicate the function of the cluster program, a step-by-step example of input and output from the program is given in the Appendix.

Input data were standardized and the means and standard deviations computed for each variable. Multiple range tests were used to identify significant mean differences. The cluster groupings reported were further tested by discriminant analysis to determine individual placement probabilities. All variables were distinguished by three terms: low, moderate, or high. Each designated significant difference. Thus, low was significantly less than moderate, which was significantly less than high.

The regions were numbered in the order reported by the cluster program. In most instances the program tended to list the groupings from lowest to highest; however, in multivariate cases the relationship was often difficult to determine visually. Region 1 for resources thus was not necessarily the same as Region 1 for supply or demand, unless it was consistently low for all variables.

A grid overlay of ¼-inch squares was fitted to an Alabama map with a scale of 40 miles to 1 inch. Each county's center of population distribution was identified by coordinates of the grid, and the two grid coordinates were included in all analyses.

Cluster analysis using only the grid coordinates always created contiguous grouping in the absence of other data. The coordinates were weighted to force additional contiguity in the examples.

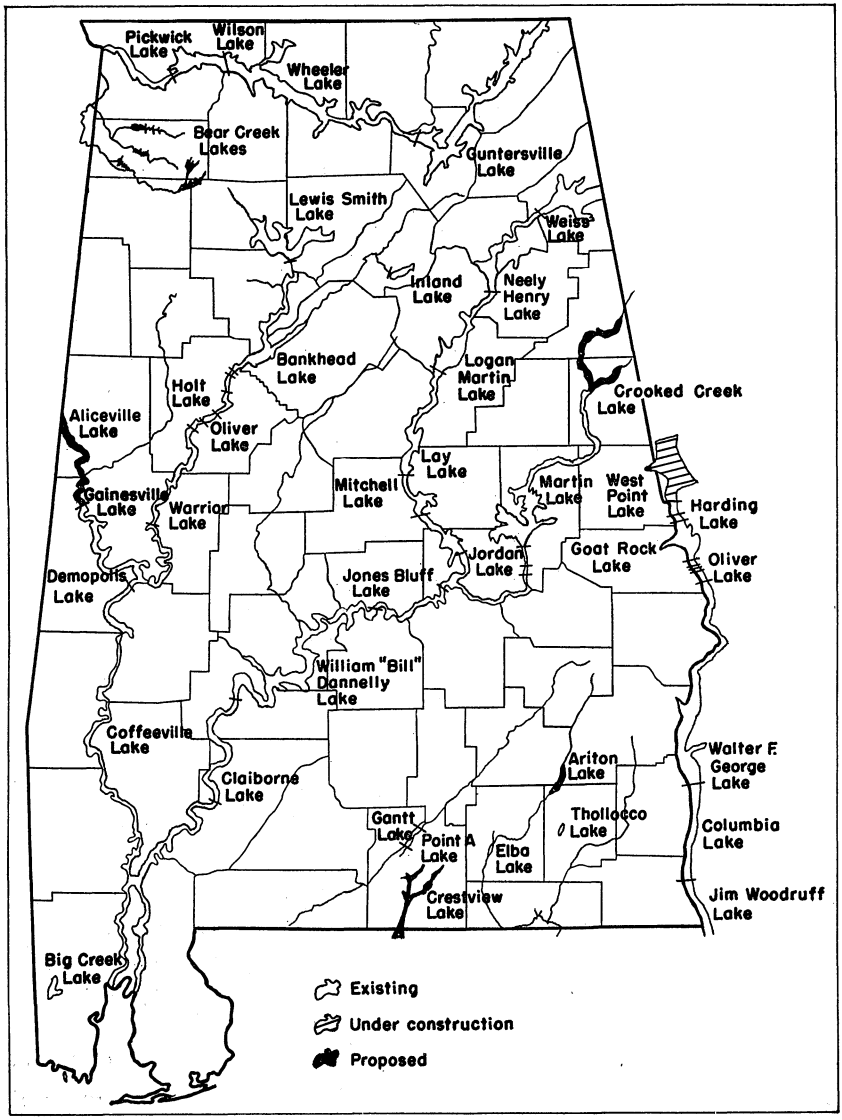


FIG. 1. Alabama rivers and impoundments.

Outdoor Recreation Resource Regions

Many resource variables were considered for the cluster analysis. Because of multicollinearity, however, two factors were used in the final regional delineation – total land area and inland water – both adjusted by population.[2] County data were used although demarcations on a more precise resource basis would be preferable. For example, many reservoirs in Alabama extend through or border on several counties. Although the entire reservoir surface may be available to the population of each county, amount of water assigned to each county may differ. No attempt was made to manipulate the data to correspond to subjective evaluations of resource availability.

Rivers and impoundments in Alabama are shown in Figure 1. Additional inland water consists of all small reservoirs over 40 acres in size.

Five outdoor recreation resource regions based on total land and inland water per capita were identified, Figure 2. Even though the factor to force contiguity of the counties was incorporated in the data, counties in Region 4 were not contiguous. The regions were characterized by the relative amounts of total land and inland water available per unit of population. Region 1, even though it contained Lake Eufaula, was relatively low in both total land and inland water. Much of the inland water available to this region was considered to be in Georgia. Region 2 had a moderate amount of total land per capita, but it was low on inland water. Region 3 was relatively low on total land and had a moderate amount of inland water per capita. The region actually contained a large amount of inland water, but rated moderate per capita because of a high concentration of

[2] Adjustment was made by dividing total land and inland water by the county population.

TABLE 1. COMPARATIVE AMOUNTS OF PER CAPITA LAND AND WATER AVAILABLE IN OUTDOOR RECREATION RESOURCE REGIONS IN ALABAMA

Region	Land	Rank[1]	Water	Rank[1]
1	low	2	low	1
2	moderate	3	low	2
3	low	1	moderate	3
4	moderate	4	high	5
5	high	5	moderate	4

[1] Ranked from lowest to highest.

population. Region 4, which was not contiguous, had a moderate amount of total land and a high amount of inland water per capita. The region generally extended through the Coosa and Tallapoosa River basins, Figure 3. Region 5 had the greatest total land per capita and was moderate with respect to inland water. Relative position of each region was shown in Table 1.

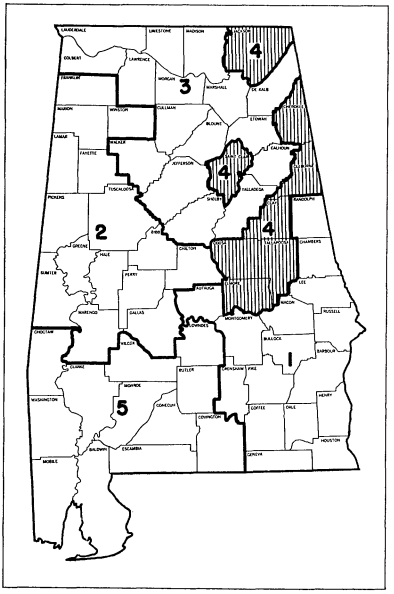


FIG. 2. Alabama outdoor recreation resource regions.

The comparative ranking for each region indicated that each could be uniquely identified on the basis of the two variables considered. Regional designation could be simplified by combining Region 3 and Region 4 into a single region. However, the resultant four-region grouping would not be the same four-region arrangement as was designated by the cluster procedure. In using the cluster procedure the intuitive answers often did not equate with the mathematical. In the four-cluster arrangement, Regions 1, 2, 3, and 5 were put into three groups and Region 4 remained intact.

In non-cluster terms there apparently were five distinct outdoor recreation resource regions in Alabama. First, a region along the Chattahoochee River that was characterized by limited amounts of land and water relative to the population. Second, a northwest region along the Mississippi border that had limited water per capita. Third, a region along the Tennessee Valley, and including Birmingham, that was low in total land. Fourth, a non-contiguous region rated high in inland water that included the Alabama Power Company lakes of Weiss, Martin, Logan Martin, Mitchell, and Lay. Fifth, a southwest region along the Mississippi border that was high in total land. It included the Gulf Coast and Mobile Bay, but these waters were not included in the analysis.

Outdoor Recreation Supply Regions

The second cluster example was based on the supply of outdoor recreational facilities available in the various counties in

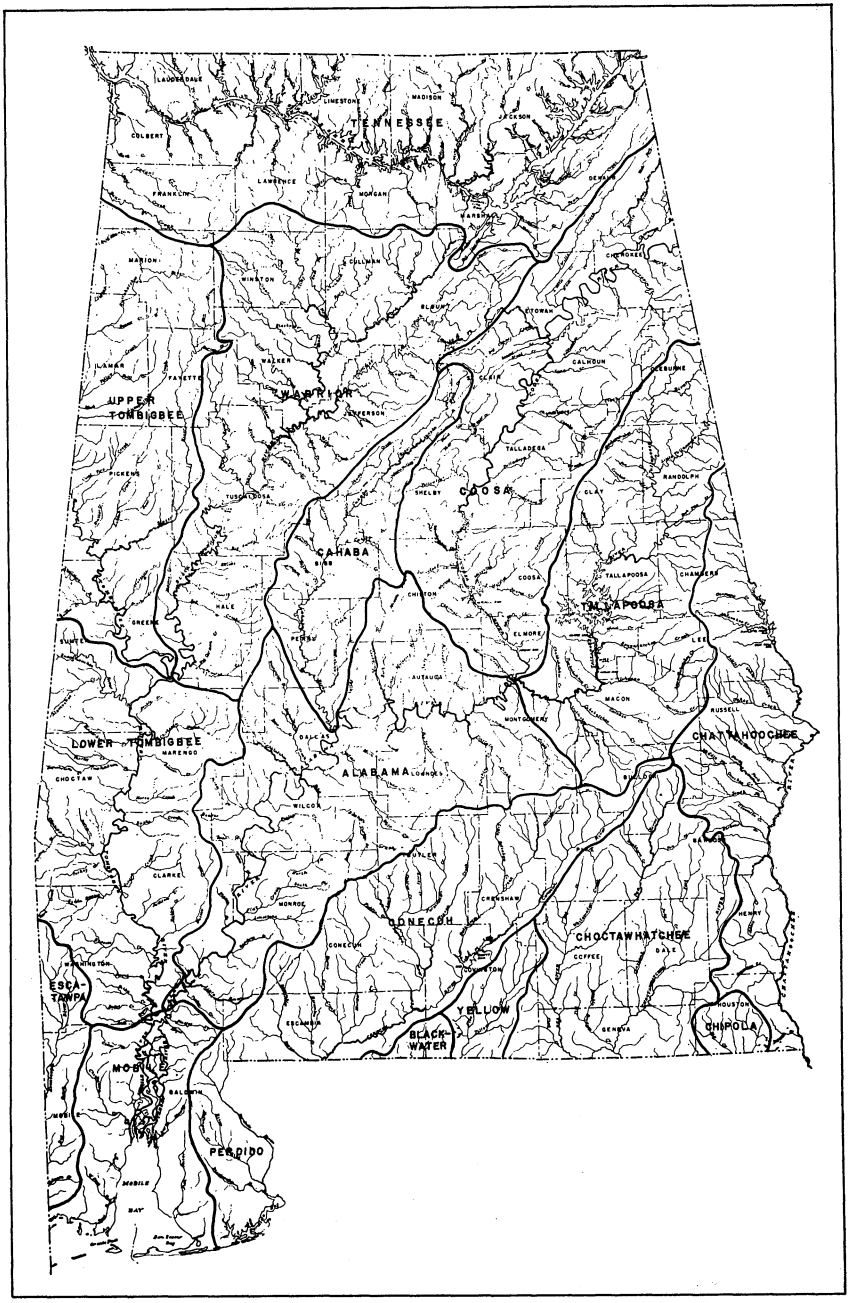


FIG. 3. River basins in Alabama.

TABLE 2. COMPARATIVE AMOUNTS OF OUTDOOR RECREATION LAND AND SITES IN OUTDOOR RECREATION SUPPLY REGIONS OF ALABAMA

Region	Land	Rank[1]	Sites	Rank[1]
1	low	1	low	2
2	moderate	2	low	1
3	moderate	3	moderate	3
4	high	4	high	4

[1] Ranked from lowest to highest.

Alabama. Supply regions and variables were reported in the same manner as resource regions.

Outdoor recreation supply normally consists of land and facilities designated for outdoor recreation use. Supply in this study indicated the amount of the outdoor recreation resources currently being used for outdoor recreation purposes. By means of factor analysis, outdoor recreation land and sites were identified as variables expressing outdoor recreation supply. Only the quantity of supply available was considered. No attempt was made to weigh the quality of existing developed facilities.

Four outdoor recreation supply regions were identified, Figure 4. Even though contiguity was forced, Region 4 counties were not together physically. The recreational supply in these counties, expressed as recreational land and sites, differed from their adjoining counties to a large extent. Comparative positions among regions are shown in Table 2.

Supply Region 1 was lowest in land and essentially tied with Region 2 for fewest sites. Region 4 was high in both land and sites. The five counties in Region 4 were Covington, with its large acreage in the Conecuh National Forest; Winston, with the Bankhead National Forest; Baldwin, with the extensive Gulf Coast recreational developments; and Jefferson and Mobile, the two major urban centers in the State.

Regions 1 and 2 had little recreational development. Region 2 had more land devoted to outdoor recreation than Region 1, but slightly fewer developed sites. Region 3, which included the Tennessee Valley as well as the upper reaches of the Coosa River basin, had more development for outdoor recreation.

Since there was little difference between Region 1 and Region 2, these were considered simultaneously. When these regions were combined, their recreational supply picture as postulated from the two variables considered was quite bleak. Combining

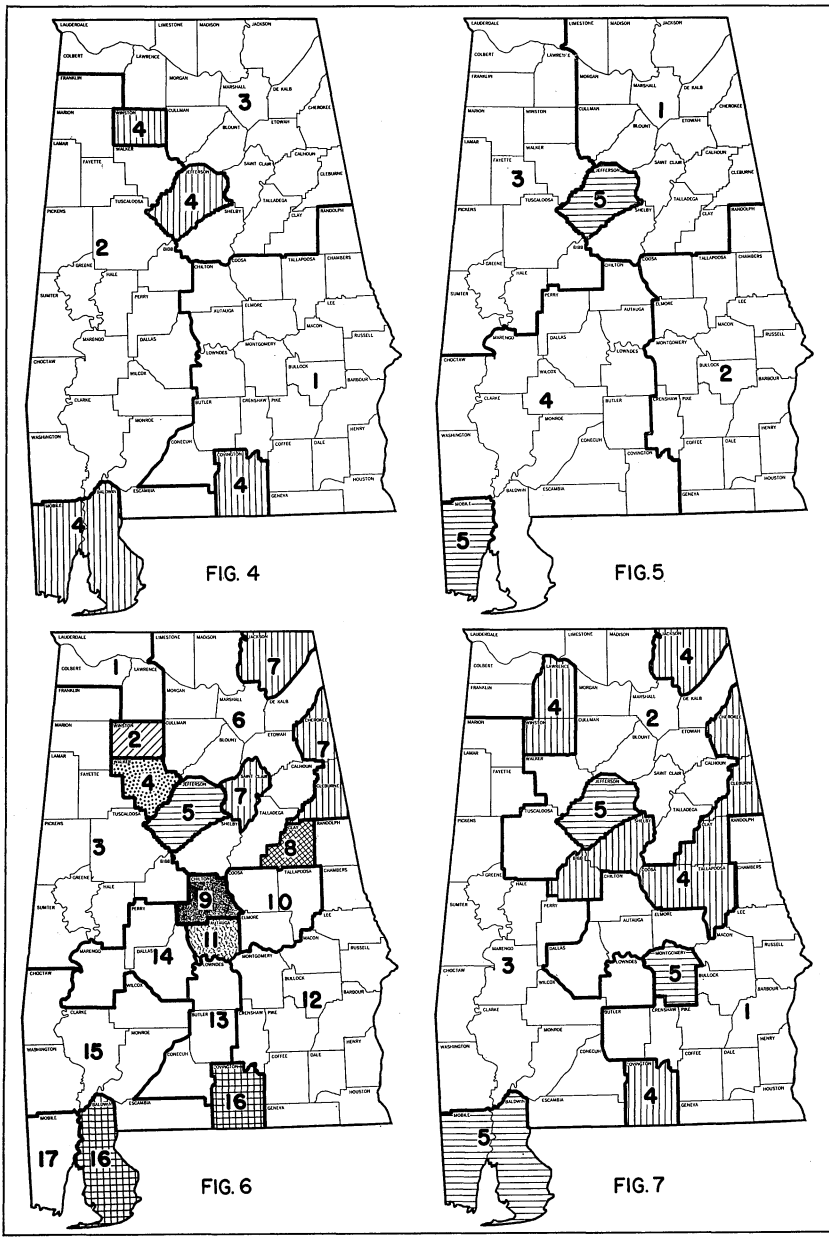


FIG. 4. Alabama outdoor recreation supply regions.

FIG. 5. Alabama outdoor recreation demand regions.

FIG. 6. Alabama outdoor recreation resource-supply-demand overlay regions.

FIG. 7. Alabama outdoor recreation resource-supply-demand regions.

these two regions resulted in the entire eastern and southern sections of the State being almost void of recreational development, in a comparative sense. Only Region 3 plus the scattered counties of Region 4 had developed facilities for outdoor recreation.

Outdoor Recreation Demand Regions

The third cluster example was an attempt to delineate outdoor recreation demand regions.

Outdoor recreation demand as an aggregate was influenced by many factors. Urban population and number of urban places in each county were the factors chosen to represent demand for cluster analysis. Other factors could have been used, such as estimated aggregate demand in each county, but the two chosen were considered adequate for example purposes.

Five outdoor recreation demand regions were identified, Figure 5, but the identification was not as precise as for the resource or supply regions. There were only two significant demand regions in the State: One composed of Mobile and Jefferson counties and the second containing the remainder of the State. The five regions selected were significantly different by discriminant analysis, but three groups overlapped in multiple range analysis. Comparative ratings for the five regions are given in Table 3. The entire western half of the State was relatively low in both urban population and urban places.

Demand for outdoor recreation services was directly related to size of the surrounding population. The demand regions designated were interpreted to indicate that the demand from Jefferson and Mobile counties equaled the demand from any other region. The clusters were reliable over a large range. Additional clusters beyond five involved subgroupings within the four groups. At and beyond the nine-cluster grouping, Jefferson

TABLE 3. COMPARATIVE AMOUNTS OF URBAN POPULATION AND URBAN PLACES IN OUTDOOR RECREATION DEMAND REGIONS IN ALABAMA

Region	Urban population	Rank[1]	Urban places	Rank[1]
1	low	4	moderate	4
2	low	3	low	3
3	low	2	low	1
4	low	1	low	2
5	high	5	high	5

[1] Ranked from lowest to highest.

County formed a single-county cluster and Mobile County moved to various arrangements with other non-contiguous counties.

The demand clusters exhibited a feature present in all contiguous cluster regions. Distance from the population center of each county to the geometric center of the cluster was less than the distance to the geometric center of any other cluster. While it was difficult to determine the geometric center of Region 4 by inspection, it was closer to the population center of Chilton County than was the geometric center of Regions 1, 2, or 3. These distances were straight line without regard to road mileage involved. On the assumption that air miles and land miles were directly proportional, outdoor recreation facilities located precisely in the center of each of the four regions would maximize recreational access to the population of the region and State.

Outdoor Recreation Resources-Supply-Demand

The simultaneous consideration of resources, supply, and demand was accomplished in two ways. The first involved a map overlay combining the regions created by single-emphasis analysis. The single-analysis regions formed into 17 combined regions, Table 4. These 17 were reduced to 10 multiple-county regions and 6 single-county regions by combining Regions 11 and 12,

TABLE 4. COMPARATIVE AMOUNTS OF COMBINED RESOURCES, SUPPLY, AND DEMAND AVAILABLE IN OUTDOOR RECREATION REGIONS

Region	No. of counties in region	Resources[1]		Supply[1]		Demand[1]	
		Land	Water	Land	Sites	Pop.	Places
1	3	L	M	M	M	L	L
2	1	M	L	H	H	L	L
3	10	M	L	M	L	L	L
4	1	L	M	M	L	L	L
5	1	L	M	H	H	H	H
6	11	L	M	M	M	L	M
7	4	M	H	M	M	L	M
8	1	M	H	L	L	L	M
9	1	M	L	L	L	L	L
10	3	M	H	L	L	L	L
11	1	L	L	L	L	L	L
12	15	L	L	L	L	L	L
13	3	H	M	L	L	L	L
14	3	M	L	M	L	L	L
15	6	H	M	M	L	L	L
16	2	H	M	H	H	L	L
17	1	H	M	H	H	H	H

[1] L is low, M is moderate, H is high.

Figure 6. These regions had the same characteristics for resources, supply, and demand. The single-county regions 2, 4, 5, 8, 9, and 17 represented Winston, Walker, Jefferson, Clay, Chilton, and Mobile counties, respectively. Each of the 16 regions differed from every other region with respect to at least one aspect of resources, supply, or demand.

The second method of simultaneous consideration of resources, supply, and demand involved incorporating the six variable factors in one cluster analysis. Using this method forced greater contiguity of the counties. Analysis of the cluster output revealed five outdoor recreation resource-supply-demand regions, Figure 7. The five regions were unique with respect to the variables considered, Table 5.

Region 1 was relatively low in resources, supply, and demand. The southeastern section of the State has Chewacla State Park in Lee County, Blue Springs State Park in Barbour County, Valley Creek State Park in Dallas County, and numerous State fishing lakes. A new State Park is planned on Lake Eufaula in Barbour County, and its development will increase the region's outdoor recreation supply. Demand will increase with additional population and the resource picture is brightened by considering West Point and Lake Eufaula reservoirs available to the area.

Region 2 in the Tennessee Valley had a relatively high number of recreational sites and a moderate demand. The expansion of Lake Guntersville State Park in Marshall County, DeSoto State Park in DeKalb County, Monte Sano State Park in Madison County, and Joe Wheeler State Park in Limestone County will greatly improve State facilities in the region. Additional population growth will undoubtedly increase demand.

Region 3 had relatively high land resources but was low for all other variables. The Corps of Engineers is developing additional water-based recreational sites on the Alabama River in this region. Little population growth is projected for this area. Increased access to available lands for hunting will enhance recreational use of the region.

Region 4 was relatively high in resources and land devoted to outdoor recreation but low in developed sites and demand. The region was non-contiguous and included the Talladega, Conecuh, and Bankhead National Forests as well as Lakes Martin, Weiss, Smith, Logan Martin, Guntersville, Wilson, and Wheeler. The counties of this region have sufficient resources

TABLE 5. COMPARATIVE AMOUNTS OF ALL VARIABLE FACTORS INCLUDED IN OUTDOOR RECREATION RESOURCE, SUPPLY, DEMAND REGIONS IN ALABAMA

Region	Resources				Supply				Demand			
	Land	Rank[1]	Inland water	Rank[1]	Recreation land	Rank[1]	Recreation sites	Rank[1]	Urban population	Rank[1]	Urban places	Rank[1]
1	low	3	low	1	low	1	low	3	low	3	low	3
2	low	2	moderate	3	low	2	high	4	moderate	4	moderate	4
3	high	5	low	2	low	3	low	1	low	1	low	1
4	high	4	high	5	high	5	low	2	low	2	low	2
5	low	1	moderate	4	moderate	4	high	5	high	5	high	5

[1] Ranked from lowest to highest.

for development of additional outdoor recreation facilities, and they are located relatively close to areas of high outdoor recreation demand.

Region 5, rated low in resources but high in supply and demand, included Mobile, Baldwin, Montgomery, and Jefferson counties. These counties have the highest population and the greatest recreational development in the State.

The five-cluster recreational resource-supply-demand regions did not have the precise identification of the overlay regions. Single counties with unique attributes for outdoor recreation were not identified. The choice of method to use depends on the degree of precision desired by the recreational planner as well as the desired use of the resulting clusters.

SUMMARY OF OUTDOOR RECREATION REGIONS

A computer program for grouping multiple data items was developed. It was designed to accept any number of observations and any number of variable measures for each observation. Capable of accepting any type of quantitative data, the program was used to determine the interrelationships between outdoor recreation resources, supply, and demand.

Five outdoor recreation resource regions were derived using this program. Variables used in the analysis were land and inland water per capita, designated as available land and water. The five regions ranged from the Chattahoochee River Basin, which was low with respect to both land and water, to the Coosa River Basin that was rated high in water and moderate in land. The waters of Lake Eufaula are primarily in Georgia and were not considered in the Chattahoochee Region.

Each of the resulting outdoor recreation resource regions was unique. The combination of resources in each region was unlike the combination of resources in any other region. Region 5 had the greatest quantity of land, Region 3 had the least. Region 4 had the most water and Region 1 the least. Region 2, which included the west-central portion of the State, was relatively low in water and had a moderate amount of land.

Four outdoor recreation supply regions were identified, using land in outdoor recreation use and number of outdoor recreation sites developed as variables. The fourth region identified included five non-contiguous counties that were high in both variables. The Tennessee Valley Region was moderate with respect

to availability of both land and sites. The southeastern part of the State was low in recreational land and inland water. The west-central region had a relatively moderate amount of land devoted to outdoor recreation but had the lowest number of recreational sites.

Variables used to identify five outdoor recreation demand regions were urban population and number of urban places. These demand regions were not as uniquely identified as the resource and supply regions. Three of the demand regions were relatively low in both variables. Region 5, which included Jefferson and Mobile counties, was relatively high in both variables. Region 1 occupied the northeastern corner of the State and it had a moderate number of urban places but a low urban population.

The demand regions were determined primarily by size of the population. Areas with a relatively low population occupied a relatively large region. If all residents of a region traveled to the center of the region in which they live, approximately the same total distance would be traveled in each region. This is not true for Region 5 where the two counties are not contiguous.

Simultaneous consideration of resources, supply, and demand was accomplished in two ways. First, the individual regions were superimposed on one map. Sixteen unique regions of outdoor recreation were created in this fashion. Some of the regions were determined primarily by resource availability, some by supply, and some by demand. Only one county in the State ranked high in terms of all variables. Mobile County had relatively high amounts of available land, moderate amounts of available inland water, high amounts of recreational land and sites, and high numbers of urban population and urban places. Jefferson County had high demand and supply but lacked resources for additional development. The remaining regions had relatively low demand. The southeastern corner of the State was low in all factors.

Simultaneous consideration of all variable factors resulted in five regions being identified. The regions ranged from those low in all factors to those high in both supply and demand. One region high in resources was identified as well as one high in available land and one high in developed facilities. The five regions identified by cluster analysis on all variables were not precisely the same as those found by overlaying the independent regions.

APPENDIX

The Willis-McCoy Cluster Program

The cluster program designed by agricultural economists at Auburn University is written in PL 1. Its use requires a PL 1 compiler. The program accepts as input any number of observations and variables subject only to the capacity of the computer used. Observations are grouped into every number of clusters from two through the maximum number specified by the user. The computer printout lists the cluster membership, the cluster centroid for each variable, the cluster standard deviation for each variable, and a distance figure that is the sum of the squares of the distance from each observation in a cluster to every other observation in the cluster summed over all clusters. Included on the printout is a ratio that is the distance for N-1 clusters divided by the distance for N clusters, and S_b which is the proportion of the total distance contained between the clusters.

The input to the program is divided into four sections: parameters, weights (optional), data, and labels (optional).

PARAMETERS. The parameters must be punched as shown below, in any order, with one or more spaces between parameters and with a semicolon following the final parameter.

A. *Mandatory parameters* – these must be included.

1. Observations = XX, where XX is the number of observations. Observations are limited only by machine capacity.

2. Variable = XX, where XX is the number of variables.

3. Maxclusters = XX, where XX is the maximum number of clusters the operator desires to have printed.

B. Optional parameters

1. Title = '___title desired___' The title may occupy up to 80 spaces, not including the quote mark (the quote marks are not printed). If no title is specified, blanks are printed.

2. DP = XX, where XX is the number of decimal places used when the weights, data, cluster means, and standard deviations are printed. If none is specified, DP = 3 is assumed.

3. FW = XX, where XX is the field width used in printing out weights, data, cluster means, and standard deviations. Default is FW = 10.

4. LL = XX, where XX is the label length. Default is LL = 20, and if a value less than 20 is specified, it is ignored.

5. Options = '. . . options list' where options list consists of a list in any order separated by commas of the options listed below.

A. Lab – labels are used

B. LRD – list raw data

C. NORM – normalize data (divide each variable for each observation by the sum of the squares of the variables for that observation).

D. STD – standardize data (divide each variable for each observation by the standard deviation of the variable).

E. WT – weight data

These options are performed in the following order (if specified):

D. Standardize

E. Weight

C. Normalize

WEIGHTS. If the option WT is specified, each variable of each observation is multiplied by the weight associated with the variable. Weights are specified as shown below.

WT(N) = XX, where N is the number of the variable the weight is associated with and XX is the value of the weight. If a weight is not specified 1.0 is assumed. The weights may be in any order on the card but must be separated by one or more spaces and followed by a semicolon.

DATA. The data for each observation consists of a key followed by the variables for the observation. The key must be a number between one and the number of observations and must be different from the key for the remaining observations. There must be one or more spaces between the key and the first variable and between variables. Data for the observations must be input in order, but without respect to location of the variable on the card. The variables may be listed on separate cards for each observation or listed continuously without respect to cards. The key is printed on the output if labels are not used. If labels are used the key must match a key on the label cards.

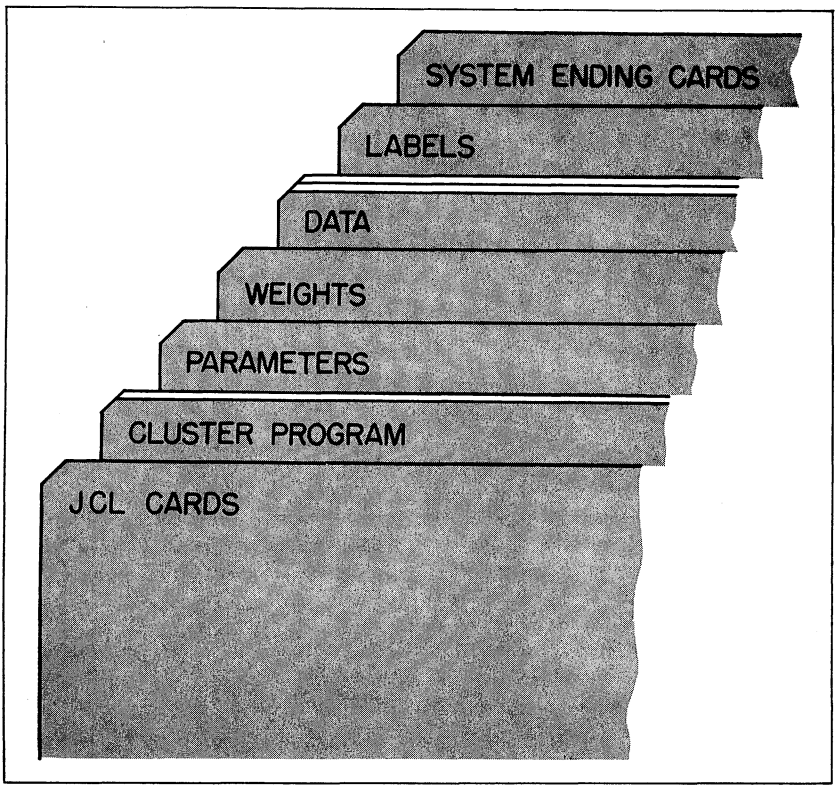
LABELS. If label cards are used they are read as a key followed by the label without quote marks. The labels may be numeric or alphabetic. The key must be numeric.

Format for Program Input

1. J C L cards
2. Program
3. Parameter card(s) (mandatory)
4. Weight card(s) (optional)
5. Data
6. Label(s) (optional)
7. System ending cards

The schematic form for data input is shown in the Appendix Figure, below.

An example program is presented below. Only one variable per observation is included, to simplify analysis, but the procedure is similar for multi-variate cases. The data used are from the preliminary 1970 census of population in Alabama counties.



APP. FIG. Schematic of job deck set-up for Willis-McCoy Cluster Program.

APPENDIX TABLE 1. CLUSTER GROUPING OF 1970 ALABAMA COUNTY POPULATION DATA

County	Number of clusters								
	2	3	4	5	6	7	8	9	10
Autauga	2	2	2	2	2	2	2	2	2
Baldwin	2	3	3	5	3	5	5	5	5
Barbour	2	2	2	2	2	2	2	2	2
Bibb	2	2	2	2	2	7	7	7	7
Blount	2	2	2	2	2	2	2	2	2
Bullock	2	2	2	2	2	7	7	7	7
Butler	2	2	2	2	2	7	2	2	2
Calhoun	2	3	3	3	3	3	3	3	3
Chambers	2	2	2	5	5	2	8	8	8
Cherokee	2	2	2	2	2	7	7	7	7
Chilton	2	2	2	2	2	2	2	2	2
Choctaw	2	2	2	2	2	7	7	7	7
Clarke	2	2	2	2	2	2	2	2	2
Clay	2	2	2	2	2	7	7	7	7
Cleburne	2	2	2	2	2	7	7	7	7
Coffee	2	2	2	5	5	2	8	8	8
Colbert	2	2	2	5	5	5	5	5	5
Conecuh	2	2	2	2	2	7	7	7	7
Coosa	2	2	2	2	2	7	7	7	7
Covington	2	2	2	5	5	2	8	8	8
Crenshaw	2	2	2	2	2	7	7	7	7
Cullman	2	3	3	5	5	5	5	5	5
Dale	2	3	3	5	5	5	5	5	5
Dallas	2	3	3	5	5	5	5	5	5
DeKalb	2	2	2	5	5	5	8	8	10
Elmore	2	2	2	2	5	2	8	8	8
Escambia	2	2	2	5	5	2	8	8	8
Etowah	2	3	3	3	3	3	3	3	3
Fayette	2	2	2	2	2	7	7	7	7
Franklin	2	2	2	2	2	2	2	2	2
Geneva	2	2	2	2	2	7	2	2	2
Greene	2	2	2	2	2	7	7	7	7
Hale	2	2	2	2	2	7	7	7	7
Henry	2	2	2	2	2	7	7	7	7
Houston	2	3	3	5	5	5	5	5	5
Jackson	2	2	2	5	5	2	8	8	10
Jefferson	1	1	1	1	1	1	1	1	1
Lamar	2	2	2	2	2	7	7	7	7
Lauderdale	2	3	3	5	3	5	5	9	9
Lawrence	2	2	2	2	2	2	2	2	2
Lee	2	3	3	5	3	5	5	9	9
Limestone	2	2	2	5	5	5	8	8	10
Lowndes	2	2	2	2	2	7	7	7	7
Macon	2	2	2	2	2	2	2	2	2
Madison	1	3	4	4	4	4	4	4	4
Marengo	2	2	2	2	2	2	2	2	2
Marion	2	2	2	2	2	2	2	2	2
Marshall	2	3	3	5	5	5	5	5	5
Mobile	1	1	4	4	6	6	6	6	6
Monroe	2	2	2	2	2	7	2	2	2
Montgomery	1	3	4	3	4	4	4	4	4
Morgan	2	3	3	3	3	3	3	9	9
Perry	2	2	2	2	2	7	7	7	7
Pickens	2	2	2	2	2	7	2	2	2
Pike	2	2	2	2	2	2	2	2	2
Randolph	2	2	2	2	2	7	7	7	7
Russell	2	2	3	5	5	5	8	5	10
St. Clair	2	2	2	2	2	2	2	2	2
Shelby	2	2	2	5	5	2	8	8	8
Sumter	2	2	2	2	2	7	7	7	7
Talladega	2	3	3	5	3	5	5	9	9
Tallapoosa	2	2	2	5	5	2	8	8	8
Tuscaloosa	1	3	3	3	4	3	3	3	3
Walker	2	3	3	5	3	5	5	5	5
Washington	2	2	2	2	2	7	7	7	7
Wilcox	2	2	2	2	2	7	7	7	7
Winston	2	2	2	2	2	7	7	7	7

Labels are the county names. The systems cards are excluded since these vary from system to system. The data appears on the 80-column input card precisely with the spacing and wording as shown below.

Observations = 67 variables =1 maxcluster = 10
title = 'population clusters' options = 'lab, std';
1 23990 2 58193 3 22010 4 13764 . . .
11 24462 12 . . .

The remaining data were punched on 13 additional cards. As an alternative the data could be listed on individual cards

1 23990
2 58193
etc., for 65 additional cards.

The labels are listed below

1 Autauga
2 Baldwin
etc., for 65 additional labels.

To utilize the data for other statistical analysis it is convenient to keypunch it on the cards in a fixed right justified format with at least one blank space between variables. As shown, the fixed format is not a necessary condition for use of the cluster program.

The program calls for a printout of the data by clusters. Immediately below each cluster grouping is a printout of the means and standard deviations for each variable. Appendix Table 1 lists the cluster groupings for the Alabama county population data. Observation of this table can indicate the stability of certain data within clusters and the tendency for groups of data to simultaneously move from cluster to cluster. Appendix Table 2 contains a listing of the amount of total distance contained

APPENDIX TABLE 2. PERCENTAGE OF TOTAL VARIATION WITHIN CLUSTER GROUPINGS FOR 1970 ALABAMA COUNTY POPULATION DATA

Number of clusters	Percentage within groupings
2	92.46
3	97.82
4	99.14
5	99.60
6	99.76
7	99.91
8	99.97
9	99.98
10	99.99

APPENDIX TABLE 3. STANDARDIZED CLUSTER MEANS AND STANDARD DEVIATIONS WITHIN CLUSTER GROUPINGS FOR 1970 ALABAMA COUNTY POPULATION DATA

Cluster	Number of clusters																			
	1		2		3		4		5		6		7		8		9		10	
	$\overline{X}$	S	$\overline{X}$	S	$\overline{X}$	S	$\overline{X}$	S	$\overline{X}$	S	$\overline{X}$	S	$\overline{X}$	S	$\overline{X}$	S	$\overline{X}$	S	$\overline{X}$	S
1	0.	1.0	−.21	.24	−.32	.11	−.34	.09	−.37	.06	−.37	.06	−.40	.03	−.41	.02	−.41	.02	−.42	.02
2			2.66	2.46	.36	.46	.14	.23	−.04	.13	−.10	.10	−.25	.06	−.31	.03	−.31	.03	−.31	.03
3					4.90	2.66	1.95	.93	.67	.37	.25	.19	.03	.09	−.16	.05	−.17	.04	−.19	.02
4							6.78	0.0	2.28	1.05	1.17	.43	.52	.17	.07	.06	.03	.05	−.10	.03
5									6.78	0.0	3.02	0.0	1.41	.17	.52	.17	.19	.07	.04	.03
6											6.78	0.0	3.02	0.0	1.41	.17	.59	.11	.19	.07
7													6.78	0.0	3.02	0.0	1.41	.17	.59	.11
8															6.78	0.0	3.02	0.0	1.41	.17
9																	6.78	0.0	3.02	0.0
10																			6.78	0.0

$\overline{X}$ is the mean of the variables in the cluster.
S is the standard deviation of the variables in the cluster.

within clusters. This quantity is printed at the beginning of each cluster grouping. Appendix Table 3 indicates the means and standard deviations within each cluster grouping. Standard range tests or analysis of variance can be used to determine if the cluster groupings are significantly different. The means and standard deviations are printed directly below the listing of the observations within the cluster. Discriminant analysis should be used to test for significant differences in multiple variable clusters.

The cluster program was written to facilitate ease of use for non-programmers. Further information regarding the program can be obtained by contacting:

Willis-McCoy Cluster Program
Department of Agricultural Economics and Rural Sociology
Auburn University
Auburn, Alabama 36830

BIBLIOGRAPHY

(1) Outdoor Recreation Resources Review Commission. 1962. Outdoor Recreation for America. ORRRC Study Rept. 27 Volumes. Washington, D.C.

(2) Hinman, R. C. 1969. The Impact of Reservoir Recreation on the Whitney Point Microregion of New York State. Cornell Univ. Water Resources and Marine Science Center. Tech. Rept. No. 18.

(3) State Park Division. 1967. Alabama Statewide Comprehensive Outdoor Recreation Plan. Ala. Dept. of Conservation. Montgomery, Ala.

(4) Williams, J. W. and R. W. Schermerhorn. 1968. Economic Analysis of the Potential for Developing Overnight Camping Facilities on or Near Major Highways in Oklahoma. Okla. Agr. Exp. Sta. Bull. B-660.

(5) Grueter, J. 1969. Simulation of a Recreational Firm: Flow Chart and Computer Program. Maine Agr. Exp. Sta. Orono, Maine. Tech. Bull. 36.

(6) McCoy, E. W. 1970. "The Willis Cluster Program – Method for Grouping Data." Proceedings of the Fall Meeting, Ala. Statistical Assoc. Huntsville, Ala.

(7) Cox, Thomas P., Bernard Siskin, and Allan Miller. 1969. A More Objective Procedure for Determining Economic Subregions: Cluster Analysis. Southern J. of Agr. Econ. pp. 37-45.

(8) State Park Division. 1970. Alabama Statewide Comprehensive Outdoor Recreation Plan. 8 Volumes. Ala. Dept. of Conservation. Montgomery, Ala.