OBSTACLES TO ACCURATE AND VALID GEOGRAPHIC ASSESSMENT OF VITAL EVENT DATA

David R. Slaby and Robert J. Casady
National Center for Health Statistics
Hyattsville, Maryland 20782

and

Henry J. Malin and Judy F. Coakley
National Institute on Alcohol Abuse and Alcoholism
Rockville, Maryland 20857

Introduction

Choropleth mapping of the less common causes of death, such as cirrhosis of the liver, produces difficulty in interpretation particularly when mortality data are viewed at the county level. In 1975 cirrhosis accounted for about 21,000 deaths among males in the United States and 11,000 deaths among females; this was approximately two percent of male deaths and one percent of female deaths for that year. Cirrhosis of the liver can be implicated with loss of life and illness and is associated with other morbidities and mortalities. Cirrhosis is also associated with economic costs to society in terms of disability days lost from work, resources devoted to clinical treatment for morbid conditions, and institutional support for long term care.

Cirrhosis is of interest from an epidemiologic perspective in that it is highly associated with heavy consumption of alcoholic beverages and is often used in measuring the incidence and prevalence of alcoholism and alcohol abuse for discrete geographic areas and population subgroups. Death from cirrhosis has been

linked with various occupation positions such as chemical workers and might be an indicator of severe environmental chemical hazards. It becomes clear that even though cirrhosis accounts for a relatively small absolute number of deaths in the U.S. population as compared to other causes of death such as heart disease, it is implicated in many morbid conditions and events which lead to death, and is therefore worthy of study.

The distribution of this cause of death is of some interest. National mortality data by year and by county are available from the National Center for Health Statistics. Population estimates from the Bureau of the Census are also available at the county level of aggregation. It is a relatively easy task to age-adjust these data to produce mortality rates by cause for every county or equivalent in the United States. The rates can be plotted to produce a county polygon choropleth map, but it is here that problems of interpretation arise due to the phenomenon of density driven rates. Counties large in area and low in population and counties with uneven population distributions are given inordinate "visual" weight on the map.

For example, Riverside and San Bernadino counties in California have large population centers located in the extreme western portions of the counties while their eastern parts contain largely desert. The resulting visual effect using this choropleth mapping technique is that there is a relatively high rate of cirrhosis mortality in an area where we suspect there are not many people. The contribution to the rate primarily comes from a small portion of the county. Our goal from a visual viewpoint is to maintain the characteristics of the "hot spot" in the county where the major proportion of the population is located while downplaying those relatively empty spaces. That is, to weight toward the population centers where most vital events occur.

In this paper we describe our attempt to disaggregate and smooth cirrhosis of the liver mortality rates. The purpose of the exercise is not to propose yet another method, but rather to generate interest and ideas in the presentation of vital event data.

The Problem

There are certain problems in using the county as the

basic geographic unit in preparing maps of mortality rates. For most causes of death the mortality rate is quite low and in many cases the population base at the county level is not sufficient to produce reliable estimates of the rates. When rates are computed at the county level and color coded to represent relative magnitudes of mortality, much of what the map reader observes is random noise. One way to overcome this rate instability is to collapse contiguous counties into larger geographic mapping units. This has been done and has resulted in aggregated units such as State Economic Areas. The aggregation of counties to larger units still does not fully solve the problem of mapping the rarer causes of mortality where in many aggregated units the death rate may be zero for a particular year. In some cases it may be necessary to aggregate to the state level to capture a population that will generate a stable rate.

A general rule of thumb for aggregation is that no mapping unit should have a population base of less than $5/\lambda$ where λ is the national mortality rate for the particular cause of death under consideration (Cochran, 1954).

However, aggregation tends to defeat our purpose because of the undesirable visual effect caused by comparing geographic units of heterogeneous size. Additionally, maps color coded based on percentile rankings of the estimated rates could still mislead the map reader by presenting little more than random noise. For example, even if the underlying mortality rates are exactly the same for each of the geographic mapping units, the estimated rates will vary. When the mapping units are categorized by arbitrary classes of rates the map reader will see only multicolored or gray-tone shaded random noise. Thus, if the maps are to be color coded in this manner it would be advisable to first test the hypothesis:

$$H_0: \quad r_1 = r_2 = \ldots = r_M$$

where r_i is the mortality rate for the i^{th} mapping unit.

The statistic $$X^2 = \sum_{i=1}^{M} N_i \, (\hat{r}_i - \hat{r})^2 / \hat{r}$$

where r_i is the estimated rate for the i^{th} mapping unit,

N_i is the population base for the i_{th} mapping unit,

and $$\hat{r} = \sum_{i=1}^{M} N_i \hat{r}_i / \sum_{i=1}^{M} N_i$$

is distributed as a chi-square with (M-1) degrees of freedom under H_0 could be used as the test statistic.

However, even if counties are aggregated to a minimum population for a specific cause, the chi-square statistic calculated, and the null hypothesis rejected we probably still would not have a visually acceptable map product. The heterogeneity of the geographic size of the mapping units would provide an element of visual "static" and the eye of the map reader would probably be drawn to the larger areas, passing over the smaller mapping units.

A Possible Approach

As a tentative step to solve this problem we could superimpose a grid system over the geographic area to be mapped and label the points of intersection p_1, p_2, ... p_N; then calculate the mortality rate for each county and associate this rate with the point representing the population centroid of the county. Denote the centroid points by c_1, c_2,...c_M. For an arbitrary intersection point p_i denote the distance between p_i and c_j by d_j and assume the centroid points have been labeled so that $d_1 \leq d_2 \ldots \leq d_M$.

We could then estimate the mortality rate at p_i by

$$r_i^* = \frac{\sum_{j=1}^{\alpha} 1/d_j \hat{r}_j}{\sum_{j=1}^{\alpha} 1/d_j}$$

where "α" is the smallest integer such that

$$\text{VAR}(r_i^*) = \sum_{j=1}^{\alpha} \left(\frac{1/d_j}{\sum_{j=1}^{\alpha} 1/d_j} \right)^2 \frac{\hat{r}_j(1-\hat{r}_j}{N_j} \leq A$$

where A is a prespecified upper bound for the variance of the r_i^*'s. After all of the r_i^*'s have been estimated, they could be ordered by magnitude and categorized for color coding. If the cell size were small enough this would emulate digital satellite data consisting of picture elements. This might be an interesting method for the presentation of this type of data.

What we actually did was somewhat different and easier. The three state area of Arizona, California, and Nevada was chosen as the test area for this exercise. It has the attributes of a gradient of population from large to small and it contains relatively empty areas. Counties in Arizona and Nevada are large and irregularily shaped; in California they range from large to small. Cirrhosis of liver mortality is relatively high in this test region compared to the rest of the United States (Figure 1).

Age-adjusted 1975 rates for male cirrhosis of liver mortality were calculated in the region for each county and the rates plotted. The major population center in each of the counties was located and plotted. A grid was then superimposed over the region and the population centers forced to the nearest grid intersection. For

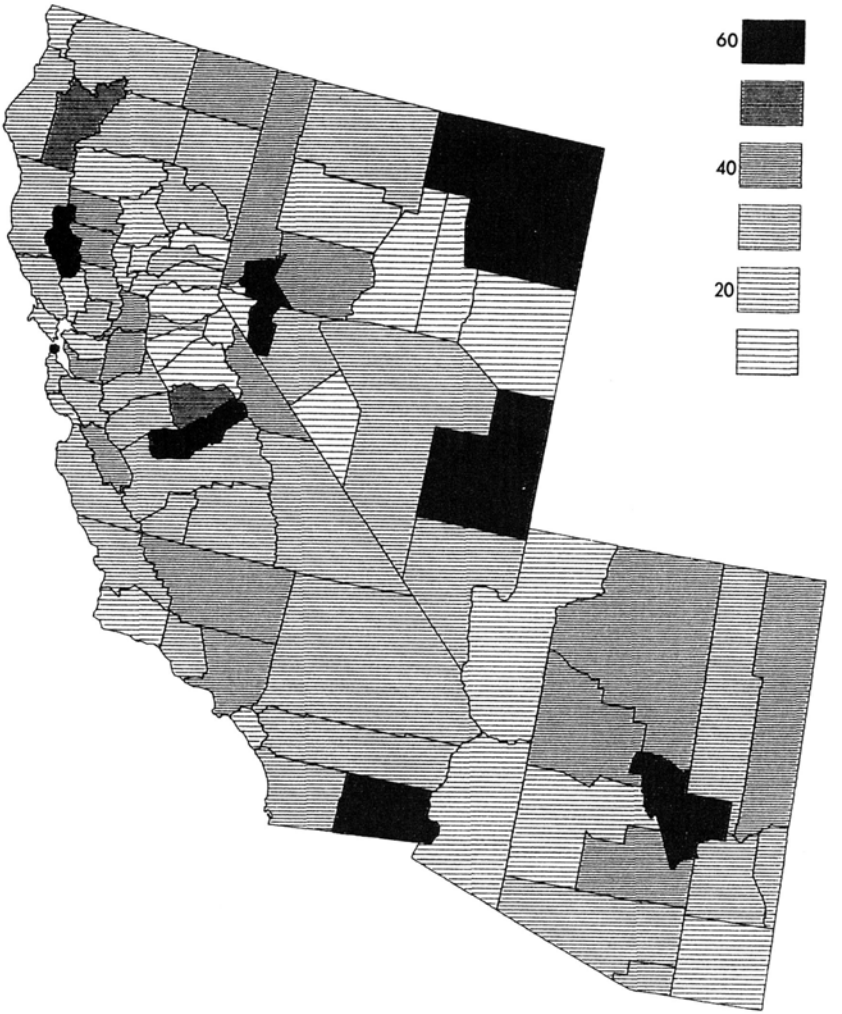

Figure 1. Age-adjusted male cirrhosis of liver mortality for Arizona, California, and Nevada by county, 1975.

some cases in California this required adding two or more small counties together and recalculating the age-adjusted rate for the grid intersection.

A weighted rate was calculated for each cell bounded by four grid intersections. This was accomplished by constructing a matrix where each odd i,j ordered pair of coordina represented a grid intersection and each even ordered pair of coordinates represented a cell. That is, the matrix points 1,1; 1,3; 3,1; and 3,3 defined the corners of the cell at 2,2. As each cell entered the rate calculation scheme, adjacent grid points were symmetrically searched for enough population to meet the minimum population criterion. The search was stopped when this criterion was met. Otherwise an additional set of rates from symmetrical intersections were added to the rate, weighted by the square root of the distance of the intersections from the center of the cell. This produced a grid of shaded cells over which the county boundries were superimposed (Figure 2).

The matrix of cell rates were then contoured (Tukey, 1979) and the county boundries superimposed over the contours (Figure 3).

Discussion

In Arizona, the gridding technique generally defined the population centers of Phoenix and Tucson. They were shifted North somewhat due to the influence of Gila County; Gila County recorded 20 cirrhosis of liver deaths for an estimated 1975 male population of 15,800. It seems appropriate that the two Arizona cities be highlighted since there were 363 deaths in Phoenix and 139 in Tucson for this cause in 1975.

Using the choropleth method, in Nevada it would appear that four counties have extremely high mortality rates for cirrhosis of liver. However, the estimated 1975 male populations for these counties ranged from 500 to 8,000; the number of deaths from 4 to 16. When the state was gridded and the population search algorithm carried out, the high rates in all these counties disappeared. However, the technique did produce an artifact at the northern edge of Nye county. If both numerators and denominators had been assigned to the

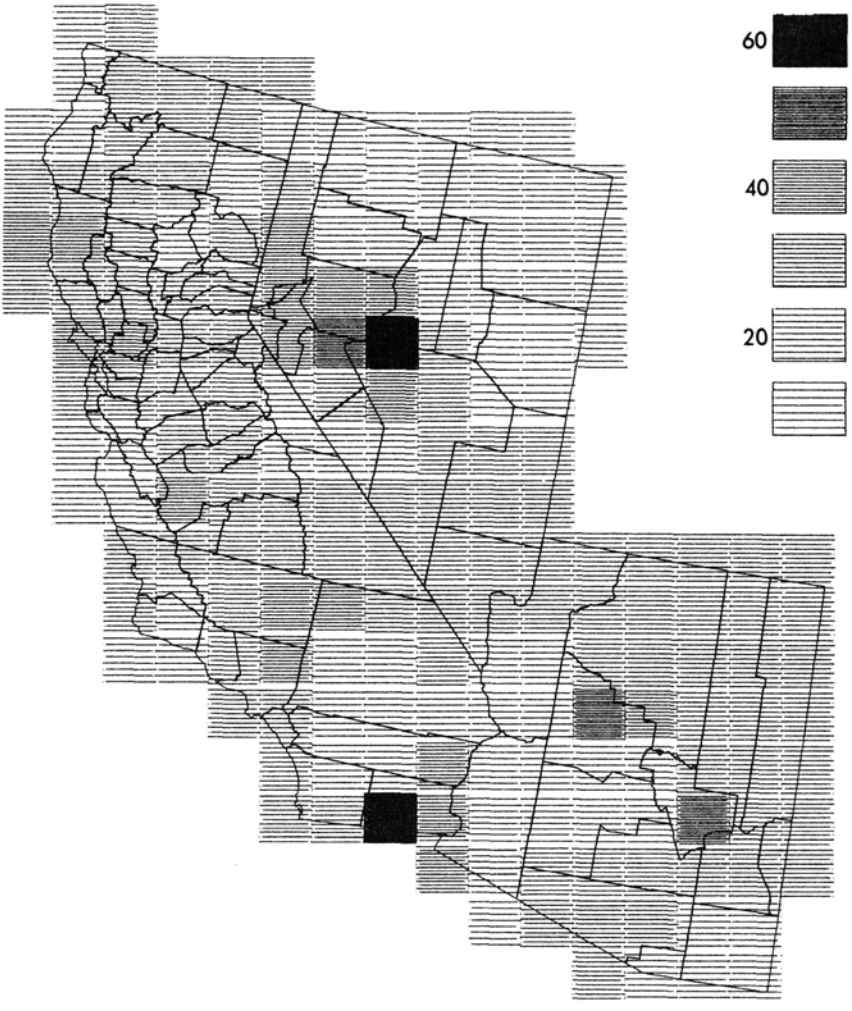

Figure 2. Age-adjusted male cirrhosis of liver mortality for Arizona, California, and Nevada by arbitrary cells, 1975.

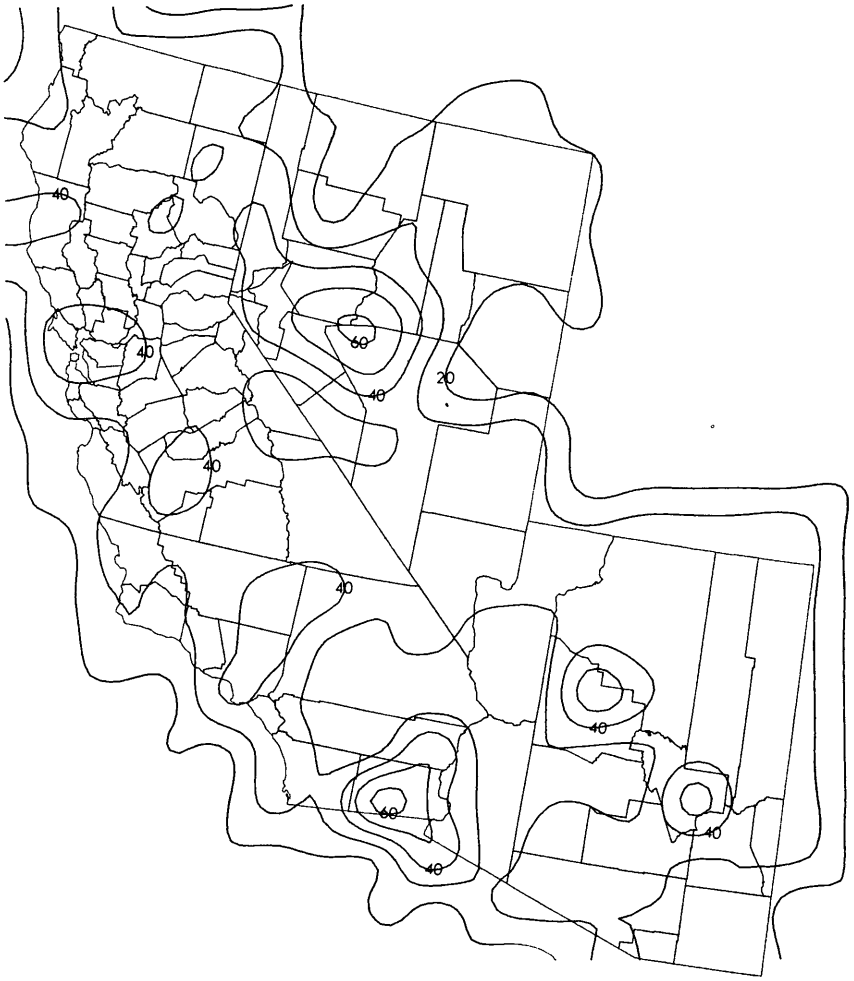

Figure 3. Contoured age-adjusted male cirrhosis of liver mortality for Arizona, California, and Nevada, 1975.

nearest grid intersections instead of an age-adjusted county rate, the minimum population found, and then an age-adjusted rate calculated, this artifact would disappear as well. This was confirmed by a hand calculation.

The Imperial County California "hot spot" is real.

It seems apparent that the choropleth technique is not appropriate for the mapping of the rarer causes of mortality because of the great heterogeneity of polygon size. Other methods that disaggregate or ignore political boundaries might more adequately display this type of data.

References

Cochran, W. G. 1954. Some methods for strengthening the X^2 common tests. Biometrics, 10, 417-451.

Tukey, J. W. 1979. Statistical mapping: What should not be plotted. In NCHS: Proceedings of the 1976 Workshop on Automated Cartography and Epidemiology. DHEW Publication No. (PHS) 79-1254. U.S. Government Printing Office. August 1979.