

**PORTIONS
OF THIS
DOCUMENT
ARE
ILLEGIBLE**

CONF-780317--2

TITLE: COMPUTER GRAPHICS FOR EXTRACTING INFORMATION FROM DATA

AUTHOR(S): Ronald K. Lohrding, Myrie M. Johnson, David E. Whiteman

MAILER

SUBMITTED TO: For presentation at Computer Science and Statistics: 11th Annual Symposium on the Interface to be held on March 6-7, 1978, in Raleigh, North Carolina

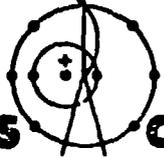
NOTICE
This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represent that its use would not infringe privately owned rights.

By acceptance of this article for publication, the publisher recognizes the Government's (license) rights in any copyright and the Government and its authorized representatives have unrestricted right to reproduce in whole or in part said article under any copyright secured by the publisher.

The Los Alamos Scientific Laboratory requests that the publisher identify this article as work performed under the auspices of the USERDA.

NOTICE

PORTIONS OF THIS REPORT ARE ILLEGIBLE. It has been reproduced from the best available copy to permit the broadest possible availability.


Los Alamos
scientific laboratory
of the University of California
LOS ALAMOS, NEW MEXICO 87545

An Affirmative Action/Equal Opportunity Employer

PEK

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

COMPUTER GRAPHICS FOR EXTRACTING INFORMATION
FROM DATA

by

RONALD K. LOHRDING, MYRLE M. JOHNSON, DAVID E. WHITEMAN

Energy Systems and Statistics
Los Alamos Scientific Laboratory
Los Alamos, NM 87545

Abstract

This paper presents computer graphics which are useful for displaying and analyzing data. Many classical and several newly developed graphical techniques in statistical data analysis are presented for small univariate and multivariate data sets. These include histograms, empirical density functions, pie charts, contour plots, a discriminant analysis display, cluster analysis, Chernoff "faces", Andrews' sine curves, three-dimensional plots, and probability plots.

Recent advances in data collection technology and computer data base management systems have made it imperative to utilize computer graphics for large data sets. Several innovative graphical techniques are presented to handle this situation.

Spatial relationships among the data (particularly geographic data) are difficult to conceptualize. Several cartographic techniques are presented which enhance the understanding of these spatial relationships within the data.

I. INTRODUCTION

The Energy Systems and Statistics Group at the Los Alamos Scientific Laboratory (LASL) is involved in several projects with energy-related data. Some of these projects have small univariate or multivariate data sets, while others have large data sets which require data management systems for efficient data manipulation. A statistically-oriented graphics package is presently under development; numerous modules have been completed. The purpose of this package is to provide graphical techniques for the initial examination of the data. This paper uses data from several projects to demonstrate some of these techniques.

In Section 2, we discuss graphical methods useful for a preliminary analysis of small data sets. In Section 3, graphical techniques which are appropriate for large data sets are presented. Finally, spatial relationships in geographic data sets are explored in Section 4. Throughout this paper, examples of computer graphics are used to illustrate the techniques. (The 35-mm color slides of computer-generated graphics shown at the conference are reproduced in black and white for this paper.)

II. PRELIMINARY DATA ANALYSIS OF SMALL UNIVARIATE AND MULTIVARIATE DATA SETS

Computer graphics for a preliminary raw data analysis may include histograms, empirical distribution function plots and probability plots. The data used in this section was collected on 17 variables for each of the 50 states plus the District of Columbia. The variables and their means and standard deviations are listed in Table I. Of particular interest is the average household BTU consumption per

TABLE I

VARIABLE		MEAN	STANDARD DEVIATION
1. HIBTU	Household BTU per capita (10^6)	87.33	21.30
2. DEGD	Heating degree day loads $\sum_{365}^{65-Y} 0_F$ ($10^3 \text{ } ^\circ\text{F}$) where $Y = \begin{cases} \text{average daily temp. if } Y \leq 65^\circ\text{F} \\ 65^\circ \text{ if } Y > 65^\circ\text{F} \end{cases}$	5.00	2.23
3. MAXT	Normal July maximum temperature ($^\circ\text{F}$)	86.41	5.96
4. PCAIR	Percent of households with air conditioning	33.73	18.44
5. POP	1971 population (10^6)	4.04	4.36
6. FZR	Percent of population with freezers	32.90	10.04
7. ONEP	Single individuals per housing unit	218.63	271.30
8. PCURB	Percent urban population	66.47	15.11
9. COML	Percent commercial sector commercial residential & commercial	36.71	3.31
10. MEDIN	Median income (10^3)	9.17	1.43
11. LOWIN	Percent of family incomes below gov't poverty levels	11.67	5.18
12. SINGLE	Percent of single family houses	71.72	11.19
13. NEWIS	Percent of houses built since 1960	25.91	7.92
14. OLDIS	Percent of houses built before 1950	53.42	12.34
15. AVEIN	Average income per capita (10^3)	3.96	.63
16. LAT	Latitude of center of the state	39.48	6.44
17. LONG	Longitude of center of the state	93.59	19.30

capita (HHBTU). The histogram in Figure 1 shows that the assumption of normality may be questionable. Two graphical tests of normality are shown in Figures 2 and 3. One test uses Lilliefors' test statistic; the other uses a test statistic developed by Lohrding. In the former, the normality assumption is tested by placing $(1-\alpha)100\%$ confidence bounds on the empirical distribution function (edf). The normal cumulative distribution function (cdf) with mean and variance estimated by the sample mean and sample variance is plotted. If the edf falls outside the bounds placed on the cdf, the assumption of normality is rejected at the α level of significance. In the latter, the normality assumption is tested by placing $(1-\alpha)100\%$ confidence bounds on the normal cdf with mean and variance estimated by the sample mean and sample variance. If the edf falls outside the bounds placed on the cdf, the assumption of normality is rejected at α level of significance. In neither test is normality rejected at the 95% level of significance. A normal probability plot and a lognormal probability plot, two additional graphical techniques which may give further insight to the structure of the data, are given in Figures 4 and 5.

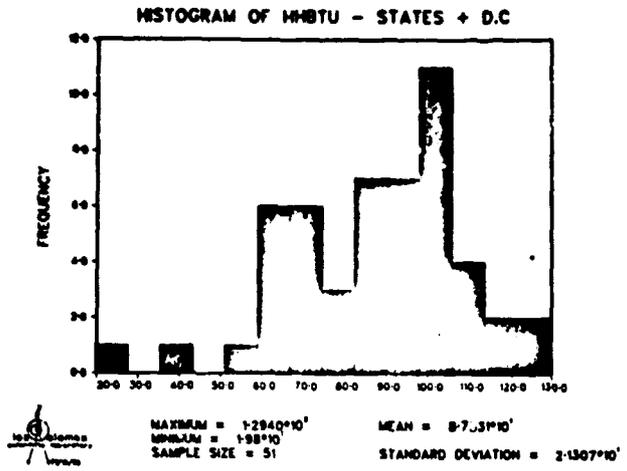


Figure 1

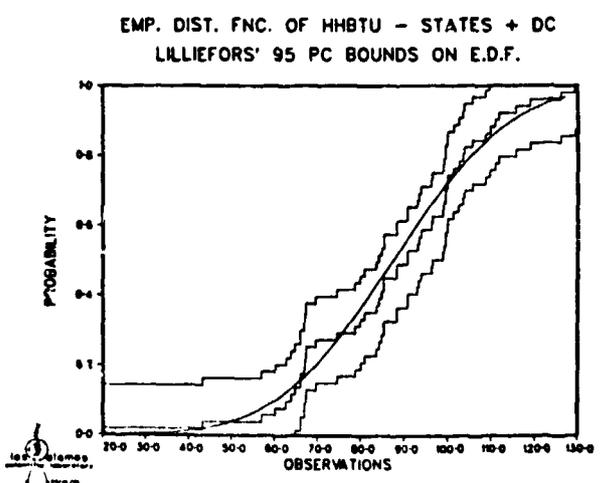


Figure 2

EMP. DIST. FNC. OF HHBTU - STATES + DC
 LOHRDING'S 95 PC BOUNDS ON NORMAL C.D.F.

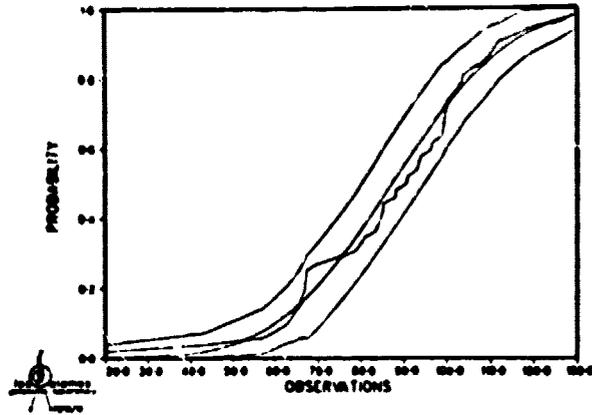


Figure 3

NORMAL PRBLT. PLOT OF HHBTU - STATES + DC

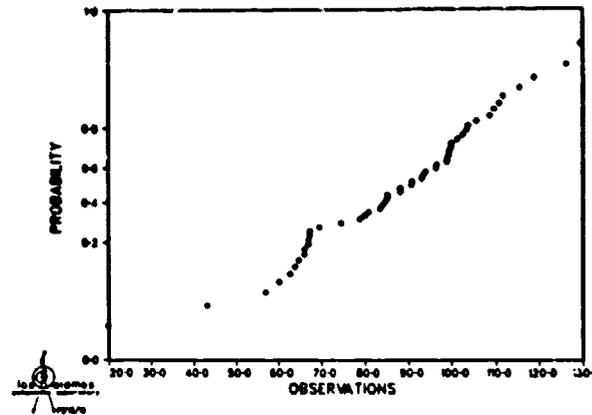


Figure 4

LOG-NORMAL PRBLT. PLOT OF HHBTU - STATES + DC

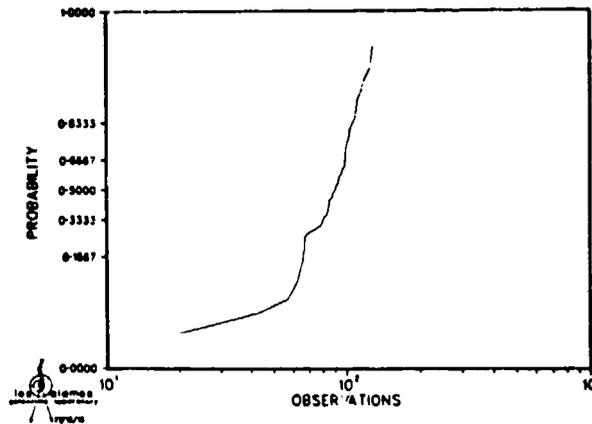


Figure 5

To describe the joint relationship of HHBTU to 26 other variables (including transformations of some of the variables), a linear multiple stepwise regression procedure is used. Seventy-five percent of the variance is accounted for by the variables degree days (DEGD) and percent urban population (PCURB). The equation of the fitted linear multiple regression model is

$$Y_i = 22.155 + 8.657 X_{2,i} + 0.328 X_{8,i}$$

where $i = 1, 2, \dots, 51$

Y_i = HHBTU for the i th state (z axis)

$X_{2,i}$ = DEGD for the i th state (x axis)

$X_{8,i}$ = PCURB for the i th state (y axis).

Figure 6 shows a three dimensional graphical representation where the fitted plane and the data points are plotted. Lines are drawn from the data points to the surface to give some indication of the deviations. In a nonlinear regression analysis, the equation of the fitted model is

$$Y_i = 33.835 - 77.607 \left(\frac{X_{2,i}^2}{X_{3,i}} \right) + 1374.90 \left(\frac{X_{2,i}}{X_{3,i}} \right)$$

where $i = 1, 2, \dots, 51$

Y_i = HHBTU for the i th state (z axis)

$X_{2,i}$ = DEGD for the i th state (x axis)

$X_{3,i}$ = MAXT (maximum temperature) for the i th state (y axis).

The fit of the data to the surface is shown in Figure 7. The two extreme points are Alaska and Hawaii.

Several techniques are available for displaying multivariate data. We first discuss a gray-level coded correlation matrix which displays the pairwise correlations between variables. The gray level scale ranges from positive to zero to negative correlations. Frequently, such a display may be useful in directing attention to interesting variable

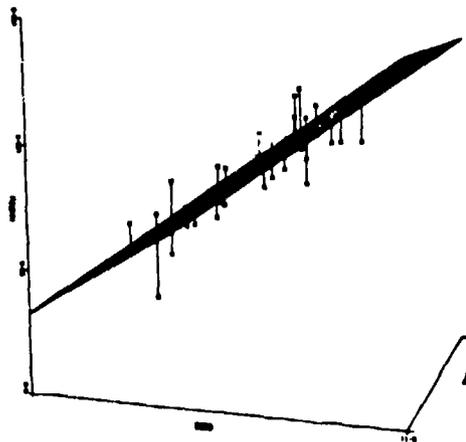


Figure 6

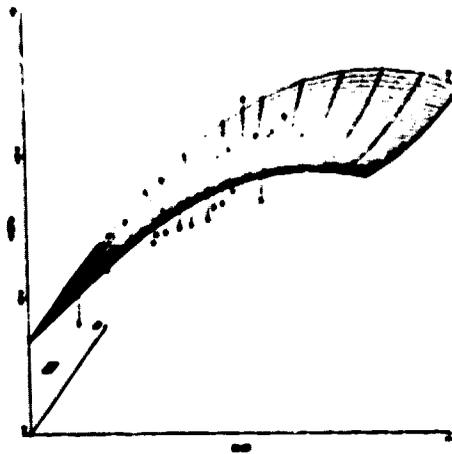


Figure 7

relationships. In Figure 8, note that HBHTU is positively correlated with DEGD, LAT, OLWIS, MEDIN, and AVEIN, negatively correlated with HAXT, PCAIR, LWVIN, SINGLE, and NEMIS, and not correlated with POP, FRZR, ONEP, PCURB, CONL and LONG.

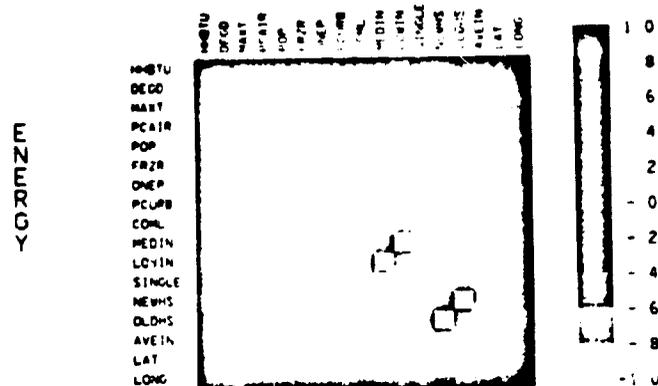


Figure 8

Another technique called Andrews' sine curves uses the standardized data as coefficients of a function involving sines and cosines of t in the range $(-\pi, \pi)$. A function involving the 17 variables was plotted for each of the 50 states plus the District of Columbia to visually cluster similar states. Relatively tight bands suggest clusters. When the original data are used, it is very difficult to separate clusters as shown in Figure 9. However, a plot of the factor coefficients from a principal components analysis in Figure 10 shows three possible clusters of states.

A useful technique in analyzing multivariate data is principal components. As pointed out in ref. 9, the first few and last few principal components are the ones of primary interest. Figure 11 is the plot of component 1 versus component 3. This plot reveals that Alaska, Hawaii, California, and New York are possible aberrant observations.

Another technique for locating possibly anomalous points involves plotting the sum of squared lengths of the projections of the observations on the last few or first few principal component coordinates on a gamma probability plot with a suitably chosen shape parameter. Figure 12 is the plot of the sum of squared projections on the first five components.

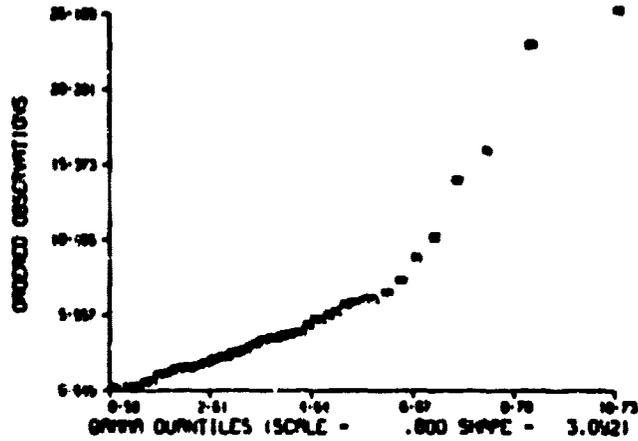


Figure 12

Figures 13, 14 and 15 show the so-called Chernoff faces for the 50 states plus the District of Columbia. Here, a facial characteristic is associated with a variable as indicated in Table II. For example, wide noses correspond to large single populations and long noses correspond to large populations. The faces for New York and California are striking because of this feature. Similarly, Alaska has a wide face because of the large IBIBTU consumption per capita, whereas Hawaii has a thin face.

TABLE II

Pacial Characteristic	Variable
1. Face Width	IBIBTU
2. Brow Length	SINGLE
3. Face Height	MAJT
4. Eye Separation	LAT
5. Pupil Position	AVEIN
6. Nose Length	POP
7. Nose Width	ONEP
8. Ear Diameter	PCURB
9. Ear Level	CONL
10. Mouth Length	DEGD
11. Eye Slant	MEDIN
12. Mouth Curvature	PCAIR
13. Mouth Level	FRZR
14. Eye Level	LOWIN
15. Brow Height	OLDIS
16. Eye Eccentricity	LONG
17. Eyebrow Angle	WENHS

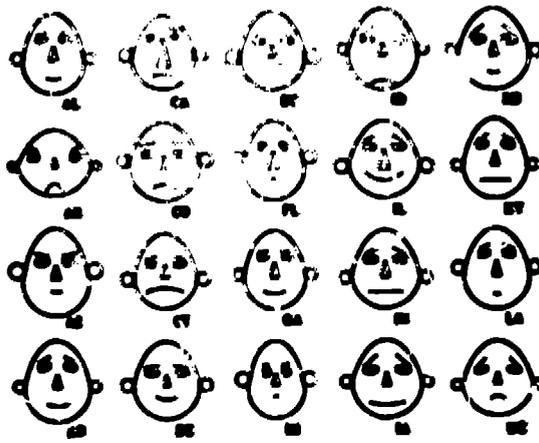


Figure 13

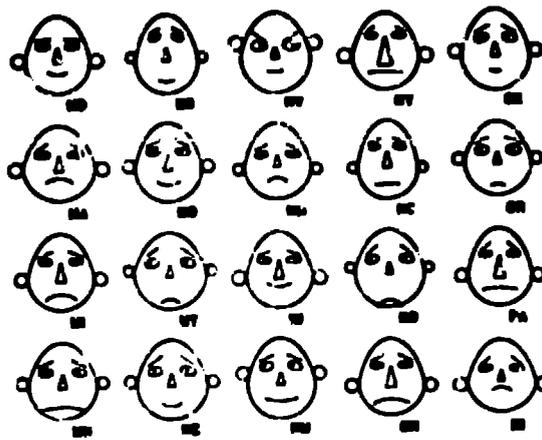


Figure 14

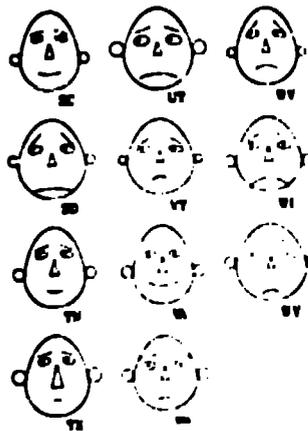


Figure 15

The dendrogram, a tree-like graph of non-overlapping hierarchical partitions, is another visual technique used in cluster analysis. A computer program containing eight clustering techniques (nearest neighbor, furthest neighbor, simple average, group average, median, centroid, Lance and Williams' flexible strategy, and Ward's method) is used. Initially, the data are standardized; both classical and robust standardization techniques are used. Regardless of the standardization and algorithm used, Alaska, California, Hawaii, and New York are distinct from the main cluster.

III. LARGE DATA SETS

In data analysis many of the ensuing problems can be attributed to the data itself--perhaps inaccurate, missing, too little, and recently too much. These large data sets not only create a tremendous storage problem, but challenge computer graphics for effective display techniques.

The analyses considered here deal with National Uranium Resource Evaluation (NURE) data. The objective of this nationwide airborne and stream sediment reconnaissance survey is to classify regions with respect to their potential mineralization. For example, in the stream sediment survey, LASL analyzes the data from five states: Wyoming, Colorado, Montana, New Mexico and Alaska. In the second year of a five-year study, LASL data bases already contain seven million words. Graphical techniques presented here include scattergrams, 3-D and 2-D density plots, a linear discriminant analysis display, contour maps, and moving windows.

Figure 16 is a scattergram of bismuth vs. thallium for all geological formations on one map line in the Lubbock-Plainview area in Texas. The data in the lower left-hand corner represent recent geological formations and most of the formations follow a linear trend except for the data on the right-hand side of the plot where Tl becomes constant with Bi increasing. These data belong to two older formations with known uranium mineralization. Figure 17 shows data for one geologic formation. Scattergrams such as this one are useful in identifying clusters representing misclassified geological formations data.

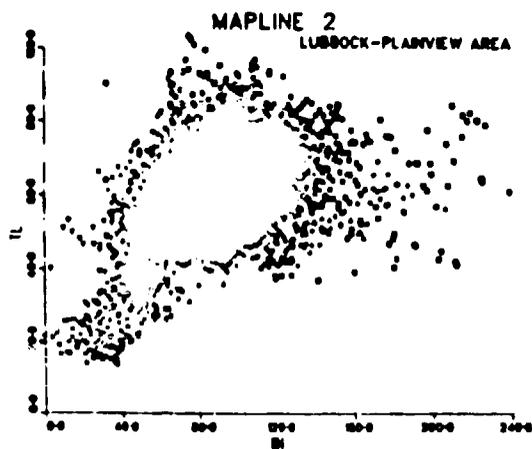


Figure 16

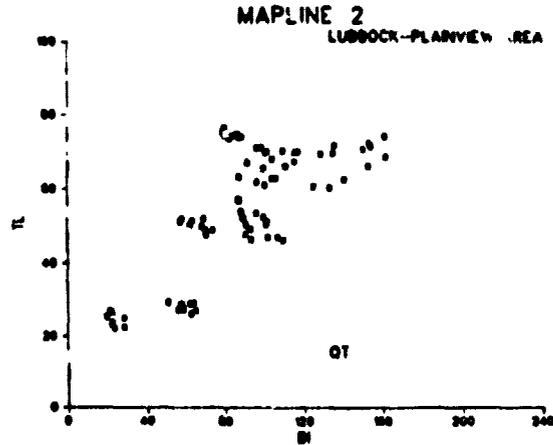


Figure 17

The probability distributions of certain random variables such as thallium signals over a given geological formation of a flight line are thought to indicate uranium concentration. A technique for computing an empirical density function (edf) used to estimate a probability density function has been developed. As many as 100 of these densities, each representing a map line or transect, can be displayed simultaneously as shown in Figure 18. Since some of the edfs may be visually obscured by other edfs, the 3-D plots have been compressed into a 2-D grid plane in a lightness-darkness plot shown in Figure 19.

Figure 20 shows a linear discriminant analysis displayed as a gray-level matrix useful in delineating between favorable and unfavorable regions of uranium mineralization. Each square represents 100 records (i.e., 100 seconds of gamma-ray signals on a map line) in the Lubbock area. The 23 rows represent 23 map lines. There are eight gray levels which are linearly spaced from light to dark over the interval [0,1]. The lighter shades represent low probability of favorable uranium mineralization while darker shades represent high probability of favorable uranium mineralization.

Contour maps of the Lubbock-Plainview area also indicate regions where the probability of finding uranium is high. An example is shown in Figure 21.

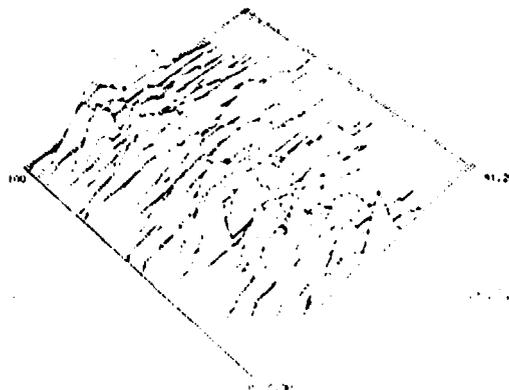


Figure 18

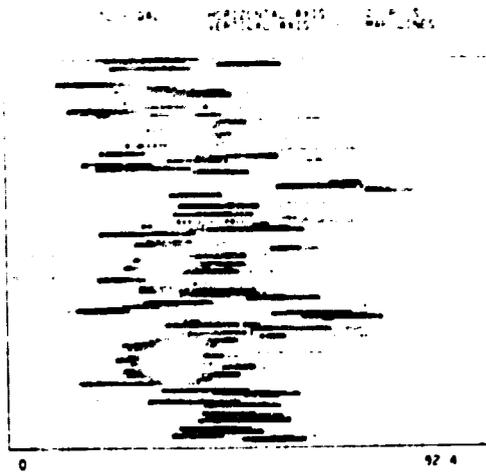


Figure 19

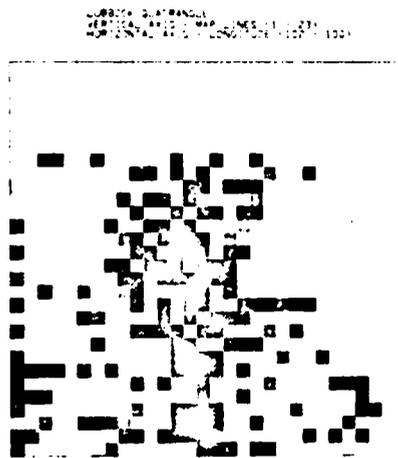
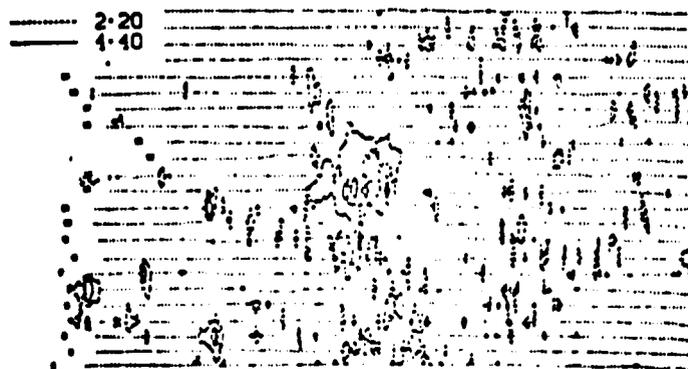


Figure 20



Bismuth-Thallium ratio.

Figure 21

Figure 22 shows a method for detecting clusters of bismuth anomalies. The map of the anomalies in the Lubbock area represents a 2-D Poisson process. A rectangular moving window 6 miles wide and 8.5 miles (or 3 map lines) high is used to identify clusters containing 5, 6, and 7 or more anomalies at a specified probability level.

Figures 16-22 represent different ways of displaying data analysis techniques. It is interesting to compare these displays as to whether they indicate the same favorable areas of potential mineralization.

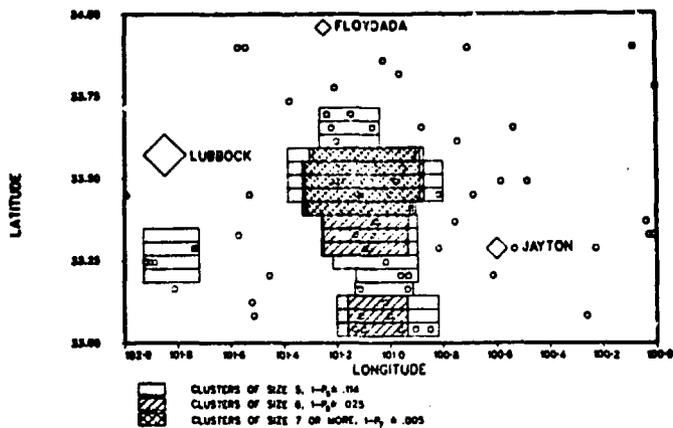


Figure 22

IV. CARTOGRAPHIC DATA SETS

Maps are very useful in displaying and communicating information contained in data with spatial/geographic relationships. The figures shown are applications of cartographic techniques and have been extracted from various on-going projects.

Figure 23 summarizes U.S. offshore oil and gas lease data from October 1954 - November 1976. The number of leases, the leasing years, the acreage and the producing acres through 1974 are given for individual states and regions as well as totals for all the leases. Of the total 2,260,000 producing acres, Louisiana has 2,116,000 acres and Texas has 118,000 acres.

Figures 24 - 26 are for a study of the impacts of electric power generation in the West. The location of existing and proposed power plants by type for the Western and Rocky Mountain regions are shown in Figure 24. The letters represent the type of plant, i.e., coal, oil, gas and nuclear. The size of the letters indicate three levels of power generation: small, 500-999 MWe, medium, 1000-1999 MWe, and large, 2000+ MWe. The Los Angeles and San Francisco areas have a number of oil-fired plants and these areas are simply shaded. Figures 25 and 26 are maps to study pollution dispersion patterns. Figure 25 shows SO₂ concentration in Southwest Wyoming for 1985 with pollution contours drawn every 0.25 $\mu\text{g}/\text{m}^3$. Figure 26 shows change in length of life due to pollution in days per person. Similar graphical displays

were done for exposure to suspended particulates, additional restricted activity days due to pollution and annual morbidity costs per person and per town.

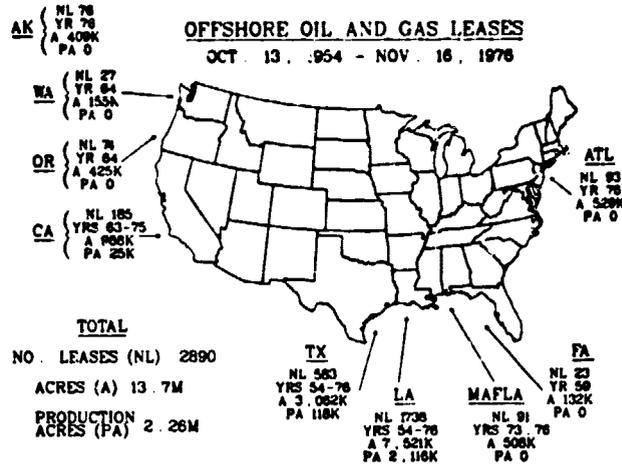


Figure 23

WSCC POWER PLANTS



Figure 24

SO₂ CONCENTRATION IN SW WYOMING, 1985.

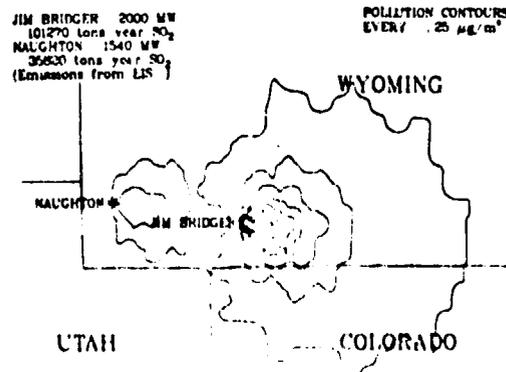


Figure 25

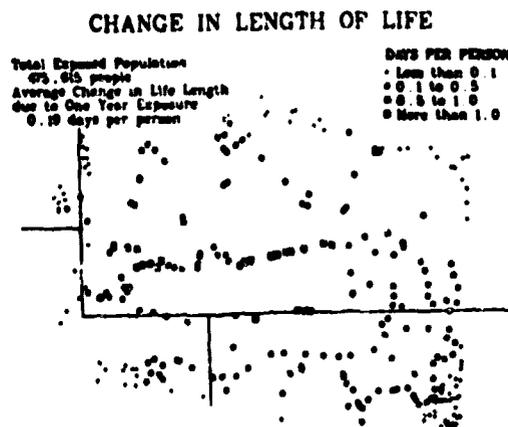


Figure 26

Computer-generated images aid in the study of the effect of various contaminants on visibility. Photographs of landscapes have been digitized on a microdensitometer in wavelengths corresponding to red, blue and green light. Data on transmitted and total radiation are calculated and converted to equivalent densities using a plume visibility code. The densities are superimposed on the image to form the pollution cloud. The color and extent of the cloud are determined by the type of pollutant, different control technologies and varying meteorological conditions.

Figures 27 - 29 are from solar feasibility studies. The first map shows heating degree days which is the average of the high and low temperatures subtracted from a 65° base temperature for the 48 contiguous states. Simply, the colder the climate, the higher the number of heating degree days. Contrast Florida with 214 and Maine with 7511. The second map shows 1977 residential gas prices in dollars per thousand cubic feet by state. Gas is generally cheaper in the southern, central and Rocky Mountain regions. Note that Maine has higher prices than nearby Vermont and New Hampshire. Figure 29 shows the pattern of economic feasibility for domestic hot water under incentives provided by the National Energy Plan of April 1977 and the House Modification of that plan.

Heating Degree Days

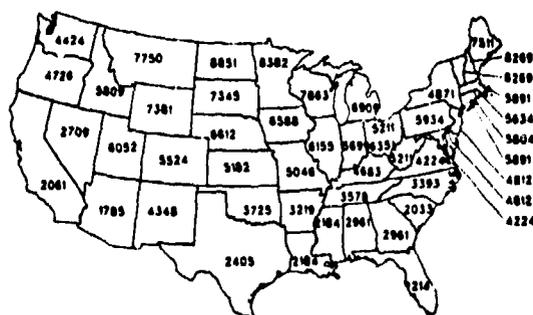


Figure 27

1977 RESIDENTIAL GAS PRICES
DOLLARS PER MCF

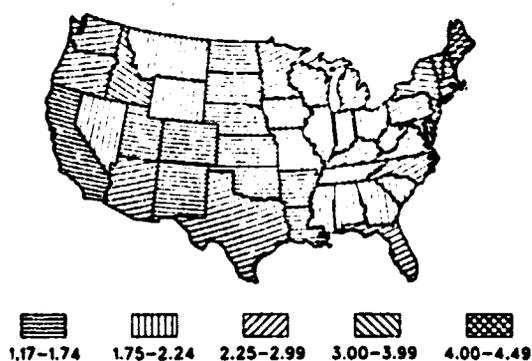


Figure 28

SOLAR FEASIBILITY - DOMESTIC HOT WATER
ALTERNATIVE SYSTEM - ELECTRIC RESISTANCE
10 YEAR LIFE CYCLE COSTING

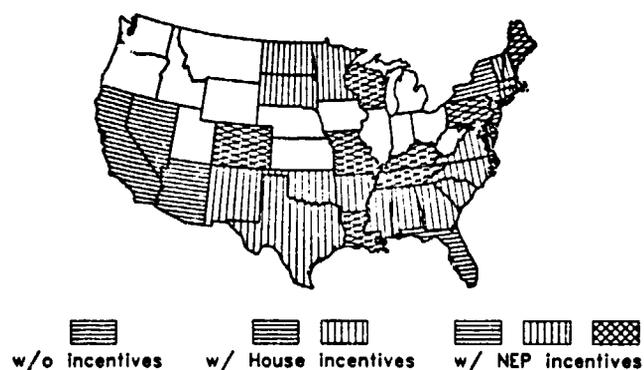


Figure 29

Figures 30 - 37 are maps displaying energy-related data from the regional studies program. Figure 30 shows the five coal export-import regions and Figure 31 is a flow map for the export of Rocky Mountain coal. The circle represents the within region total and the thickness of the arrows represents relative amounts of export to the other four regions. Bar charts and pie charts are useful in displaying energy totals for regions or states. Production, consumption, export and negative export (import) figures are displayed in Figures 32 and 33 using shaded bars. Figure 34 uses varying sized circles to indicate production levels by region. Sections of a circle are shaded differently to indicate coal, oil and natural gas, hydro, nuclear and other and uranium production. Figure 35 shows county air quality maintenance area data for the Rocky Mountain region. An interactive composite geo-information mapping system known as GMAPS provides map data on such items as wilderness areas, ecosystem trends, locations of natural resources, etc. for selected regions in the U.S. Figure 36 shows types of coal fields and Figure 37 is a composite of coal fields with oil shale basins in the Rocky Mountain region.

COAL EXPORT-IMPORT REGIONS

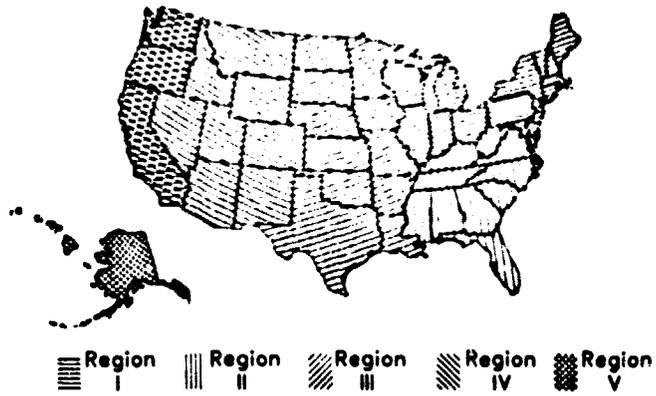


Figure 30

ROCKY MOUNTAIN COAL

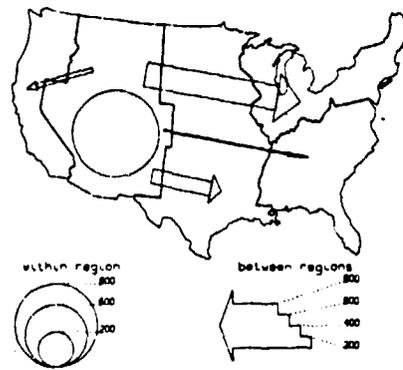


Figure 31

1975 REGIONAL ENERGY TOTALS

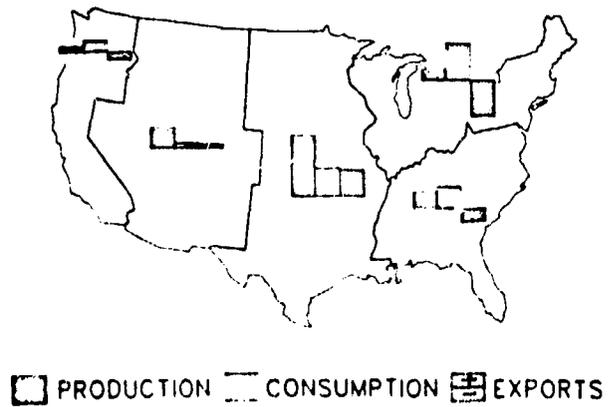


Figure 32

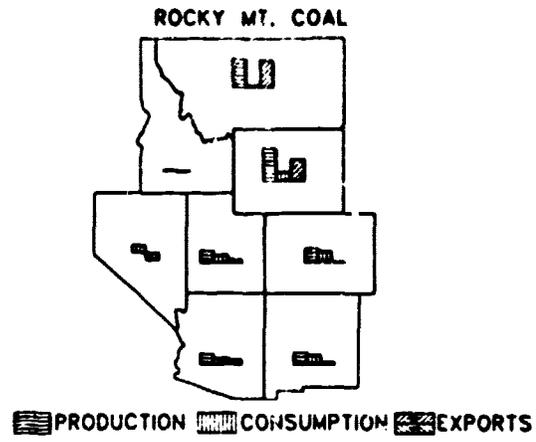


Figure 33

1975 REGIONAL ENERGY PRODUCTION

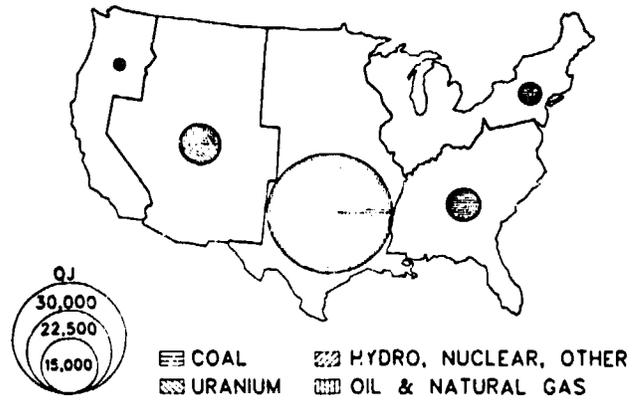


Figure 34

**AIR QUALITY MAINTENANCE AREAS
ROCKY MOUNTAIN STATES**

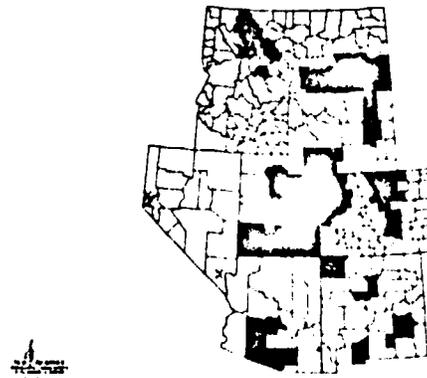


Figure 35



Figure 36

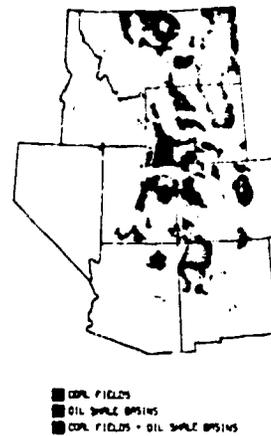


Figure 37

V. SUMMARY

The computer-generated graphic products described in this paper represent a variety of techniques for displaying and analyzing small univariate and multivariate data sets, large data sets and cartographic data sets. Computer graphics are useful tools for communicating information efficiently and effectively.

ACKNOWLEDGMENTS

We wish to thank Katherine Campbell and Mona Wecksung for the use of some of their slides, and Thomas Bement for assisting us on the cluster analysis. Melvin Prueitt supplied the 3-D plotting program, PICTURE.

BIBLIOGRAPHY

1. D.F. Andrews, "Plots of High Dimensional Data," *Biometrics*, 28, 125-136 (March 1972).
2. T.R. Bement, D.V. Susco, D.E. Whiteman, R.K. Zeigler, "National Uranium Resource Evaluation

Program," Los Alamos Scientific Laboratory Report LA-6501-PR (May 1977).

3. L.A. Bruckner, M.M. Johnson, G.L. Tietjen, "The Analysis of Lease, Production and Revenue Data from Offshore Oil and Gas Leases," Los Alamos Scientific Laboratory paper presented at the Second ERDA Statistical Symposium, Oak Ridge, TN (October 1976).
4. K. Campbell, "INGPROC - An Image Processing Program for CDC 6600 and 7600 Computers," Los Alamos, NM (November 1974).
5. H. Chernoff, "The Use of Faces to Represent Points in K-Dimensional Space Graphically," Journal of the American Statistical Association, 68, 361-368 (June 1973).
6. W.J. Conover, Practical Nonparametric Statistics (John Wiley & Sons, Inc. New York, 1974).
7. W.J. Conover, T.R. Bement, R.L. Inen, "On a Method for Detecting Clusters of Possible Uranium Deposits," submitted to Technometrics.
8. A. Ford and H.W. Lorber, "Methodology for the Analysis of the Impacts of Electric Power Production in the West," Los Alamos Scientific Laboratory Report LA-6720-PR (January 1977).
9. R. Gnanadesikan and J.R. Kettenring, "Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data," Biometrics, 28, 81-124 (March 1972).
10. R. Gnanadesikan, Methods for Statistical Data Analysis of Multivariate Observations (John Wiley & Sons, Inc., New York, 1977).
11. G.N. Lance and W.T. Williams, "A General Theory of Classificatory Sorting Strategies. I. Hierarchical Systems," Computer J., 9, 373-380 (1967).
12. E.M. Leonard, M.D. Williams, J.P. Mutschlechner, "The Visibility Issue in the Rocky Mountain West," Los Alamos Scientific Laboratory Report LA-UR-77-2558 (September 1977).
13. R.K. Lohrding, "Statistical Analysis and Display of Energy-Related Data," Los Alamos Scientific Laboratory paper presented at the ERDA - Wide Conference on Computer Support of Environmental Science and Analysis, Albuquerque, NM (July 1975).
14. R.K. Lohrding, "Comparative Power Studies of Some Tests of Normality," Los Alamos Scientific Laboratory Report LA-5101-MS (November 1972).
15. R.A. Waller, E.A. Nonash, and J. Lohrenz, "Some Computerized Graphic Technical Applications for Federal Mineral Lease Management Support," Los Alamos Scientific Laboratory paper presented at the First Computer Symposium, Reston, VA (March 1977).
16. J.H. Ward, Jr., "Hierarchical Grouping to Optimize an Objective Function," J. Amer. Statist. Assoc., 58, No. 301, 236-244 (1963).
17. M. Wecksung, R. Wiley, and K. Turner, "GNAPS User's Manual," Los Alamos Scientific Laboratory Report LA-6975-M.
18. M.B. Wilk, R. Gnanadesikan, M.J. Huyett, "Estimation of Parameters of the Gamma Distribution

Using Order Statistics," *Biometrika*, 49, Nos. 3 and 4, 525-545 (1962).

19. M.B. Wilk, R. Gnanadesikan, N.J. Huyett, "Probability Plots for the Gamma Distribution," *Technometrics*, 4, No. 1, 1-20 (February 1962).
20. J.W. Wood, "A Computer Program for Hierarchical Cluster Analysis," *Newsletter of Computer Archaeology*, 2, No. 4, 1-11 (June 1974).