



Adult Leukemia: A Spatial Analysis

Steve Selvin; Deane W. Merrill

Epidemiology, Vol. 13, No. 2. (Mar., 2002), pp. 151-156.

Stable URL:

<http://links.jstor.org/sici?sici=1044-3983%28200203%2913%3A2%3C151%3AALASA%3E2.0.CO%3B2-4>

Epidemiology is currently published by Lippincott Williams & Wilkins.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/lww.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

The JSTOR Archive is a trusted digital repository providing for long-term preservation and access to leading academic journals and scholarly literature from around the world. The Archive is supported by libraries, scholarly societies, publishers, and foundations. It is an initiative of JSTOR, a not-for-profit organization with a mission to help the scholarly community take advantage of advances in technology. For more information regarding JSTOR, please contact support@jstor.org.

Adult Leukemia: A Spatial Analysis

Steve Selvin and Deane W. Merrill

Abstract: A simple and direct analysis of the spatial distribution of adult leukemia data results from a geopolitical map transformed to have a uniform density of the population at risk. Geographic displays and statistical assessments then lead to a series of informative descriptions of the observed spatial pat-

tern for six sex- and age-specific categories of non-Hispanic white individuals. Using a statistical/graphical approach, no consistent pattern of disease is observed; however, indications emerge of isolated nonrandom influences among young females (ages 10–29 years). (EPIDEMIOLOGY 2002;13:151–156)

Key words: density-equalizing transformation, spatial analysis, adult leukemia.

The study of the spatial distribution of leukemia has a long and rich history (reviewed by Alexander and Boyle¹). The vast majority of this research focuses on childhood leukemia. These studies fall into three broad categories: ad hoc observations that frequently generate post hoc hypotheses,² space-time analyses,³ and statistical analyses of routinely collected population-based data. Population studies of leukemia can be further separated into two general kinds: those generated from data surrounding putative point sources of exposure⁴ and those generated from surveillance data.⁵ The use of surveillance data is particularly plagued by two inherent problems: sparseness of cases and the spatial heterogeneity of human populations. Various methods have been proposed to analyze such data (reviewed by Moore and Carpenter⁶). The present study, based on surveillance data, differs from most previous studies in two respects. The analysis concerns adult rather than childhood leukemia, and the scale of measurement is based on human population distributions rather than geographic distances.

The spatial distribution of cases of disease is so dominated by the influences of the spatial distribution of the population at risk that a direct plot on a geopolitical map is rarely useful in the study of disease patterns. Human populations tend to concentrate in specific areas and, therefore, regardless of the other factors, human disease similarly tends to concentrate in these same areas. An approach to the display and statistical evaluation of a spatial disease pattern with the influence of a heteroge-

neous population at risk removed involves redrawing the geopolitical boundaries. These boundaries are transformed so that the density of the relevant population is equal over the entire study area; spatial patterns detected on such a density-equalized map then result from influences other than the distribution of the population at risk.⁷

This transformation is accomplished by a computer algorithm that expands densely populated subareas and contracts sparsely populated subareas, producing a map with equal density of the population at risk. For this analysis of adult leukemia, the 441 census tracts of two California counties (Contra Costa and Alameda Counties) serve as the subareas for such a transformation, producing six sex- and age-specific equal-density maps. A map redrawn to reflect variables other than geographic distance is called a cartogram. A population-based cartogram can be created with commercial software⁸ for small, simple maps; and a detailed description of a computer implementation⁹ of a specific density-equalizing algorithm¹⁰ is available for detailed maps divided into a large number of subareas (approximately 400 in the presented analysis). The algorithm used to produce the density-equalized map is a finite approximation to a theoretical cartogram¹⁰ providing a unique solution that “minimally distorts” the map.

A map transformed to have an equalized population density directly reflects disease risk and, at the same time, is easily analyzed and interpreted. The transformed map visually displays a population-free distribution of cases of disease; also, the newly created distribution has a simple statistical structure, namely a bivariate uniform distribution, when the probability of disease is constant among the members of the population at risk. Taking advantage of this simple statistical structure, “two-di-

Address correspondence to: Steve Selvin, School of Public Health, University of California, Berkeley, CA 94720; selvin@stat.berkeley.edu

Submitted March 28, 2001; final version accepted October 23, 2001.

Copyright © 2002 by Lippincott Williams & Wilkins, Inc.

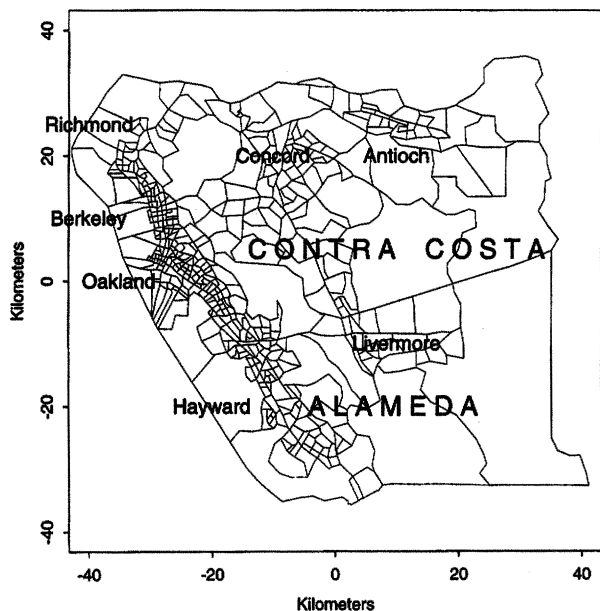


FIGURE 1. A geopolitical map of the U.S. census tracts of Contra Costa and Alameda Counties (San Francisco Bay area, CA); 1990.

mensional" quantile plots are used to assess statistically the sex- and age-specific spatial patterns of adult leukemia incidence. Furthermore, these density-equalized maps are used to pinpoint graphically the regions of highest risk for this disease on geopolitical maps.

Methods

Data

We abstracted incident cases of adult leukemia (*International Classification of Diseases, Adapted* diagnosis codes 204.0 to 208.9¹¹) for two San Francisco Bay Area counties, Contra Costa and Alameda, from data collected as part of the National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) program. These two counties, made up of 441 census tracts, include 1.75 million people who vary widely in demographic characteristics. We restricted abstracted cases of leukemia to non-Hispanic white individuals who were diagnosed during the years 1989, 1990, and 1991 and who reported ages and census tracts of residence. The counts of "non-Hispanic white" individuals result from an adjustment of the 1990 U.S. census data using the classification by race and the classification of Hispanic/non-Hispanic individuals. The sex- and age-specific counts of the populations at risk within the 441 census tracts, each containing roughly 4,000 individuals, also originated from the 1990 U.S. census enumeration.

The census tracts are displayed in Figure 1 for the two-county region. Incident cases of adult leukemia are analyzed by sex and age (10–29, 30–49, and 50–69

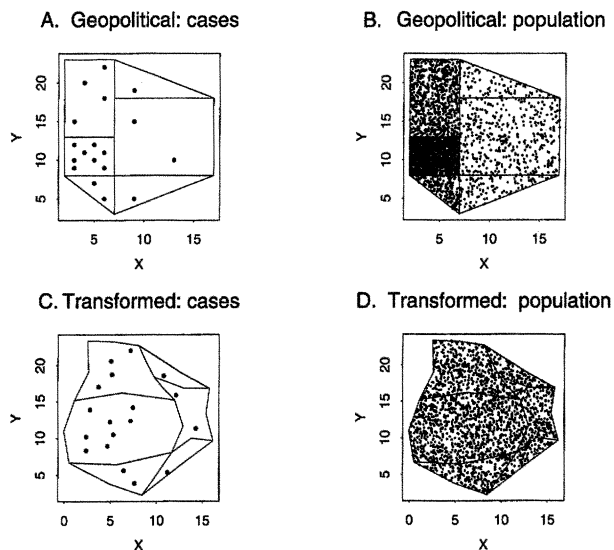


FIGURE 2. A comparison of the distributions of hypothetical cases of disease and populations at risk displayed on geopolitical and density-equalized maps.

years). For each of the six sex and age strata, the cases are located at a random point within the 1990 U.S. census tract of residence (untransformed), because more exact geographic detail (eg, addresses) is not available from the SEER public-use database. The lack of geographic location introduces a slight but unimportant bias.

We describe the analysis for females 10–29 years of age in detail as an illustration of the six stratum-specific analyses. For the other five age strata, we present only the statistical results and geopolitical maps.

Density-Equalized Transformation

To illustrate a density-equalizing map transformation, we constructed a simple hypothetical region consisting of six subareas. This hypothetical study area is constructed to have six equal-risk subareas, as described in Table 1. Figure 2A displays the 18 cases of "disease" on a geopolitical map. As expected, the majority of cases are found in the same areas as the majority of the population at risk (Figure 2B). Figure 2C displays the same "cases" on a density-equalized map, in which the subareas are transformed to have equal densities of populations (75.8 persons per square unit area). The displayed distribution of the "cases" appears random, because risk is the same for all six areas and the number of individuals at risk is now identical for any two areas of equal size (Figure 2D). Therefore, both visual and statistical spatial comparisons are possible over the entire density-equalized map, and the geographic subareas are no longer relevant; that is, comparisons are no longer influenced by the varying population at risk.

Quantile Plot

Again using a density-equalized map, the proportion of the total study area associated with each transformed disease data point (x_i, y_i) is

$$\hat{P}_i = \left[\frac{(EY - y_i) - B_j(EX - x_i)}{(EY - Y_j) - B_j(EX - X_j)} \right]^2$$

$i = 1, 2, 3, \dots, n = \text{number of points}$

where

$$B_j = \frac{Y_j - Y_{j+1}}{X_j - X_{j+1}}$$

Note that uppercase letters denote the coordinate points describing the boundary of the study area in terms of an m -sided arbitrary polygon ($j = 1, 2, \dots, m$). The boundary point (X_j, Y_j) is the vertex of the triangle containing the observed data point (x_i, y_i) formed by connecting the centroid (EX, EY) to the two consecutive points on the study area polygon boundaries, namely (X_j, Y_j) and (X_{j+1}, Y_{j+1}) . That is, the triangle containing (x_i, y_i) is described by the three points (X_j, Y_j) , (EX, EY) , and (X_{j+1}, Y_{j+1}) . A formal definition of \hat{P}_i is given in the Appendix.

The ordered proportions $\hat{P}_1, \hat{P}_2, \hat{P}_3, \dots, \hat{P}_n$ associated with the n transformed disease locations $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$ are the basis of a typical quantile plot. The \hat{P}_i -values form an estimate of a cumulative distribution function, which is an estimate of the uniform cumulative distribution function when no spatial pattern exists. A plot of the estimated values \hat{P}_i ordered from low to high against the values $P_i = i/n$ randomly deviates from a straight line (intercept = 0 and slope = 1) when no spatial pattern of disease exists. The n differences between the estimated and the expected values (ie, $|\hat{P}_i - P_i|$) can be statistically evaluated with the classic Kolmogorov test. Both the graphic display and a rigorous statistical test provide a useful assessment of the observed spatial pattern of disease, addressing the question of whether or not the distribution is random. In addition, a similar strategy, based on estimating two cumulative distributions, could be used to explore differences between the spatial patterns of disease observed in different groups (Kolmogorov-Smirnov test).

Areas of High Incidence

For disease locations distributed over a defined area, various smoothing techniques exist to estimate a two-dimensional incidence density.¹² Additionally, these techniques provide estimates of the contours of the estimated density, specifically locating areas of high incidence. When contours are estimated from cases of disease plotted on a density-equalized transformed map, they pinpoint the areas of high risk independent of the

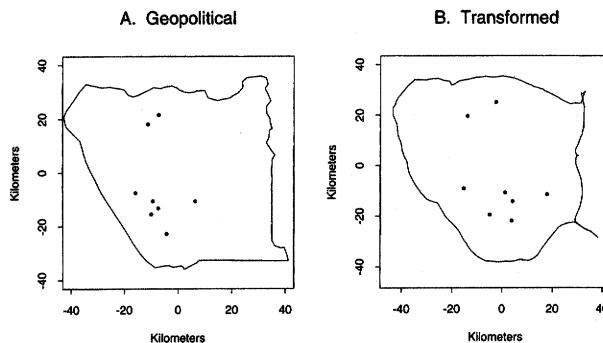


FIGURE 3. The geopolitical and density-equalized maps of the eight cases of adult leukemia among non-Hispanic white females of ages 10–29.

distribution of the population at risk (a continuous “two-dimensional rate”). Regions located on a density-equalized map can then be directly translated (inverse-transformation) to the corresponding geopolitical regions and displayed in terms of natural distances on the original map.

Results

The locations of the eight incident cases of leukemia observed among young females (10–29 years) are plotted on a geopolitical map of the two-county area (Figure 3A). Figure 3B displays transformed locations of the eight leukemia cases on an equal-density map in which each of the 441 transformed census tracts has the same density of individuals at risk for this analytic subgroup. The boundaries of these census tracts are not shown because they are not relevant on a density-equalized map.

The quantile plot associated with the eight cases of leukemia is displayed in Figure 4A constructed from the values of the cumulative distribution functions (observed and expected values) given in Table 2. The Kolmogorov test based on the maximum distance ob-

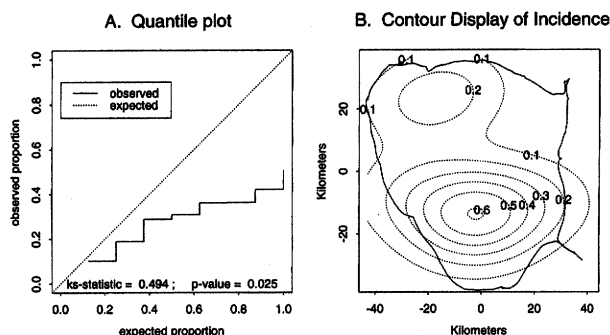


FIGURE 4. The quantile plot and the estimated contours of the incidence density based on the eight cases of adult leukemia among non-Hispanic white females of ages 10–29.

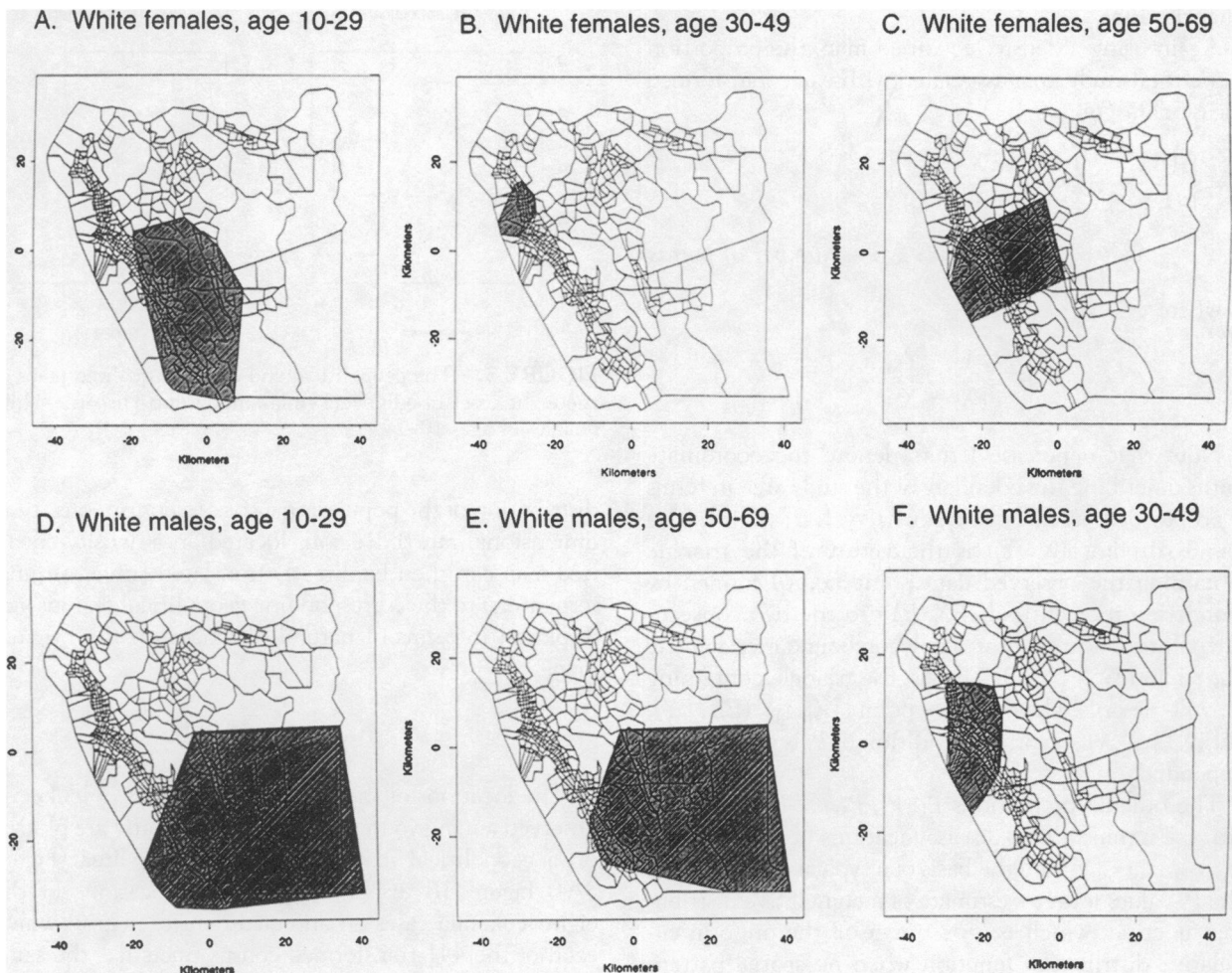


FIGURE 5. Six sex- and age-specific geopolitical maps of adult leukemia in Contra Costa and Alameda Counties, CA (shaded areas indicate the areas of highest risk).

TABLE 1. Description of the Hypothetical Study Area by Subarea

Subarea	Area (Per Square Unit)	Population	Density (Per Square Unit)	Cases	Rate/1,000
1	50.0	4,000	80	4	1.0
2	25.0	8,000	320	8	1.0
3	12.5	2,000	160	2	1.0
4	25.0	1,000	40	1	1.0
5	100.0	2,000	20	2	1.0
6	25.0	1,000	40	1	1.0
Total	237.5	18,000	75.8	18	1.0

served between the step function (data) and the values expected (theoretical uniform distribution) produces a *P*-value of 0.025. That is, the maximum deviation of 0.494 would be observed by chance alone with probability 0.025 when no spatial pattern exists.

Using standard smoothing techniques,¹³ contours of the disease-incidence density on a density-equalized map are estimated, as displayed in Figure 4B. The area of "highest risk" is somewhat arbitrarily defined as the

lowest-level contour producing a single region of high incidence of disease on the density-equalized map. In less technical terms, this selection amounts to locating the highest single "peak" of disease incidence occurring within the studied region. For example, the contour labeled 0.3 (Figure 4B) indicates the highest-risk area for this population subgroup. As a last step, this highest-risk area is untransformed and plotted on a geopolitical map (Figure 5A). Because any census tract identified on the

TABLE 2. Cumulative Distribution Functions \hat{P}_i and P_i Associated with the Eight Cases of Adult Leukemia on Hispanic Females Ages 10–29 Years.

	\hat{P}_i	P_i	$P_i - \hat{P}_i$
1	0.103	0.125	0.022
2	0.189	0.250	0.061
3	0.289	0.375	0.086
4	0.309	0.500	0.191
5	0.362	0.625	0.263
6	0.363	0.750	0.387
7	0.421	0.875	0.454
8	0.506	1.000	0.494

transformed map corresponds to the same census tract on the geopolitical map, the shaded area (Figure 5A) on the geopolitical map indicates the area of highest risk for adult leukemia in young females independent of the distribution of the population at risk.

This back-transformation can be further refined to transform more exactly the identified areas, a process that is mathematically complex and requires a sophisticated computer algorithm. There are a number of approximate approaches, however, leading to locating areas on a geopolitical map once they are identified on a transformed map.

The six sex- and age-specific spatial distributions of adult leukemia display no consistent pattern (Figure 5 and Table 3). Three of the six maps (males ages 10–29 and 30–49, and females ages 10–29) show a tendency for cases to occur in the southern portion of the two-county study area. Based on the Kolmogorov test, the pattern of incidence among females ages 10–29 is the only significantly nonrandom ($P = 0.025$) pattern (Table 3), but this observation is based on an incidence density estimated from only eight cases of leukemia. Such an estimate is extremely sensitive to a variety of biases and is particularly influenced by extreme values.

Discussion

Maps of disease risk such as in Figure 5 dramatically display the pattern of disease but contain essentially the same information found in a set of rates for a series of defined geographic areas. Much has been written recently about the use of maps and geographic information systems in epidemiology,¹⁴ some of it critical. The

primary purpose of these leukemia maps is to provide a graphic picture of the distribution of disease incidence. As with tables of rates, such maps of disease risk occasionally contain useful information on associations with other risk factors although with no direct evidence of causality. Additional information on how such analyses can be performed and applied to public health practice is presented in our earlier paper.⁷

In general, a statistical approach that reduces a two-dimensional distribution to a one-dimensional summary incurs a loss of information. Consequently, certain spatial configurations of cases are not easily detected with specific spatial summaries (low statistical power). It is, however, likely that many such patterns would be noticed by inspection or identified by other statistical techniques. Not surprisingly, the degree of statistical power associated with any summary of geographic disease data depends largely on the postulated spatial pattern underlying the observed disease.

A density-equalized transformed map produces a graphic display of the distribution of disease incidence, free of the confounding influences of a heterogeneous distribution of the population at risk. Although this transformation is not the only way to estimate an incidence density of a specific disease, it allows simple and direct assessments of the impact of random variation on the observed spatial pattern using elementary statistical techniques. Furthermore, the process remains useful and rigorous when relatively few observations are available.

Appendix

Definition of \hat{P}_i

A concentric polygon is naturally defined as a sub-polygon contained entirely within a polygon with the same centroid and all sides parallel to the sides of the original polygon. The quantity denoted \hat{P}_i is the proportion of the total study area contained within the boundary of a subpolygon that intersects the point (x_i, y_i) and is concentric with respect to the entire study area. Therefore, for the k th ordered concentric polygon (low to high), the quantity k/n is the proportion of points expected to be within its boundaries when the n points (x_i, y_i) are uniformly distributed over the entire polygon.

TABLE 3. The Analytic Results of Sex- and Age-Specific Cases of Leukemia (ICDA Diagnoses 204.0–208.9).

Sex/Age	N	k Statistic*	P-Value
White females, ages 10–29	8	0.494	0.025
White females, ages 30–49	26	0.171	0.386
White females, ages 50–69	71	0.071	0.868
White males, ages 10–29	21	0.148	0.691
White males, ages 30–49	50	0.105	0.603
White males, ages 50–69	114	0.056	0.867

ICDA = International Classification of Diseases, Adapted.

* Kolmogorov statistic (maximum deviation).

An exact concentric polygon can only be constructed by the process described in the text when the angles between successive boundary points [i e, (X_j, Y_j) and (X_{j+1}, Y_{j+1})] relative to the centroid of the study area polygon (EX, EY) are a monotonic function of j . The outline of most transformed geopolitical maps can be smoothed (detail removed) so it is at least approximately monotonic, typically leading to a useful but slightly biased Kolmogorov test.

References

1. Alexander FE, Boyle P, eds. *Methods for Investigating Localized Clustering of Disease*. IARC Scientific Pub. No. 135. Lyon, France: International Agency for Research on Cancer, 1996.
2. Heath CW, Manning MD. Leukemia outbreak? *Lancet* 1964;1:1394.
3. Knox EG. Epidemiology of childhood leukemia in Northumberland and Durham. *Br J Prev Med* 1964;18:17-24.
4. Gardner MJ. Review of reported increases of childhood cancer rates in vicinity of nuclear installations in the UK. *J R Stat Soc Ser A* 1989;152:307-335.
5. Chen R, Iscovich J, Goldbourt U. Clustering of leukemia cases in a city in Israel. *Stat Med* 1996;16:1873-1887.
6. Moore DA, Carpenter TM. Spatial analytic methods and geographic information systems: use in health research and epidemiology. *Epidemiol Rev* 1999;12:143-161.
7. Selvin S, Merrill DW, Erdmann C, White M, Ragland K. Breast Cancer Detection: maps of two San Francisco Bay Area counties. *Am J Public Health* 1998;88:1186-1191.
8. Jackal CB. Using ArcView to create continuous and non-continuous area cartograms. *Cartogr Geogr Inform Syst* 1997;24:101-109.
9. Close ER, Merrill DW, Holmes H. Implementation of a new algorithm for density-equalizing map projections. Report LBL-35838, Lawrence Berkeley Laboratory, July 1995.
10. Gusein-Zade SM, Tikunov VS. A new technique for construction continuous cartograms. *Cartogr Geogr Inform Syst* 1993;20:167-173.
11. National Cancer Institute. SEER Public-Use CD-ROM, 1973-1994, release May 1997. Bethesda, MD: National Cancer Institute, 1997.
12. Bowman AW, Azzalini A. *Applied Smoothing Techniques for Data Analysis*. Oxford: Clarendon Press, 1997.
13. S-PLUS: 1996 MathSoft, Inc [computer program]. Version 3.4 Release 1 for Sun SPARC. SunOS 5.3; 1996.
14. Mayer JD. The role of spatial analysis and geographic data in the detection of disease causation. *Soc Sci Med* 1983;17:1213-1221.