LBL-36630 Rev.
UC-405

Use of Density Equalizing Map Projections (DEMP) in the Analysis of Childhood Cancer in Four California Counties

D.W. Merrill,[a*] S. Selvin,[b] E.R. Close,[a] and H.H. Holmes[a]

[a]Information and Computing Sciences Division
Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

[b]School of Public Health
140 Warren Hall
University of California
Berkeley, California 94270

April 1995

*Address correspondence to: D.W. Merrill, MS 50B-2239, 510-486-5063, dwmerrill@lbl.gov

This work was supported by the Director, Office of Epidemiologic Studies, Office of Health, Office of Environment, Safety and Health, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

DISCLAIMER

USE OF DENSITY EQUALIZING MAP PROJECTIONS (DEMP) IN THE ANALYSIS OF CHILDHOOD CANCER IN FOUR CALIFORNIA COUNTIES

SUMMARY

In studying geographic disease distributions, one normally compares rates of arbitrarily defined geographic subareas (e.g. census tracts), thereby sacrificing the geographic detail of the original data. The sparser the data, the larger the subareas must be in order to calculate stable rates. This dilemma is avoided with the technique of Density Equalizing Map Projections (DEMP). Boundaries of geographic subregions are adjusted to equalize population density over the entire study area. Case locations plotted on the transformed map should have a uniform distribution if the underlying disease rates are constant. The density equalized map portrays both individual cases and rates, and can be understood by untrained observers. Simple statistical techniques can be used to test the uniformity of the transformed map.

The present report describes the application of the DEMP technique to a sizeable "real-world" data set: 401 childhood cancer cases occurring between 1980 and 1988 in four California counties. In an earlier analysis of the same data, the California Department of Health Services (DHS) calculated rates for 101 communities and found no significant geographic variability.

The DHS 1980-88 population estimates are no longer available, so in the present analysis 1980 Census data were used; the geographic units were 262 census tracts. A *k*'th nearest neighbor analysis, corrected for boundary effects and for within-tract variability, provides strong evidence for geographic non-uniformity in tract rates ($p < 10^{-4}$). No such effect is observed for artificial cases generated under the assumption of constant rates. Pending re-analysis with 1980-88 population estimates, no epidemiologic conclusions can be drawn at this time.

Now that the validity of the DEMP technique has been demonstrated, it can become a valuable tool for routine surveillance activities, especially if coupled to data bases containing the necessary population data and map files for all regions of the United States.

INTRODUCTION

Density equalizing map projections (DEMP), also known as cartograms or anamorphoses, have long been used for display of thematic data. The DEMP technique is appropriate for analyzing disease distributions because on a density equalized map, population density is constant. Therefore cases should occur randomly and uniformly under the null hypothesis of equal risk.

Normally, one analyzes geographic disease distributions by comparing rates of different subareas. Relative to conventional methods, the DEMP technique has several advantages:

- A density equalized map portrays both individual events and rates, and can be understood by untrained observers. The full geographic detail of the data is preserved, and arbitrary grouping of subareas is avoided.

- The DEMP technique avoids the calculation of unstable rates. The technique is appropriate, and even works best, when the number of cases is small.

- No *a priori* knowledge is required for testing the null hypothesis of equal risk. Hence the DEMP technique is appropriate for automatic analysis of routinely collected surveillance data. The uniformity of the density equalized map can be tested with simple statistical techniques.

- Testing a model other than the null hypothesis can be performed by equalizing the map according to expected cases, rather than population at risk. The same method can be used to incorporate geographic variation of age, race, and other risk factors.

FOUR-COUNTY CHILDHOOD CANCER STUDY

The present study analyzes 401 childhood cancer cases occurring between 1980 and 1988, in children of ages 0 through 14, in four California counties: Fresno, Kings, Kern and Tulare. The data were originally collected and analyzed by the California State Department of Health Services (DHS) to investigate a reported childhood cancer cluster in McFarland, California.[2,3] DHS subdivided the four counties into 101 communities. Six communities had rates that fell outside 95% confidence limits (three with more cases than expected and three with fewer cases than expected). The result is consistent with the null hypothesis of uniform underlying rates. One community (McFarland) had an elevated rate outside the 99% confidence limit, exactly what would have been expected from chance alone.

DATA PREPARATION

DHS kindly provided to LBL the case records from the previous analysis.[3] LBL analyzed all 401 cases as a single data set, without regard to year of diagnosis, age category, gender or race. A correct analysis should have used the 1980-88 population at risk for children 0-14. Unfortunately, these data were lost at DHS and are no longer available. In the present analysis LBL used the 1980 Census population for children of ages 0-17, obtained from LBL SEEDIS (Socio-Economic Environmental Demographic Information System).[4] 1980 Census tract boundary files, originally purchased from National Planning Data Corporation, were also obtained from SEEDIS.

DENSITY EQUALIZATION

Development of a practical DEMP algorithm has been a goal of the LBL PAREP project since 1985. Computer algorithms developed elsewhere are mostly suitable for graphic display purposes, but not for the requirements of statistical analysis. Earlier LBL algorithms addressed this problem but were too slow for practical use. In 1993 Gusein-Zade and Tikunov described a new algorithm[5] which was independently implemented and tested at LBL.[6] New features were added in January 1995.[1]

Unnecessary geographic detail was removed and the 262 tracts were subdivided into 1212 triangles. The resulting tract boundaries (solid), triangle boundaries (dotted), and 401 case locations are shown in Figure 1. Target populations were assigned to each of the 1212 triangles, assuming population density within each tract to be uniform. The triangles were converted to hexagons by subdividing each triangle face; then the 1212 hexagons were density equalized in ten equal steps (*i.e.*, with step sizes equal to 1/10, 1/9, ... 1/2, 1/1).

Figure 2 shows the tract boundaries (solid), the hexagon boundaries (dotted), and the locations of the 401 cases after density equalization. The same transformation was applied to artificial cases which were generated assuming equal risk. The artificial cases are uniformly distributed on the density equalized map, proving that the DEMP transformation does not introduce artificial clustering.[1] The same transformation was also applied to 401 points plotted at random in the same tracts, respectively, as the real cases. This was necessary for statistical analysis: geographic detail below the tract level must be removed from the case data, to correspond to the tract level population and map data. To equalize population density within tracts would require population data and map files with greater geographic detail.

Figure 1. Filtered and triangulated map, with 401 cases.

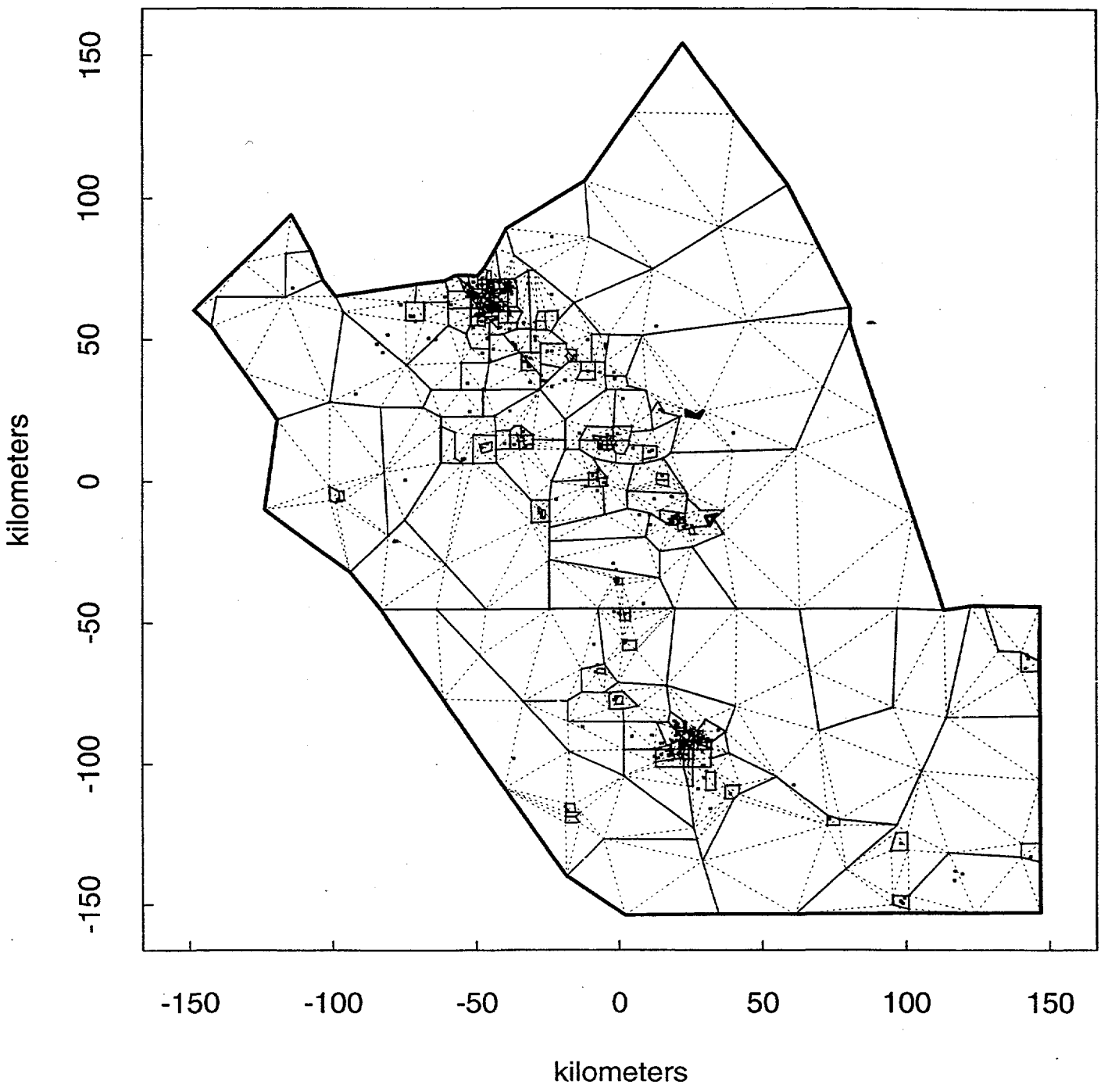
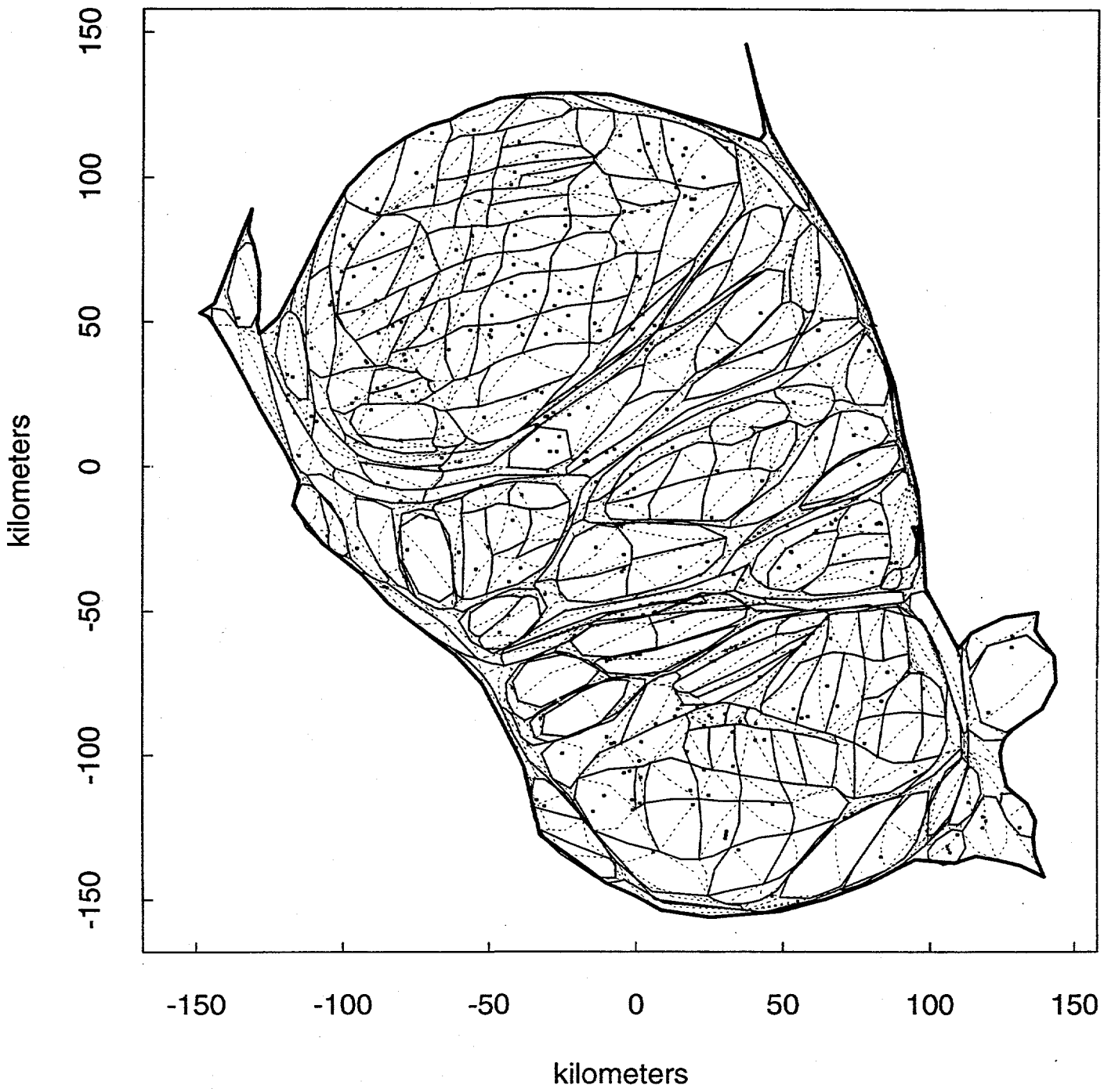


Figure 2. Tracts, hexagons and 401 cases, density equalized.



VISUAL ANALYSIS

The density equalized map, with internal boundaries removed, is shown in Figure 3. Each case is randomly plotted at a random location in its own tract. Figure 3 portrays visually not only the individual cases but also the variation of rates. With the aid of the legend box one can roughly assess the statistical significance of dense or sparse regions. For example, just below the center of the map, seven or eight cases occur in a small region having one-fifth the area of the legend box; *i.e.* where only one case is expected. To the northeast, only two cases occur in a broad region where about ten are expected.

The sensitivity of this visual technique (but not its statistical significance) can be enhanced by plotting each case at more than one random location within its tract. In Figure 4 each case is plotted 20 times; the number of observed cases in any region is 1/20 the number of points. Expected cases are estimated as in Figure 3. In Figure 4, a larger number of dense and sparse regions are evident, including some snakelike regions corresponding to sparsely populated rural tracts. The prominent empty regions are areas where no cases were observed; most of these are not significant because their areas are small, corresponding to only a few cases expected.

STATISTICAL ANALYSIS

To test the hypothesis of spatial randomness, simple statistical techniques can be applied to the density equalized map. We illustrate with a *k*th nearest neighbor analysis, which is a generalization of a nearest neighbor analysis presented earlier.[1]

k'th Nearest Neighbor Analysis

On the density equalized map, each of the n (= 401) cases has a nearest neighbor distance which is the distance to the nearest of all the n-1 other cases. To test for spatial randomness among the n cases, we calculate the mean nearest neighbor distance d_1, which is the average of these n distances. In the k'th nearest neighbor analysis we consider not only the distance d_1 to the nearest neighbor (k=1), but also the distance d_2 to the next nearest (k=2), and d_k for all values of k up to n-1. The different values of k are not independent tests, but different choices of a variable parameter. The analysis is sensitive to either small-area or large-area effects, depending on the value of k.

Figure 3. Density equalized map, with one point per case.

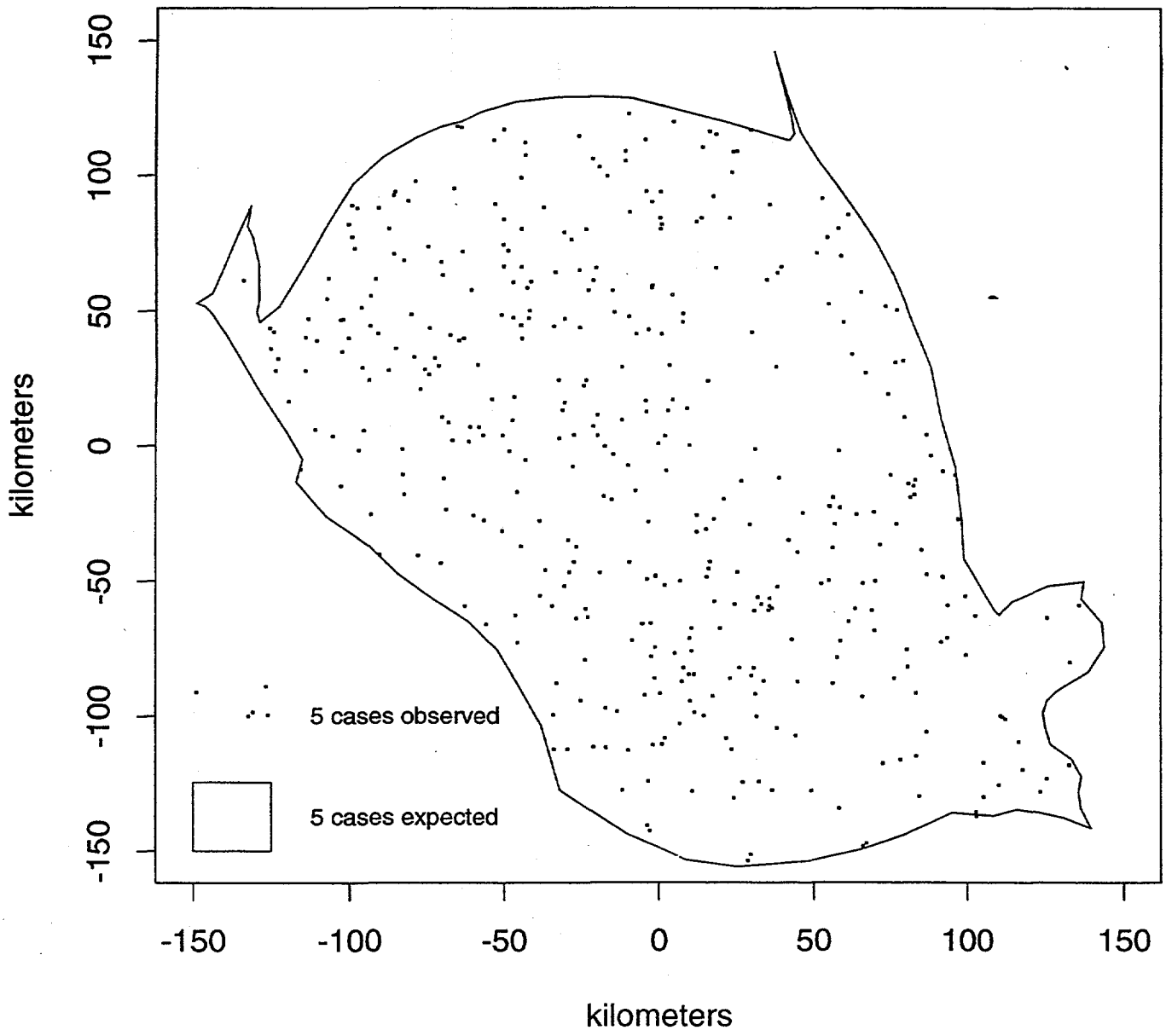
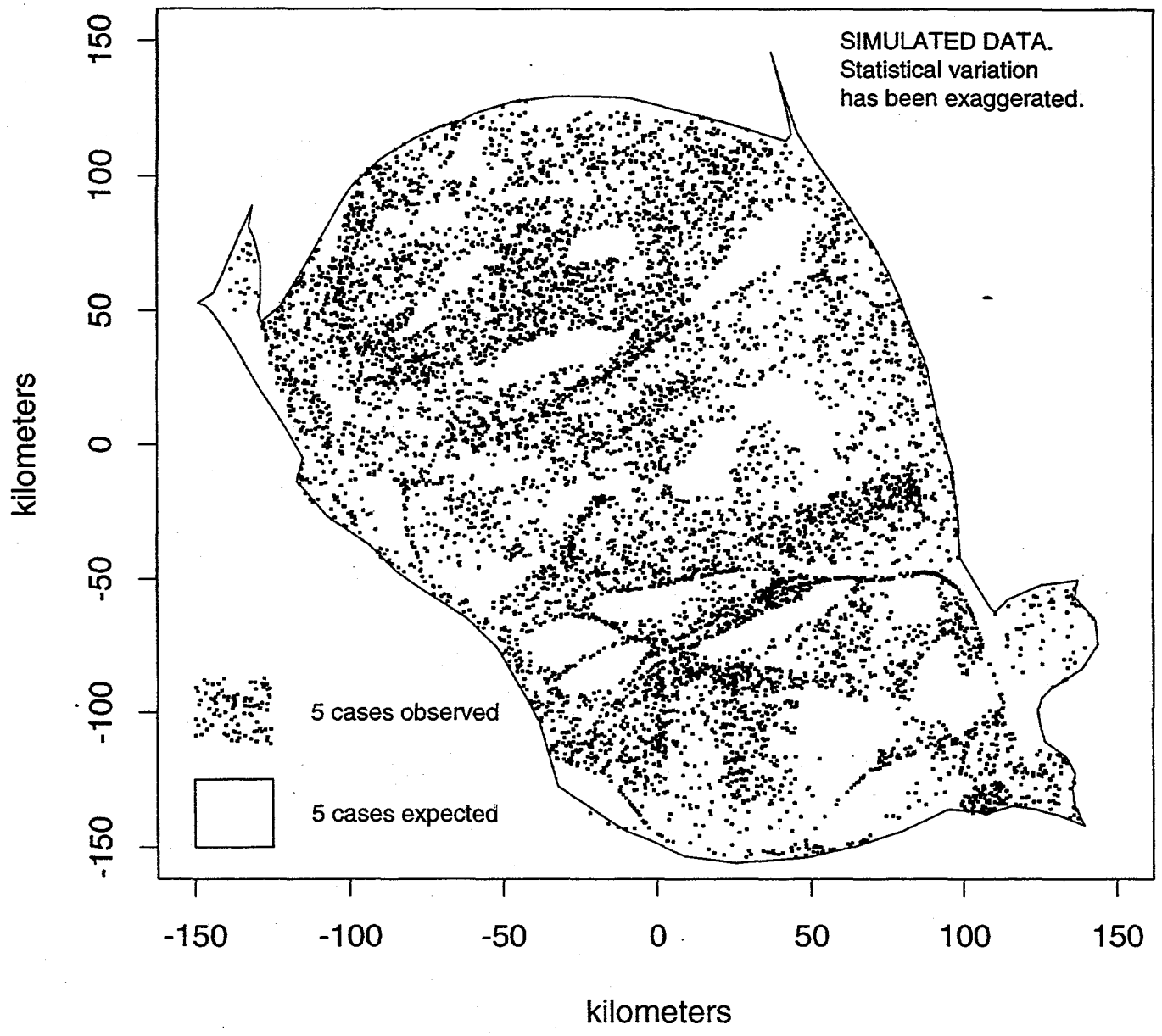


Figure 4. Density equalized map, with 20 points per case.



The expected value of d_k (and the standard error and all higher moments) can be derived under the assumption that points are spatially distributed at random. The formula for the expected value of d_k is given by Cressie.[7] For comparison with theory it is convenient to make d_k dimensionless; that is, to express d_k in units which are equal to the square root of A/n, where A is the area of the region containing n cases. With this convention the expected value of d_1 (for example) is 0.5 for points distributed at random. An observed value of d_1 less than 0.5 indicates mutual attraction, or clustering; a value greater than 0.5 indicates mutual repulsion, or anti-clustering.

Boundary Bias

For any finite study area, the observed value of d_k is biased upward relative to the theoretical value. This occurs because cases near the boundary of the study area have reduced probability of having close nearest neighbors, and so their k'th nearest neighbor distances are biased upward. The boundary bias becomes increasingly important as k increases.

As n increases the bias in d_k becomes smaller, but so does the standard error of d_k. At least for $k=1$, it happens that the bias and the standard error of d_1 have the same functional dependence on n, so that the ratio of the bias to the standard error is independent of n. For the area under investigation, that ratio is approximately 1; for $k=1$ and any value of n the bias of d_1 is equivalent to a one standard deviation effect.[1]

In testing for spatial randomness among the n cases, it is necessary to correct for the boundary bias. One simple method is to generate artificial random cases outside the boundary of the density equalized map, with the same density as that observed inside. The artificial cases are included as nearest neighbor candidates when calculating d_k for the cases inside the boundary. With the external random cases included, k is not limited to n-1 and can be as large as desired provided the extended area is sufficiently large. For very large k (about 2000 in our analysis), the k'th nearest neighbor of *every* case is a random external point, so no clustering effects are observed for k greater than this value.

In the present analysis, the same DEMP transformation was applied to three separate data sets: (a) 401 artificial case locations randomly generated assuming equal risk (b) the 401 actual case locations (c) the 401 cases plotted at random locations within their respective tracts. For correction of the boundary bias, each data set was augmented (after density equalization) with a set of random points outside the boundary. To reduce the effects of random variation, all three data sets (a), (b) and (c) were replicated 20 times. The random points in all 60 samples were completely independent of each other.

Results

In Figure 5, as a function of k, we present the ratio of the observed mean d_k to the theoretical mean for (b) the actual case locations (solid line) and (c) the cases plotted at random locations within their respective tracts (dotted line). The dashed lines represent the expected value (1.0) and the 95% confidence interval (1.96 s.d.) of variability for a single sample.

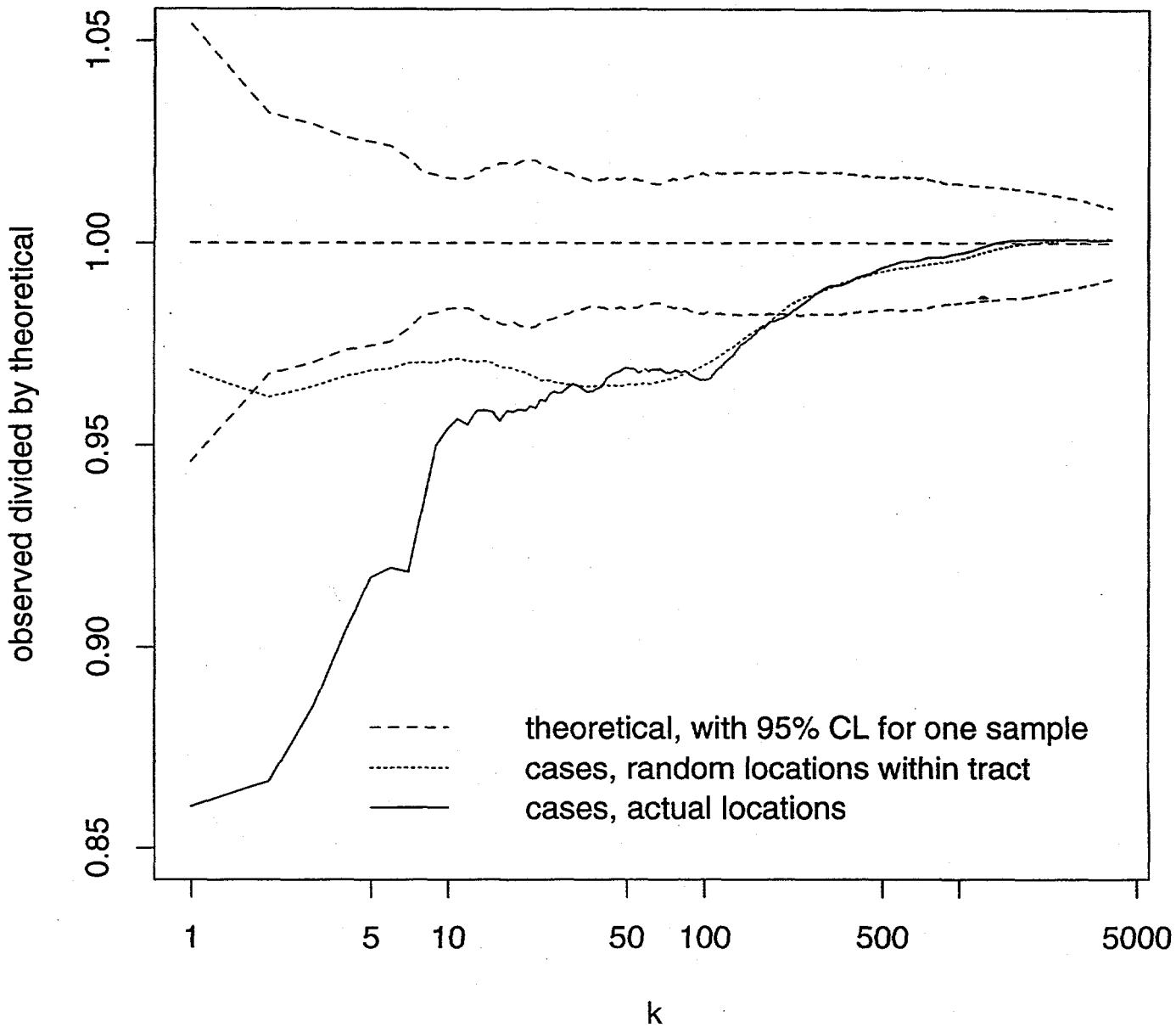
For all values of k, the 20-sample mean value of d_k calculated for the random cases (a) was consistent with the theoretical mean value given by Cressie.[7] Because the DEMP transformation, case simulation and boundary correction are performing as expected, we compare results from (b) and (c) directly with the theoretical value. As a function of k, the confidence limits in Figure 5 were estimated empirically from the variance among the 20 independent samples.[8] At least for $k=1$, that variance agrees with the variance calculated from the jackknife method.[1]

For $k=1$, samples (b) and (c) produced mean values of d_1 equal to .426 and .481, corresponding to .85 and .96 respectively in Figure 5. The single-sample standard error of 0.015 observed in sample (a) corresponds to 95% confidence limits in Figure 5 of 0.94 and 1.06. Samples (b) and (c) show clustering effects equivalent to -4.9 s.d. ($p < 10^{-6}$) and -1.3 s.d. ($p = .10$) respectively. Not surprisingly, clustering is insignificant in sample (c) because the nearest ($k=1$) neighbor of a case is frequently in the same tract as the case, and in sample (c) the locations within each tract have been randomized.

As k increases, the distance from a case to its k'th nearest neighbor becomes large compared with the dimensions of a single tract, so the randomization in (c) becomes insignificant. For values of k greater than about 40, the values of d_k from sample (b) and (c) differ only due to random variation. For values of k greater than about 2000, only randomly generated events contribute to d_k, so samples (a), (b), and (c) are all in agreement.

The most interesting finding is that sample (c), for all values of k between 2 and about 200, shows evidence of significant clustering ($p < .025$). This effect, unlike that in (b), is due entirely to observed rate differences among different tracts. The effect is most marked for k between 20 and 100, where the evidence for non-uniformity of tract rates is about 4 standard deviations ($p < 10^{-4}$).

Figure 5. Mean k'th nearest neighbor distance.



DISCUSSION

The LBL analysis of the four county data set provides strong statistical evidence ($p < 10^{-4}$) for non-uniformity of tract rates of childhood cancer. However, the observed effects may be entirely due to biases in the input data. The use of 1980 population data in conjunction with 1980-1988 case data is a serious defect that must be remedied before epidemiologic conclusions can be discussed. It is very likely that different tracts experienced different rates of population growth, which would dramatically affect the results presented here.

In addition, confounding bias can result from differential Census undercount of certain social groups (*e.g.* Hispanics or migrant workers) coupled with non-uniform geographic distribution of those groups. To some extent this can be checked by stratifying the analysis on those social characteristics.

Finally, purely random errors in either the cases or population or both can produce statistically significant effects. Evidence of geographic nonuniformity of rates, however statistically significant, is not meaningful unless a systematic pattern is detected and unless all sources of bias can be ruled out.

The caveats just stated concern the validity not of the DEMP methodology but rather of the input data. These caveats would apply equally to any analysis of the same data.

CONCLUSIONS

A major accomplishment described in this report is the successful density equalization (in Figure 2) of the complex and highly non-uniform map in Figure 1. Analysis of artificially generated cases showed that significant biases were not introduced.[1] Figure 3 portrays individual cases and rates in a single display that can be understood by untrained observers. In Figures 3 and 4, "hot spots" can be identified and their statistical significance estimated.

Various simple statistical techniques can be applied to density equalized maps. The mean k'th nearest neighbor distance provides a sensitive and unbiased measure of overall non-uniformity of rates, provided one corrects for boundary effects and for non-uniform population distribution within individual tracts. Figure 5 presents evidence of geographic nonuniformity of tract rates, with statistical significance equivalent to four standard deviations ($p < 10^{-4}$). However, no epidemiologic conclusions can be drawn at this time due to deficiencies in the population data that were used.

The four-county problem with 262 subareas required about 20 hours on a SPARC 10 work station. Opportunities exist for improved implementations, now that a reliable algorithm is available. The DEMP technique can become a valuable tool for routine surveillance activities, especially if automatically coupled to data bases containing the necessary population data and map files for all regions of the United States.

ACKNOWLEDGMENTS

The authors thank the California Department of Health Services for providing the case data from the Four County Childhood Cancer Study. Raymond Neutra, Peggy Reynolds, and Enid Satariano provided useful guidance and assistance.

We have enjoyed an active electronic correspondence with Vladimir Tikunov and Sabir Gusein-Zade, the Russian authors of the DEMP algorithm we have implemented and modified. Michael Mohr wrote many of the programs used to manipulate map files. Early programmatic support was provided through the efforts of Carl Quong and Robert Goldsmith.

This work was supported by the Director, Office of Epidemiologic Studies; Office of Health; Office of Environment, Safety and Health; U.S. Department of Energy under Contract No. DE-AC03-76-SF00098.

REFERENCES

(For electronically published references, WWW URL's are Uniform Resource Locators in the World Wide Web.)

1. Merrill, D., Selvin, S., Close, E.R. and Holmes, H.H. Use of density equalizing map projections (DEMP) in the analysis of childhood cancer in four California counties. Report LBL-36630, January 1995. WWW URL: http://cedr.lbl.gov/pdocs/cdc9501/lbl36630.html.

2. Satariano, E., Reynolds, P., Smith, D. and Goldman, L. The Four County Study of childhood cancer incidence: Interim report I. Environmental Epidemiology and Toxicology Branch, California Department of Health Services, May 1990.

3. Reynolds, P., Satariano, E. and Smith, D. The Four County Study of Childhood cancer incidence: Interim report II. Environmental Epidemiology and Toxicology Program, California Department of Health Services, October 1991.

4. SEEDIS (Socio-Economic Environmental Demographic Information System). WWW URL: http://cedr.lbl.gov/mdocs/seedis/seedis.html (1995).

5. Gusein-Zade, S.M. and Tikunov, V.S. A New Technique for Constructing Continuous Cartograms; *Cartography and Geographic Information Systems*, **20(3)**, 167-173 (1993).

6. Close, E.R., Merrill, D.W. and Holmes, H.H. Implementation of a new algorithm for Density Equalizing Map Projections (DEMP). Report LBL-35738, December 1994. WWW URL: http://cedr.lbl.gov/pdocs/tr940401/all.html.

7. Cressie, N. *Statistics for Spatial Data.* John Wiley & Sons, Inc., New York, 1991.

8. Merrill, D. Graphics presented at the 1995 CDC/ATSDR Symposium on Statistical Methods; Atlanta GA, January 25-26, 1995. WWW URL: http://cedr.lbl.gov/pdocs/cdc9501/25graphics.html.

FIGURE CAPTIONS

1. Filtered and triangulated map, with 401 cases. Unnecessary geographic detail has been removed, and the 262 tracts (solid lines) have been subdivided into 1212 triangles (dotted lines). Locations of the 401 cases are indicated by dots.

2. Tracts, hexagons and 401 cases, density equalized. This is the same map as in Figure 1, after density equalization. Prior to density equalization, the 1212 triangles were converted to hexagons by subdividing all triangle faces. The map scale is not true distance, but can be interpreted as "equivalent kilometers" since the maps of Figures 1 and 2 have the same area. Population density is everywhere equal in Figure 2.

3. Density equalized map, with one point per case. Each case is plotted once at a random location within its own tract. On the average, five cases should occur in any region (of any shape) having area equal to that of the legend box. For example, just below the center of the map, seven or eight cases occur in a small region where about one case is expected. To the northeast, only two cases occur in a broad region where about ten are expected.

4. Density equalized map, with 20 points per case. Each case is plotted at 20 random locations within its own tract. On the average, five cases (100 points) should occur in any region that has area equal to the legend box. The statistical significance of Figure 4 is no greater than that of Figure 3; however, relatively dense and sparse regions are more easily identified. The empty regions have no observed cases; they are generally not significant because they have small areas, corresponding to only a few cases expected. The thin snakelike patterns are sparsely populated rural areas.

5. The mean k'th nearest neighbor distance, divided by the theoretical value, is plotted as a function of k for the actual case locations (solid line) and for cases plotted at random locations within their respective tracts (dotted line). The 95% confidence interval about the expected value of 1.0 (dashed lines) was calculated from random cases which were generated assuming equal risk. For values of k between 20 and 100, the evidence for non-uniformity of tract rates is about 4 standard deviations ($p < 10^{-4}$).