BIVARIATE CONSTRUCTION OF EQUAL-CLASS THEMATIC MAPS

Geoffrey Dutton

Laboratory for Computer Graphics
and Spatial Analysis
Harvard Graduate School of Design
48 Quincy Street
Cambridge, Mass. 02138

ABSTRACT

Shaded choropleth maps normally divide their data into discrete classes, rather than symbolizing them as a continuous distribution. In practice, five to ten categories are normally used, rarely more than twenty. Class intervals for data can be established which have equal value ranges, equal numbers of observations, or unequal sizes; the latter variety may be derived by factoring, clustering or statistical grouping procedures. This paper presents a generalization of equal-membership classification, in which class intervals for a mapped variable are fashioned to include equal amounts of a second, related variable. This usually yields unequal ranges for the mapped variable, but the ranges are easier to interpret as they equalize some selected quantity. Maps employing such classifications may give superior insights into the joint distribution of the mapped and classifying variables. Following a discussion of value classification issues, the concept of bivariate classification is introduced and discussed; a procedure to accomplish it is described, and a Fortran implementation of it is given. A series of maps illustrate the results of bivariate classification for a group of related statistics.

1.0 ELEMENTS OF SHADED MAPS

Shaded maps communicate three essential elements of information to map readers, and require map makers to make decisions regarding all three. These elements are:

1. A geographic region, subdivided into zones;
2. Attributes of the zones, grouped into classes;
3. Symbolism shading each class with a certain pattern or color.

Map makers must usually decide upon each of these elements, although in particular cases one or more of the elements may be given. Frequently, element 1 is not a matter of choice, as both the region and its subdivisions are determined by the thematic data which one has and wishes to map.

Computer cartographers usually define their own value classifications and shading for maps they produce. Usually these are specified independently; Once the number of levels is decided, a method of classification is then chosen (which may be automatic or manual) which creates class breaks at certain points along the Z-value range. All zones with values lying between two adjacent breakpoints are thus grouped into the same category, and will be symbolised on the

map with the same texture, darkness and color. Symbolism for each class is then defined which may vary its appearance systemmatically according to Z-value, or it may not; although rarely desirable, two levels can be defined to have the same shading, and these levels need not be adjacent ones in the value range. Most map makers will make each level as distinct as possible, and will choose shading patterns with a progression of tones to represent the overall range of Z-values. Occasionally this is accompanied by changes in the design of shading patterns, for instance by using dot patterns to represent lower levels, line patterns for middle levels and crossed line patterns to represent upper levels. Such variations in pattern and texture help readers discriminate among a larger range of classes.

2.0 CLASSIFYING CHOROPLETH MAPS

Despite a rich cartographic literature on the subject, only a small number of value classification techniques are employed in published thematic maps. Most such maps are choropleth (shaded zone) representations, a large portion of which display continuous (as opposed to categorical) thematic data group their observations into a small number of classes, usually with equal value ranges. Few other methods (aside from logarithms, roots and related arithmetic transformations) commonly are employed to communicate thematic regularities of spatial data in shaded maps.

Although many methods have been developed to analytically classify thematic variables, their currency remains limited and largely academic. This is due both to complexity of methodology (such as cluster and principal components analyses), and to the non-intuitive nature of their results, which are too often presented without adequate discussion of the classification procedures employed.

Most statistical classification techniques are designed to maximize value homogeneity of resulting classes. Some methods attempt to do so in the spatial domain, others in the value domain, and some in both. Criteria such as proximity, compactness and contiguity are used to maximize spatial homogeneity, whereas tests such as chi-squared and f-ratio are used to segregate values into classes without regard to their spatial properties.

The main alternative to maximizing intra-class homogeneity is to attempt to maximize the equality of classes. Such techniques are both simple to apply and straightforward to interpret. The two predominant ones are (a) equal value ranges and (b) equal value frequencies. By keeping either the class intervals or the class memberships constant, a map helps its readers interpret spatial patterns because the "importance" of each class is made equal by a simple rule.

There is another form of equal classing, normally employed for dot-distribution mapping, but rarely in choropleth maps. This method strives to equalize the number of objects represented by each symbol, such as the number of people, acres of agricultural land or volume of business transactions. This implies that the data are extensive in nature, rather than intensive -- that is, individual observations must be capable of being summed to a meaningful total. While this excludes many forms of data (e.g., densities, per capita ratios and other intensive measures), in many cases the components of such data (such

as population and area) can be used, as will be shown. It is thus conventional cartographic wisdom to use shading to portray intensive quantities, (densities or ratios); when shading intensity varies, it serves as a natural visual analog for such data. For this reason cartographers are critical of maps of extensive quantities in choropleth form; when shading density symbolizes extensive quantities, variations in the area of zones can perceptually bias interpretation.

It is nevertheless possible to portray absolute data on a choropleth map, and do so without areal bias. One way to do this is to use the third dimension, creating a surface. One can make the height of the surface over a zone proportional to either an absolute value or an intensity. The former procedure, however, produces the same problems as does shading absolute values; large zones can visually dominate small ones even if their thematic values are less. However, if the height is made proportional to the areal density of some quantity, such as population, then the volume thus created is proportional to the size of that quantity in that zone. While map readers may not always be able to estimate the relative size of volumes (especially if their shapes change), such a three-dimensional representation is at least honest, in that:

1. Area of zones can be accurately portrayed in the x-y plane;
2. Heights of zones are proportional to densities;
3. Volumes of prisms are therefore meaningful, the integral of density over area.

The result is in one sense a bivariate map as both population size and population density can be read from it.

Not every cartographer has access to software for plotting 3-d prism maps, and such maps are difficult to construct manually. Furthermore, clients still tend to prefer 2-d choropleth maps over 3-d ones, especially in application areas such as social science and marketing. It would thus be convenient and practical to offer their advantages in two dimensions (as conventional shaded maps), which the procedure described below can begin to accomplish.

3.0 WHAT ARE EQUAL CLASSES?

An equal-interval classification is normally regarded as one in which the range of data values (or some standardized range, such as zero to one hundred percent) is partitioned into exactly equal parts. One assigns shading to such a set of intervals, as one would apply hypsometric tints to a contour map. Once the number of classes and the overall data range are selected, the sizes and limits of all intervals are completely specified. Consequently, the particular distribution of data values (aside from their lower and upper limits) has no effect on the results. This is why equal-interval classes are usually a poor choice when mapping skewed value sets, such as population densities.

When very non-uniform statistics are mapped into equal intervals, one or two classes (often the lower ones) may contain the bulk of the observations, and thus not communicate their variation. Likewise, some classes (usually in the middle range) may end up devoid of

observations, reducing the information content still further. The greater the number of classes and the fewer the number of zones to be symbolized, the more likely this is to occur.

Figure 1 is a shaded map of U.S. state population densities, classified into 5 equal-sized intervals, each spanning a range of 200 persons per square mile. Due to the non-uniformity of this value distribution, fully 38 states, or 78 percent, fall into the lowest class. The resulting map is a caricature of the data, with little graphic utility. The main order of business for such a map is to communicate the spatial distribution of population. Toward this end, it might be useful to modify class intervals to reflect the properties of that distribution. That is, why not equalize the contents of classes rather than their extents? To do this, we can reclassify densities in any of three fairly obvious ways, grouping them into classes which contain:

1. Equal numbers of observations,
2. Equal areas, or
3. Equal populations.

The first option should be familiar to thematic cartographers, being the so-called "quantile" method of classification, more generally known as Histogram Equalization. In maps where unit areas are constant, such as images or other gridded data, this procedure is tantamount to the following one, equal-area classification, yielding the same results. For choropleth maps of most real geographies, however, they are not the same, and thus option 1 is a special case of option 2. The results of a quantile classing of the data mapped in figure 1 is the map shown as figure 2. Here each class contains about 10 states, identifying those states ranked in the bottom fifth of population density, second fifth, etc. Note that the classes are of variable extent, but that in terms of membership they are equal.

3.1 Equal-area Classification

To standardize the areas occupied by each class in the general case, a second set of values -- the area of each zone -- must be available and incorporated into the classification process. This turns the procedure from a univariate to a bivariate one, a subroutine for accomplishing which is listed in Appendix I. To equalize class area, this processor reads in data records containing both population density and area for each zone in a region, sorts the records by increasing population density and simultaneously computes total land area. This total is divided by the number of classes to be portrayed, yielding its average, the area which each class should ideally cover. Then, proceeding up the sorted list of densities, zone area is accumulated until this average size has been achieved; the value of population density given for this observation then defines the upper limit of the first class. Repeating this procedure until the highest density is reached computes a set of class breaks which partition the data into classes of roughly equal area. These breaks are then used to classify and map population densities. The more observations which exist, the more accurate this process can be. With the U.S. states, especially if Alaska is included, equal-area classing can be somewhat imprecise, the amount of error depending on what quantity is being

mapped. Figure 3 presents a map of densities classed to have equal area. This, rather than figure 2, approximates a Histogram Equalization for the data, given that land area varies from state to state.

3.2 Equal-population Classification

Approximate as it may be, this simple procedure for deriving equal-area classes is remarkably general; one soon realizes that attributes other than area can be equalized just as easily, and this leads to consideration of option 3, equal-population classification. Specifically, each class can be fashioned to include an equal number of people, simply by substituting population counts for land area when classifying population densities. The results of this substitution are displayed as figure 4. Each of the five classes in that map contain forty to fifty million persons, specifying the range of densities at which respective fifths of the U.S. population live, at the state level.

3.3 Equal-income Classification

If thematic classes can be fashioned to standardize the number of people which each contains, clearly they can likewise be computed to partition any extensive (countable) quantity of interest. To demonstrate that the technique of class equalization is independent of the data it manipulates, consider the map displayed in figure 5. In this map, population densities are again aggregated into 5 levels, but classified so that each one contains equal amounts of disposable personal income. Although population densities and income may be empirically correlated, these two statistics measure different phenomena and have no common factors. Yet, one expects that urban regions will generate income more rapidly than rural ones, even if the specific causalities are not understood. Figure 5 is an attempt to spatially simplify this relationship. The constriction of top-level symbolism in it indicates a concentration of wealth in the most dense states.

4.0 FURTHER EXPLORATIONS

The final three maps, figures 6, 7 and 8 further illustrate bivariate classification, changing the thematic variable to income per capita by state for the U.S. in 1979. While population densities have a highly skewed, almost lognormal distribution, per capita incomes are nearly normally distributed, slightly skewed to the left, as the histograms show in the last 3 maps. Consequently, an equal-interval classification is much more appropriate for this data; figure 6 shows the result of this, in which each of six classes covers a span of 1000 dollars per capita. While far more informative than figure 1, figure 6 still obscures a certain amount of data. As the U.S. is predominantly a middle-class nation, important distinctions may be blurred in the middle income ranges by using equal-interval classes.

Classifying these data to equalize population-per-class (figure 7) yields a different pattern of symbolism. This map might be considered to be "more representative" of the distribution of income levels, as the population represented by each class is now nearly equal, being about 33 million people. The fact that the largest number of states is found in the bottom class indicates that there is probably a positive correlation between income levels and population size, if not population density. (This insight may help one to re-interpret figure 5, in which population density is grouped into classes with equal aggregate income.) Notice also that figure 7 seems to draw firmer distinctions than figure 6; for instance, Colorado and Nebraska both occupy class 3 in the equal-interval map, but in the equal-population map Nebraska has gravitated down to level 2 and Colorodo has risen to level 4, as level 3 now has a much smaller value range. The author leaves the reader with the exercise of comparing figure 8 to figures 7 and 6. Examine the differences in class breaks; what generalizations can be made about an equal-income classification of per capita income in comparing it with the equal-interval and equal-population maps?

5.0 CONCLUDING COMMENTS

Communicating the patterns in which living communities group themselves is thematic cartography's essential challenge and its special competence. Maps which display single variables may be easier to interpret than bivariate ones, but can never really express the richness of relationships which characterize living systems. Although all maps are simplifications of reality, the use of a second variable in classifying a coverage can deliberately add information without greatly burdening the map reader, if carefully employed. It would certainly be easy to abuse bivariate classification, by attempting to associate completely unrelated variables. It is equally tempting and easy to compute spurious correlations between variables with statistical analysis packages. As in all data processing, Garbage In, Garbage Out. One may nevertheless have special and valid reasons for looking at associations among variables which other analysts may not regard as having any important relationship; one person's data can be another person's trivia. In any case, the burden of proof should be on the mapmaking analyst, who must provide a framework within which bivariate (or any) maps may be interpreted. The framework may be a theoretical model, hypotheses concerning interesting empirical regularities, or it may simply be a set of maps and measures designed to describe certain spatially distributed thematic variables. At the very least, viewing thematic maps classified in several different ways can inform one of more nuances of spatial structure than any single map can communicate.

REFERENCES

The literature of thematic cartography in general and classification techniques in particular is quite extensive. No specific references have been cited in this paper, however, due to the exploratory nature and apparent novelty of the material presented. While the technique for bivariate classification it presents is not complicated, and may even be regarded as fairly obvious once explained, the author has never encountered it in print. This is somewhat surprising, and may constitute an admission of ignorance; if any reader is aware of similar work, reported in the literature or unpublished, the author would welcome learning of it.

APPENDIX I

FORTRAN Procedure for Computing Bivariate Classifications

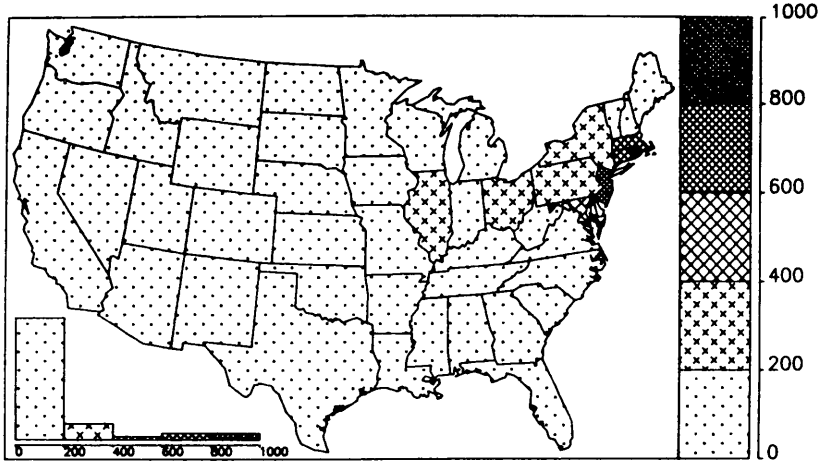
```
      SUBROUTINE CLASSY (V1,V2,NVAL,BREAKS,BINS,NCLASS,V2SUM,BAD)
C
C   DOES CLASSIFICATION OF VALUE LIST ACCORDING TO EQUAL AMOUNTS
C   OF A SECOND QUANTITY, RETURNING CLASS BREAKS.  IT IS ASSUMED
C   THAT BOTH VARIABLES ARE REAL, RATIO QUANTITIES.
C
C     GEOFFREY DUTTON, HARVARD LABORATORY FOR COMPUTER GRAPHICS
C     AND SPATIAL ANALYSIS;  AUGUST 1982.
C
C     V1     - VARIABLE LIST TO CLASSIFY
C     V2     - CLASSIFYING VARIABLE LIST
C     NVAL   - NUMBER OF OBSERVATIONS (FOR V1 AND V2)
C     BREAKS - VECTOR OF CLASS BREAKS
C     BINS   - HISTOGRAM OF CLASSIFIER
C     NCLASS - NUMBER OF CLASS LEVELS
C     V2SUM  - SUM OF CLASSIFYING VARIABLE
C     BAD    - INVALID DATA FLAG
C
      DIMENSION BINS(NCLASS)
      DIMENSION V1(NVAL), V2(NVAL), BREAKS(NCLASS)
      DIMENSION INDEX(5000), V1SORT(5000)
C
C             FIRST CREATE INDEX POINTERS
C             AND SUM WEIGHTING VARIABLE
C
      V2SUM = 0.0
      DO 10 I = 1,NVAL
        INDEX(I) = I
        VAL = V2(I)
        IF (VAL .NE. BAD) V2SUM = V2SUM+VAL
        V1SORT(I) = V1(I)
   10 CONTINUE
C
C             ZERO OUT THE BINS ARRAY
C
      DO 15 I = 1,NCLASS
        BINS(I) = 0.0
   15 CONTINUE
```

```
C
C               SORT THE PRIMARY VARIABLE AND AN INDEX TO IT
C
      CALL SORT (V1SORT,INDEX,NVAL)
C
C               V2BAR IS HOW MUCH OF V2 EACH CLASS SHOULD GET
C
      V2BAR = V2SUM/FLOAT(NCLASS)
C
C               NOW ASSIGN VALUES TO CLASSES, VERY SIMPLY
C
      KLASS = 0
      VSUM2 = 0.0
C
      DO 20 I = 1,NVAL
        J = INDEX(I)
        VAL = V2(J)
        IF (VAL .NE. BAD) VSUM2 = VSUM2+VAL
        LEVEL = VSUM2/V2BAR
        IF (VAL .NE. BAD) BINS(LEVEL+1) = BINS(LEVEL+1)+VAL
        IF (LEVEL .LE. KLASS) GOTO 20
C
C               NEW LEVEL REACHED; SOME OF V2 SPILLS INTO NEXT ONE
C
        J0 = INDEX(I-1)
        BREAKS(LEVEL) = (V1(J)+V1(J0))/2.
        KLASS = LEVEL
   20 CONTINUE
C
      RETURN
      END
```

FIGURES

The following maps of the U.S.A. by state demonstrate the use of bivariate classification. These illustrations were produced by the POLYPS program, the 2-d choropleth mapping module of the Laboratory's ODYSSEY Geographic Information System. POLYPS was written by Scott Morehouse.

STATE POPULATION DENSITIES, 1975

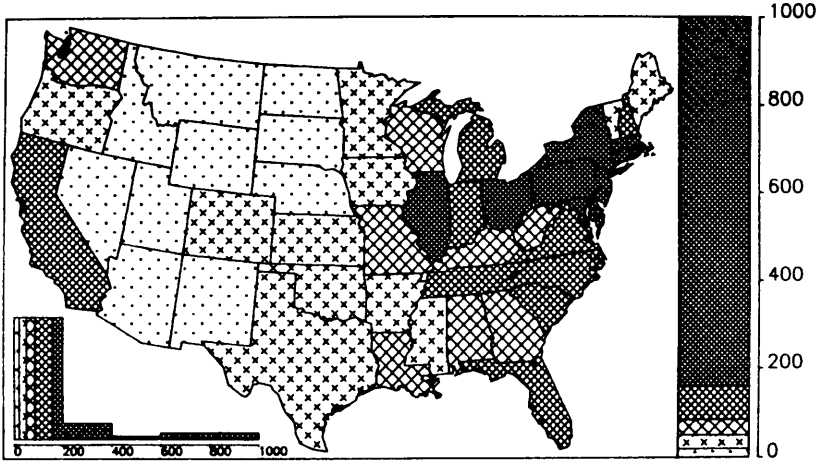


AN EQUAL–INTERVAL CLASSIFICATION

Class contents are: 200, 200, 200, 200, 200 persons/sq. mile
Class breaks are: 200, 400, 600, 800 persons/sq. mile
Class Memberships are: 38, 5, 1, 2, 2 states.

Figure 1

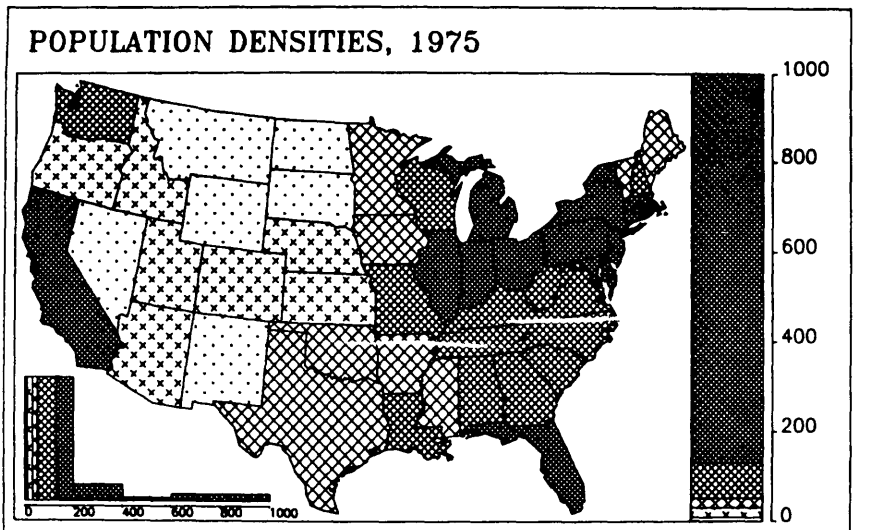
POPULATION DENSITIES, 1975



EQUAL–MEMBERSHIP (QUANTILE) CLASSIFICATION

Class contents are: 10, 11, 8, 9, 10 states
Class Breaks are: 20, 51, 85, 160 persons/sq. mile;
Class Memberships are: 10, 11, 8, 9, 10 states;

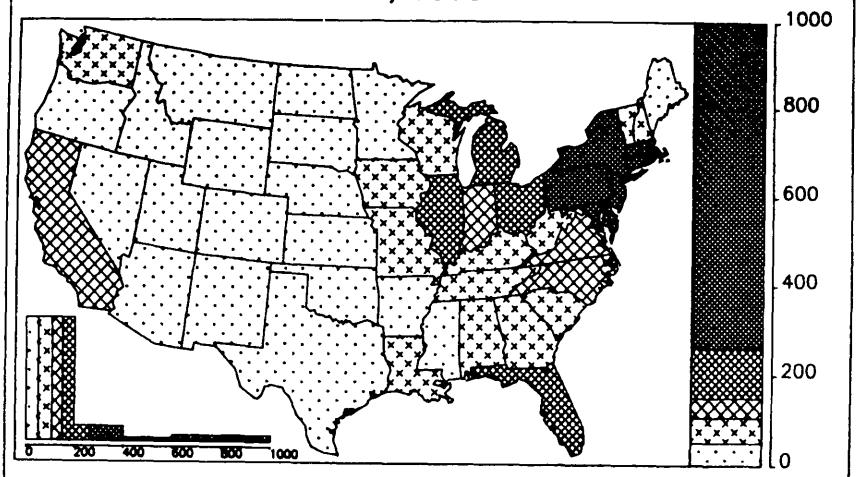
Figure 2



EQUAL–AREA CLASSIFICATION:

Class contents are: 522, 636, 606, 577, 526 square miles;
Class Breaks are: 9.5, 31, 52, 130.5 persons/sq. mile;
Class Memberships are: 6, 7, 8, 13, 14 states

Figure 3

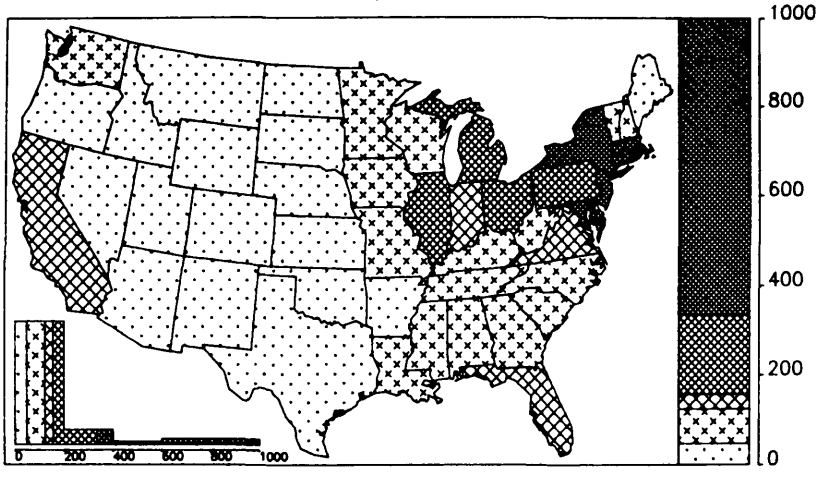


AN EQUAL–POPULATION CLASSIFICATION:

Class contents are: 38.7, 39.4, 34.8, 43.5, 39.4 million persons;
Class Breaks are: 50.5, 106.5, 150, 263 persons/sq. mile;
Class Memberships are: 22, 13, 4, 4, 8 STATES;

Figure 4

POPULATION DENSITIES, 1975

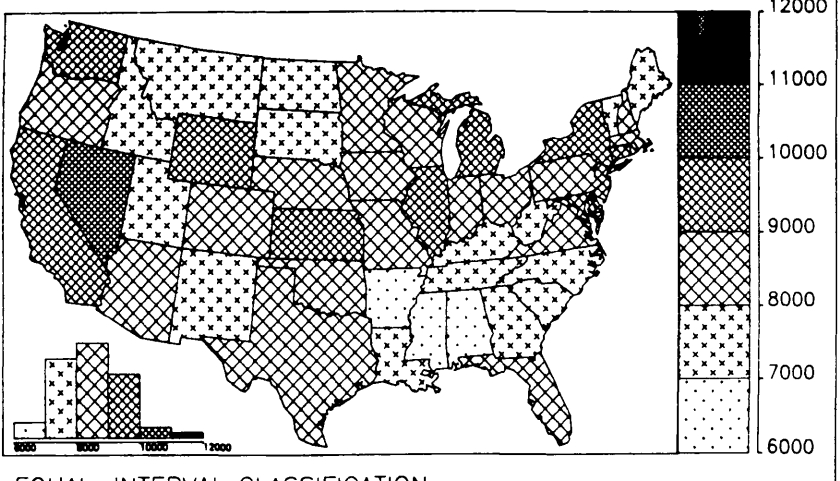


AN EQUAL–INCOME CLASSIFICATION

Class contents are: 333, 341, 348, 395, 359 billion dollars;
Class breaks are: 48, 124, 156.5, 335 persons/sq. mile;
Class Memberships are: 20, 16, 4, 5, 6 states.

Figure 5

DISPOSABLE INCOME PER CAPITA, 1979

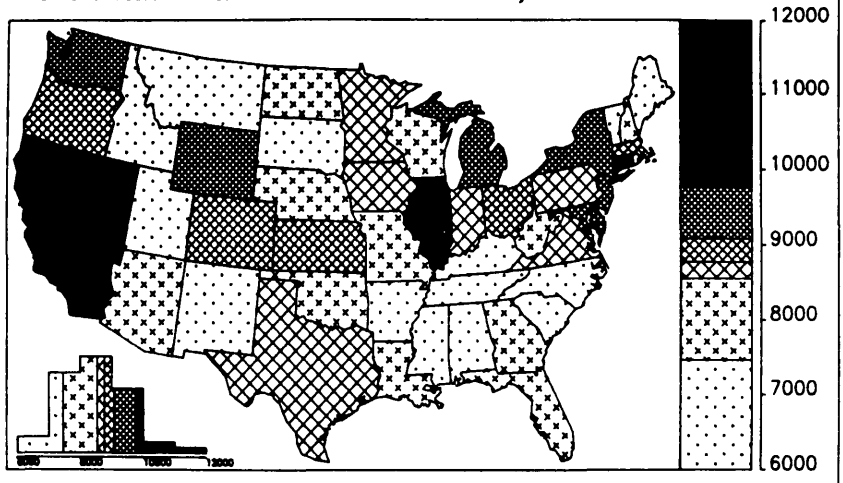


EQUAL–INTERVAL CLASSIFICATION

Class contents are: 1000, 1000, 1000, 1000, 1000, 1000 $/person
Class breaks are: 7000, 8000, 9000, 10000, 11000 $/person;
Class memberships are: 3, 15, 18, 12, 2, 1 states;

Figure 6

DISPOSABLE INCOME PER CAPITA, 1979

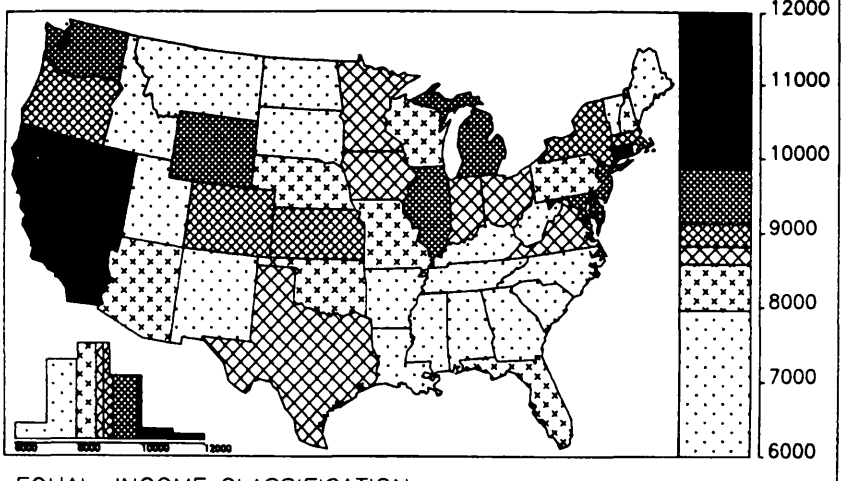


EQUAL–POPULATION CLASSIFICATION:

Class contents are: 30, 30.7, 36.2, 23.6, 39.2, 36.1 million people
Class breaks are: 7462, 8546, 8768, 9077, 9762 $/capita;
Class memberships are: 14, 12, 6, 5, 8, 6 states

Figure 7

DISPOSABLE INCOME PER CAPITA, 1979



EQUAL–INCOME CLASSIFICATION

Class contents are: 297, 288, 317, 280, 319, 275 billion $
Class Breaks are: 7953, 8574, 8808, 9124 9868$/PERSON;
Class memberships are: 14, 12, 6, 5, 8, 6 states.

Figure 8