Title
Transformations of Maps to Investigate Clusters of Disease

Author
Selvin, S.

Publication Date
1984-10-01

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

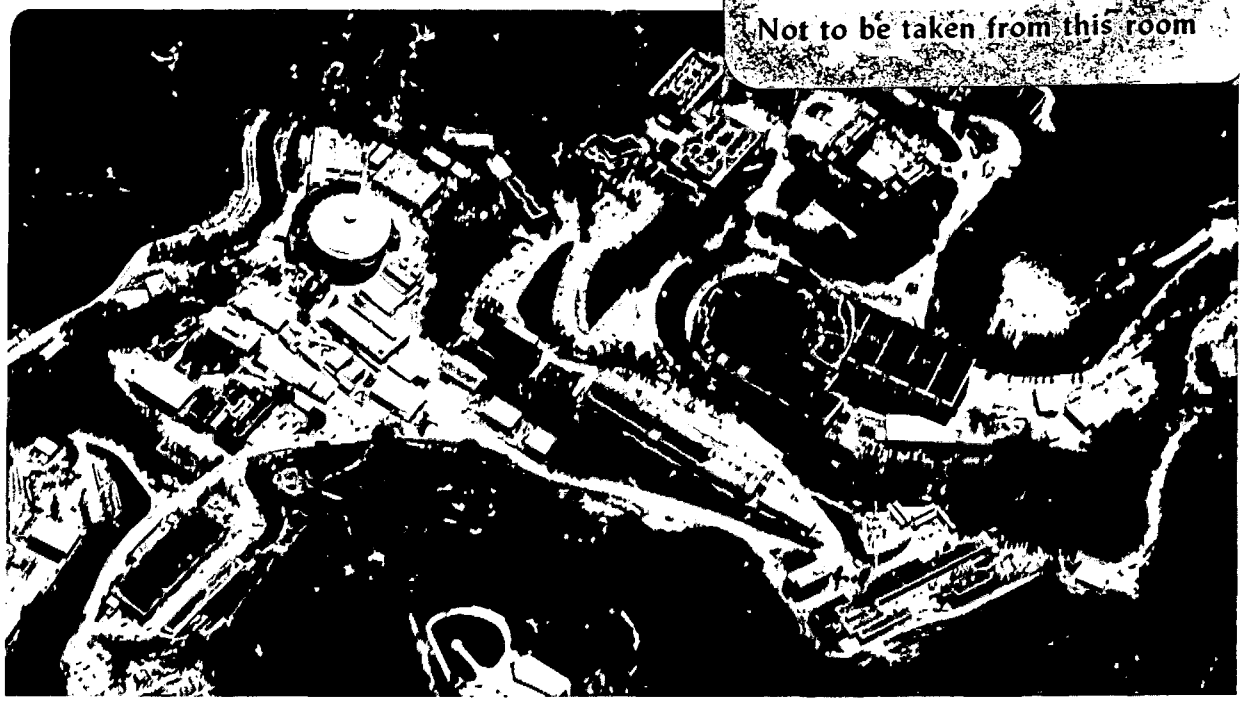
Computing Division

To be presented at the Annual Meeting of the American Public Health Association, Anaheim, CA, November 13, 1984; and to be published in the American Journal of Epidemiology

TRANSFORMATIONS OF MAPS TO INVESTIGATE CLUSTERS OF DISEASE

S. Selvin, D. Merrill, S. Sacks, L. Wong, L. Bedell, and J. Schulman

October 1984



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

DISCLAIMER

Transformations of Maps to Investigate Clusters of Disease

S. Selvin, D. Merrill, S. Sacks,
L. Wong, L. Bedell, and J. Schulman

To be presented at the annual meeting of the
American Public Health Association;
Anaheim, CA; November 13, 1984

This work was supported by the Director,
Office of Energy Research,
Office of Health and Environmental Research of the
U.S. Department of Energy under Contract Number
DE-AC03-76SF00098.

ABSTRACT

A procedure is presented to display and analyze epidemiologic data with the use of density-equalized maps. The algorithm for generating these maps is discussed in terms of several simple examples. Two specific methods for statistically analyzing these maps are given in detail, followed by an application of maps and methods to six sets of age-, race-, sex-, site-specific cancer incidence data. The data were obtained from the Surveillance, Epidemiology and End Results (SEER) project for San Francisco city/county (1978-1981) and combined with 1980 U.S. Census data.

INTRODUCTION

Distribution of disease is at the foundation of epidemiology as noted by A. Lilienfeld [1], who begins his text:

> Epidemiology may be defined as the study of the distribution of a disease or a pathological condition in human populations and the factors that influence this distribution.

Recent epidemiologic investigations have focused on geographic distributions of disease (e.g. [2]-[4]). Much of this recent work employs computer-generated maps that identify the highest and lowest site-, race-, sex-specific age-adjusted cancer mortality rates for the U.S. Further epidemiologic interpretation or extensive statistical analysis is generally not attempted. A major limitation of dealing with disease plotted on geopolitical maps stems from the fact that geographic subunits such as states, counties, or census tracts are not defined in terms of the population-at-risk to the disease. Large, sparsely populated areas tend to dominate a geopolitical map, whereas epidemiologic interest should be focused on areas with highly dense populations. The visual impact of a county map of the U.S., for example, is disproportionately influenced by the large and sparsely populated counties of the Rocky Mountain area and insufficiently influenced by the counties of the East Coast. Rigorous statistical analysis of maps based on geopolitical boundaries is complicated for essentially the same reason -- the geographic subunits often represent extremely different populations-at-risk. Statistical analysis of maps usually consists of a significance test to identify those high rates that were unlikely to have occurred due to random fluctuation. Often these statistical tests are based on the dubious assumption that disease with low frequency is described, at least approximately, by a Poisson distribution. Another way to deal with random fluctuations resulting from variation associated with small populations is to combine these less populated areas into larger geographic units (e.g., State Economic Areas [5]). This strategy involves a somewhat arbitrary combining of geographic units and decreases the specificity of any geographic analysis.

Ideal data for geographic analysis should contain the location of the cases of disease under study. Exact location data are rarely available for practical reasons and because of concern for confidentiality. The most detailed cancer incidence data available, on a large scale, come from cancer registry data (e.g., Surveillance, Epidemiology and End Results (SEER) [6]), where often the census tract of the cancer incident case is recorded. Corresponding age-sex-race specific population denominators are available from the decennial census (specifically the 1970 Second Count and the 1980 Summary Tape File 2A). As for geographic coordinates, 1970 census tract boundaries produced by Lawrence Berkeley Laboratory are available from the U.S. Census Bureau; 1980 census tract boundaries can be purchased from private sources.

Although census tract information is useful as an indicator of location, direct geographic analysis is still complicated by the lack of equality of risk among tracts. Presented here is a new approach using an old technique. The old technique involves drawing cartograms, i.e. maps scaled according to criteria other than the usual geopolitical boundaries. Here, the criterion is the equalization of population density, and the process will be referred to as a density equalizing map projection (DEMP). The goal of the transformation is to produce a map which depicts the distribution of disease uninfluenced by geopolitical boundaries but which preserves approximate spatial relationships. Such a map can be used for display purposes, or

it can be statistically analyzed to rigorously assess any observed pattern.

Such cartograms, produced by laborious non-computational methods, were described and used in public health applications as early as the 1920's [7]-[14]. Tobler discussed theoretical aspects of such transformations [15] and in 1974 developed a DEMP computer algorithm [16]-[17].

In the next section, we present a DEMP algorithm developed at Lawrence Berkeley Laboratory, followed by two suggested methods for statistical analysis of incidence data. In the final section, sample data for six cancer sites, from the SEER (Surveillance, Epidemiology and End Results) project, are analyzed with the use of density equalized maps.

ALGORITHM

The DEMP algorithm, as the name suggests, transforms boundaries so that all geographic units (e.g. census tracts) in a map have areas proportional to their population, i.e. equal population density. A pre-chosen density is set and the boundary of each tract is either expanded or contracted, creating a new map formed of these "equal density" tracts. If a number of tracts have equal population density, then a phenomenon which occurs with equal likelihood with respect to each individual will occur randomly over the entire map and the tract boundaries can be ignored. Specifically, the distribution of a disease plotted on a map with uniform population density is free from the interfering bias caused by unequal population densities inherent in a geopolitically defined map.

To illustrate the DEMP algorithm we start with the simplest possible example. Consider the two concentric circles in the left part of Figure 1 (a bull's eye). Suppose that the inner circle has an area $A=10$ square units; the outer circle encloses an area $A+B=30$ square units, so that the doughnut-shaped area lying between the two dashed circles has an area $B=20$ square units. Suppose also that the populations associated with A and B are equal.

To achieve an equal population density, area A can be increased from 10 to $A'=20$ square units while area B is maintained at 20 square units (total area = $A'+B$ = 40). The change in the radius of circle A (dr_a) and the corresponding change in the outer radius (dr_{a+b}) to achieve equal density are:

$$dr_a = r_{a'} - r_a = \sqrt{20/\pi} - \sqrt{10/\pi} = 2.52 - 1.78 = 0.74$$

and

$$dr_{a+b} = r_{a'+b} - r_{a+b} = \sqrt{40/\pi} - \sqrt{30/\pi} = 3.57 - 3.09 = 0.48$$

where $r_{a'}$ and $r_{a'+b}$ are the radii of the transformed circles, shown in the right part of Figure 1. In general, for circles of arbitrary radius, the transformed area of A is

$$A' = A+dA = \pi r_{a'}^2$$

where dA represents the necessary increase in A. Equivalently,

$$r_{a'} = r_a\sqrt{1+dA/A} = r_a M_A^{1/2}$$

where $M_A = 1+dA/A = (A+dA)/A = A'/A$ is the areal magnification factor applied to A. The necessary increase in the radius of A is

$$dr_a = r_a[M_A^{1/2}-1].$$

To keep the area of B the same, increasing the outer circle to compensate only for the increase in A, the transformed area of $(A+B)$ is

$$A'+B = A+dA+B = \pi r_{a'+b}^2$$

and the new radius is

$$r_{a'+b} = r_{a+b}\sqrt{1+dA/(A+B)} = r_{a+b}(1+(M_A-1)r_a^2/r_{a+b}^2)^{1/2}.$$

The necessary change in the radius is

$$dr_{a+b} = r_{a'+b} - r_{a+b} = r_{a+b} f \text{ , where}$$

$$f = [\sqrt{1+dA/(A+B)} - 1] = [(1+(M_A - 1) r_a^2 / r_{a+b}^2)^{1/2} - 1] .$$

The result has the following expected properties:

dr_{a+b} is zero if $M_A = 1$ (no change in A).

The radius r_{a+b} increases for $M_A > 1$ (magnification of A) and decreases for $M_A < 1$ (demagnification of A).

The formula for dr_{a+b} reduces to the expression for dr_a, for the special case $r_a = r_{a+b}$, i.e. when the circle A and the outer circle coincide.

The change in radius, dr_{a+b}, becomes small at large distances ($r_{a+b} \gg r_a$).

For the example of Figure 1,

$$dr_a = \sqrt{10/\pi}\,[\sqrt{2} - 1] = .74 \text{ , and}$$

$$dr_{a+b} = \sqrt{30/\pi}\,[(1+(2-1)(10/\pi)/(30/\pi))^{1/2} - 1] = .48 .$$

Computer maps are normally made up of polygons. These polygons are represented as a series of discrete line segments defined by a series of coordinate points (x_i, y_i). Instead of operating on radii of circles, a computer implementation to form equal-density areas requires a calculation in terms of xy-coordinates of points describing polygon boundaries. The expression for f given here for the case of concentric circles applies also to arbitrary points (x_i, y_i), provided that r_a and r_{a+b} are suitably redefined. For a fuller discussion, see the Appendix.

Figure 2 shows the application of the DEMP algorithm to a geometric configuration slightly more complicated than two concentric circles. Suppose the circle A has a population density 16 times that of the area $\bar{A}$, circle B has a population density 1/20 that of area $\bar{B}$, and the areas $\bar{A}$ and $\bar{B}$ (not including A and B respectively) have equal population density. The lower half of Figure 2 results from applying the DEMP algorithm to increase the area A by a factor of 16, to decrease the area B by factor of 20 and to leave the areas $\bar{A}$ and $\bar{B}$ unchanged. The new projection now has equal population density over the entire figure. That is, distances between occurrences plotted on this DEMP are no longer influenced by the original density inequality inherent in the top figure. If a phenomenon occurs at random (with probability proportional to the population at risk) within areas A, $\bar{A}$,B, and $\bar{B}$, then the distribution of points of occurrence will be uniformly distributed over the entire two-dimensional region.

The two maps in Figure 3 show the result of applying the DEMP algorithm to the contiguous states of the U.S. Obviously, the large Rocky Mountain states play only a small role with respect to any phenomenon related to population, such as the frequency of disease. California and the states in the northeast are enlarged, while states such as Montana, North Dakota and Idaho all but disappear. Although the transformed shape of the U.S. is not very familiar, a phenomenon that has the same probability of affecting each individual will be uniformly distributed over the transformed map. Such a Poisson random variable will be "fairly" depicted and the

contributions from each state to the visual impact will be proportional to the population-at-risk (surface area of the map).

Figures 4, 5 and 6 illustrate applications of DEMP to the 150 census tracts of the city and county of San Francisco. Figure 4 is a standard geopolitical representation of the 1980 census tracts with selected census tracts variously shaded.

In Figure 5 the same map of San Francisco has been transformed to yield equal population density for white males 35 to 54 years of age. Note that areas of high population density (dotted shading) are expanded, and areas of low population density (diagonal shading) are decreased. Two areas, namely Golden Gate Park (the long rectangle in the west) and an industrial tract in the extreme southeastern corner, completely disappear because both have no permanent residents and, therefore, no relevance to population-related phenomena.

Figure 6 is a similar DEMP for the 35 to 54 year old black male population of San Francisco. A predominantly black neighborhood (solid shading in Figure 4) shows an expectedly large increase in size from Figure 4 to Figure 6. The equal-density map of black males 35 to 54 years old reveals that these persons live predominately in three neighborhoods of the city.

Figure 7 shows thirteen cases of a hypothetical "disease" with a rate of 26 cases per 100,000 population (among white females 35 to 54 years old) and no assumed geographic pattern. The clustering in the north-eastern part of San Francisco is due to the high population density in that part of the city.

The DEMP produces a distribution of "cases" on a map with equalized population density (Figure 8). As expected, the distribution of this "disease" has no defined pattern, and this fact is reflected in the distribution plotted on the DEMP of San Francisco. In the next section, statistical analysis is used to assess chance variation as a likely explanation for such an observed configuration.

STATISTICAL ANALYSES

In this section, two statistical methods are used to illustrate ways in which data plotted in a DEMP can be analyzed. Both methods provide rigorous ways of identifying non-random patterns of disease. The null hypothesis underlying these two techniques is that the cases of disease are randomly distributed among the individuals residing in the area under investigation. The end product, as in most statistical analyses, is a probability ("p-value") associated with the likelihood of finding the observed distribution, given this null hypothesis.

A reduction (when compared to randomness) in the average distance (between all possible pairs) among the cases of a specific disease indicates one type of non-random pattern. For example, if cases of a disease are associated with a point source of pollution or the life-style of a specific neighborhood, a cluster of cases would be expected. Of course there are many ways to measure the degree of clustering in a set of observations [18], each with its own properties. A simple measure is the average squared distance among the cases. If k cases occur, then the average squared distance is

$$\bar{d} = \sum_{i<j}\sum d_{ij}^2 / \binom{k}{2}$$

where d_{ij} is the distance between case i and case j. If $\bar{D}$ is the random variable associated with the measured value $\bar{d}$, then under the null hypothesis, the expectation and variance of $\bar{D}$ are

$$E(\bar{D}) = 2(EX^2 + EY^2) \text{ and}$$

$$Var(\bar{D}) = \{[4W - (E\bar{D})^2]k - 4[W - 4(EXY)^2 + 4EX^2EY^2] + 3(E\bar{D})\} / k(k-1) \text{ , with}$$

$$W = EX^4 - 2EX^2Y^2 + EY^4 \text{ , where}$$

$$EX^iY^j \text{ is estimated by } \sum_{k=1}^{n}(x_k - \bar{x})^i (y_k - \bar{y})^j / n \text{ , and}$$

$(x_i, y_i), i = 1,2,...n$ are the centroids of the n census tracts. For example, for San Francisco white males 35 to 54 years old,

$$E(\bar{D}) = 15.71 \text{ and}$$

$$Var(\bar{D}) = (90.0k + 162.5) / k(k-1)$$

are the expectation and variance calculated from the centroids of the $n = 150$ transformed census tracts. The units of the numerical constants depend on the units of $\bar{D}$, which are arbitrary and proportional to map area. So as to maintain approximate comparability among the $\bar{d}$ of the various age-sex-specific data sets, the total area of each transformed map is kept equal to that of the original map (see Appendix).

The average squared distance $\bar{d}$ among the k cases (under the null hypothesis) has a nearly symmetric distribution for San Francisco and can be accurately approximated by a normal distribution. The approximate normality of the distribution of $\bar{D}$ stems from the Central Limit Theorem and the fact that the transformed map of San Francisco is nearly square. It should be noted that the distribution of average distance

$$\sum_{i<j}\sum d_{ij} / \binom{k}{2}$$

is less symmetric than that of the average squared distance $\bar{d}$ and, therefore, not as accurately approximated by a normal distribution as is $\bar{D}$.

The statistical analysis of $\bar{D}$ is accomplished by calculating

$$z = (\bar{d} - E\bar{D}) / \sqrt{Var(\bar{D})}$$

and $p = P(Z<z)$, where Z has an approximate standard normal distribution. The observed value of z provides a summary assessment of the likelihood that the observed $\bar{d}$ could have occurred by chance under the null hypothesis. Small values of p are an indication of clustering.

Alternatively, the distribution of $\bar{D}$ can be investigated with the use of computer simulation. By selecting a random sample of k census tracts, each with selection probability proportional to the tract population, and calculating $\bar{d}$ many times (say, 200 times) an empiric null distribution can be derived. In other words, the distribution of $\bar{D}$ based on k random "cases" can be found. The observed $\bar{D}$ can then be assessed using this computer simulated distribution. Both approaches give similar results when applied to data for San Francisco (see next section).

The probability (power) of detecting a decrease in average squared distance, when it exists among a set of cases, can be approximately calculated. If one postulates that the average squared distance among the cases is systematically decreased by a factor f, then the probability that the data measurably reflect this decrease is a function of sample size alone. Given the fact that San Francisco is approximately square, the approximate variance of $\bar{D}$ is

$$Var(\bar{D}) = r^2 (2k+3) / 90\binom{k}{2}$$

where r is the length of the side of the "square" San Francisco. The probability of detecting a decrease in squared distance is then

$$\text{power} = P(Z<z')$$

where Z has a standard normal distribution and

$$z' = (1-f) / 3\sqrt{Var(\bar{D})} - Z_{1-\alpha} / f \ .$$

Note that $Z_{1-\alpha}$ is the $1-\alpha$'th percentile of a standard normal distribution where α is typically set at 0.05, making $Z_{1-\alpha} = Z_{.95} = 1.645$. The confidence level α is the probability of detecting an apparent decrease in f if none exists (type I error); β = (1-power) is the probability of failing to detect a true decrease in f (type II error). Table 1 and Figure 9 show the results of the power calculation for f = 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0. These power curves indicate that even small sample sizes are sufficient to statistically detect clustered observations. For example, a 50% decrease (f =0.5) in average squared distance will be detected (α=0.05) with close to .90 (β = .10) probability for a sample size of k =10. Of course, this power calculation depends on an extremely simple, almost simplistic, statistical structure but does tend to indicate the efficacy of using a measure of intra-case distance for the study of the distribution of cases of a disease.

Another approach to analyzing geographic data on a transformed map is to estimate the parameters of a polynomial representation of the distribution of cases [19]. One such polynomial is

$$z_i = a + b_1 x_i + b_2 y_i + b_3 x_i^2 + b_4 y_i^2 + b_5 x_i y_i$$

where z_i is the number of cases in census tract i, having its centroid at (x_i, y_i). The coefficients for this model can be estimated by ordinary least squares methods. For a random Poisson distribution of cases, the estimated coefficients $\hat{b}_1, \hat{b}_2, \hat{b}_3, \hat{b}_4$ and $\hat{b}_5$ are randomly distributed about zero. On the other hand, a systematic geographic pattern of disease will be associated with the coordinates x_i and/or y_i, increasing the likelihood that large values of the coefficients will occur. The evaluations of the proposed model comes from the comparison of two quantities, namely

$$S_1 = \Sigma_{i=1}^{n} (z_i - \hat{z}_i)^2 / n$$

and

$$S_0 = \Sigma_{i=1}^{n} (z_i - \bar{z})^2 / n \text{ , where}$$

$$\hat{z}_i = a + b_1 x_i + b_2 y_i + b_3 x_i^2 + b_4 y_i^2 + b_5 x_i y_i \text{ ,}$$

and $\bar{z} = k/n$ is the average number cases per tract. The degree by which S_1 is reduced relative to S_0 measures the usefulness of some or all of the b-coefficients in "explaining" the distribution of the disease. The quantity

$$f = [(S_0 - S_1)/5] / [S_1/(n-6)]$$

has an approximate F-distribution when the cases occur at random with respect to the xy-coordinates. For small numbers of cases, however, the significance probability derived from f is not extremely accurate since the quantity $(z_i - \hat{z}_i)$ is not normally distributed. Nevertheless, the comparison of a series of these f-values gives a relative measure of randomness or lack thereof among a series of diseases. Fitting a polynomial to a distribution of cases for San Francisco has the added feature that the terms of the model are essentially orthogonal (since San Francisco is approximately square), thereby increasing the estimation precision.

ILLUSTRATIVE DATA

To illustrate two approaches for analyzing the geographic distribution of disease, data from the Surveillance, Epidemiology and End Results (SEER) program for residents of San Francisco are used. Incidence cases among whites (1978-1981) for six cancers (stomach, colon, rectum, Hodgkin's disease, chronic lymphocytic leukemia and acute granulocytic leukemia), two age groups (35-54 and 55-74), and both sexes were selected. The total of 1225 cancer cases among the white population was reduced to 1128 since 97 cases (8%) were not assigned to a valid 1980 census tract. The number of cases for each site, the incidence rate per 100,000 person-years, and the results of the statistical analysis are given in Table 2. Four age- and sex-specific DEMP's were generated from tract-level 1980 census data (Summary Tape File 2A) and corresponding geographic base maps. The tract-level SEER data, 1980 census population data, and geographic base maps were all taken from the SEEDIS information system at Lawrence Berkeley Laboratory [20]. In the SEER data, the exact location of the residence of each case has been suppressed to preserve confidentiality. We used instead the geographic centroid of the census tract of residence; the error thereby introduced is not significant when the number of tracts is much larger than the number of cases.

Figures 10 through 15 show the location of the cases for white females 35 to 54 years old, for each of the six cancer sites, on a population-adjusted tract map of San Francisco. The centroid of the cases is calculated, and a circle is constructed with this centroid as the center, so that 50% of the k cases are contained within the circumference. (No circle is drawn if $k<3$.)

These circles are analogous to a confidence region in that both location and variability are simultaneously depicted. The center of the circle indicates location, and the size of the circle indicates the degree of dispersion. For example, the circle generated by the distribution of colon cancers (wf: 35-54) (Figure 11) is relatively large and centered in the middle of San Francisco, whereas the circle associated with stomach cancers (wf: 35-54) (Figure 10) is off-center and somewhat reduced in size.

The average squared distances $\bar{d}$ among the cancer cases are given in Table 2. The expected value $E(\bar{D})$ for white females 35-54 years old is 13.31, which differs slightly from the other three age-sex cases (wm: 35-54 = 15.71, wm: 55-74 = 15.81 and wf: 55-74 = 15.88). In white females 35-54 years old, stomach cancer (k = 6, $\bar{d}$ = 5.9 and p = 0.04), Hodgkin's disease (k = 2, $\bar{d}$ = 1.5 and p = 0.15), and chronic lymphocytic leukemia (k = 2, $\bar{d}$ = 0.6 and p = 0.12) show the smallest values of $\bar{d}$.

The "p-values" given in the parentheses (Table 2) result from assessing $\bar{d}$ with the use of simulation techniques. None of the values differs much from the "p-values" calculated using the normal distribution as an approximation to the distribution of $\bar{D}$ (with perhaps the exception of colon cancer among women 35-54). Using an empirically generated distribution will yield useful results under most circumstances. The "normal approximation" can be used only when the distribution of $\bar{D}$ is symmetric or nearly so, which depends on the shape of the region being analyzed. Regions that are approximately circular or square produce distributions of $\bar{D}$ that are close to symmetric.

These data are presented primarily for illustrative purposes and no epidemiologic interpretation is given. In the case of Hodgkin's disease and chronic lymphocytic leukemia, each with only two observations, any inferences would be very tenuous.

The lack of reliability comes from the fact that small samples can be highly influenced by biases (e.g., misclassification, wrong diagnosis, incorrect residence, etc.). Apart from these biases, the significance probabilities are accurate (even for only two cases) and do indeed represent the likelihood that the observed squared distances are due solely to chance variation.

An analysis of cancer cases using estimated polynomials indicates the possibility of several non-random distributions among the 24 age-sex-site combinations for whites in San Francisco. Stomach cancer in women 35-54 ($k=6$, $p<0.01$), colon cancer in both men 55-74 and women 55-74 ($k=251$, $p<0.01$ and $k=291$, $p<0.01$), and rectal cancer in women 55-74 ($k=109$, $p<0.01$) show some systematic patterns in their geographic distributions. A more epidemiologically focused analysis will be necessary to verify and explain the observations noted here.

Two points are worth reiterating. First, using a density-equalized map projection as an analytic tool produces a valid analysis even for extremely small numbers of cases. Second, routinely collected registry data such as the SEER data can be explored using density-equalized maps as long as the census tract of residence is recorded.

As presently implemented, the DEMP algorithm is still too slow for the systematic analysis of large data sets. Innovative approaches are being explored in the hope of overcoming these technical difficulties. In view of the growing availability of low-cost computer resources, the authors believe that the methods described here will see widespread use in the epidemiologic analysis of routinely collected mortality and disease incidence data.

Table 1. Power calculations for f =0.5 through 1.0 and sample sizes k = 5 through 50, with the significance level set at 0.05.

	f=0.5	f=0.6	f=0.7	f=0.8	f=0.9	f=1.0
k = 5	0.30	0.19	0.12	0.09	0.06	0.05
k = 10	0.87	0.58	0.32	0.17	0.09	0.05
k = 20	1.00	0.95	0.69	0.35	0.14	0.05
k = 25	1.00	0.99	0.81	0.43	0.16	0.05
k = 30	1.00	1.00	0.89	0.51	0.18	0.05
k = 35	1.00	1.00	0.94	0.58	0.20	0.05
k = 40	1.00	1.00	0.96	0.64	0.23	0.05
k = 45	1.00	1.00	0.98	0.70	0.25	0.05
k = 50	1.00	1.00	0.99	0.75	0.27	0.05

Table 2. Incident cases, rates and statistical analysis of six cancer sites by age and sex -- San Francisco SEER data for whites (1978-1981).

		Cases	Rate	$\bar{d}$	"p-value"	Fit
Stomach Cancer						
ages 35-54	wm	18	9.3	19.7	0.95(.97)	0.88
	wf	6	4.3	5.9	0.04(.05)	<0.01
ages 55-74	wm	78	50.3	16.1	0.85(.86)	0.19
	wf	49	32.8	18.5	0.97(.98)	0.03
Colon Cancer						
ages 35-54	wm	34	17.6	14.0	0.15(.18)	0.34
	wf	29	19.4	11.6	0.14(.34)	0.94
ages 55-74	wm	251	162.0	16.5	0.85(.84)	<0.01
	wf	291	146.1	15.5	0.20(.26)	<0.01
Rectal Cancer						
ages 35-54	wm	26	13.7	15.4	0.45(.54)	0.52
	wf	20	13.3	14.4	0.71(.86)	0.90
ages 55-74	wm	141	91.0	16.9	0.92(.90)	0.94
	wf	109	50.7	15.6	0.37(.41)	<0.01
Hodgkin's Disease						
ages 35-54	wm	10	5.0	20.0	0.90(.91)	0.74
	wf	2	2.7	1.5	0.15(.11)	0.30
ages 55-74	wm	4	2.5	15.2	0.47(.49)	0.86
	wf	3	1.5	9.1	0.22(.25)	0.83
Chronic Lymphocytic Leukemia						
ages 35-54	wm	3	1.6	25.3	0.87(.87)	0.80
	wf	2	1.4	0.6	0.12(.04)	0.73
ages 55-74	wm	12	7.7	21.8	0.97(.95)	0.06
	wf	12	6.0	15.9	0.65(.65)	0.17
Acute Granulocytic Leukemia						
ages 35-54	wm	4	2.1	22.0	0.83(.86)	0.71
	wf	4	2.7	15.9	0.67(.70)	0.78
ages 55-74	wm	16	10.3	20.5	0.97(.98)	0.11
	wf	12	6.0	15.7	0.47(.49)	0.59

Note: rate given is per 100,000 person-years at risk.

REFERENCES

1. Lilienfeld AM. Foundations of Epidemiology. New York, N.Y.: Oxford Univerisity Press, 1976.

2. Mason TJ, McKay FW, Hoover R et al. Atlas of Cancer Mortality for U.S. Counties: 1950-1969. US DHEW Publication No. (NIH) 75-180. Washington, DC: US GPO.

3. Blair A, Fraumeni JF and Mason TJ. Geographic patterns of leukemia in the United States. J Chronic Dis 1980;33:251-60.

4. Selvin S, Levin L, Merrill D, and Winkelstein W. Selected epidemiologic observations of cell-specific leukemia mortality in the United States. Am J Epidemiol 1983;117:140-52.

5. Mason TJ, McKay FW, Hoover R et al. Atlas of Cancer Mortality Among Nonwhites: 1950-1969. US DHEW Publication No. (NIH) 76-1204. Washington, DC: US GPO.

6. Surveillance, Epidemiology, and End Results: Incidence and Mortality data: 1973-77. US DHEW Publication No. (NIH) 81-2330. Washington, DC: US GPO.

7. Karsten KG. Charts and Graphs, Chapter 52. 1923.

8. Wallace, JW. Population map for health officers. AJPH 1926;16:1023.

9. Gillihan AF. Population maps. Am J Pub Health 1927;17:316-9.

10. Raisz E. The rectangular statistical cartogram. Geogr Rev 1934;24:292-6.

11. Levison ME and Haddon W. Publ Hlth Rep 1965;80:55.

12. Forster F. Use of a demographic base map for the presentation of areal data in epidemiology. Brit J Prev Soc Med 1966;20:165-171.

13. Tyroler HA and Smith HL. Epidemiology and planning for the North Carolina regional medical program. AJPH 1968;58:1058-67.

14. Dean AG. Population-based spot maps: an epidemiologic technique. AJPH 1976;66:988-9.

15. Tobler WR. Geographic area and map projections. Geog Rev 1963;53(1):59-78. Annals NY Acad Sci 1973;219:215-20.

16. Tobler WR. Cartogram programs. Dept of Geography, Univ Mich, Ann Arbor. Unpublished manuscript, 110 pp, 1974.

17. Tobler WR. Cartograms and Cartosplines. In proceedings of 1976 Workshop on Automated Cartography and Epidemiology. DHEW Publ 79-1254, USGPO Aug 1979, pp 53-57.

18. Everitt, B. Cluster Analysis. London: Heinemann Educational Books Ltd., 1974.

19. Ripley BD. Spatial Statistics. New York, N.Y.: John Wiley & Sons, 1981.

20. McCarthy JL, Merrill DW, Marcus A, Benson Wh, Gey FC, Holmes H and Quong C. The SEEDIS Project: A Summary Overview of the Social, Economic, Environmental Demographic Information System. Lawrence Berkeley Laboratory, Univ of California Report PUB-424, 1982.

APPENDIX

Here we describe the DEMP algorithm used in this paper.

For an arbitrary polygon having area equal to $area_A$, choose a convenient center of expansion, for example the geographic centroid of the polygon A. The center of expansion normally lies within A, but this is not necessary.

Next consider an arbitrary point i having coordinates (x_i, y_i) relative to the A expansion center, at an angle

$$\Theta_i = arctan(y_i / x_i)$$

and a distance

$$r_i = [x_i^2 + y_i^2]^{1/2} .$$

The point i normally lies outside A, but this is not necessary. For consistency with earlier notation, we define $r_{a+b} = r_i$.

Depending on the configuration of the polygon A, its expansion center, and the point i, a line at angle Θ_i from the expansion center to point i may intersect the boundary of polygon A n times, at distances we call $r_{an} = r_{an}(\Theta_i)$. In the simplest case the expansion center lies inside A, point i lies outside, and the line intersects the boundary just once $(n=1)$. For consistency with earlier notation, we define, in this case, $r_a = r_{a1}$. More generally, we define

(a) center inside A, point i outside, n odd: $r_a^2 = +r_{a1}^2 - r_{a2}^2 + r_{a3}^2 - \cdots + r_{an}^2$

(b) center inside A, point i inside, n even: $r_a^2 = +r_{a1}^2 - r_{a2}^2 + r_{a3}^2 - \cdots - r_{an}^2 + r_i^2$

(c) center outside A, point i outside, n even: $r_a^2 = -r_{a1}^2 + r_{a2}^2 - r_{a3}^2 + \cdots + r_{an}^2$

(d) center outside A, point i inside, n odd: $r_a^2 = -r_{a1}^2 + r_{a2}^2 - r_{a3}^2 + \cdots - r_{an}^2 + r_i^2$

With these definitions for r_{a+b} and r_a, and with x_i and y_i expressed relative to the A expansion center, a transformation which multiplies the area of polygon A by a factor M_A and leaves all other areas unchanged is:

$$dx_i = x_{i'} - x_i = x_i f$$

$$dy_i = y_{i'} - y_i = y_i f \text{ , where}$$

$$f = [(1+(M_A-1)r_a^2 / r_{a+b}^2)^{1/2} - 1] \text{ , and}$$

$$M_A = [(pop_A)/(pop_{total})] \,/\, [(area_A)/(area_{total})] ,$$

where pop_{total} and $area_{total}$ are the population and (original) area of the entire map.

The form of f is identical to that derived earlier for the simple case of concentric circles. Regardless of the choice of the expansion center of polygon A, the transformation changes the area but not the shape of polygon A, and the shapes but not the areas of all other polygons.

First the areal magnification factors M_A are calculated for all polygons. Every polygon of a multi-polygon area (e.g. the state of Michigan) is assigned the same M_A based on the combined population and area of its component polygons. To remove uninhabited areas (e.g. deserts, rivers, lakes, etc.) by shrinking them to zero area, describe them as polygons with $pop_A = 0$, which implies $M_A = 0$. Otherwise ignore them, which is equivalent to describing them as polygons with $M_A = 1$.

Now the described transformation for polygon A is applied to every point (x_i, y_i) in the entire map; then another polygon A is selected and a second transformation is applied to every point (x_i, y_i); this process is repeated for all polygons A in the entire map. The order in which the individual polygons A are selected for magnification or contraction affects the shapes but not the areas of polygons in the final map.

The correct normalization of maps depends upon the statistical analyses to be performed. The stated definition of M_A implies that after all transformations are complete,

$$area'_{total} = area_{total} ,$$

i.e. the total area of the map is unchanged. This normalization is appropriate for the discussion in this paper, in particular the comparison among sexes and races of the four values of $E(\bar{D})$, or of the various values of $\bar{d}$ in Table 2.

If one wishes instead to equalize population density over all sexes and ages, for example for sex-age comparisons of case densities (rates), all x_i and y_i in each age-sex-specific map must be multiplied by the constant factor

$$[(pop_{total}) / (area_{total})]^{1/2}$$

either before or after the DEMP transformation, where pop_{total} is the total age-sex-specific population. (Defining $M_A = pop_A / area_A$ would yield correct areas but would result in undue distortion).

Yet another normalization would be appropriate if one wished to equalize density of site-specific cases, for example to detect clustering by analyzing the distribution of nearest-neighbor distances. In this case all x_i and y_i in each age-, sex-, site-specific map must be multiplied by

$$[\, k \,/ (area_{total})]^{1/2}$$

before or after the DEMP transformation, where k is the age-, sex-, site-specific number of cases.

The algorithm defined here exactly describes the correct transformation of individual points. Computationally, the only difficult task is calculation of the intersection distances $r_{ap}(\Theta_i)$. However, the transformed areas are not exactly correct, because the straight line segments of polygon boundaries must be be transformed into curves, which in turn must be approximated by transformed polygons. An accurate representation of curves requires a large number of points, and the execution time of the DEMP algorithm increases as the square of the number of points in the map. Map subregions cannot be processed separately, or the external transformed boundaries will not coincide. Therefore, an implementation that is both accurate and fast requires judicious approximations, including selective

insertion and removal of points in the polygon boundaries. This necessitates transforming the map back and forth between the usual polygon representation and a node-and-string or DIME format.

Development of an improved implementation is in progress. Details will be discussed elsewhere.

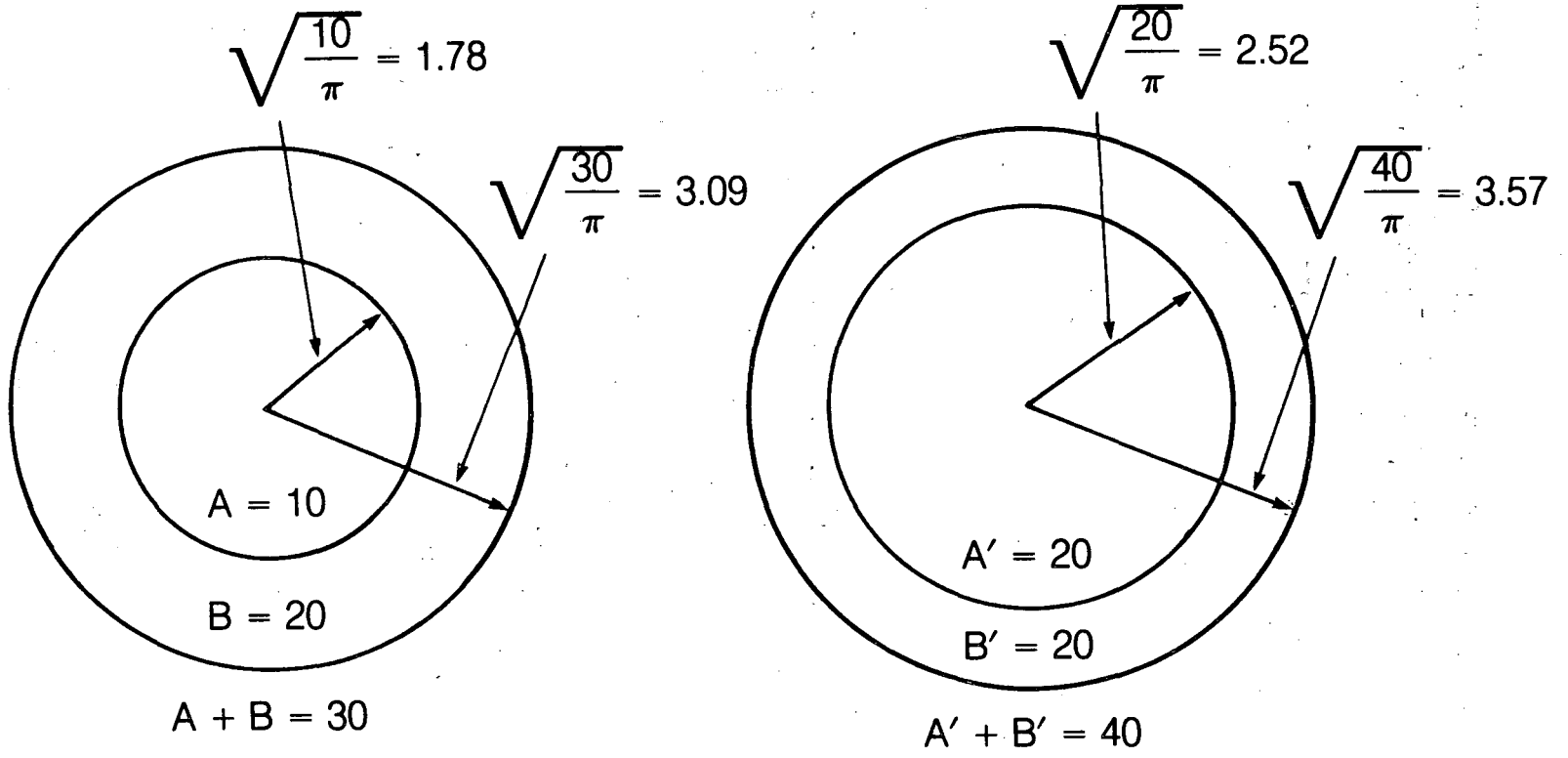
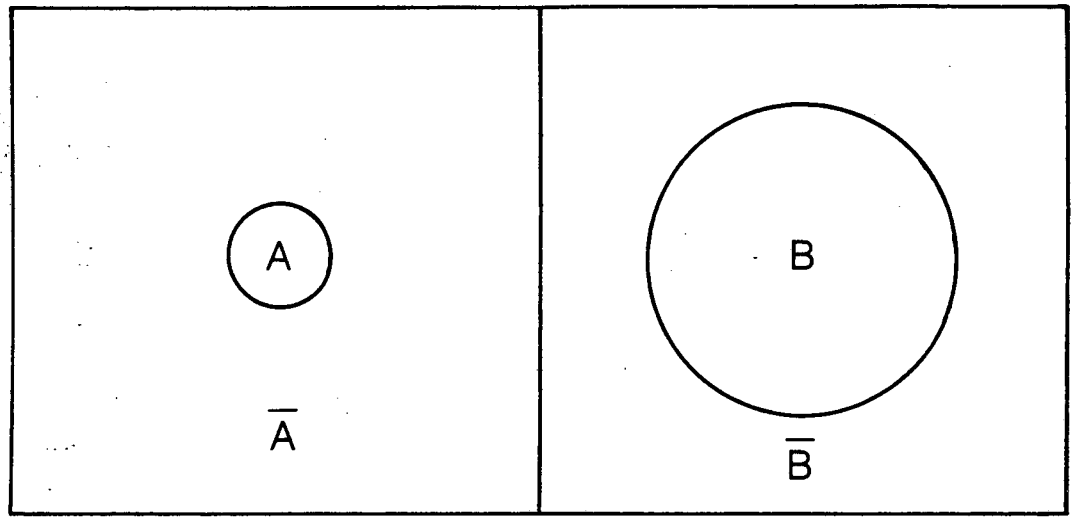


Fig. 1

XBL 849-8745

Another simple application of the map transformation algorithm.



$A' = 16A$, $\overline{A}' = \overline{A}$ $\quad$ $B' = B/20$, $\overline{B}' = \overline{B}$

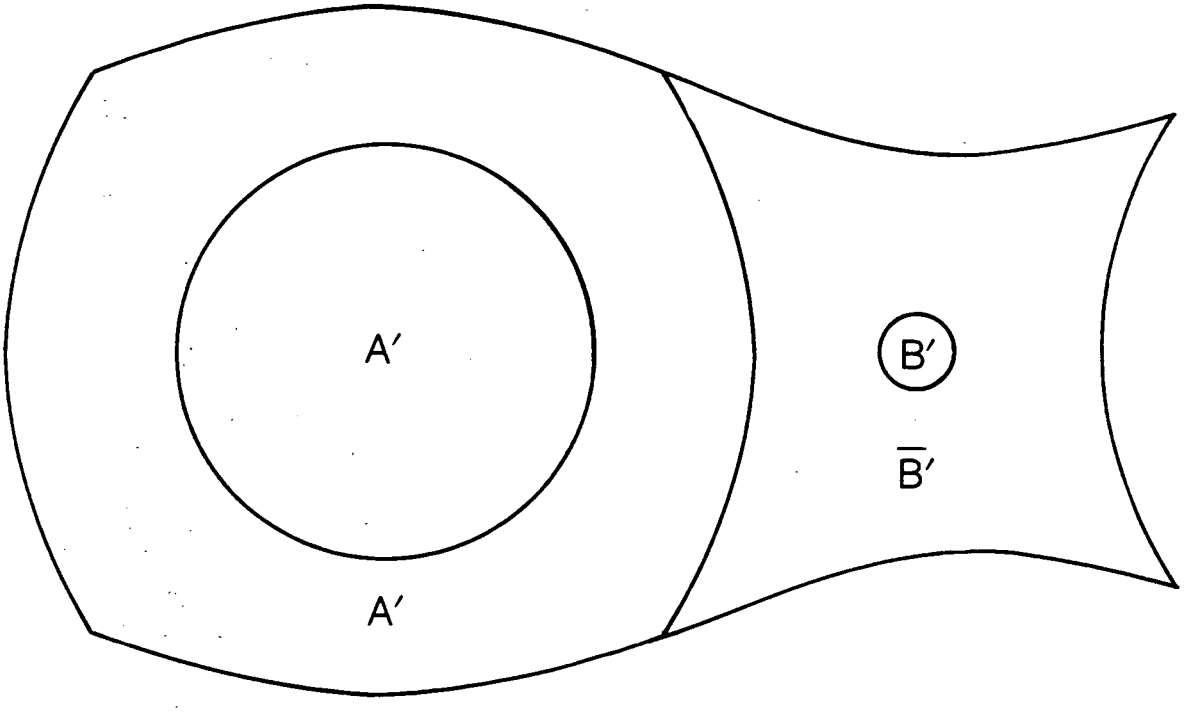


Fig. 2

XBL 848-8621

Transformed by population — United States.

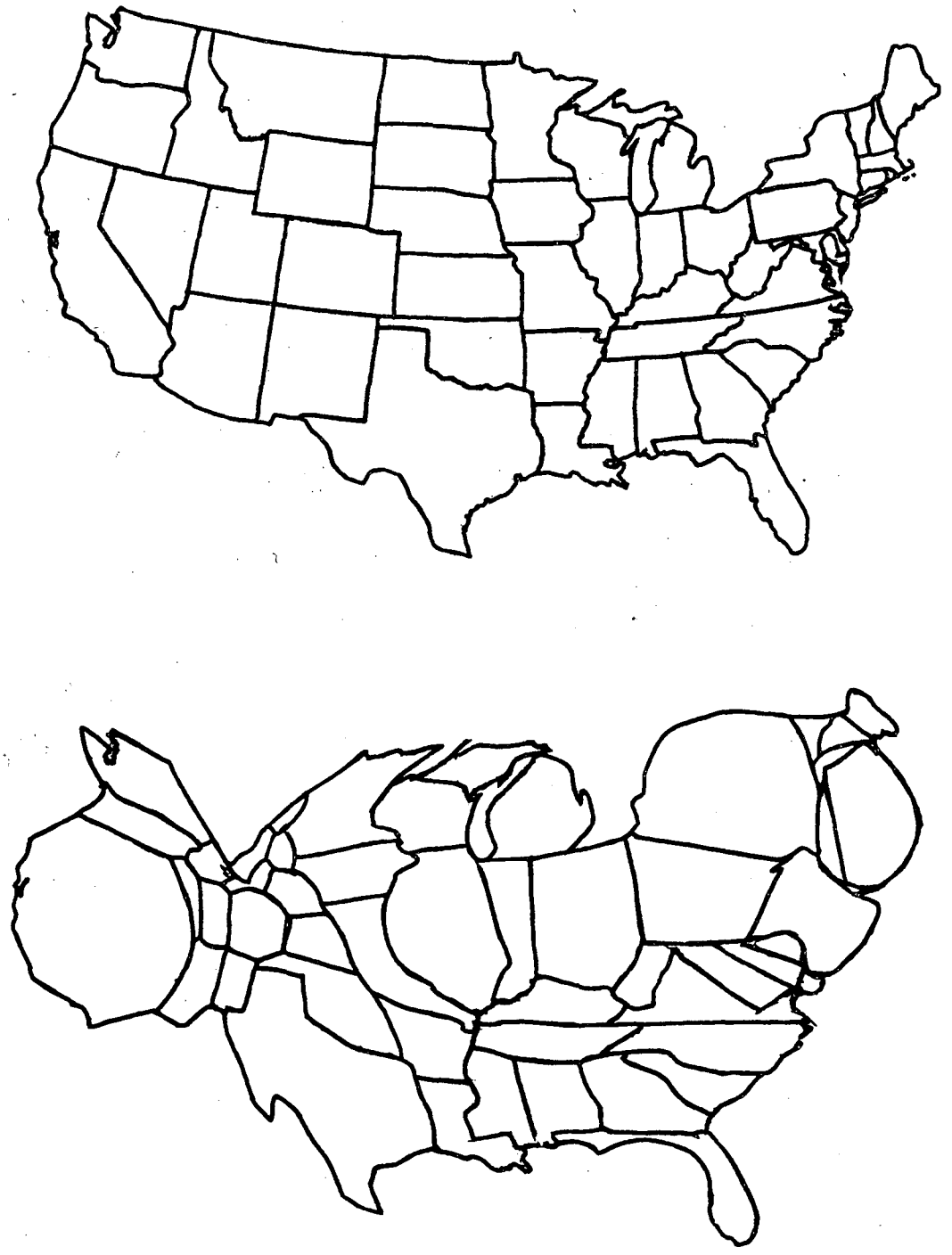


Fig. 3

XBL 848-8622

San Francisco
1980 Census Tracts

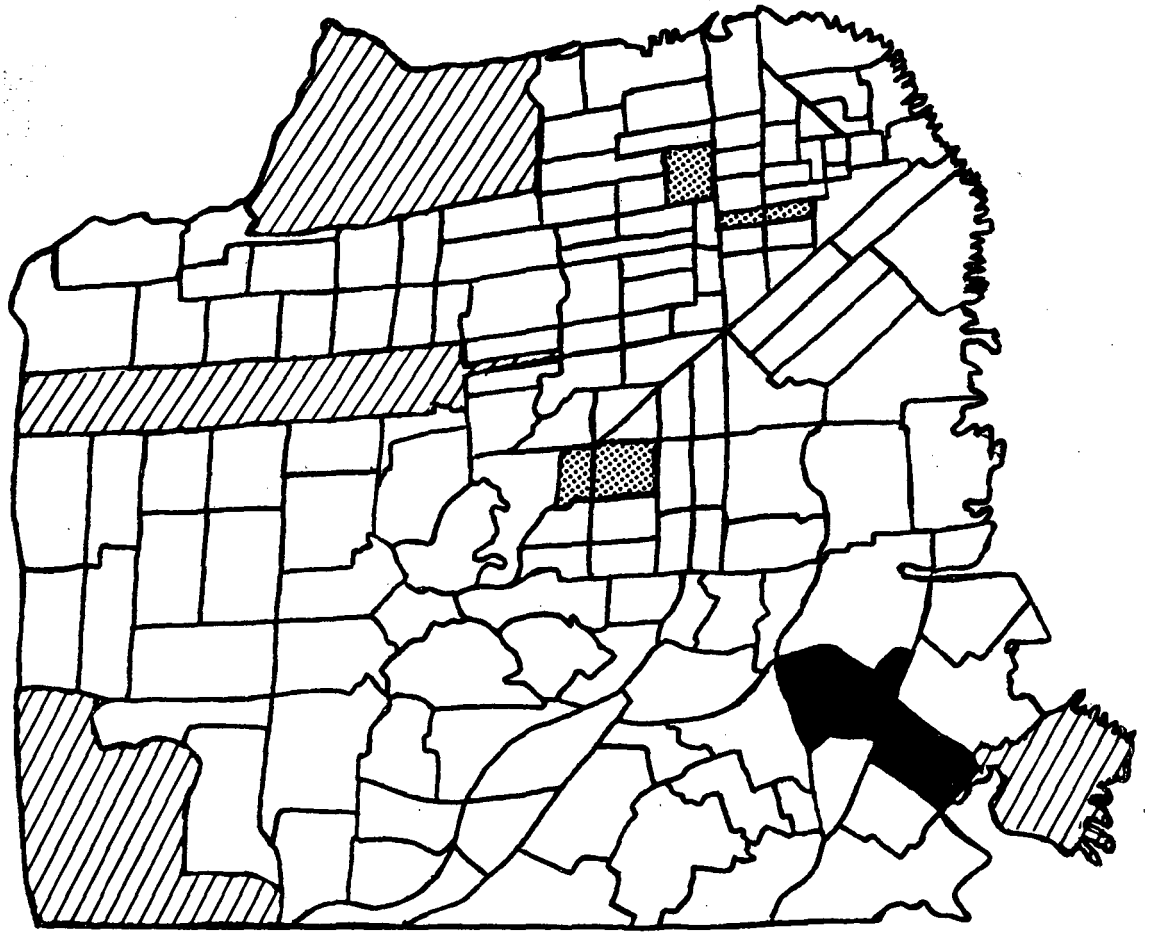


Fig. 4

XBL 848-8623

San Francisco
1980 White Male Population
ages 35 – 54

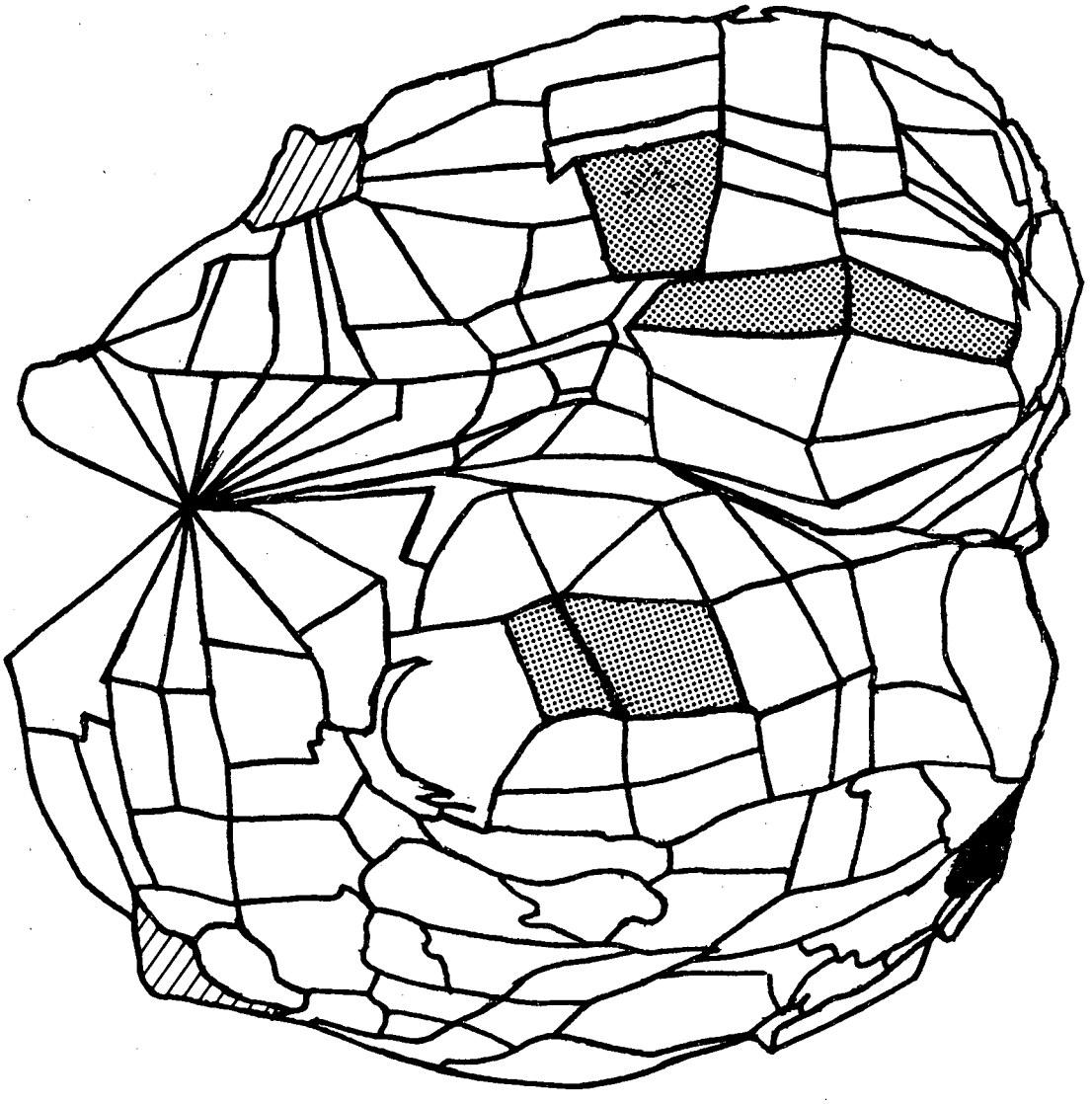


Fig. 5

XBL 848-8624

San Francisco
1980 Black Male Population
ages 35 – 54

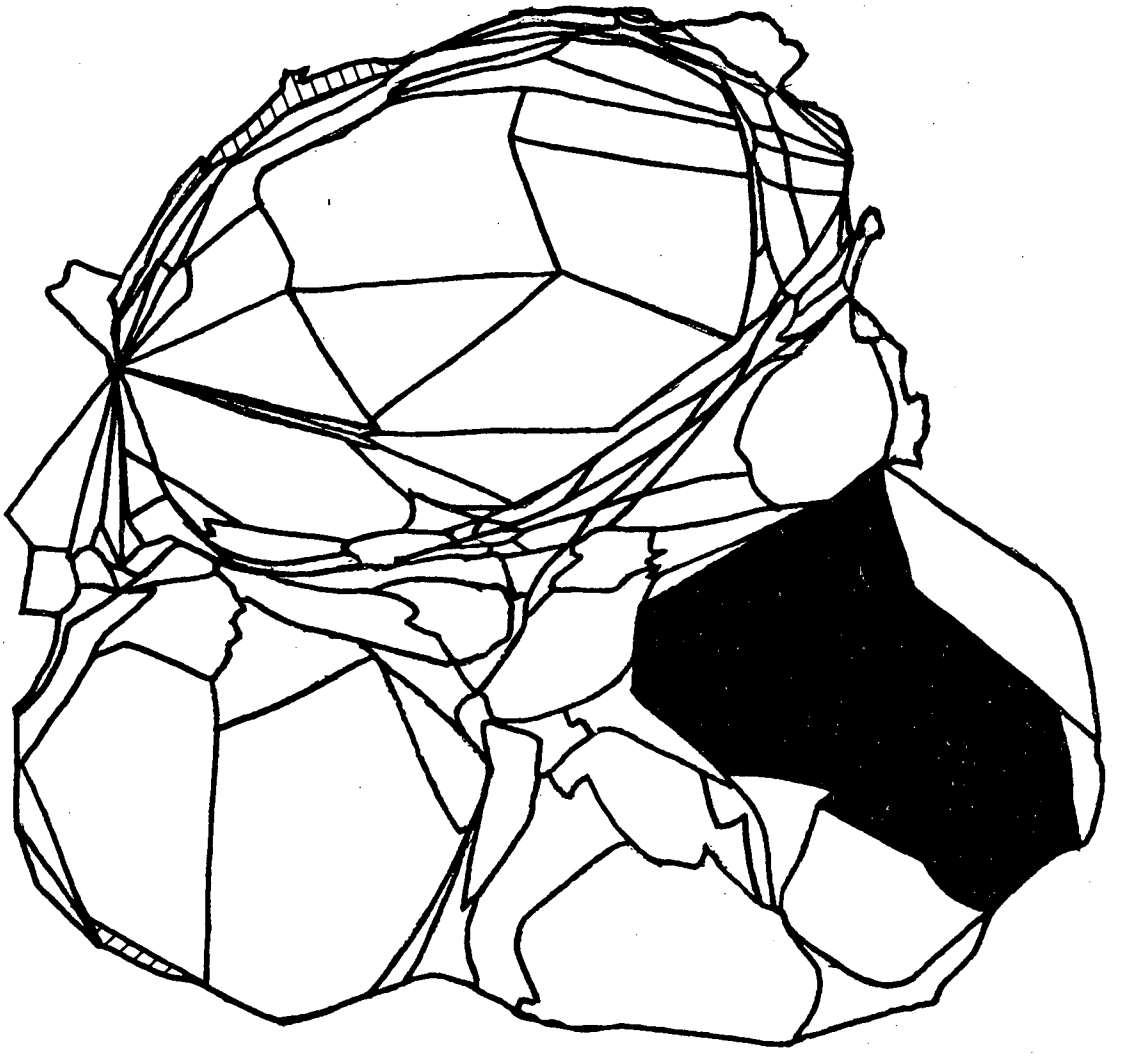


Fig. 6

XBL 848-8625

Hypothetical Cases (n = 13)
Non-transformed San Francisco map

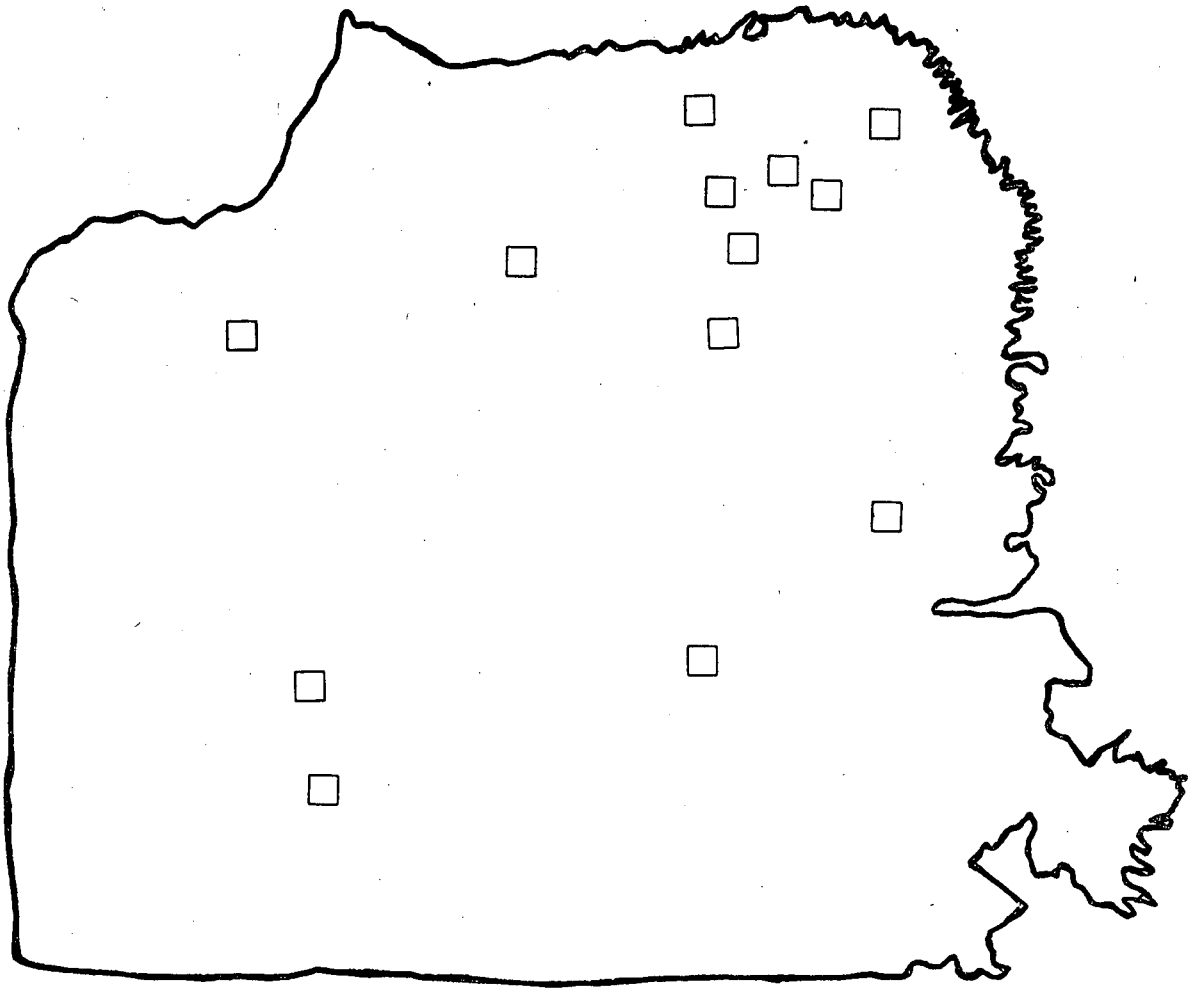


Fig. 7

XBL 848-8626

Hypothetical Cases (n = 13)
Transformed San Francisco map

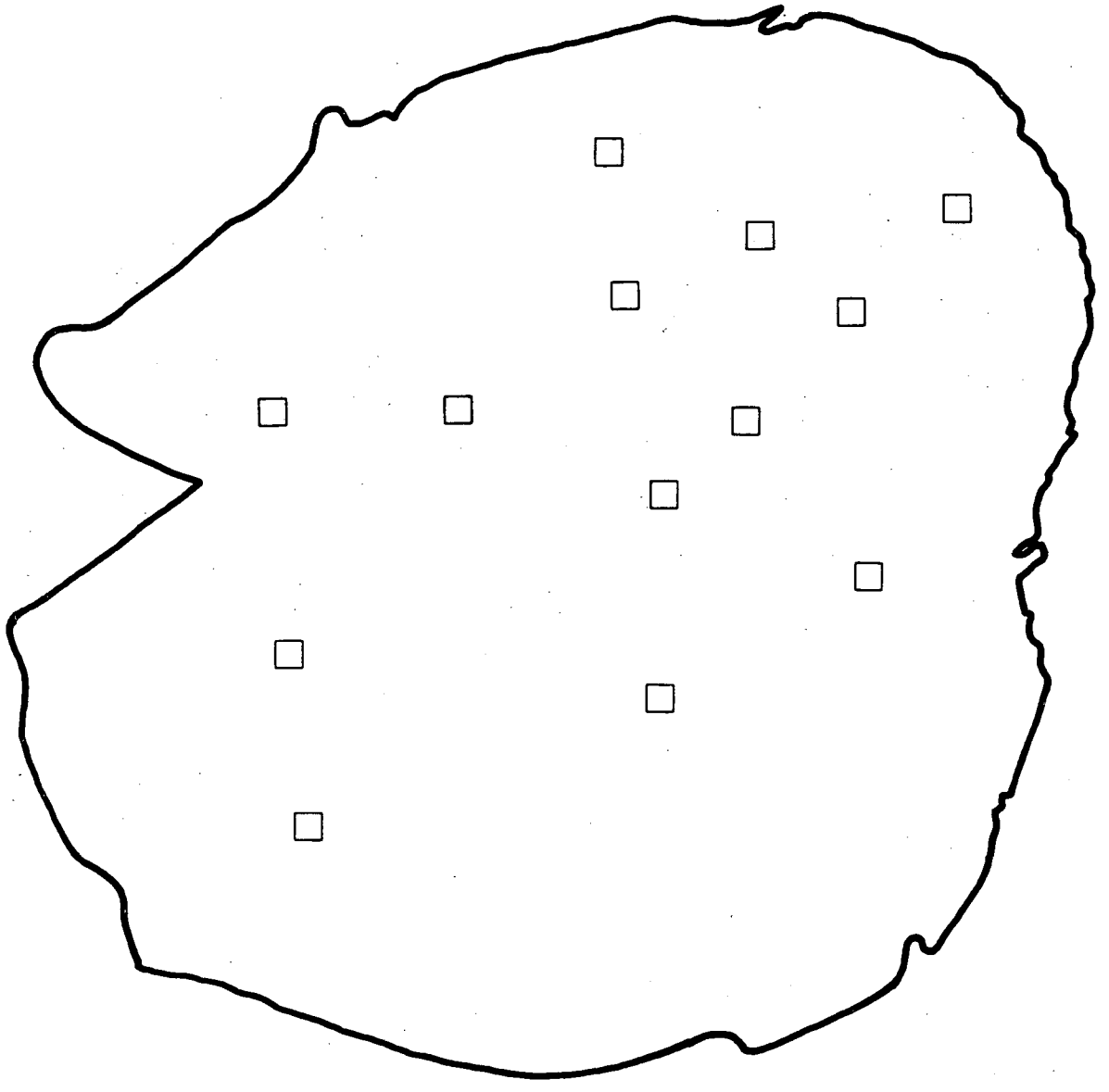


Fig.8

XBL 848-8627

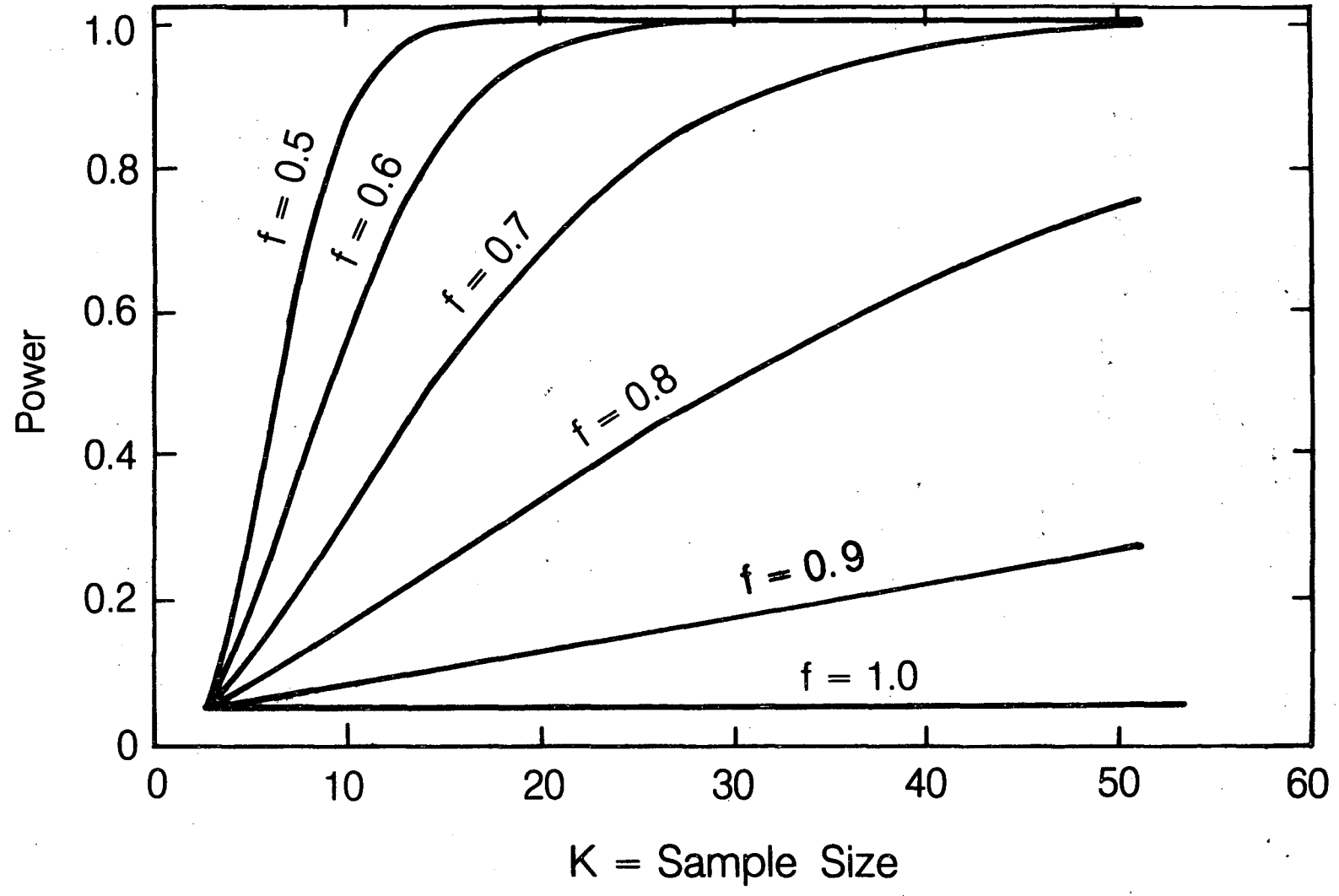
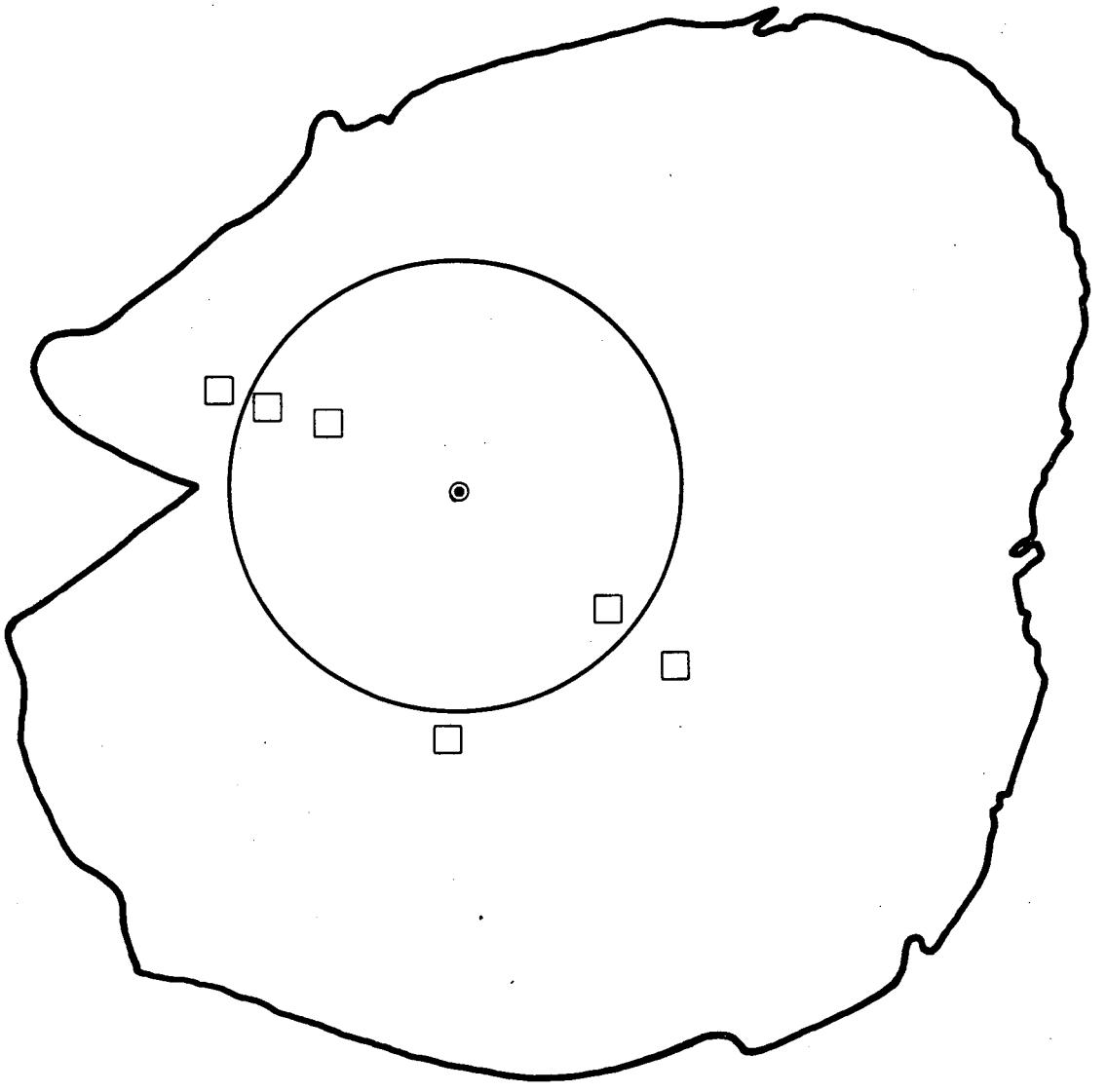


Fig. 9

XBL 848-8628

Stomach Cancer
White Females, ages 35 – 54
San Francisco, 1978 – 81

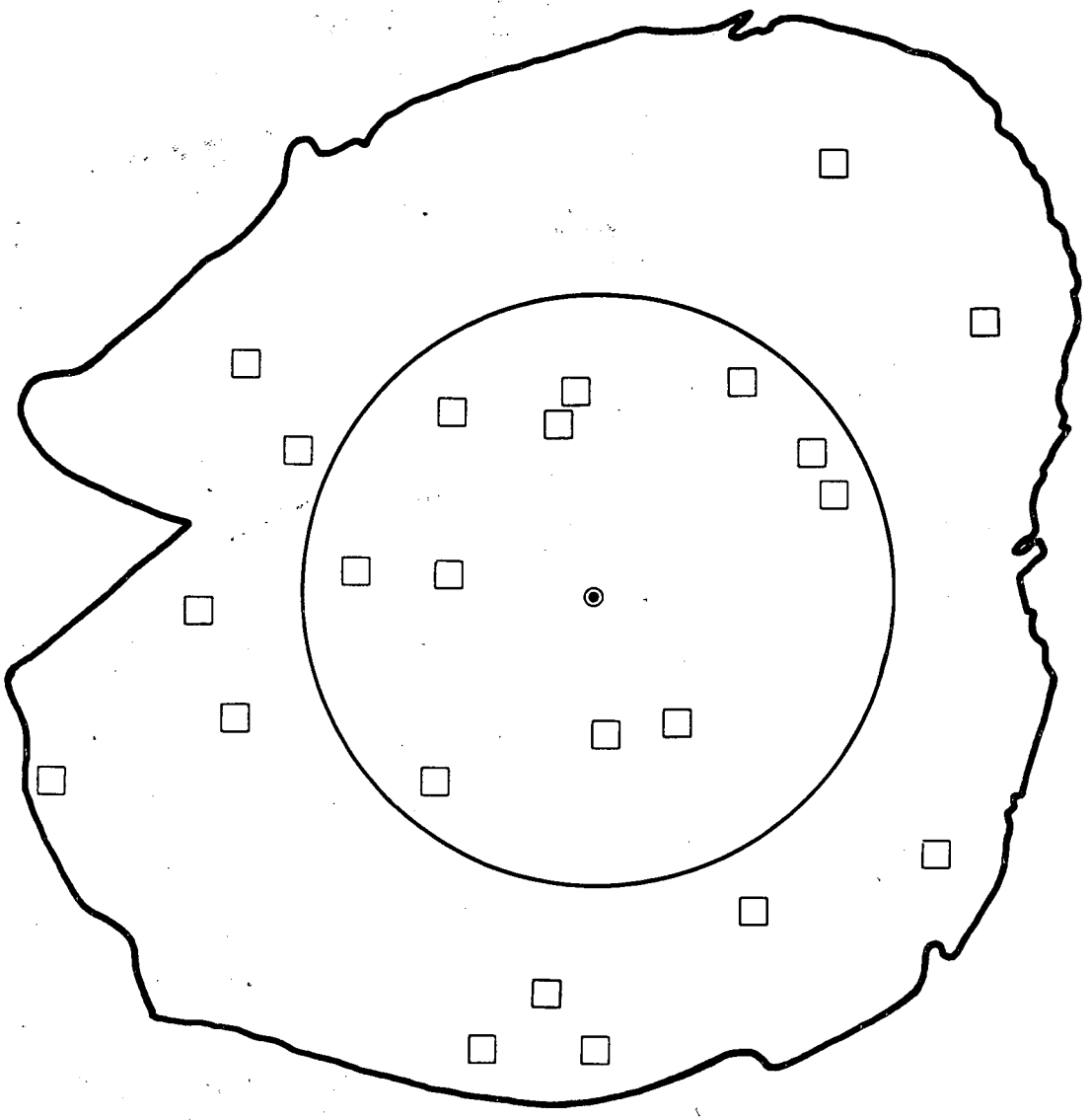


□ = case(s)
⊙ = centroid of cases

Fig. 10

XBL 848-8629

Colon Cancer
White Females, ages 35 – 54
San Francisco, 1978 – 81

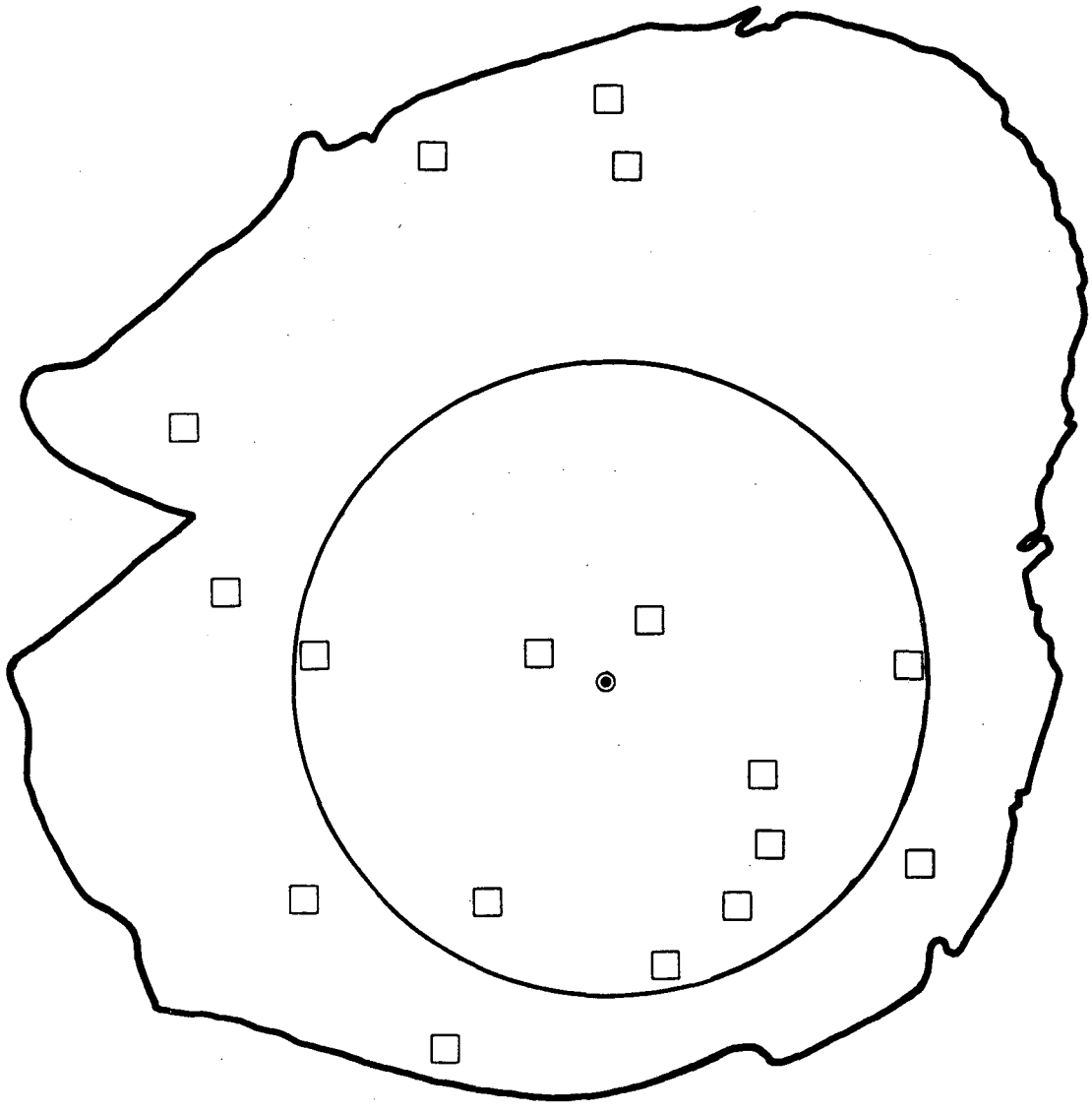


□ = case(s)
◉ = centroid of cases

Fig. 11

XBL 848-8630

Cancer of the Rectum
White Females, ages 35 – 54
San Francisco, 1978 – 81



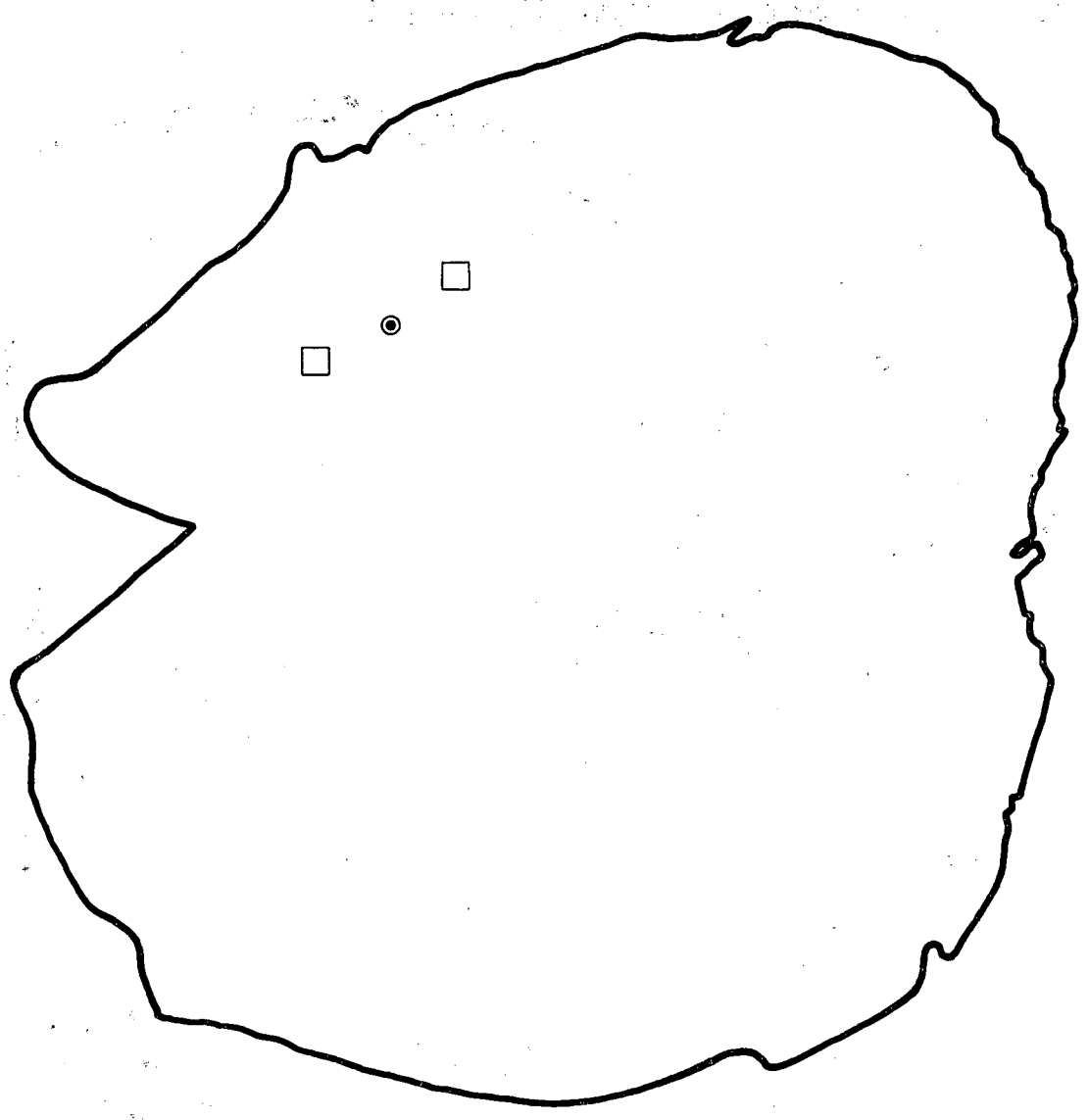
□ = case(s)

◉ = centroid of cases

Fig. 12

XBL 848-8631

Hodgkin's Disease
White Females, ages 35 – 54
San Francisco, 1978 – 81



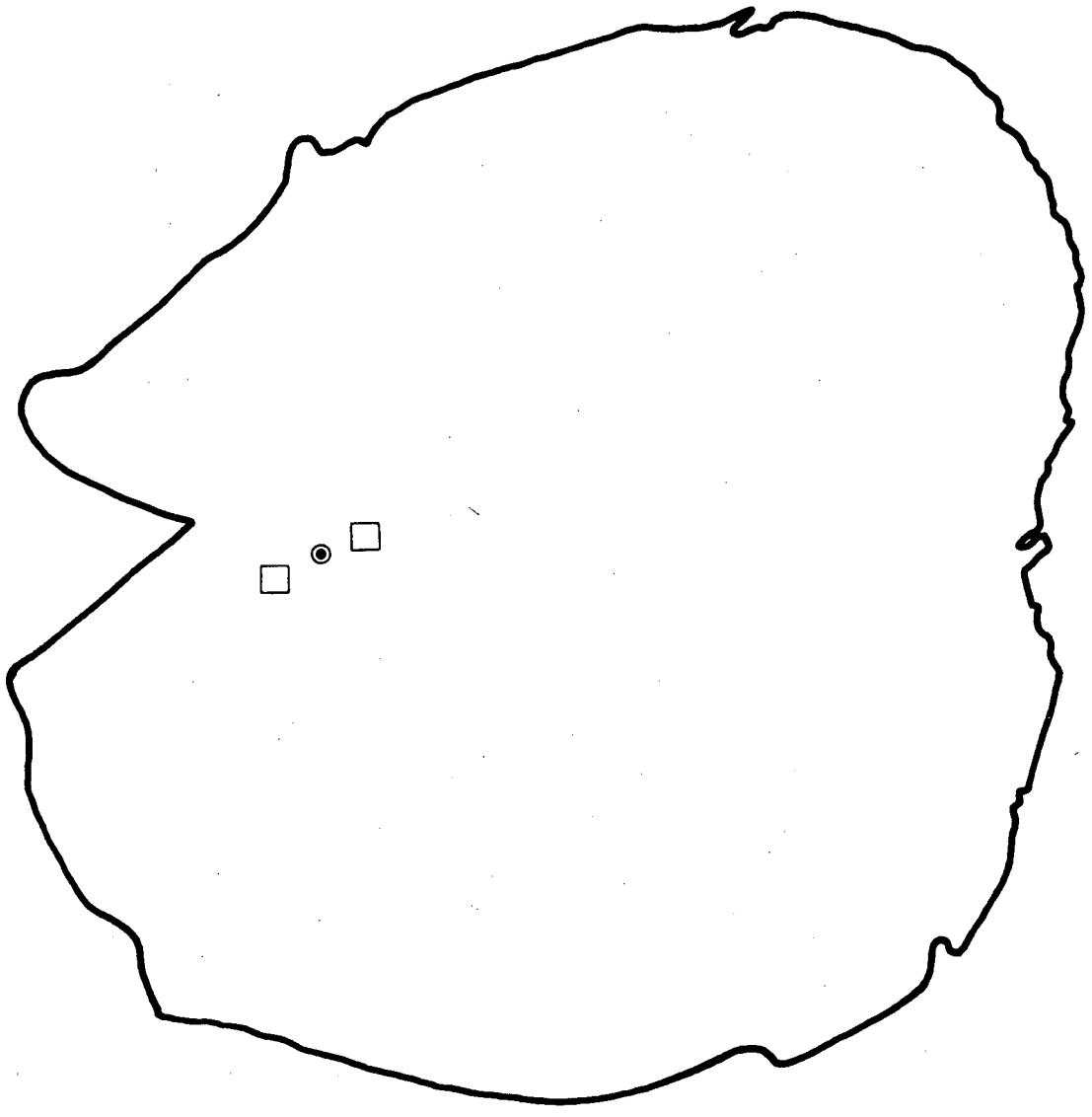
□ = case(s)

◉ = centroid of cases

Fig. 13

XBL 848-8632

Chronic Lymphocytic Leukemia
White Females, ages 35 – 54
San Francisco, 1978 – 81



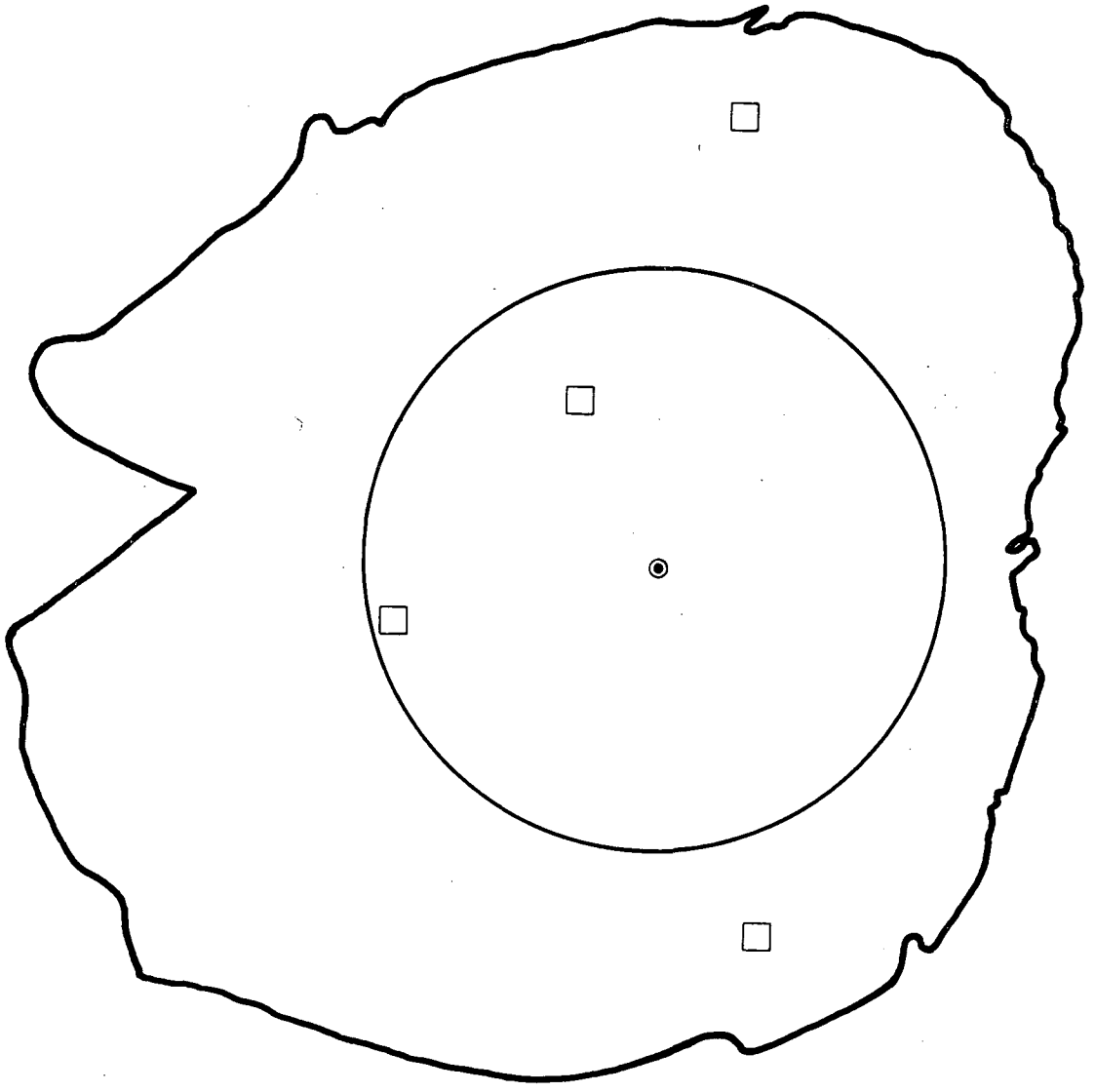
□ = case(s)

◉ = centroid of cases

Fig. 14

XBL 848-8633

Acute Granulocytic Leukemia
White Females, ages 35 – 54
San Francisco, 1978 – 81



□ = case(s)

◉ = centroid of cases

Fig. 15

XBL 848-8634

TECHNICAL INFORMATION DEPARTMENT
LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720