LBL-11084
UC-13 c.2

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

ENERGY & ENVIRONMENT DIVISION

SELECTION OF MODAL BEA REGIONS FOR URBAN AND COMMUNITY IMPACT ANALYSIS

Henry Ruderman, Flora Fung, and Rudolph J. Beran

June 1980

ENERGY
AND ENVIRONMENT
DIVISION

Prepared for the U.S. Department of Energy under Contract DE-AC03-76SF00098

DISCLAIMER

LBL-11084

SELECTION OF MODAL BEA REGIONS FOR URBAN AND COMMUNITY IMPACT ANALYSIS

Henry Ruderman and Flora Fung
Energy Analysis Program
Energy and Environment Division
Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

and

Rudolph J. Beran
Department of Statistics
University of California
Berkeley, California 94720

This work was supported by the Assistant Secretary of Environment, Office of Technology Impacts, Regional Assessments Division of the U.S Department of Energy under Contract No. DE-AC03-76SF00098.

June, 1980

SELECTION OF MODAL BEA REGIONS FOR URBAN AND COMMUNITY IMPACT ANALYSIS

INTRODUCTION

The Department of Energy is required to prepare urban and community impact analyses for all proposed major policy and program initiatives. The general procedure for preparing such an analysis is contained in Executive Order 12074 and OMB Circular A-116. Its purpose is to identify the likely effects of these programs and policy initiatives on cities, counties and other communities. The analysis will insure that potentially adverse impacts of proposed Federal activities will be identified during the decision making process.

The Department of Energy has promulgated criteria for determining whether one of its proposed activities must undergo an urban and community impact analysis. These criteria are based in part on the anticipated social and economic impacts of the activity on the community and the nation. Impacts include both direct and indirect effects on employment, population, income, cost of living, and state and local government finances. Differential impacts on central cities, suburban areas, nonmetropolitan communities, areas with high and low unemployment, and minority and low income communities are to be analyzed quantitatively wherever possible. An evaluation of aggregate effects on the U.S. economy in terms of employment, personal income, prices and fiscal conditions and on the major industrial sectors of the economy is required.

In the study reported here LBL has used cluster analysis to select communities for the impact analysis. Variables for clustering were selected from a larger data base of environmental, energy, demographic, social and economic data for BEA regions. They were selected for their relevance to quantities used in the impact analysis. Standard methods were used to cluster the BEA regions into groups with similar characteristics. A total of six separate clusters were found. These were analysed using a density estimate approach to determine which BEAs were representative of the cluster as a whole. Communities and urban areas within these BEAs were then selected for further detailed analysis.

Section 2 of this report is a summary of the results of the study. It contains a list of the clusters found and indicates which BEA regions are the representative of the group. A map is included to indicate the geographic distribution of the clusters. There is also a discussion of the characteristics of the clusters and how they are relevant to the impact analysis. The succeeding sections describe the technical details of data selection and transformation, clustering methods, and density-contour calculations. This is followed by a more detailed analysis of the results. Included in this section are an evaluation of the sensitivity of the results to the choice of variables and clustering algorithms, a discussion of generalizing the impact analysis to other BEAs, and a discussion of the limitations of the method. The conclusions of this study are presented in the final section.

SUMMARY OF RESULTS

Clustering analysis was performed on a set of fifteen variables for each of the BEA regions of the country. These data summarize the demographic, economic, financial and energy production characteristics of the regions. Six clusters were found which contain 161 of the 173 BEA re ions that cover the country; the other 12 could not be classified. The six clusters exemplify a broad range of characteristics: there are distinct differences between the racial compositions, income levels, population growth rates, unemployment rates, and resources for energy development in different clusters. Figure 1 is a map showing the BEAs in each cluster.

Several BEAs from each cluster were selected which are representative of the cluster as a whole. These modal BEAs lie close to the maximum of the estimated density in the cluster. The six clusters and their modal BEAs are shown in Table I. Five of the clusters are composed primarily of rural BEAs where we expect the largest impacts of synfuels development. They are distinguished by differences in their economic and demographic characteristics as well as their resources for energy production. Modal BEAs in each of the clusters contain fossil or

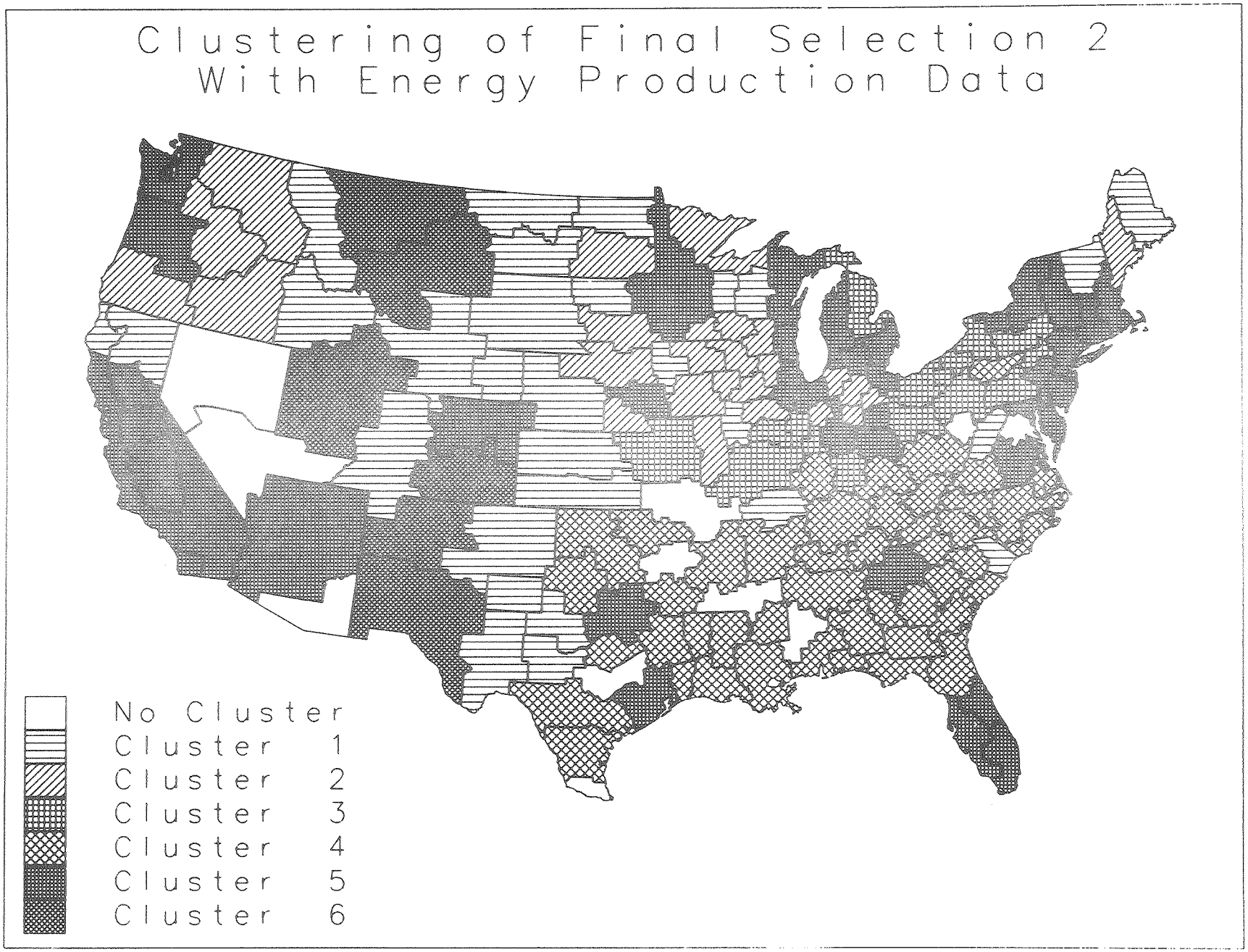
Clustering of Final Selection 2
With Energy Production Data
No Cluster
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6

FIGURE 1

biomass resources needed for producing synthetic fuels. These BEAs are suitable for studying the impacts of a variety of technologies in a single community, or the impacts of a single technology over a range of communities.

A sensitivity analysis of the clusters to a variety of clustering methods indicated that the clusters that we found are not unique. Only two of the clusters are distinct; the other four appear to be compact regions in a single large diffuse cluster. The modal BEAs, however, tend to group together as the clustering method is varied. They are relatively stable compared to the clusters, and thus form a valid set of regions for impact analysis.

The regions on which the impact analysis is performed are chosen from the modal BEAs. This selection is based on the values of the variables used in the clustering, the density estimate at the BEA, and other factors not included in the clustering analysis but that are of importance in assessing the impacts of synthetic fuels development.

Because we are unable to find a unique set of clusters, we are limited in the extent to which we can generalize the impact analysis results. We believe that the results can be extended only to those BEAs that consistently cluster with the ones on which the analysis was done. For the data we used, the results are applicable to only 34 of the 173 BEA regions.

Two additional clustering analyses were performed on subsets of the data. The first, which included only the energy production variables, resulted in four clusters. The largest consisted of regions with little or no energy production or reserves; the other three consisted of oil and gas producing regions, coal producing regions, and regions which produce both. The second clustering analysis used the remaining demographic, economic and financial variables. Four of the five clusters found represent rural areas in the South, South West and Pacific Northwest, Great Plains, and Rocky Mountain regions. The fifth cluster contains the medium and large cities of the Northeast, Midwest and Pacific Coast. These clusters can form the basis for further impact analyses. The clusters without energy production could be especially

useful in studies of the impacts of increased reliance of renewable resources and conservation.

In addition, a set of analytic tools has been developed to validate the results of this study. Of special importance are the techniques using density estimates for evaluating clusters and selecting regions for further study. These tools will be of general use for impact analysis projects that involve classification and selection of study areas with the intent of generalizing the results.

ANALYTIC METHODS AND DATA

Our general approach to clustering and selection was to choose robust analytic methods. We did not want to use methods that relied heavily on the normality of the distribution of data values or were very sensitive to outliers. For example, we use modal values to characterize clusters rather than means. We did, however, apply transformations to the data to make their distributions more normal when this improved the quality of the clusters.

Data

Most of the data used in this study were selected from a data base of over 300 items for each of the 173 BEA regions that span the country*. The data base was compiled from a number of secondary sources by Urban Systems Research and Engineering, Inc. under the sponsorship of DOE[1]. BEA regions are groups of counties surrounding a central market area, and thus are defined on economic and demographic criteria. For most of the items used in this study, the BEA level data were calculated by simply aggregating the county level datas taken from the secondary sources. The general criterion that USR&E used for choosing items to

* The 1972 definition of BEA regions was used in this study because the data were compiled on that basis.

Table I

Modal BEA Regions

Code	Region	Density	Code	Region	Density
			Cluster 1		
123	Lubbock, TX	0.461	86	Wausau, WI	0.308
125	Abilene, TX	.416	92	Grand Forks, ND	.286
110	Wichita, KS	.367	124	Odessa, TX	.279
88	Eau Claire, WI	.315			
			Cluster 2		
76	South Bend, IN	0.926	105	Waterloo, IA	0.633
75	Fort Wayne, IN	.884	89	La Crosse, WI	.614
82	Rockford, IL	.868	78	Peoria, IL	.609
69	Lima, OH	.770			
			Cluster 3		
54	Louisville, KY-IN	1.000	64	Columbus, OH	0.703
57	Springfield, IL	.731	10	Erie, PA	.657
16	Harrisburg, PA	.716			
			Cluster 4		
119	Tulsa, OK	0.887	137	Mobile, AL	0.719
28	Greenville, NC	.804	20	Roanoke, VA	.697
49	Nashville, TN	.770	25	Greensboro-Winston Salem -High Point, NC	.696
			Cluster 5		
107	Omaha, NB-IA	0.923	62	Cincinatti, OH-KY-IN	0.733
70	Toledo, OH-MI	.849	21	Richmond, VA	.729
85	Appleton-Oshkosh, WI	.844	157	Portland, OR-WA	.681
60	Indianapolis, IN	.797			
			Cluster 6		
147	Colorado Springs, CO	0.278	145	El Paso, TX	0.226
146	Albuquerque, NM	.269	151	Salt Lake City, UT	.207

incorporate into the data base was their relevance to the level and impact of energy activities. The items chosen include information on energy supply and demand, transportation, socio-economic and demographic characteristics, land use, pollutant emissions and environmental quality. The reference year is 1975, although some data items are more recent (population and income, for example) and some are older (e.g., land use data).

The data base was augmented by information from other data bases available at LBL. Estimates of minority population for 1975 were taken from unpublished data from the Bureau of the Census. Data on wholesale sales for 1972 from the 1977 City - County Data Book and marketed value of farm production from the 1974 Census of Agriculture were extracted from the LBL SEEDIS system. These data were at the county level, so it was necessary to aggregate them to BEA regions.

A major deficiency in the data base is a lack of data on oil shale reserves. Oil shale plays an important role in the President's synthetic fuels program.

From the data available to us we selected nearly forty items as possible variables to cluster on. The items were categorized into subclasses as indicated in Table II. As can be seen from the table, the categories correspond closely to the types of impacts that will be examined in the final analysis. In most cases we tried both the extensive (total) and intensive (per capita or per unit area) form of the variable to determine which was more useful in discriminating among clusters.

After the data base was received from USR&E, a preliminary check on the data was made by summing the values for each BEA to get a national total for each data item, and then comparing the sum with published national data. Many discrepancies were found, most of which could be attributed to incorrect units, e.g. acres instead of square miles for areas. There were a few items, however, that we could not correlate the values in the data base with published data. We spoke to the staff at USR&E about the discrepancies we found in the data items we planned to use in this study. They supplied us with information on the original sources of the data. By checking the sources we could clear up many of

the discrepancies or convince ourselves that the data were reasonable. The remaining data items were not used.

Because the data items are expressed in many different units, it was necessary to standardize the data. This procedure gives each of the variables approximately equal weight in the clustering analysis. We standardized using robust estimators of location and standard deviation -- the median and the interquartile range divided by 1.348, respectively.

$$Z_{ji} = \frac{X_{ji} - M_i}{(Q_{3i} - Q_{1i}) / 1.348}$$

wh X_{ji} and Z_{ii} are the raw and standardized values of variable i for region j; M_i is the median and Q_{1i} and Q_{3i} are the quartiles for variable i.

All of the extensive data items were subjected to a logarithmic transformation of the form

$$X'_{ji} = \log(X_{ji} + \epsilon_i),$$

where ϵ_i is a small quantity, typically ten percent of the smallest non-zero value of X_{ji}. (The ϵ_i is necessary so that the logarithm is finite when the smallest value of the variable is zero.) The effect of transforming the data is to emphasize the differences between small values of the variables and de-emphasize the differences between the large ones. Since the distribution of the extensive variables is sharply peaked toward low values, the transformed data are more normally distributed. Some of the intensive variables that were peaked toward low values were also transformed. The effect of the logarithmic transformation on the population data is illustrated in Figure 2. The logarithmic transformations were applied before standardizing the data.

Histograms and scatter plots of the original, transformed and standardized data were examined for outliers and unusual distributions or correlations. Variables whose distributions could not be explained were

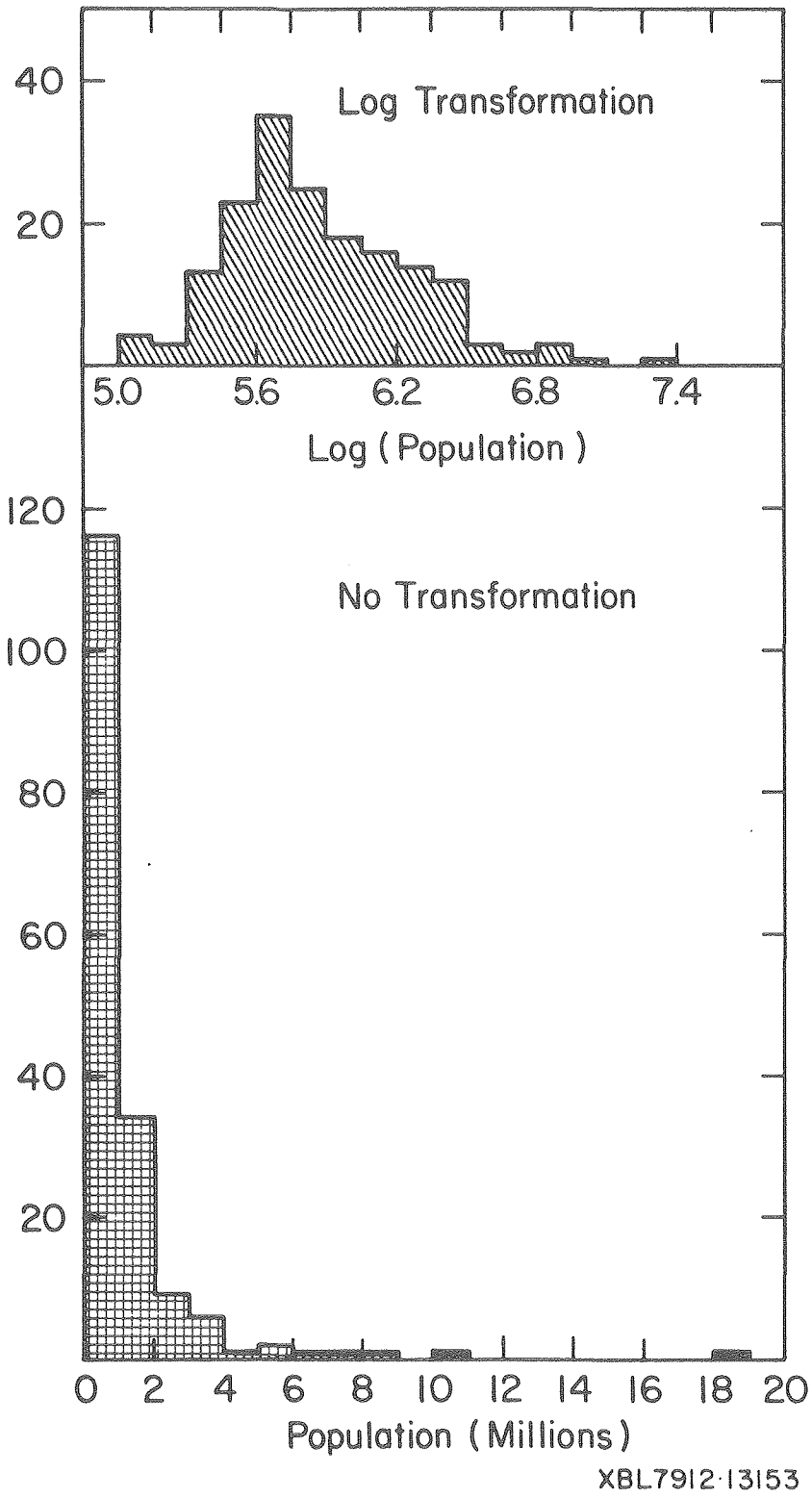

FIGURE 2

Distribution of BEA population with and without logarithmic transformation

Table II

Variables for Clustering

Extensive	Intensive
	Demographic
Area	n.a.
Population*	Population density*
n.a.	Percent non-white*
n.a.	Percent population change, 1970-75*
	Urban - Rural
Central city population	Percent population in central city*
Oullying area population	Percent population in otlying area*
Urban land area	Percent urban land
Urban vehicle miles traveled	Per capita urban VMT
Rural vehicle miles traveled	Per capita rural VMT
Single family dwellings	Percent single family dwellings
	Employment
Civilian labor force	Labor participation rate*
Total Employment	Unemployment rate*
	Personal Income
n.a.	Median household income
n.a.	Percent annual change 1969-74
n.a.	Percent below poverty level*
n.a.	Percent above $15,000*
	Economic
Total retail sales	Per capita retail sales
Total wholesale sales	Per capita wholesale sales
Value added by manufactures	Per capita value added
Electric generating capacity	Per capita capacity
Commercial landings and takeoffs	Per capita landings and takeoffs
	Energy Consumption
Residential and commercial	Per capita consumption
Industrial	Per capita industrial
Transpotation	Per capita transportation
Coal	Per capita coal
Petroleum	Per capita petroleum
Natural gas	Per capita natural gas
Electricity	Per capita electricity

Energy Production

Total coal	n.a.
Crude petroleum*	n.a.
Natural gas*	n.a.
Strippable coal reserves*	n.a.
Underground coal reserves*	n.a.

Government Finance

Local government - General expenditure	Per capita expenditure
Intergovernmental transfers	Per capita transfers
n.a.	Per capita debt*

* Used in final clustering
n.a. - not applicable

---

not used in the clustering analysis.

Clustering Techniques

The clustering routines we use are called joining algorithms [2,3] . A joining algorithm involves two choices: a rule for determining the distance between any two clusters; and a rule for selecting which clusters are to be amalgamated at each stage of the algorithm. The algorithm starts by regarding each data point in p-dimensions as a separate cluster. At each subsequent step of the algorithm, the closest pair of clusters is found and joined to form a single new cluster. The process continues until a single cluster consisting of all the points is obtained. Useful clusters are obtained from the penultimate steps of the algorithm.

Figure 3 illustrates the clustering process for nine BEA regions. In the first step BEAs 2 and 3 are amalgamated to form a single cluster. In the next step this cluster is joined by BEA 4 to form a larger cluster. The algorithm continues joining BEAs 8 and 9, then 6 and 7, until all BEAs have been merged into a single cluster. The vertical scale can be related to the distance at which two clusters are joined. The dashed horizontal line shows the level at which we decide the the number of

clusters. In this case we have determined that there are two clusters, one containing five BEAs and the other three; one BEA is an outlier.

In our work, the distance between two points, $\{X_{ji};\ 1 \leq i \leq p\}$ and $\{X_{ki};\ 1 \leq i \leq p\}$, was measured in the L_q metric:

$$d_{jk} = [\sum_{i=1}^{p} |X_{ji} - X_{ki}|^q]^{1/q}.$$

Of the three values of q tried (q = 0.5, 1.0 and 2.0), the value q = 0.5 give the best clustering. A likely explanation for this phenomenon is the relative insensitivity of the L_q metric to outliers when q is small.

We considered joining algorithms based on three different rules for selecting which clusters to amalgamate next:

(i) Single linkage (or nearest neighbor) - Join the two clusters that have the shortest distance between pairs of points in the two clusters.

(ii) Average linkage (or centroid) - Join the two clusters that have the shortest distance between the average points (centroids) of the clusters.

(iii) Maximum distance (or farthest neighbor) - Join the two clusters that have the shortest maximum distance between pairs of points in the respective clusters.

It is known from experience that single linkage tends to produce too many sausage-like clusters, while the maximum distance algorithm favors ball-like clusters. The average linkage algorithm is generally regarded as a reasonable compromise which avoids the extremes of the other two algorithms. However, it can be misled when the clusters really are sausage-like or ball-like. Average linkage was the primary algorithm used in our clustering of BEAs; sensitivity of the clusters obtained to the choice of algorithm was then explored by repeating the analysis with other algorithms.

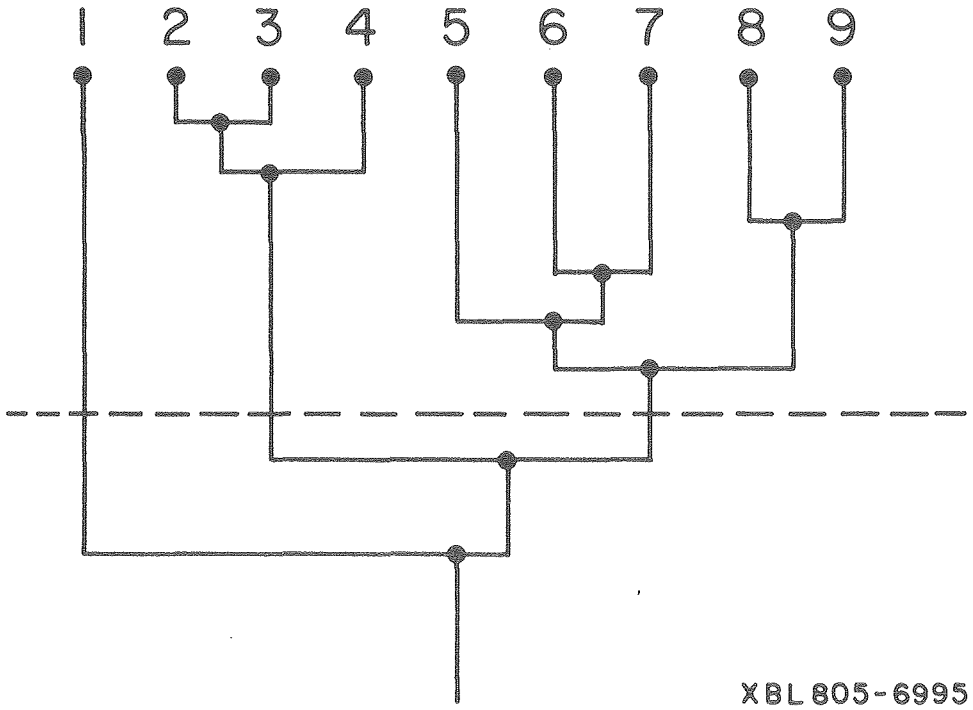

FIGURE 3

Tree diagram of the clustering example discussed in the text.

The clustering algorithms used are part of the BMDP statistical analysis package [4]. The final runs were done with the BMDP-P2M program - Cluster Analysis of Cases. A very useful feature of this program was its ability to print out the clustering tree and the amalgamation distances.

Density Estimates

The validity of clusters and the modal BEAs can be determined by estimating the probability density at points within the cluster. Valid clusters contain peaks in the density, and modal BEAs lie near these peaks. The probability density at a point in the p-dimensional space of variables is given by the weighted sum of the number of BEA regions lying within a neighborhood of the point [5]. The weight is a decreasing function of distance between the point at which the density is being evaluated and the point representing the BEA. The weighting of each point and the size of the neighborhood are incorporated in a suitably chosen window function. If the window is too narrow, then the density has many sharp peaks; if it is too broad, then the separation between clusters may be lost. We have investigated several window functions to determine which one is most useful for defining clusters.

We use the Epanechnikov window which is of the form

$$W(t) = \begin{cases} \frac{(q+1)}{2q}(1-|t|^q) & \text{if } |t| \leq 1 \\ 0 & \text{if } |t| > 1 \end{cases}$$

The parameter q defines the shape of the window (see Figure 4b). Most of our density estimates were calculated using the Euclidean distance, q = 2. The window is normalized so that

$$\int_{-1}^{1} W(t)dt = 1.$$

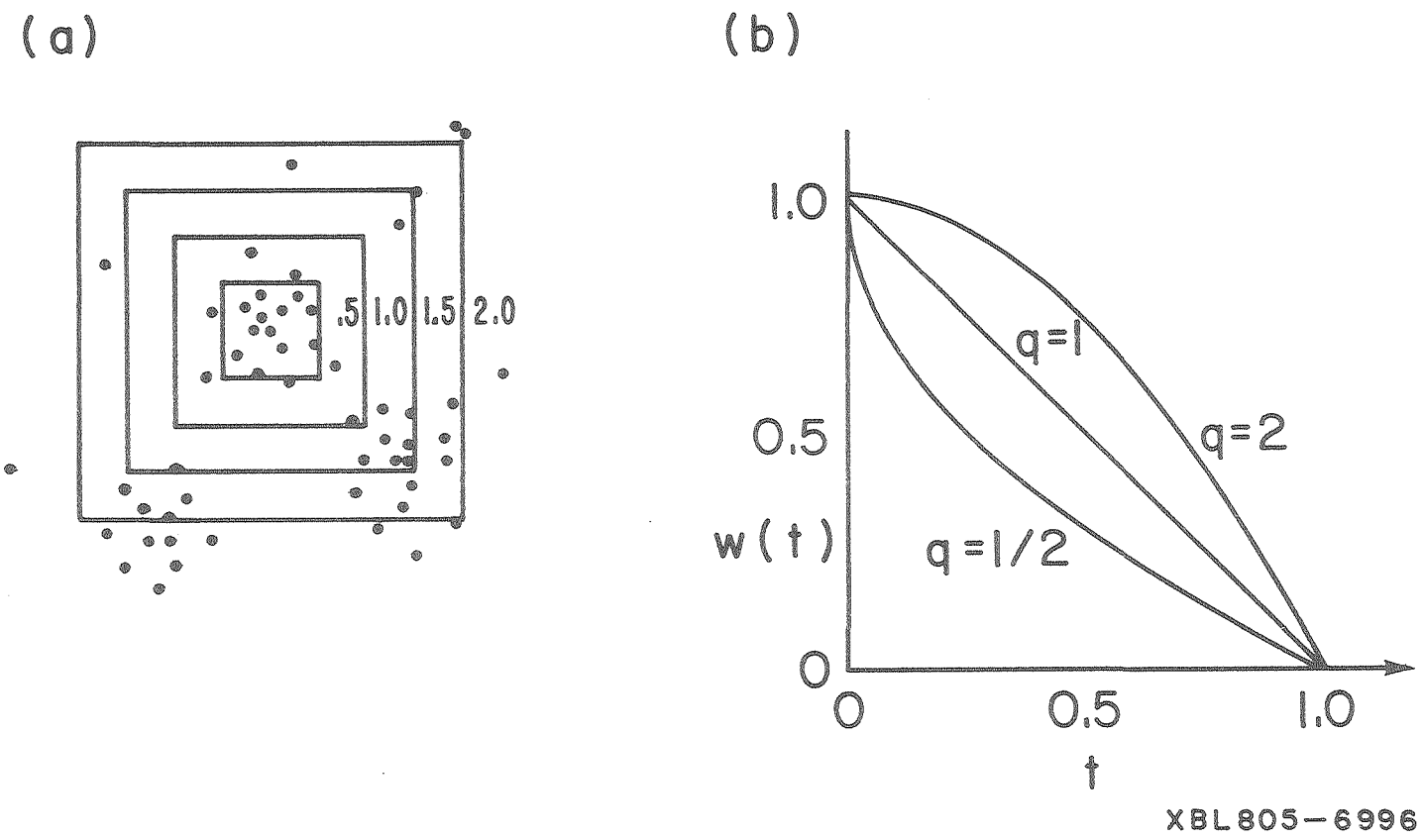

FIGURE 4

(a) Effect of window size on clustering.

(b) Effect of the parameter q on window shape.

The window density estimate at a point x in the p-dimensional space of variables is given by a sum over the n BEAs of the product of the window for each variable.

$$f_n(x) = \frac{1}{nC_n^p} \sum_{i=1}^{n} \prod_{j=1}^{p} W\left(\frac{x_j - X_{ij}}{C_n}\right).$$

In this equation X_{ij} is the value of variable j in BEA i, and C_n is a parameter which determines the bandwidth of the window. In practice, several values of C_n are tried, with the goal of obtaining a stable density estimate (i.e., a reasonable compromise between variance and bias in the estimate). The smallest C_n yielding a stable density estimate is used in the subsequent analysis of the clusters. Typically, high-dimensional data is more spread out in space than low-dimensional data, hence requires a larger value of C_n.

Figure 4a illustrates how window density estimates are used for evaluating clusters. In this two-dimensional example there are three clusters plus a few outlying BEAs. A window centered on the upper cluster has been drawn with four different values of C_n. The window with $C_n = 0.5$ is too narrow to encompass the entire cluster. When C_n is greater than 1.0, the window begins to include BEAs in other clusters or outliers. In this example the proper value of C_n is about 1.0. Figure 4b shows the shape of the window for different values of q.

Another useful technique for validating clusters is to examine the density estimates along the line (in p-dimensional space) between two BEAs. If they are in different clusters there should be a distinct minimum between them. If the BEAs are in the same cluster there may be a maximum in the density or a shallow minimum. Again, this effect will become more or less pronounced as C_n is varied.

We define the modal BEAs in a cluster as those with the largest density values. For one or two variables the window density estimate can be plotted and its maximum located for each cluster. Modal BEAs can be selected by examining these density-contour plots. In our case, with many variables, we evaluate the density at each of the BEAs and rank the results to determine the BEAs in each cluster that have the largest

densities.

If the largest density value achieved over the BEAs in a cluster is relatively high (exceeds 0.4 of the maximum density value achieved over the entire set of BEAs, say), we have reason to believe that the cluster is real and that the modal BEAs for that cluster are typical members of the cluster. These modal BEAs are particularly useful in the interpretation of the cluster. On the other hand, if the largest density value achieved over the BEAs in a cluster is relatively small (less than 0.1 of the maximum density value achieved over the entire set of BEAs, say), it is clear that the cluster is diffuse or nonexistent and cannot be usefully summarized by modal values.

Calculation of the data density thus permits critical assessment of the quality of the clusters obtained from a joining algorithm. Modal BEAs selected on the basis of relatively large density values are far more reliable and intelligible than those obtained by the traditional centroid calculation. (To see this, consider a crescent-shaped cluster. The centroid can lie outside the cluster, so it lacks useful meaning in this case.) Generally speaking, the centroid is overly sensitive to the extreme points in a cluster. Similarly, the density values are more trustworthy than sums of squares calculations in identifying a cluster as tight or diffuse.

Density estimates can also be used to compare the results of different clustering runs. If we let

$$p_{ik} = \begin{cases} [f_n(x_i)]^{1/2} & \text{when BEA } i \text{ is in cluster } k \\ 0 & \text{otherwise,} \end{cases}$$

then we can define the similarity between cluster k and cluster l found in two separate runs by

$$S_{kl} = \frac{\sum_i p_{ik} p_{il}}{[\sum_i p_{ik}^2 \sum_i p_{il}^2]^{1/2}},$$

where the sums are taken over all BEA regions. Note that this

similarity function is defined so that it has a value of one when the two clusters contain the same regions with the same probability distribution, and it has a value of zero when the clusters have no BEAs in common.

Clustering and Selection of Modal BEAs

A preliminary clustering analysis was performed on each of the 14 data sets shown in Table II. The clustering tree printed by the computer program was examined to decide upon the clusters. The number of clusters found varied from three to eight. Information from the clustering tree and from the density estimates were used to judge the validity of the clusters. For some data sets several functional forms for the variables were tried before deciding which gave the best clusters. In general, the extensive data gave better clusters than the intensive data. Density estimates were also used to determine similarities among clusters of different data sets. There was a good deal of similarity among the extensive data sets (except energy production). We believe this is because these data are strongly correlated with population so that the clustering reflected population clusters. Details of these results are not included in this report because of space limitations.

With the knowledge gained from the preliminary clustering and keeping in mind that the purpose of the study was to investigate the impacts of synthetic fuels production, fifteen variables were selected for the final analysis. These are indicated by asterisks in Table II. The population and energy production variables were extensive; the remainder were intensive. The energy production variables were chosen to indicate the presence or absence of physical resources needed for synfuels production. Others were chosen to quantify the urban vs. rural character of the BEAs or their "assimilative capacity" -- the amount of human and financial resources available to support development. Energy consumption variables were not included because we felt they were not relevant to a study that focuses on energy production.

Clustering on Energy Production Variables

We performed clustering on the energy production variables and the rest of the variables separately before combining the two data sets for an overall clustering.

For the energy production variables, the BEAs fall into four clusters with no outliers. The largest cluster containing 66 regions represents BEAs with little or no energy production or reserves. The second cluster contains regions with little or no oil and gas production, but having substantial coal reserves. Cluster 3 has oil and gas production but no coal reserves, while Cluster 4 has both production and reserves. The four clusters are mapped in Figure 5. The algorithm used was not able to put into separate clusters regions with strippable and underground coal. Most likely this is due to the fact that many regions have both.

The density estimates show that the first cluster is extremely compact as is expected since most of the BEAs in it have no production or reserves. The other three clusters have maximum densities relative to Cluster 1 of 0.186, 0.299, and 0.279, respectively. These densities are acceptable in light of the tightness and large number of BEAs in Cluster 1. These clusters are also relatively compact with more than five BEAs in each one having a density within five percent of the maximum.

Clustering on Demographic and Economic Variables

A second clustering was performed on the 11 variables which quantify the demographic, economic and financial status of the regions. The clustering tree indicated that there were five clusters plus 12 BEAs that did not fall into any of the clusters. The clusters are mapped in Figure 6. The first cluster is comprised almost entirely of BEAs in the rural South. Cluster 2 contains rural areas in the South West and the Pacific Northwest. Cluster 3 contains the medium and large sized cities located primarily in the Northeast, Midwest, and the Pacific Coast. The rural areas of the Great Plains and the Rocky Mountain regions make up Clusters 4 and 5, respectively.

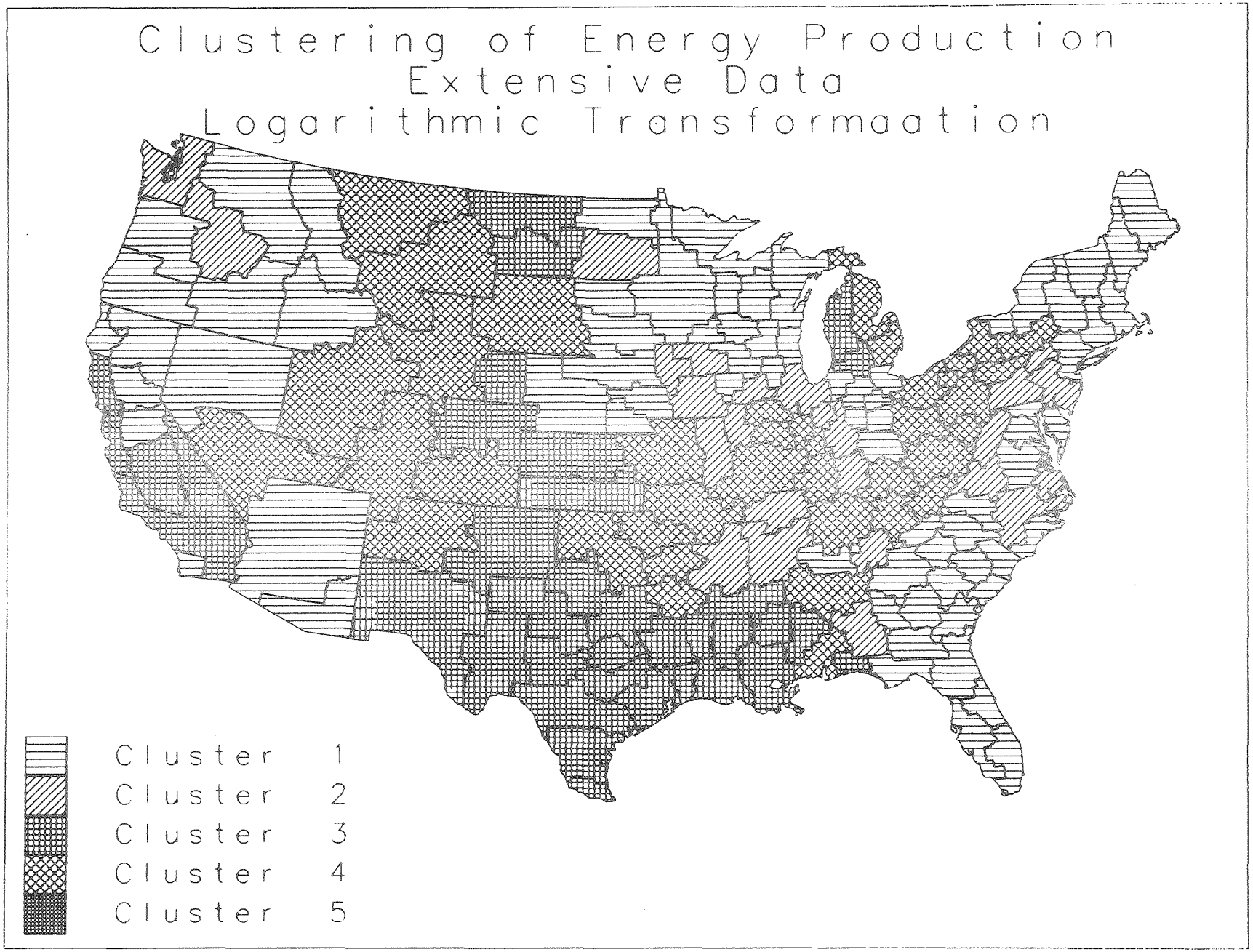
Clustering of Energy Production
Extensive Data
Logarithmic Transformaation
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5

FIGURE 5

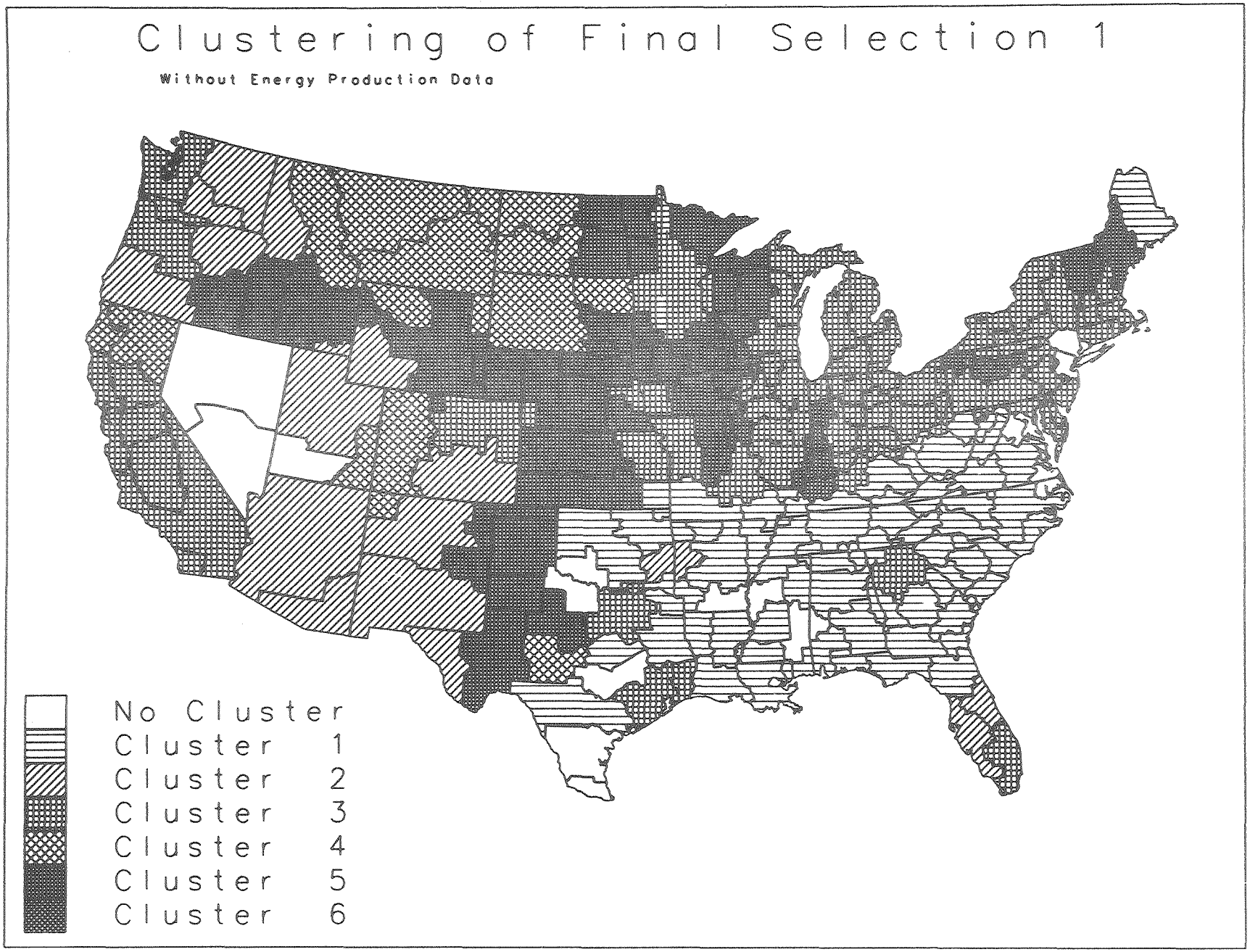
Clustering of Final Selection 1
Without Energy Production Data
No Cluster
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6

FIGURE 6

Clusters 1, 3 and 5 have high densities and appear to be relatively compact. Clusters 2 and 4 are much smaller, containing only 12 and 11 BEAs, respectively. Both the low density estimates and the form of the clustering tree indicate that these are more diffuse.

The modal BEAs in different clusters show distinct differences in the values of the variables. Cluster 1 is characterized by BEAs that are about average in population density, but are above average in percent of non-white population and rate of population growth. They are relatively poor, with an above average per capita debt. Cluster 2 differs from Cluster 1 in having a low population density, high unemployment, and a lower per capita debt. Cluster 3, with its more urban BEAs, shows a higher population and population density, smaller than average population growth, low unemployment, and relatively high income. The BEAs in Cluster 4 are the most rural, having little or no urban areas, extremely low population densities, and low incomes and per capita debts. Cluster 5 is similar to Cluster 1 except that its BEAs are somewhat higher in income, are growing less rapidly, and have a lower percentage of non-whites. These differences give us a broad range of social, economic and demographic conditions under which the impacts of synfuels development can be analyzed.

Clustering on All Variables

The final clustering analysis was performed by combining the energy production data with the demographic, economic and financial data. Our initial analysis showed that there were five clusters. However, for the purposes of synfuels analysis, it seemed appropriate to split the cluster that had many of the coal producing BEAs into two separate clusters -- one containing the eastern and mid-western coal regions, and the other containing the western coal regions. An examination of the density estimates shows that this sixth cluster is indeed distinct. This split also permits us to examine the impacts of oil shale development on two different types of communities. The modal BEAs of these clusters are listed in Table I. The clusters are mapped in Figure 1. In the remainder of this section we will discuss the validity of the clusters;

an interpretation with reference to synthetic fuels production is given in a later section.

Four of the six final clusters have relative densities above 0.85. These appear to be fairly compact. Cluster 1 is somewhat more diffuse, with a maximum density of 0.461. Cluster 6, which consists of six western BEAs originally included in Cluster 3, has a maximum density of 0.278. There are also 12 BEAs that do not fall into clusters.

A closer look at the density estimates shows that the six clusters are not well separated. We examined the density estimates along a line between the modal BEAs in pairs of clusters for several values of the width parameter C_n. For large widths ($C_n > 2$) there is no minimum between modal BEAs in different clusters, while for small widths ($C_n < 1$) in many cases there is a minimum between modal BEAs in the same cluster. We chose an intermediate value of $C_n = 1.5$ for which the density estimates between modal BEAs in the same cluster is always greater than 0.8. Clusters 1 and 6 show good separation from the rest, but there is no minimum between clusters 2 and 3, 2 and 4, 3 and 4, and 3 and 5. Thus it appears that clusters 2, 3, 4, and 5 are not distinct. Their modal BEAs represent relatively compact sub-clusters of a single large cluster. The results of the sensitivity analysis bear out this conclusion.

Examining the results of clustering with and without the energy production variables, we see that only Cluster 4 (rural southern BEAs) did not change significantly. The other clusters were split and recombined by the energy production data to form new clusters. BEAs in the very rural cluster (Cluster 4) using the demographic, economic and financial variables only, for example, were incorporated into Clusters 1 and 6 when energy production data were included. Similarly, BEAs in the urban cluster went primarily into Clusters 1 and 2 with a few in Clusters 3 and 4. The energy production data therefore are important factors in distinguishing between clusters.

DISCUSSION OF RESULTS

Interpretation of Clusters

In this section we examine the final clusters in terms of the values of the variables that differentiate between them. We interpret our results for their appropriateness to an analysis ot the demographic, social and economic impacts of synthetic fuel development.

We characterize clusters by examining the data values for the modal BEAs in each cluster. The reasons that modal values are used rather than mean values over the cluster are discussed above in the section on density estimates. One thing to be aware of in this approach, or any other method that chooses a few cases to represent a multi-dimensional distribution, is the possibility that for some of the data items the modal BEAs may not be typical of the rest of the BEAs in the cluster. Also, by looking at the values of the individual data items, we may be missing the correlations between items that may actually be the determining factors in the clustering and density estimates.

Cluster 1

The BEAs that comprise this cluster are mainly located in the Great Plains and Eastern Rocky Mountain regions. The modal BEAs are in Western Texas, Oklahoma and Kansas. They are areas of low population and low population density. Population densities are in the range 15 to 25 persons per square mile. They have small central cities with little or no suburban population. During the early 1970's the population of these regions increased five percent as compared to a national average of 6.7 percent. They are somewhat above average in labor force participation rate and have a very low unemployment rate of about four percent. (The national unemployment rate was 6.5 percent.) Personal income and per capita local government debt are not distinguishing characteristics.

The BEAs in this cluster tend to be oil and gas producers and may have some strippable coal reserves. There are also extensive oil shale deposits in this area. Some of the BEAs, especially in Kansas and Nebraska, are major grain producers. We can thus examine the impacts of a range of technologies in these BEAs. Their assimilative capacities are nearly as low as those in Cluster 6, but they have not experienced as much growth in recent years. The severity of impacts in this cluster should be less than in Cluster 6.

Cluster 2

This cluster contains BEAs in the rural areas of the Pacific Northwest and the Midwest. The modal BEAs are located in Indiana, Ohio and Illinois. They are relatively small in area, so that even though they are below average in population their population density is high, typically 100 to 200 persons per square mile. The central cities are small. The population of the modal BEAs had been increasing slowly at a rate of less than four percent over a five year period. They have a low unemployment rate and a high rate of labor force participation. Personal income is well above average, with only six to eight percent of the families below poverty level.

Although these BEAs do not have oil and gas production nor coal reserves, they do contain agricultural and forest lands that could be utilized for synthetic fuel production. Many are located in the corn belt and thus would be suitable for large scale ethanol production. In this cluster we could investigate how stable, long-standing communities having relatively large financial and institutional resources are affected.

Cluster 3

BEAs in cluster 3 are located in the coal producing regions of the Midwest and Northern Appalachia. The modal BEAs are in the Midwest. They are above average in both population and population density, and their population has been increasing slowly. Population densities are in the range 80 to 180 persons per square mile, but they have increased by less than five percent during the early 1970's. As in the previous

two clusters, their central cities are small. Their labor force participation rates and unemployment rates are slightly below average, but their income is above average. BEAs in this cluster have more than fifteen percent of the families with incomes above $15,000 per year. The per capita debt tends to be high (about $400 per person). They are characterized by having large reserves of strippable and underground coal. They may also produce some oil and gas.

These BEAs are well suited for an analysis of coal development to produce liquid and gaseous fuels. They are located close to the major industrial regions of the Northeast and Midwest. In terms of assimilative capacity, they are similar to BEAs in Cluster 2. Thus a comparison of the results from these two clusters will point up the differences between synfuels production from biomass and coal in similar communities.

Cluster 4

The BEAs in this cluster are concentrated in the rural South. The modal BEAs have somewhat above average populations and population densities. This cluster is the only one that is characterized by having more than ten percent non-white persons. The population of the modal BEAs increased by about nine percent during the 1970-75 period. The labor force participation rates are slightly above average, and unemployment is low (5.0 - 6.5 percent). However, they are poor areas with an above average per capita debt. Typically fifteen percent of the families are below poverty level, and the per capita debt is about $400. Some of the BEAs in this cluster have oil and gas production and coal reserves.

Cluster 4 is the most suitable for investigating changes in minority population and employment. Both coal conversion and wood waste plants can be located in this region. In comparison to the previous two clusters, these BEAs have a lower assimilative capacity and have been growing more rapidly. Thus the impacts of synfuels development should be felt more strongly.

Cluster 5

BEAs containing the midsized and larger cities fall into this cluster; however, the smaller BEAs seem to be typical. Although the population an population density are high, their central cities are relatively small compared to their outlying areas. Most of them have been growing slowly. They typically have a high labor force participation rate and low unemployment. Per capita income is highest of any cluster, as is per capita debt. Less than ten percent of the families are below poverty level, whereas about twenty percent have incomes of more than $15,000. Most BEAs in this cluster are not energy producers.

Many of the BEAs in Cluster 5 could be used for case studies of biomass conversion facilities. There are abundant wood wastes in the Pacific Northwest and agricultural residues in California and some of the eastern and midwestern BEAs. These would show examples of impacts on relatively wealthy areas, with well developed infrastructures.

Cluster 6

This cluster is similar to Cluster 3 in that the BEAs have large coal reserves as well as oil and gas production. The BEAs are located in the sparsely populated Rocky Mountain States. Typical population densities are less than 20 persons per square mile. Population in this region has been increasing rapidly, with a gain of nearly fifteen percent from 1970 to 1975. The central cities and outlying areas are small. In general this is a poor region; unemployment rates are in the range 7 - 9 percent, and many families have below poverty level incomes. The typical per capita debt of $250 is low, however.

We expect to see some of the largest impacts of synfuels development in BEAs in this cluster. They are especially suitable for studies of coal and oil shale development in areas of very low assimilative capacity. Their human and economic resources are limited, and they lack the institutional infrastructure for dealing with problems of rapid development. Moreover, some have experienced rapid growth in the past few years, so that their resources are already strained.

Five of the six final clusters are comprised primarily of rural BEAs. Even in Cluster 5, which contains urban regions, the modal BEAs are the ones with the smaller cities. Since we expect synfuels development to impact most strongly the smaller communities, our clustering technique has identified a variety of regions in which significant impacts could occur.

Modal BEAs in each of the clusters contain some resources suitable for producing synthetic fuels. Coal for liquefaction and gasification is abundant in Clusters 1, 3, 4 and 6. Clusters 1 and 6 also contain large reserves of oil shale. Most of the modal BEAs can grow crops or have wood wastes or other biomass that can be converted to alcohol. Using our results, the impacts of a variety of technologies in a single community can be investigated, or the impacts of a single technology studied over a range of communities.

BEA Regions Used for Impact Analysis

Six BEA regions were selected as typical areas in which synfuels development might take place. Staff members from the National Laboratories, the Office of Technology Impacts, and Science Applications, Inc. participated in the selection. Modal BEAs in each cluster were considered in decreasing order of estimated density. Each BEA was evaluated in terms of the variables used in the clustering and other considerations such as detailed information on the existence of an adequate resource base, the current status of resource development, and demographic and social factors not included in the clustering. The types of synfuels plants that were appropriate for the region were identified. A region that was suitable for more than one type of plant was given more consideration since it could be used for a comparative assessment. The selected BEAs on which the impacts analysis was performed are listed in Table III.

Table III

BEA Regions Selected for Analysis

Cluster	Selected Region	Regions to Which Analysis May Be Generalized
1	92 Grand Forks, ND	1 Bangor, ME 3 Burlington, VT 86 Wausau, WI 88 Eau Claire, WI 152 Idaho Falls, ID 153 Butte, MN 169 Redding, CA 170 Eureka, CA
2	105 Waterloo, IA	59 Lafayette-West Lafayette, IN 106 Des Moines, IA
3	57 Springfield, IL	10 Erie, PA 11 Williamsport, PA 12 Binghamton, NY-PA 56 Terra Haute, IN 58 Champaign-Urbana, IL
4	49 Nashville, TN	50 Knoxville, TN 51 Bristol, VA-TN 52 Huntington-Ashland, WV-KY-OH 53 Lexington, KY 55 Evansville, IN-KY 119 Tulsa, OK 120 Oklahoma City, OK 137 Mobile, AL
5	157 Portland, OR-WA	21 Richmond, VA 85 Appleton-Oshkosh, WI 107 Omaha, NB-IA
6	146 Albuquerque, NM	145 El Paso, TX 147 Colorado Springs, CO

Sensitivity Analysis

Two major sensitivity analyses were performed to examine the effects of using different definitions of distance and different clustering algorithms. An implicit sensitivity analysis of the clustering was done while selecting the variables to use by considering various combinations

of data elements and transformations. No formal analysis was done; the combination finally used was chosen by comparing the clustering trees for each case. In addition, the effect of changing the width and shape of the density estimate window on how the clusters are interpreted was studied.

The analysis of the sensitivity of the clusters to changing the definition of distance in the average linkage joining algorithm was performed by varying the values of the parameter q in Eq. (3). Three values of q were tried, q = 0.5, 1.0 and 2.0. The latter corresponds to the usual Euclidean distance. Larger values of q are more sensitive to outliers.

Clustering with the Euclidean distance resulted in three major groups with 30 to 40 BEAs in each. One of these clusters resembles Cluster 1 plus a small admixture of BEAs from Cluster 2. The second contains only BEAs that are in Cluster 4, while the third contains BEAs that are in Clusters 2, 3, 4, and 5. The remaining regions showed little tendency to cluster at small distances.

With the intermediate value of the metric parameter (q = 1.0), six clusters were found. However, the structure of the clustering tree is not as clean as for the original q = 0.5 case. Clusters 1 and 6 still show up, but the other four clusters contain BEAs that are in the original Clusters 2, 3, 4, and 5. In addition there are twenty outliers.

This analysis indicates that Cluster 1 is the most stable. Cluster 6, which is small and diffuse, is seen for small values of q. The other four clusters change configuration, indicating that they are not distinct. Since the clustering tree showed the most uniform growth and the smallest number of outliers for q = 0.5, this value was used in the analysis.

The sensitivity to the choice of clustering method was studied by comparing the results from the BMDP average linkage algorithm to those of five other algorithms. These were two additional versions of the average linkage algorithm (centroid sorting and group average) [6], single and complete linkage [7], and the median method of Gower [8] . The

program KLUSTER, obtained from Dan Moore of Lawrence Livermore Laboratory, was modified to read our data formats and to use the q = 0.5 distance function.

The results showed considerable differences between the six algorithms studied. Even the three average linkage algorithms showed differences in detail (the order in which the regions were amalgamated) as well as in overall structure (the number and composition of the clusters). The single linkage algorithm produced one large cluster which was formed by incorporating small clusters and individual BEAs. Gower's method resulted in two large and one small cluster. The large clusters were also formed by amalgamating many smaller clusters. Neither of these two methods seems suitable for clustering the type of data we have.

The three average linkage and the complete linkage algorithms gave better results. The clustering trees showed many small clusters that gradually combined to form four to six larger clusters. Cluster 1 was distinct in these four methods. Portions of the other clusters containing about five to twenty BEAs could be consistently identified in the clustering trees, but they amalgamated differently to form different larger clusters. These results reinforced the conclusion based on examining the density estimates that Clusters 2, 3, 4 and 5 are not sharply separated.

We conclude from thesesensitivity analyses that the clusters we have found are not unique. Their number and composition depend on the joining algorithm and the definition of distance used. Since Clusters 1 and 6 appear in most of the cases we examined, we believe that these are real and distinct clusters. BEAs in Clusters 2, 3, 4, and 5, on the other hand, form different constellations of clusters depending on the clustering method. Although the clusters are not unique, this does not preclude the use of them or their modal BEAs in the impact analysis. This does effect, however, to what extent the results of the impact analysis can be generalized. The question of generalizing the results

is discussed in the following section.

Generalization of Results

One reason for performing a clustering analysis to select BEA regions for the Urban and Community Impact Analysis is to be able to generalize the results obtained for a community in one BEA to communities in similar BEAs. Generalizing to other BEA's requires some criteria for determining how similar two BEAs are. One standard approach is to perform an analysis of variance to determine the within cluster and between cluster sums of squares. If the average within cluster sum of squares is much smaller than the average between cluster sum of squares, one is justified in generalizing the results for one BEA to other BEAs in the cluster.

We do not use this approach because it is very sensitive to outliers as well as being most applicable to normal distributions. Moreover, it uses the centroids rather than the modal BEAs to characterize clusters. Instead, we use the density estimates as one method of determining the extent of the cluster. If the clusters are well separated, then we can choose some minimum density such that any BEA with a density estimate greater than this minimum is in the core of the cluster. To do this we examine the distribution of density along a line in the p-dimensional space between the modal BEAs of two clusters. The minimum along this line gives the density that separates the two clusters. The results of an impact analysis on a modal BEA can with some precautions be generalized to other BEAs in the core of the same cluster.

A second method which can be used if the clusters are not well separated is to perform a sensitivity analysis on the clustering. We cluster using several methods and list those BEAs that cluster tightly with the modal BEA in each cluster. We can then generalize the results to the regions that appear on these lists for most of the clustering methods.

An examination of the density estimates showed that we could not find a minimum value which could separate clusters -- clusters 2, 3, 4, and 5 do not appear to be distinct. Instead, we selected BEAs that, under different clustering methods, most often clustered with the BEA chosen for analysis. Table III lists the regions to which the analytic results may be generalized. For some of the clusters the selected BEAs cluster consistently with only two or three other regions. Our impact analysis is therefore applicable to only 34 of the 173 BEA regions in the country.

In generalizing the results of our analysis, factors specific to synfuels impacts must be considered. First, the impacts will actually be determined at a sub-BEA level, most likely at the county or community level. Second, the fossil fuel, biomass, and other physical resources in the region under study must be adequate to support the development. This second consideration is technology specific. Before the generalizing process can begin, the minimum resource levels necessary for economically viable operation of plants using each of the technologies must be determined. These levels must be applied uniformly to the counties or communities in all clusters. If the facility siting within the BEAs is consistent (i.e., it is located in similar portions of the BEAs), generalizing within clusters is justified. A final consideration is the effects of recent development on the assimilative capacity of the region. A community that has already experienced boomtown effects may have its assimilative capacity greatly reduced relative to other similar communities in the cluster. With the procedures outlined above, we sh uld generate generalizations which are accurate and empirically justified within the clusters, and consistent between case studies and clusters.

Limitations of the Study

Clustering is more of an art than a science. During the course of the analysis judgments must be made concerning the data, the algorithms used, and the validity and interpretation of the results. To have confidence in a clustering analysis, the sensitivity of the results to

these judgments has to be evaluated. Since we are more interested in selecting typical BEA regions rather than classifying all of them into nique clusters, the sensitivity analysis is simplified. In the latter case the clusters must remain stable when the judgmental factors are changed. In our case there must be groups of BEA's that cluster together, and the selected BEA's must stay near the mode of cluster as the factors are varied.

The first decision to be made is usually to select which data items to include or exclude from the analysis. Often this decision is constrained by the availability of data or the resources required to collect and analyze them. Moreover, there may be alternative ways of representing the data - population vs. population density or total vs. per capita income are two examples - and a choice has to be made between them. Once the data are in hand, transformations may be applied which, in effect, change the relative importance of the different variables in assigning the BEA's to different clusters. Transformations can not be avoided when the data items are not commensurate. We have chosen to standardize the data using robust estimates of location and variance, so that problems with units are avoided and the items are approximately equally weighted. Also, logarithmic transformations were used on some of the variables to make their distributions of more normal.

There are many methods for clustering data and, in general, they give different results. The nearest neighbor joining algorithm tends to give long, sausage shaped clusters, whereas the furthest neighbor tends to give compact, ball-like clusters. Furthermore, the choice of a distance function within the clustering algorithm may affect the results. The reality of clusters identified by a particular algorithm must be assessed critically. Comparisons with the results of other algorithms and interpretation of the clusters are essential critical steps. Identification of clusters and modal BEA's on the basis of a single clustering technique is not sufficient.

The clustering methods we use start by joining the two closest BEA regions to form the first cluster. At each succeeding step either a new cluster is formed or another BEA is added to an existing cluster. The

process terminates when all the BEA's have been put into one big cluster. If we visualize the clustering process as a tree with the initial BEA regions as leaves and clusters as branches which finally join together to form the trunk, then we must decide which of the main branches are the clusters that we want. Thus the number of clusters we end up with is also a matter of judgment. However, since the computer program gives information on the distance between the clusters it is joining at each step, the decision is not completely arbitrary.

The use of density estimates is more straight forward, but there is still some room for choice. In particular, the width and shape of the window function can have some effect on the number of clusters found or how we classify a BEA that lies between two clusters. Since we use the density to select BEA's close to the mode of the cluster, the choice of window function is not critical.

A final step where judgment is needed is in interpreting the clusters in terms of the objectives of the study. We need to determine which variables, or combinations of variables, contribute the most to distinguishing between clusters, and the to decide if they are important to the final analysis. This step is made difficult by the high dimensionality of the data. Our approach is to determine the distinguishing factors by examining the differences among the modal BEA's. By using factor or discriminant analysis, a more quantitative determination of these factors might be made if they can be considered normally distributed. Until better methods of displaying relationships in multidimensional space are available, it is possible that significant factors may be overlooked.

This discussion points up the need for a great deal of effort to ensure the validity of the results of this study. Because of time limitations, however, we were only able to test a few of the sensitivities. These tests indicate that the clusters we found were not unique; different choices would lead to different clusters. We have more confidence in our selection of modal BEAs since they depend on finding peaks in the density estimate which is independent of the clustering method.

CONCLUSIONS

This study demonstrates that BEA regions can be grouped in a meaningful way using clustering analysis techniques. Density estimates can be used to select modal BEA regions in the clusters. Modal BEAs in different clusters differ in their demographic, economic, and energy production characteristics, thereby rendering them suitable for analyzing differences in the impacts of synthetic fuels development. However, one must go beyond the clustering analysis to choose the regions and communities for the impact analysis. The judgement of the analyst is needed to weigh the clustering results against other factors that could not be included in the data base. Further judgement is required to determine the extent to which the impact analysis can be generalized to other communities in similar regions.

The clusters found in our analysis are not the only ones that could be found. Varying the clustering algorithm or the distance function would result in a different set of clusters. If a different set of variables were chosen or they were transformed differently, we would again expect to find different clusters. The average linkage joining algorithm with a distance function that is insensitive to outliers gave the best clusters. Transforming the data so that their distributions were more normal also improved the clusters.

For the impact analysis it is less important that we arrive at a unique set of clusters than that we can select regions for analysis with some confidence. The sensitivity analyses indicate that the modal BEAs, which are chosen because they lie in the neighborhood of peaks in the density estimates, tend to cluster together as the clustering method is varied. Thus the modal BEAs are more stable than the clusters and form a valid set of regions on which to perform the impact analysis.

The clustering method we chose resulted in six clusters. Two of the clusters appear to be distinct, but the others appear to be part of a larger cluster that has several regions of higher density. Five of the clusters are composed primarily of rural BEAs where we expect the largest impacts of synfuels development. They are distinguished by differences in their economic and demographic characteristics as well as their

resources for energy production. Modal BEAs in each of the clusters contain fossil or biomass resources needed for producing synthetic fuels. These BEAs are suitable for studying the impacts of a variety of technologies in a single community, or the impacts of a single technology over a range of communities.

Because we are unable to find a unique set of clusters, we are limited in the extent to which we can generalize the impact analysis results. We believe that the results can be extended only to those BEAs that consistently cluster with the ones on which the analysis was done. For the data we used, the results are applicable to only 34 of the 173 BEA regions in the country. In generalizing the results to these BEAs, several factors must be considered: the economic and demographic structure of the communities where the impacts actually occur, the existence of an adequate resource base to support synfuels development, and the effects of recent development within the communities or nearby areas.

Two additional clustering analyses were performed on subsets of the data. The first, which included only the energy production variables, resulted in four clusters. The largest consisted of regions with little or no energy production or reserves; the other three consisted of oil and gas producing regions, coal producing regions, and regions which produce both. The second clustering analysis used the remaining demographic, economic and financial variables. Four of the five clusters found represent rural areas in the South, South West and Pacific Northwest, Great Plains, and Rocky Mountain regions. The fifth cluster contains the medium and large cities of the Northeast, Midwest and Pacific Coast. These clusters can form the basis for further impact analyses. The clusters without energy production could be especially useful in studies of the impacts of increased reliance of renewable resources and conservation.

Clustering analysis is not the appropriate method for classifying BEA regions into a few disjoint clusters based on the type of data we are using because the differences between BEAs appear to be more significant than their similarities. It is thus not valid to perform an impact analysis on a few selected regions and generalize the results to

the rest. Cluster analysis appears to be more useful for finding a larger number of small groups of closely related BEAs. More case studies will then have to be performed, but there will be more confidence in generalizing their results.

ACKNOWLEDGEMENTS

This study was sponsored by the Regional Assessments Division of the Office of Technology Impacts in the U.S. Department of Energy. Ted Harris and Jerry Hinkle of RAD were very supportive of this effort, as was Al Cooke of Science Applications, Incorporated. Dick Stone answered many of the questions we had about the data base that was compiled by Urban Systems Research and Engineering. Several people in the other National Laboratories that participated in the Urban and Community Impact Analysis showed a continuing interest in our analysis and results. In particular, several stimulating conversations were held with Dee Wernette of Argonne and Anne Stroupe of Los Alamos. At LBL, Robin Mogavero assisted in acquiring, converting and running the computer programs for performing the sensitivity analysis. Deane Merrill and Bill Benson gave advice and assistance in the use of SEEDIS (the LBL Socio-Economic - Environmental - Demographic Information System) for extracting data and displaying the results. Will Siri, the Manager of the Energy Analysis Program at LBL, and Nancy Schorn, his Administrative Assistant, were invaluable in keeping the project running smoothly.

REFERENCES

1. Urban Systems Research and Engineering, Inc., "Energy/Environment Data Study", Cambridge, MA (May 1979).

2. Hartigan, J. A., Clustering Algorithms, John Wiley & Sons, New York (1975)

3. Everitt, B., Cluster Analysis, Heinemann Educational Books, Ltd., (1977)

4. Dixon, J. W. and Brown, M. B. (Eds), BMDP Biomedical Computer Programs P-Series, University of California Press, Berkeley, CA (1977)

5. Rosenblatt, M., Ann. Stat. Vol. 3, No. 1, pp 1-14, (1975).

6. Lance, G. N., and Williams, W. T., "A General Theory of Classificatory Sorting Strategies, 1. Hierarchical Systems," The Computer Journal, Vol. 9, No. 4, pp. 373-380, (Feb. 1967).

7. Johnson, S. C., "Hierarchical Clustering Schemes," Psychometrika, Vol. 32, No. 3, pp. 241-254, (Sep. 1967).

8. Gower, J. C., "A Comparison of Some Methods of Cluster Analysis," Biometrics, Vol. 23, No. 4, pp. 623-637, (Dec. 1967).