NASA

Technical Memorandum 85009

Experimental Philosophy Leading to a Small Scale Digital Data Base of the Conterminous United States for Designing Experiments with Remotely Sensed Data

M. L. Labovitz, E. J. Masuoka, P. W. Broderick, T. R. Garman, R. W. Ludwig, G. N. Beltran, P. J. Heyman, and L. K. Hooker

APRIL 1983

National Aeronautics and
Space Administration

Goddard Space Flight Center
Greenbelt, Maryland 20771

TM 85009

EXPERIMENTAL PHILOSOPHY LEADING TO A SMALL SCALE DIGITAL DATA BASE OF THE CONTERMINOUS UNITED STATES FOR DESIGNING EXPERIMENTS WITH REMOTELY SENSED DATA

M. L. Labovitz
E. J. Masuoka

Geophysics Branch
Earth Survey Applications Division
NASA/Goddard Space Flight Center
Greenbelt, MD

P. W. Broderick
T. R. Garman
R. W. Ludwig
G. N. Beltran
P. J. Heyman
L. K. Hooker

Department of Geology
University of Maryland
College Park, MD

April 1983

GODDARD SPACE FLIGHT CENTER
Greenbelt, MD 20771

EXPERIMENTAL PHILOSOPHY LEADING TO A SMALL SCALE DIGITAL DATA BASE OF THE CONTERMINOUS UNITED STATES FOR DESIGNING EXPERIMENTS WITH REMOTELY SENSED DATA

ABSTRACT

Research using satellite remotely sensed data, even within any single scientific discipline, has often lacked a unifying principle or strategy with which to plan or integrate studies conducted over an area so large that exhaustive examination is infeasible, e.g., the U.S.A. However, such a series of studies would seem to be at the heart of what makes satellite remote sensing unique, that is the ability to select for study from among remotely sensed data sets distributed widely over the U.S., over time, where the resources do not exist to examine all of them. What we do lack is the previously noted strategy to aid in the development of formal testable hypotheses and the selection of study locations so as to minimize the number of samples subject to the ability to construct desired inferences.

Using this philosophical underpinning and the concept of a unifying principle, we have constructed an operational procedure for developing a sampling strategy and formal testable hypotheses. We believe the procedure to be applicable across disciplines, when the investigator restates the research question in symbolic form, i.e. quantifies it.

The procedure is set within the statistical framework of general linear models. The dependent variable is any arbitrary function of remotely sensed data and the independent variables are values or levels of factors which represent regional climatic conditions and/or properties of the earth's surface. These factors are operationally defined as maps from the U.S. National Atlas (U.S.G.S., 1970). Eighty-five maps from the National Atlas, representing climatic and surface attributes, were automated by point counting at an effective resolution of one observation every 17.6 km (11 miles) yielding 22,505 observations per map. The maps were registered to one another in a two step procedure producing a coarse, then fine scale registration. After registration, the maps were iteratively checked for errors using manual and automated procedures. The "error-free" maps were anotated with identification and legend information and then stored as card images, one map to a file.

A sampling design will be accomplished through a regionalization analysis of the National Atlas data base (presently being conducted). From this analysis a map of "homogeneous regions" of the U.S.A. will be created and samples (Landsat scenes) assigned by region.

While designed for use with remote sensing experiments, the data base, the method of analyzing it and the philosophy behind it are general enough to serve as a framework for other studies being conducted over large portions of the U.S.A.

EXPERIMENTAL PHILOSOPHY LEADING TO A SMALL SCALE DIGITAL DATA BASE OF THE CONTERMINOUS UNITED STATES FOR DESIGNING EXPERIMENTS WITH REMOTELY SENSED DATA

MOTIVATION AND OBJECTIVE

To date satellite remote sensing has been characterized by repetitive coverage of large areas resulting in large volumes of data. As such this data set is unique and still unexamined for its potential utility in the detection of patterns over large regions, over time. Conceptually, the types of patterns we wish to detect in the remotely sensed data are those we can relate to the variation in regional climatic conditions, and/or properties of the earth's surface. These patterns are often intimately associated with the solutions to research problems in several disciplines. For example, classifiers and decision rules are applied to remotely sensed data in a variety of disciplines. An important question to address then becomes, does the goodness of fit of a particular classifier vary as a function of location and time of year? Other examples of patterns of interest could be important inputs into setting parameters of a remote sensing system or the strategy for its use. These would include the amount of data compression achievable over a given location, or, with pointable systems, the number of times that an area should be imaged over a given time period.

There would be little dispute that the answer is yes to the question, do such variables vary with location and time of year? However, there is also little information in such an answer. Therefore in order to examine the variables fo. such patterns, the hypotheses are equivalenced to a series of analyses involving, for example, changes in frequency distributions of digital numbers with location (frequency distributions are fundamental attributes of classifiers), or measures of relatedness of neighboring pixels such as the autocorrelation function (related to data compress), spatial and textural measures. To complete the search for patterns in these functions of the remote sensing data, the functions are modeled as functions of suitably operationalized ancillary factors, i.e. attributes over space and time of the ground location.

In such an analysis scheme the patterns of interest are computationally intensive, and the problem of detecting patterns is further complicated by the vast quantities and the cost of data. Therefore, any analysis scheme must be automated (hence quantitative) and use only a portion or sample of the available data.

While a great deal of research with satellite remotely sensed data has been performed in a number of disciplines, little thought has been given a priori on how to integrate studies at different locations. Even less thought has been given to selecting a series of study sites, a priori, based upon some unifying conception. The approach of combining "grab" samples to infer any pattern might be called a "bottom up" design and is a very inefficient, potentially biased sampling design. The alternative is what might be called a "top-down" experimental design. The objective of this paper is to present a data base and philosophy which we are using in just such a "top down" approach as the rationale for selecting samples (Landsat scenes) and for generating testable hypotheses about variation in remotely sensed data.

THE ANALYSIS SETTING

Methods for selecting representative subsets of individuals and discerning patterns of variation are formalized within the realm of statistics and we shall motivate the data base within such a framework. More specifically, the target population is the conterminous United States. The individual elements of the population are the 80 m pixels described by Landsats 1, 2 and 3. The variables of interest are the reflected energies occurring in the four spectral bands, recorded by the satellite at each pixel. We might then set about collecting a simple random sample of pixels. However, such a sampling scheme is neither practical nor efficient for discerning patterns. An alternative is to group pixels first and then sample within the group. Reasons are as follows:

1. The geographic or spatial position of a pixel is a fundamental property of the data set.
2. Some of the derived variables are functions of groups of pixels.
3. A pixel is not precisely the same location for every pass of the satellite.

4. As Landsat data for any given pixel is only available by acquiring a nominal scene, it would be reasonable to use the path-row system of nominal scenes as a grouping factor.

Clearly we are describing a multistage sampling strategy with important stages being initially the selection of representative sets of scenes and secondly, pixels within a scene. Further, we have regressed a portion of the sampling question to another scale – the selection of scenes. A strategy for this selection is developed below.

The adoption of the path-row scheme may also be viewed as a stratification of the population. In general, the rationale for imposing a stratification upon a population is that the researcher believes, that when the elements of the population are grouped according to a set of attributes other than the random variable(s) under consideration, the within group variation of the random variables is less than that for any combination of two or more of the groups.

Therefore, by judicious selection of grouping of factors, the researcher can formulate and test hypotheses about the population. Indeed, in this experimental environment, a stratification is equivalent to a hypothesis. Note that the discovery of a stratification which satisfactorily reduces the ratio of the within group variance to the total variance is precisely what we mean by the discernment of patterns.

Such an experiment can be effected operationally by modeling the value of the random variables as a general linear model. Under this approach the strata are known as treatments which are combinations of specific values of the grouping factors. Once the treatments are established, the random variables are expressed as linear functions of the effects of grouping factors and the interactions between these factors. For example:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$$

is a general linear model,

where:

Y_{ijk} is the random variable measured on the kth element of the i, j treatment (stratum);

μ is an overall mean;

α_i is the effect or contribution to the random variable from the ith level of factor A, which occurs at I levels (i = 1, . . . , I);

β_j is the contribution to the random variables from the jth level of factor B, which occurs at J levels (j = 1, . . . , J);

$\alpha\beta_{ij}$ is an interaction arising from factor A occurring at the ith level and factor B occurring at the jth level;

ϵ_{ijk} is an error term.

In this model, every combination of a specific i and j represents a treatment and from our previous argument defines a particular subpopulation or stratum. Analyses of variance (ANOVA) theory (Scheffe, 1959) provides the method for partitioning variation in the random variables among factors and test statistics which can be used to determine which factors are explaining significant amounts of variation.

In addition to representing a set of hypotheses to be tested, a particular stratification can be used to determine the total number and allocation of samples. For example, the total number of samples should be proportional to the total number of strata with the number of samples allocated to a particular stratum equal to the proportion of the total population represented by the stratum times the total number of samples.

OPERATIONALIZING THE APPROACH:

The Factors

Clearly, the problem of setting up a stratification in order to formulate hypotheses or select Landsat scenes (samples) is now recast as the selection of factors and the description of the distribution of their values (levels) over the population. In setting up a stratification for use with

remote sensing reasonable factors to use include climatic variables, properties of the land surface and their interactions. These factors were operationally defined as the maps of the National Atlas produced by the U.S. Geological Survey (USGS, 1970). We selected 85 maps from the National Atlas (see Table 1) and transformed them into data sets we could manipulate. This quantification was achieved by placing transparent grids on each map and recording a code for the value of the map occurring at each row column interaction, a procedure commonly called point counting. All maps were on an Albers equal area projection base (Maling, 1973), so square grids were used to point count all the maps.

The maps which were point counted possessed three scales. Maps having a scale of 1 to 7,500,000 or 1 to 17,000,000 were point counted so that an observation was taken every 17.6 km (11 miles) across the U.S.A. A third set of maps displaying monthly climatic variables was mapped at a scale of 1 to 34,000,000. For these maps grids were used so that an observation was recorded every 35.2 km (22 mi.) across the U.S.A. Since these climatic maps possess broad contours, we were fairly confident in doubling the maps in both an east-west and north-south direction to achieve the 17.6 km data collection increment. Some of the contour lines on the "quadrupled" maps are saw-toothed, but considering the degree of smoothing, the standards for map accuracy as well as the reproduction methods, one might argue that these generated contour lines are clearly within the level of uncertainty of the maps.

Each point counted map consists of a rectangular array 154 rows by 258 columns, yielding 39,732 points. Of these points, 17,227 are coded as outside the study area (for example, the 1st and 154th rows are completely outside the study area). These points lie either in open ocean, Canada or Mexico. For the remaining 22,505 points, 1185 points are coded as inland water. Included in the inland water category are not only interior lakes, but also bays and bodies of water on the landward side of islands, water between barrier islands and water between peninsulas and the coast.

Table 1
Maps in National Atlas Data Base

Map	Title	Sheet Number(s)	Map Code	Scale (1:x million)	Grid Spacing (Obs. Per cm)	Resolution* (1 Pt. Per X km)
1	Major Forest Types	154–155	1	7.5	3.9	17.6
2	United State General Reference (States Only)	2–3	1B	7.5	3.9	17.6
3	Geology	74–75	2	7.5	3.9	17.6
4	Potential National Vegetation	90–91	3	7.5	3.9	17.6
5	Tectonic Features	70–71	4	7.5	3.9	17.6
6	Classes of Land-Surface Form	62–63	6	7.5	3.9	17.6
7	Annual Solar Radiation	93	7A	17	8.8	17.6
8	January Solar Radiation	93	7B	34	8.8	35.2
9	April Solar Radiation	93	7C	34	8.8	35.2
10	October Solar Radiation	93	7D	34	8.8	35.2
11	July Solar Radiation	93	7E	34	8.8	35.2
12	Mean Annual Sunshine	96	7F	17	8.8	17.6
13	Mean Annual Pan Evaporation	96	7G	34	8.8	35.2
14	Mean May–October Evaporation, Percent of Annual	96	7H	34	8.8	35.2
15	Topographic Relief	59	8A	17	8.8	17.6
16	Physiographic Divisions	60	8E	17	8.8	17.6
17	Mean Monthly Sunshine, January	94	9A	34	8.8	35.2
18	Mean Monthly Sunshine, February	94	9B	34	8.8	35.2
19	Mean Monthly Sunshine, March	95	9C	34	8.8	35.2
20	Mean Monthly Sunshine, April	95	9D	34	8.8	35.2
21	Mean Monthly Sunshine, May	95	9E	34	8.8	35.2
22	Mean Monthly Sunshine, June	95	9F	34	8.8	35.2
23	Mean Monthly Sunshine, July	95	9G	34	8.8	35.2
24	Mean Monthly Sunshine, August	95	9H	34	8.8	35.2
25	Mean Monthly Sunshine, September	94	9I	34	8.8	35.2
26	Mean Monthly Sunshine, October	94	9J	34	8.8	35.2
27	Mean Monthly Sunshine, November	94	9K	34	8.8	35.2
28	Mean Monthly Sunshine, December	94	9L	34	8.8	35.2

Map	Title	Sheet Number(s)	Map Code	Scale (1:x million)	Grid Spacing (Obs. Per cm)	Resolution* (1 Pt. Per X km)
29	Mean Monthly Minimum Temperature, January	106	10A	34	8.8	35.2
30	Mean Monthly Minimum Temperature, February	106	10B	34	8.8	35.2
31	Mean Monthly Minimum Temperature, March	107	10C	34	8.8	35.2
32	Mean Monthly Minimum Temperature, April	107	10D	34	8.8	35.2
33	Mean Monthly Minimum Temperature, May	107	10E	34	8.8	35.2
34	Mean Monthly Minimum Temperature, June	107	10F	34	8.8	35.2
35	Mean Monthly Minimum Temperature, July	107	10G	34	8.8	35.2
36	Mean Monthly Minimum Temperature, August	107	10H	34	8.8	35.2
37	Mean Monthly Minimum Temperature, September	106	10I	34	8.8	35.2
38	Mean Monthly Minimum Temperature, October	106	10J	34	8.8	35.2
39	Mean Monthly Minimum Temperature, November	106	10K	34	8.8	35.2
40	Mean Monthly Minimum Temperature, December	106	10L	34	8.8	35.2
41	Mean Monthly Average Temperature, January	102	11A	34	8.8	35.2
42	Mean Monthly Average Temperature, February	102	11B	34	8.8	35.2
43	Mean Monthly Average Temperature, March	103	11C	34	8.8	35.2
44	Mean Monthly Average Temperature, April	103	11D	34	8.8	35.2
45	Mean Monthly Average Temperature, May	103	11E	34	8.8	35.2
46	Mean Monthly Average Temperature, June	103	11F	34	8.8	35.2
47	Mean Monthly Average Temperature, July	103	11G	34	8.8	35.2
48	Mean Monthly Average Temperature, August	103	11H	34	8.8	35.2
49	Mean Monthly Average Temperature, September	102	11I	34	8.8	35.2
50	Mean Monthly Average Temperature, October	102	11J	34	8.8	35.2
51	Mean Monthly Average Temperature, November	102	11K	34	8.8	35.2
52	Mean Monthly Average Temperature, December	102	11L	34	8.8	35.2
53	Surface Water, Minimum Annual Runoff	120	12B	34	8.8	35.2
54	Surface Water, Maximum Annual Runoff	120	12C	34	8.8	35.2
55	Surface Water, Coefficient of Variation	120	12D	34	8.8	35.2
56	Population Density, Population Per Square Mile by County	418	13C	17	8.8	17.6
57	Percent of Population Urban, by County	418	13D	17	8.8	17.6
58	Mean Annual Precipitation	97	14A	17	8.8	17.6

Table 1 (continued)

Map	Title	Sheet Number(s)	Map Code	Scale (1:x million)	Grid Spacing (Obs. Per cm)	Resolution* (1 Pt. Per X km)
59	Mean Annual Maximum Rainfall in 24 Hours	97	14C	34	8.8	35.2
60	Mean Annual Maximum Rainfall in One Hour	97	14D	34	8.8	35.2
61	Mean Monthly Precipitation, January	98	14G	34	8.8	35.2
62	Mean Monthly Precipitation, February	98	14H	34	8.8	35.2
63	Mean Monthly Precipitation, March	99	14I	34	8.8	35.2
64	Mean Monthly Precipitation, April	99	14J	34	8.8	35.2
65	Mean Monthly Precipitation, May	99	14K	34	8.8	35.2
66	Mean Monthly Precipitation, June	99	14L	34	8.8	35.2
67	Mean Monthly Precipitation, July	99	14M	34	8.8	35.2
68	Mean Monthly Precipitation, August	99	14N	34	8.8	35.2
69	Mean Monthly Precipitation, September	98	14Q	34	8.8	35.2
70	Mean Monthly Precipitation, October	98	14P	34	8.8	35.2
71	Mean Monthly Precipitation, November	98	14T	34	8.8	35.2
72	Mean Monthly Precipitation, December	98	14R	34	8.8	35.2
73	Major Land Uses	158–159	15	7.5	3.9	17.6
74	Monthly Maximum Temperature, January	104	16A	34	8.8	35.2
75	Monthly Maximum Temperature, February	104	16B	34	8.8	35.2
76	Monthly Maximum Temperature, March	105	16C	34	8.8	35.2
77	Monthly Maximum Temperature, April	105	16D	34	8.8	35.2
78	Monthly Maximum Temperature, May	105	16E	34	8.8	35.2
79	Monthly Maximum Temperature, June	105	16F	34	8.8	35.2
80	Monthly Maximum Temperature, July	105	16G	34	8.8	35.2
81	Monthly Maximum Temperature, August	105	16H	34	8.8	35.2
82	Monthly Maximum Temperature, September	104	16I	34	8.8	35.2
83	Monthly Maximum Temperature, October	104	16J	34	8.8	35.2
84	Monthly Maximum Temperature, November	104	16K	34	8.8	35.2
85	Monthly Maximum Temperature, December	104	16L	34	8.8	35.2

*Maps listed as having a resolution of 35.2 km were doubled as described in the text, so that the final effective resolution for all maps is 17.6 km.

The maps were registered to one another in two steps. A coarse registration was accomplished by use of reference points on the base maps to position the transparent point counting grids. In the second registration step each map was registered to a truth mask. The truth mask was a separately point counted map containing only three map symbols:

1. outside the study area,
2. study area excluding inland water, and
3. inland water.

The application of this truth mask to each map registered the boundaries of the study area and the inland water bodies.

Errors within the study area were detected using a combination of automated and manual checks. Firstly, an automated procedure was used to flag every singleton cell and absolute differences in the numeric codes of neighboring cells greater than a selected tolerance. Next, either a shade print or an image produced on the Interactive Digital Image Manipulation System (IDIMS) was compared to the original map. Errors were noted, corrected and the resulting map was reviewed again.

PHYSICAL ARRANGEMENT OF MAP DATA

The maps are available on magnetic tape, each map composing a file. The contents of a file are given in Table 2. The first record of each file gives the map number as per Table 2 and the page of the National Atlas, on which the map can be found. For maps covering more than one page, the page number is the first page. The next record has the words MAP TITLE in the first nine columns followed by the title in columns 12 through 80. The next set of records contains the map legend, one record per map symbol. The symbol numeric code is an integer value right justified in columns 1 to 4. This value is followed in columns 6 through 80 by the definition of the symbol. The end of every legend is denoted by the numeric code 0 which is the code for "outside study area" in each map. In the legend the map symbols are arranged in order of increasing numeric values until the

Table 2
Setup of Records Within a Map File

Record 1, Map Identification Number

Column 1

MAP NO. AAA FROM PAGE [or SHEET] III OF THE NATIONAL ATLAS, 1970 EDITION

AAA is a right justified alphanumeric field

III is a right justified integer field

Record 2 +, Title Record

Column 1

MAP TITLE: followed by title and information applicable to entire map.

Note the Title Record may be continued on subsequent cards and continuation will be indicated by **** in columns 77–80.

Record 3 To 3 + NC, Legend Records

A legend record will occur for the number of codes (NC) in each map. For these records, the numeric code associated with each map symbol is right justified in the first four columns. This code is followed by a blank and the meaning of the code. For maps in which the code represents an interval of values, the information following the numeric code is as follows:

Column 6

$DDDD_1$ TO $DDDD_2$

where: $DDDD_i$ are real or integer values of the end points.

Note a Legend Record may be continued on subsequent cards and continuations will be indicated by **** in column 77–80.

Note 0 is the code for outside study area in all maps and is the east numeric code in the legend.

Records 3 + NC + 1 To 3 + NC + 2002, Rows of Map

Thirteen records per row.

First twelve, 20 fields of I4.

Thirteen record, 18 fields of I4.

0 code. Thus either the decrease in value or the 0 could be used as a flag to indicate the end of the legend. Both the MAP TITLE and individual legend records may be continued with **** in columns 77-80, denoting the continuation. The map number, title, and legend are made up of EBCDIC records of length 80. The next records are the rows of the map. Each of these records consists of 20 integer values, with each integer value right justified in a field of length four. Thus each map row is composed of 13 records with the entire map, exclusive of the header information, requiring 2002 records. Like the earlier records, the map rows are EBCDIC. As stated above, the first and last rows of each map are completely outside the study area, so, for example, the first 13 map records possess 258 0 values.

Three examples of the maps from the data base are given in Figure 1.

USE OF THE DATA BASE

While at this time we have not concluded analysis of this data base, we outline below, in general terms our analysis strategy (see Figure 2). A further treatment of this topic will be given in a future report.

The first step in the analysis consists of variable reduction. For each location in the grid there are 85 values, which are far too many to be considered together when selecting samples or for constructing a general linear model. The variable reduction can be accomplished in two steps.

Conceptually, we can portray the cross tabulation of the factors in the data base as a hyperdimensional data cube, an example of which for three dimensions is given in Figure 3. The cells of the cube contain the number of grid points possessing the combination of factor levels represented by the cell. The first step of variable reduction is qualitative, in that the number of values that any map possesses can be reduced by examining the data cube and map legends and deleting those levels and combination of levels which are unimportant or trivial for the objectives of the research being conducted. A natural extension of this step is the elimination of entire maps. However, most of

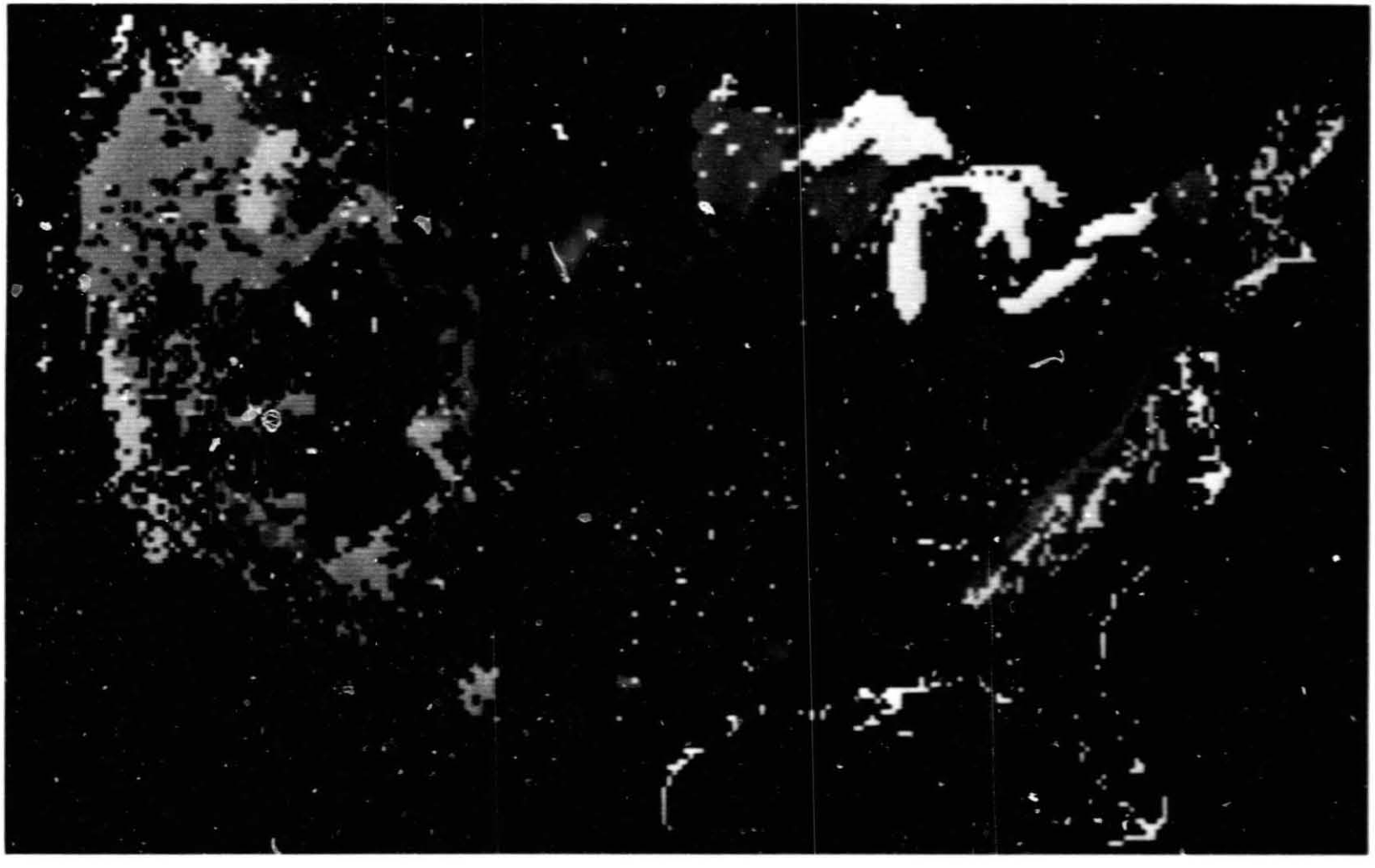

Figure 1(a). Examples from National Atlas Data Base - Geology (Sheets 74-75)

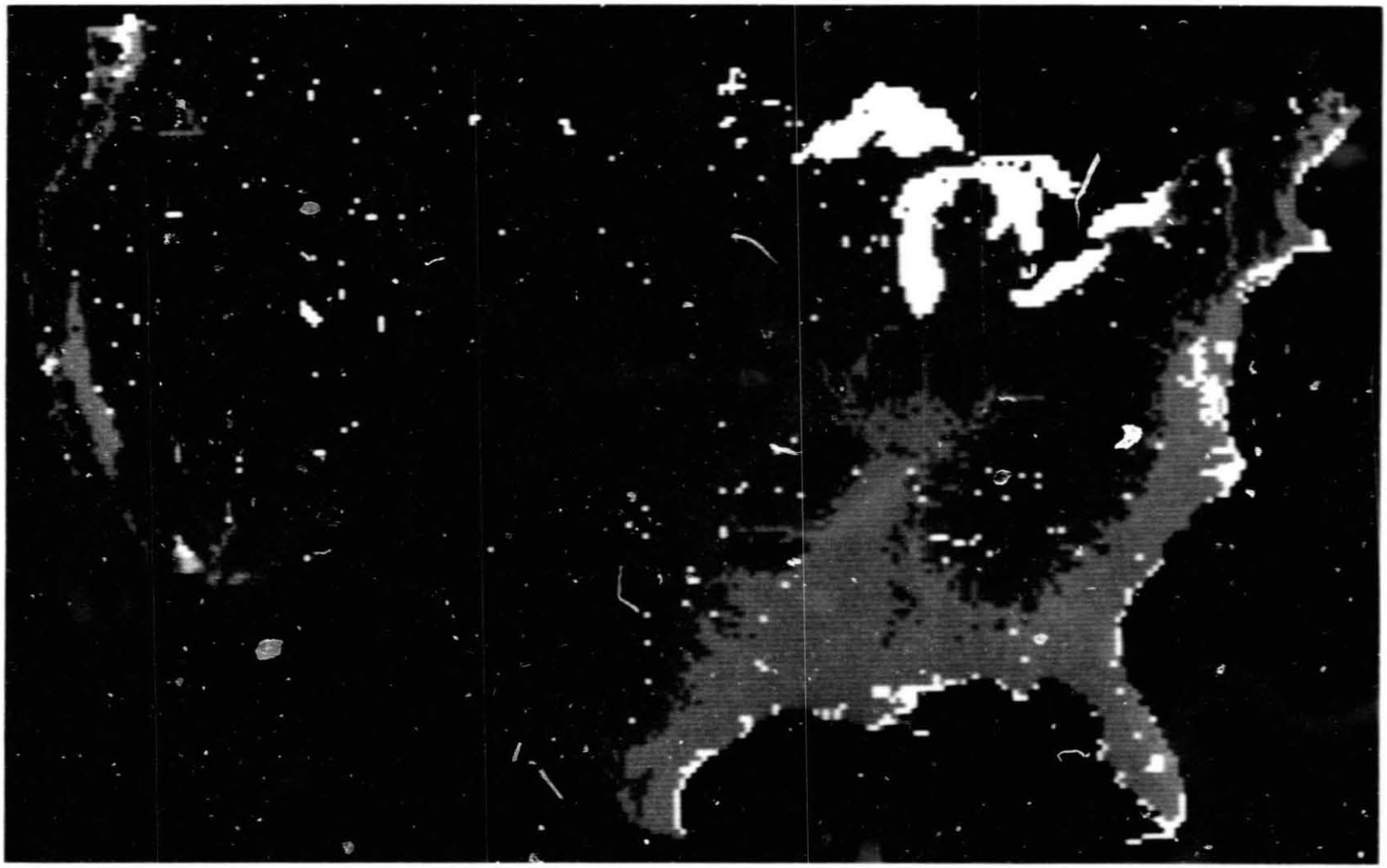

Figure 1(b). Examples from National Atlas Data Base – Potential Nature Vegetation (Sheets 90–91)

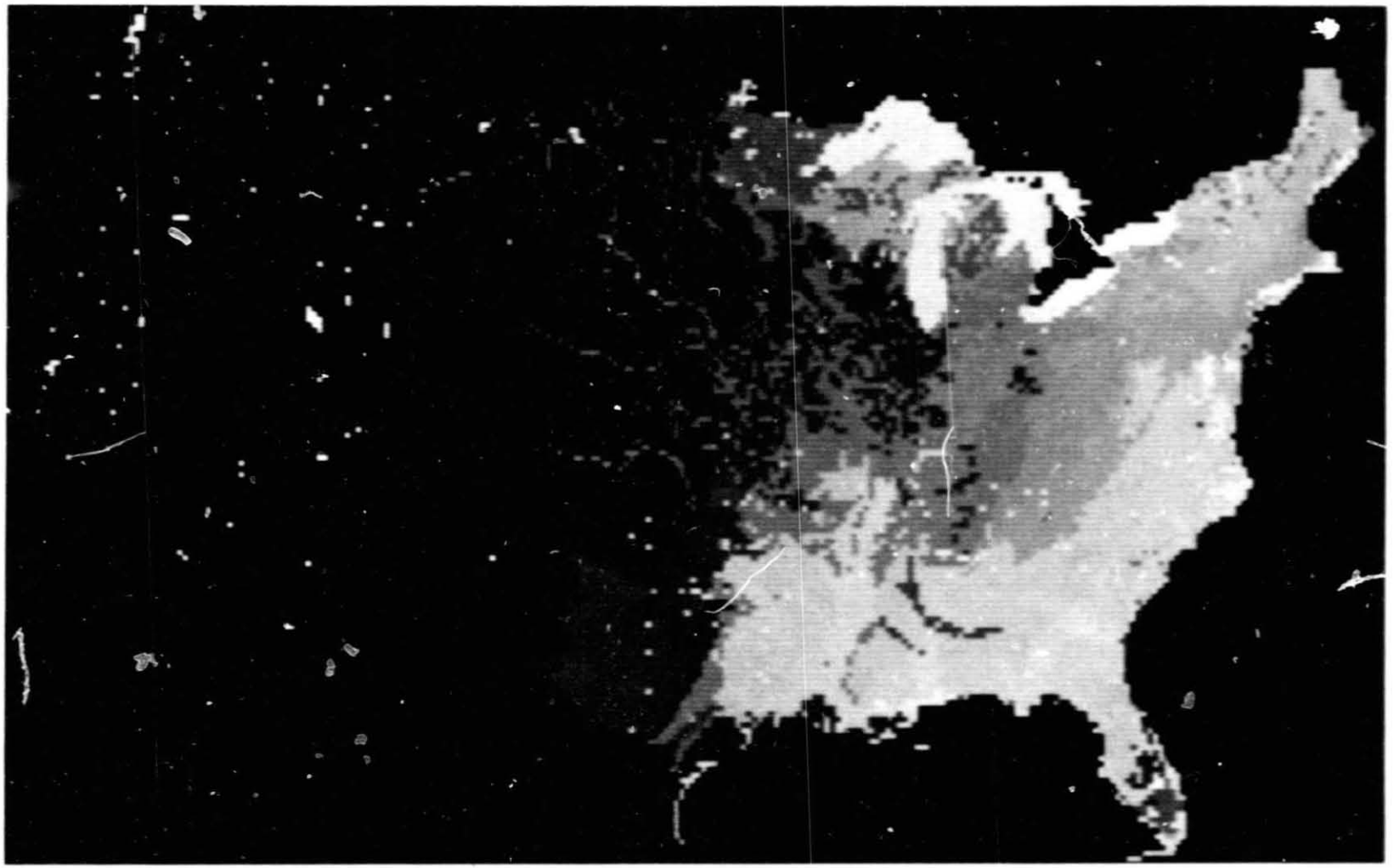

Figure 1(c). Examples from National Atlas Data Base – Topographic Relief (Sheet 59)

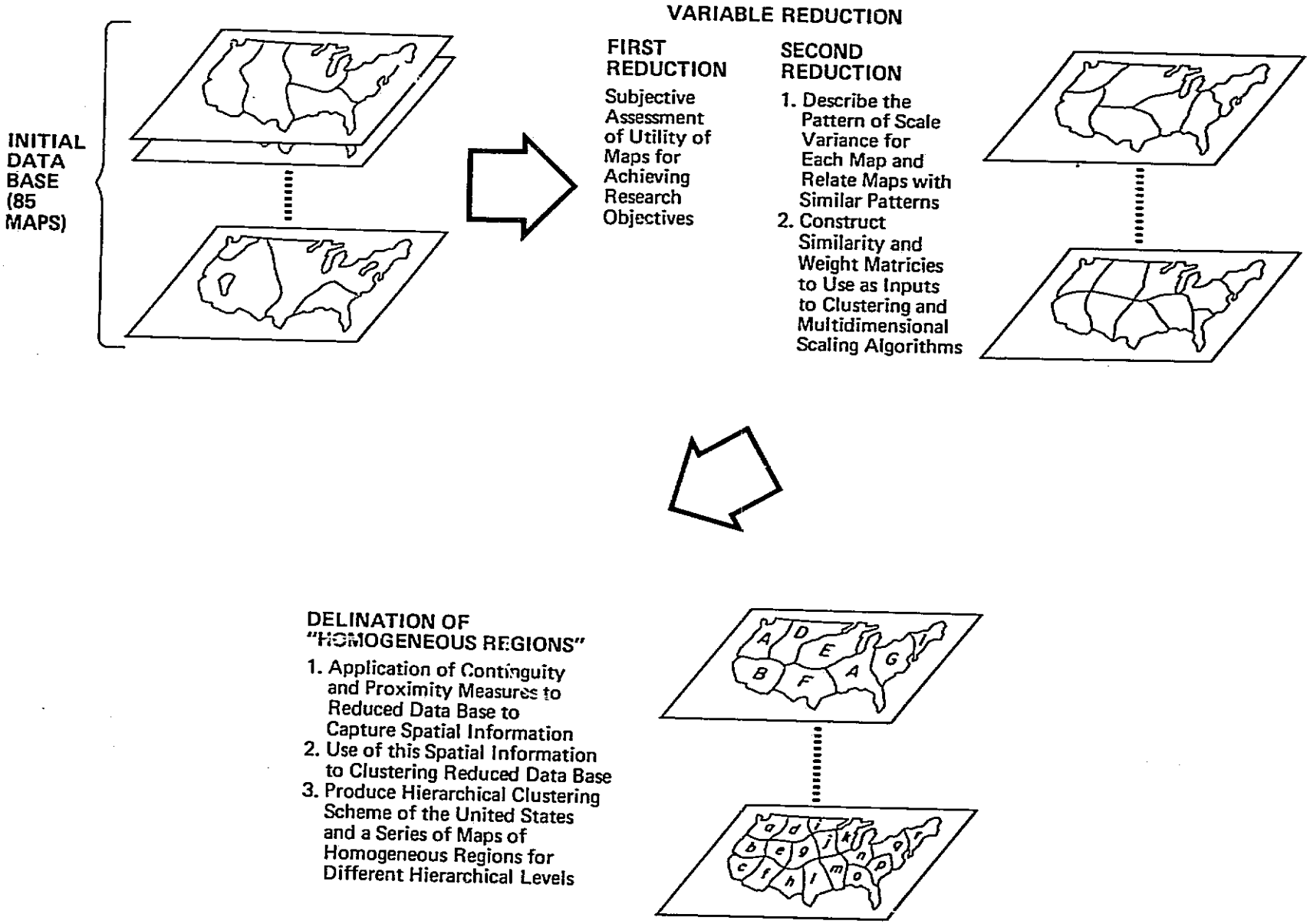

Figure 2. Generalized Strategy for Analysis of National Atlas Data Base

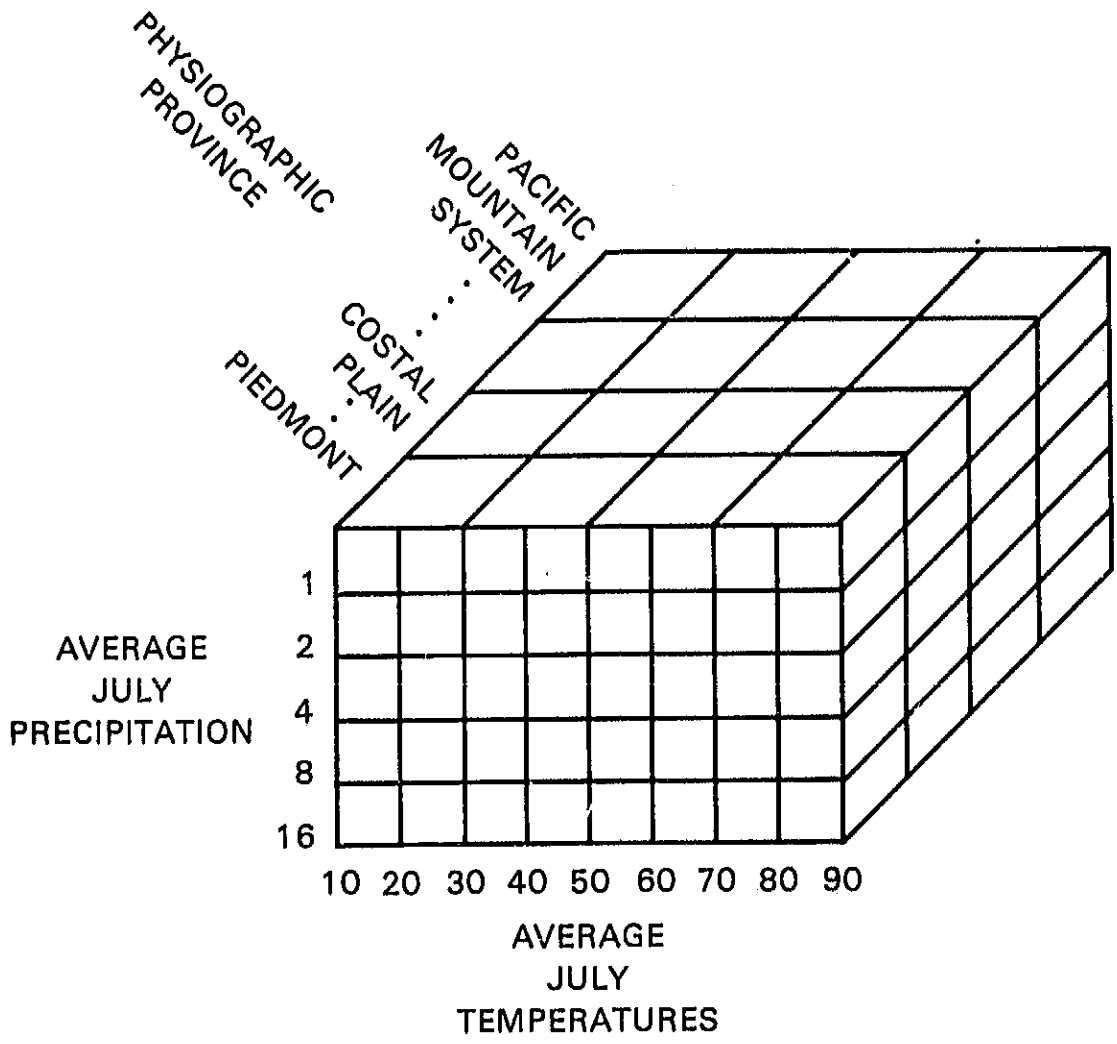

Figure 3. Example of Conceptual Crossing of Factors from National Atlas Data Base

the maps should pass through this initial sieve unaltered. In the next step a variety of quantitative methods will be used to construct subsets of the 85 maps such that members of a subset have similar patterns of variation or are closely correlated with one another. These methods are included under:

1. ANOVA related techniques (Greig-Smith, 1964) used to assign variation according to scale, and
2. measures which result in the construction of similarity matricies (Everett, 1974), which are inputs for non-parametric methods or procedures for the analysis of multilevel data sets, such as multidimensional scaling (Kruskal, 1964).

Once the subsets of maps are constructed, maps representative of each subset can be randomly selected. We will cluster this reduced data set, incorporating information from a series of continguity or proximity measures (Cliff and Ord, 1973) into clustering algorithms. The outcome of these analyses will be a hierarchical series of clusters representing "homogeneous regions" of the U.S.A. From among these "homogeneous regions," the proportion of the population covered by each region may be calculated and an experimental design may be constructed. The results could be used in turn to make an estimate of the total number of samples needed and their spatial distribution.

SUMMARY AND CONCLUSIONS

We have developed a data base composed of 85 maps for the U.S.G.S. National Atlas. The data base is motivated by the need for a rationale to select representative areas of the U.S.A. for examination by remote sensing. The maps included in the data base are those which might reasonably be believed to influence remote sensing and display properties of the land surface and climatic conditions over the conterminous U.S.A. The maps were point counted or transformed so that an observation was made every 17.6 km (11 mi.) across the U.S.A. This resulted in a grid of 154 rows by 258 columns for each map yielding 39,732 points of which 22,505 are within the study area. The maps were registered to another and then iteratively error checked using manual and automated approaches. The "error-free" maps were placed on magnetic tape one map per file, each map precede in the file by identification information and a legend of map symbols. In future work we will use the data base to create a map of "homogeneous" regions. Such a map will represent a stratification of the population. The validity of this stratification will be examined by random sampling scenes from the strata, sampling pixels from the scenes, and then testing the significance of the grouping factors (which formed the stratification) using techniques from the field of general linear models, particularly analyses of variance.

We have stressed the utility of the data base for the selection of areas the size of scenes or greater. A second, and perhaps a third stage, of sampling is needed to complete the data collection.

We have not yet investigated strategies for these additional stages. This recognition of the presence of multiple sampling scales is related to the recognition that the grouping factors which will define the homogeneous regions will be unlikely to account for all the variation in the remotely sensed data. However, the amount of variation explained by these factors is still unknown and will be important in designing smart or onboard satellite processing as well as in analyzing previously collected data. Such work will be important for questions pertaining to optimizing temporal and spectral resolutions. Finally, although we have considered the data base solely within a remote sensing context, it and the philosophy behind it clearly have utility for other studies where the researcher desires to make inferences about large geographic regions.

REFERENCES

Cliff, A. D. and J. K. Ord, 1973, "Spatial Autocorrelation," Monographs in Spatial and Environmental Systems Analysis, Pion Ltd., London, U.K., 178 p.

Everett, Brian, 1974, "Cluster Analysis," Heineviann Educational Books, Ltd., London, U.K., 122 p.

Greig-Smith, P., 1964, "Quantitative Plant Ecology," Butterworths, London, U.K.

Kruskal, J. B., 1964, "Nonmetric Multidimensional Scaling: A Numerical Method," Psychometrika, V. 29, pp. 115-129.

Maling, D. H., 1973, "Coordinate Systems and Map Projections," George Philip and Son Limited, London, U.K., 255 p.

Scheffe, H., 1959, "The Analyses of Variance, " John Wiley & Sons, New York, NY

U.S.G.S., 1970, "The National Atlas of the United States of America," United States Department of the Interior, Geological Survey, Washington, D.C., 417 p.

BIBLIOGRAPHIC DATA SHEET

1. Report No. 85009	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Experimental Philosophy Leading to a Small Scale Digital Data Base of the Conterminous United States for Designing Experiments with Remotely Sensed Data		5. Report Date April 1983 6. Performing Organization Code
7. Author(s) M.L. Labovitz, E.J. Masuoka, P.W. Broderick, T.R. Garman, R.W. Ludwig, G.N. Beltran, P.J. Heyman, L.K. Hooker		8. Performing Organization Report No.
9. Performing Organization Name and Address Geophysics Branch, Code 922 Earth Survey Applications Division NASA/Goddard Space Flight Center Greenbelt, MD 20771		10. Work Unit No. 11. Contract or Grant No. 13. Type of Report and Period Covered TM
12. Sponsoring Agency Name and Address NASA/Goddard Space Flight Center Greenbelt, MD 20771		14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

Research using satellite remotely sensed data, even within any single scientific discipline, has often lacked a unifying principle or strategy with which to plan or integrate studies conducted over an area so large that exhaustive examination is infeasible, e.g., the U.S.A. However, such a series of studies would seem to be at the heart of what makes satellite remote sensing unique, that is the ability to select for study from among remotely sensed data sets distributed widely over the U.S., over time, where the resources do not exist to examine all of them. What we do lack is the previously noted strategy to aid in the development of formal testable hypotheses and the selection of study locations so as to minimize the number of samples subject to the ability to construct desired inferences.

Using this philosophical underpinning and the concept of a unifying principle, we have constructed an operational procedure for developing a sampling strategy and formal testable hypotheses. We believe the procedure to be applicable across disciplines, when the investigator restates the research question in symbolic form, i.e. quantifies it.

The procedure is set within the statistical framework of general linear models. The dependent variable is any arbitrary function of remotely sensed data and the independent variables are values or levels of factors which represent regional climatic conditions and/or properties of the earth's surface. These factors are operationally defined as maps from the U.S. National Atlas (U.S.G.S., 1970). Eighty-five maps from the National Atlas, representing climatic and surface attributes, were automated by point counting at an effective resolution of one observation every 17.6 km (11 miles) yielding 22,505 observations per map. The maps were registered to one another in a two step procedure producing a coarse, then fine scale registration. After registration, the maps were iteratively checked for errors using manual and automated procedures. The "error-free" maps were anotated with identification and legend information and then stored as card images, one map to a file.

A sampling design will be accomplished through a regionalization analysis of the National Atlas data base (presently being conducted). From this analysis a map of "homogeneous regions" of the U.S.A. will be created and samples (Landsat scenes) assigned by region.

While designed for use with remote sensing experiments, the data base, the method of analyzing it and the philosophy behind it are general enough to serve as a framework for other studies being conducted over large portions of the U.S.A.

17. Key Words (Selected by Author(s))	18. Distribution Statement
National Atlas, Data Base, Experimental Design, General Systems Approach	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified		

*For sale by the National Technical Information Service, Springfield, Virginia

GSFC 25-44 (10/77)