LBL-36630
UC-405

# Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

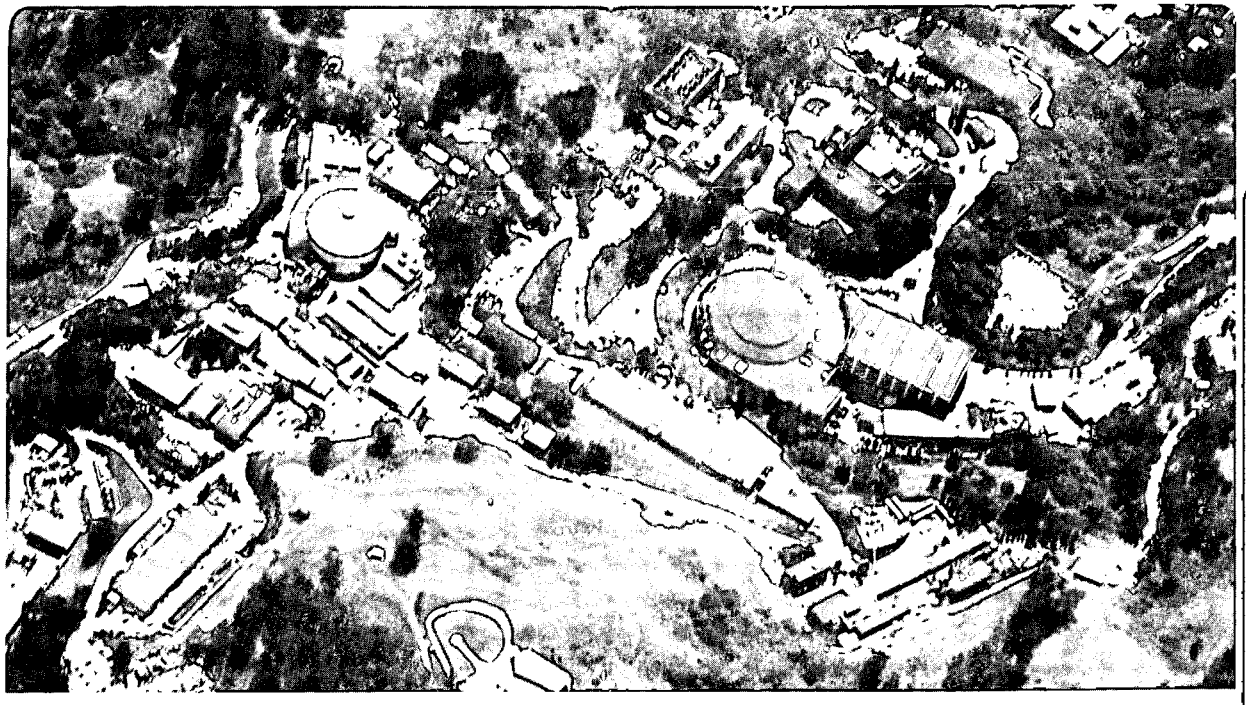
## Information and Computing Sciences Division

To be presented at the 1995 CDC/ATSDR Symposium on Statistical Methods: Small Area Statistics in Public Health: Design, Analysis, Graphic and Spatial Methods, Atlanta, GA, January 25–26, 1995, and to be published in *Statistics in Medicine*

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

D.W. Merrill, S. Selvin, E.R. Close, and H.H. Holmes

January 1995





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

## DISCLAIMER

LBL-36630
UC-405

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

D.W. Merrill, S. Selvin, E.R. Close, and H.H. Holmes

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

January 1995

Address all correspondence to Deane W. Merrill, Information and Computing Sciences Division, 50B-2239, Lawrence Berkeley Laboratory, 1 Cyclotron Road, Berkeley, CA 94720. Tel: (510) 486-5063. Fax: (510) 486-6363. Internet: dwmerrill@lbl.gov. WWW home page: http://cedr.lbl.gov/~merrill/index.html.

The electronic version of this document is publicly available at WWW URL http://cedr.lbl.gov/pdocs/cdc9501/cdc9501.html. Future revisions will be incorporated in the electronic version.

This work was supported by the Assistant Secretary for Environment, Safety and Health, Office of Epidemiology and Health Surveillance, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

this page intentionally left blank

## ABSTRACT

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. On the transformed map, the statistical analysis of the observed distribution is greatly simplified. Even for sparse distributions, the statistical significance of a supposed disease cluster can be reliably calculated.

The present report describes the first successful application of the DEMP technique to a sizeable "real-world" data set of epidemiologic interest. An improved DEMP algorithm [GUSE93, CLOS94] was applied to a data set previously analyzed with conventional techniques [SATA90, REYN91]. The results from the DEMP analysis and a conventional analysis are compared.

ACKNOWLEDGMENTS

The authors are grateful to the California Department of Health Services for permission to use the case data from the Four County Childhood Cancer Study. In particular, Raymond Neutra, Peggy Reynolds, and Enid Satariano provided useful guidance and assistance.

Via the Internet, we have enjoyed a active and useful 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.

Programmatic support in the early years of DEMP development was provided through the efforts of Carl Quong at Lawrence Berkeley Laboratory and Robert Goldsmith in the Department of Energy.

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

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES (CONTINUED)

# INTRODUCTION

Density equalizing map projections (DEMP), also known as cartograms or anamorphoses, have long been used for display of thematic data, but practical computerized implementations were unavailable until recently. 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.

The usual technique for analyzing geographic disease distributions is the comparison of rates from different subareas. Relative to conventional methods, the DEMP technique has the following advantages:

- Like a conventional map, the density equalized map is a graphic representation which can be understood without statistical analysis. But only on the density equalized map can one easily see effects occurring in small densely populated areas.

- The DEMP technique avoids the calculation of unstable rates for small subareas where the number of cases is small.

- The full geographic detail of the data can be used.

- The DEMP analysis is appropriate, and even works best, in the analysis of rare diseases where the number of cases is small.

- Systematic effects across broad regions of the map are easily detected, without the need for arbitrary grouping of subareas.

- A number of rigorous, simple well-developed statistical techniques are available for analyzing the density equalized map.

- 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.

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

## FOUR-COUNTY CHILDHOOD CANCER STUDY

For the present study, a four-county area in California (Fresno, Kings, Kern and Tulare) was selected because of the availability of small-area health data. The data, which were kindly provided through a collaborative agreement between Lawrence Berkeley Laboratory (LBL) and the California State Department of Health Services (DHS), were originally collected by DHS to investigate a reported childhood cancer cluster in the community of McFarland, California. The data, which consist of 401 childhood cancer cases occurring between 1980 and 1988, were previously analyzed by DHS [SATA90, REYN91].

The first DHS report [SATA90] examined childhood cancer rates by cancer site, age, sex, race/ethnicity (Anglo, Hispanic and other), county, and land use (rural versus urban, and agricultural versus non-agricultural). The calculation of population at risk is described in detail. Observed rates were found to be consistent with rates reported in other studies; the only significant departure from uniformity was that rates among children in urban non-agricultural areas are slightly higher than those in rural non-agricultural areas. The urban non-agricultural rates are comparable to urban rates elsewhere in California. Rates in agricultural areas are not elevated.

The second DHS report [REYN91] examined the differences among geographic areas in greater detail. The four-county study area was subdivided into 101 communities. The community boundaries and case locations are shown in Figure 1. For each community the observed number of cases was compared with the number expected, assuming the underlying cancer rate to be uniform. The cancer incidence rate ratios (and 95% confidence limits) are shown in Figure 2. Six of the 101 communities had rates that fell outside 95% confidence limits (three with more cases than expected and three with fewer cases than expected). The locations of the three high rate communities and the three low rate communities are shown in Figure 3. 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.

The purpose of the present study is to test the applicability of density equalizing map projections (DEMP) for analyzing disease distributions in the vicinity of suspected environmental hazards. The DEMP technique can be applied in any area where adequate small-area data (cases, population, and map files) exist. The present report describes the first successful application of the DEMP technique to a sizeable "real-world" data set.

Map 5.

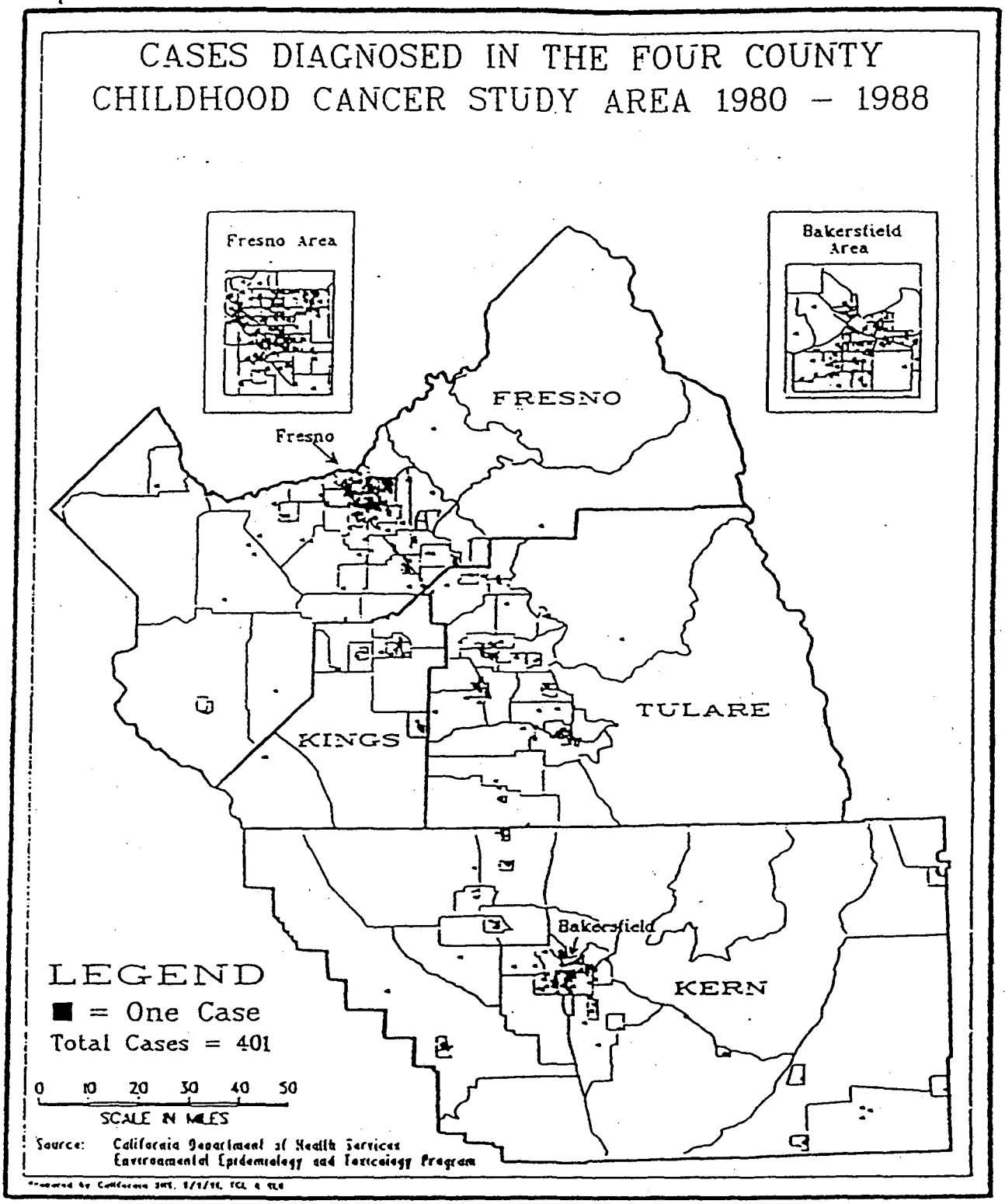
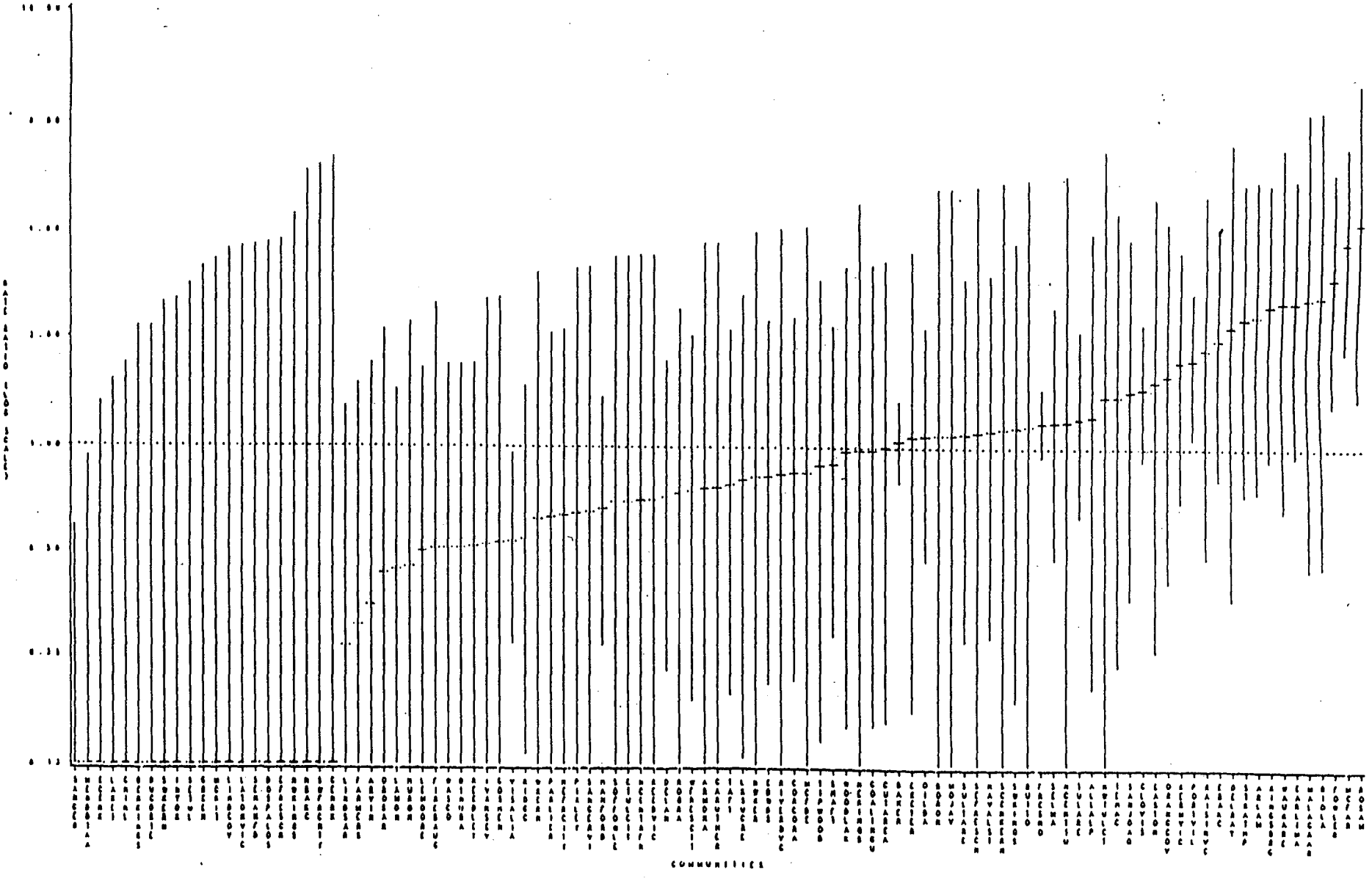



Figure 1.

Figure 2. *Childhood cancer incidence rate ratios (and 95% CI) for Four County communities compared to the overall Four County rate.*
Four County Cancer Study, 1980-88, DHS.



Source: California Department of Health Services, Environmental Epidemiology and Toxicology Program, October 24, 1991.

Map 6.

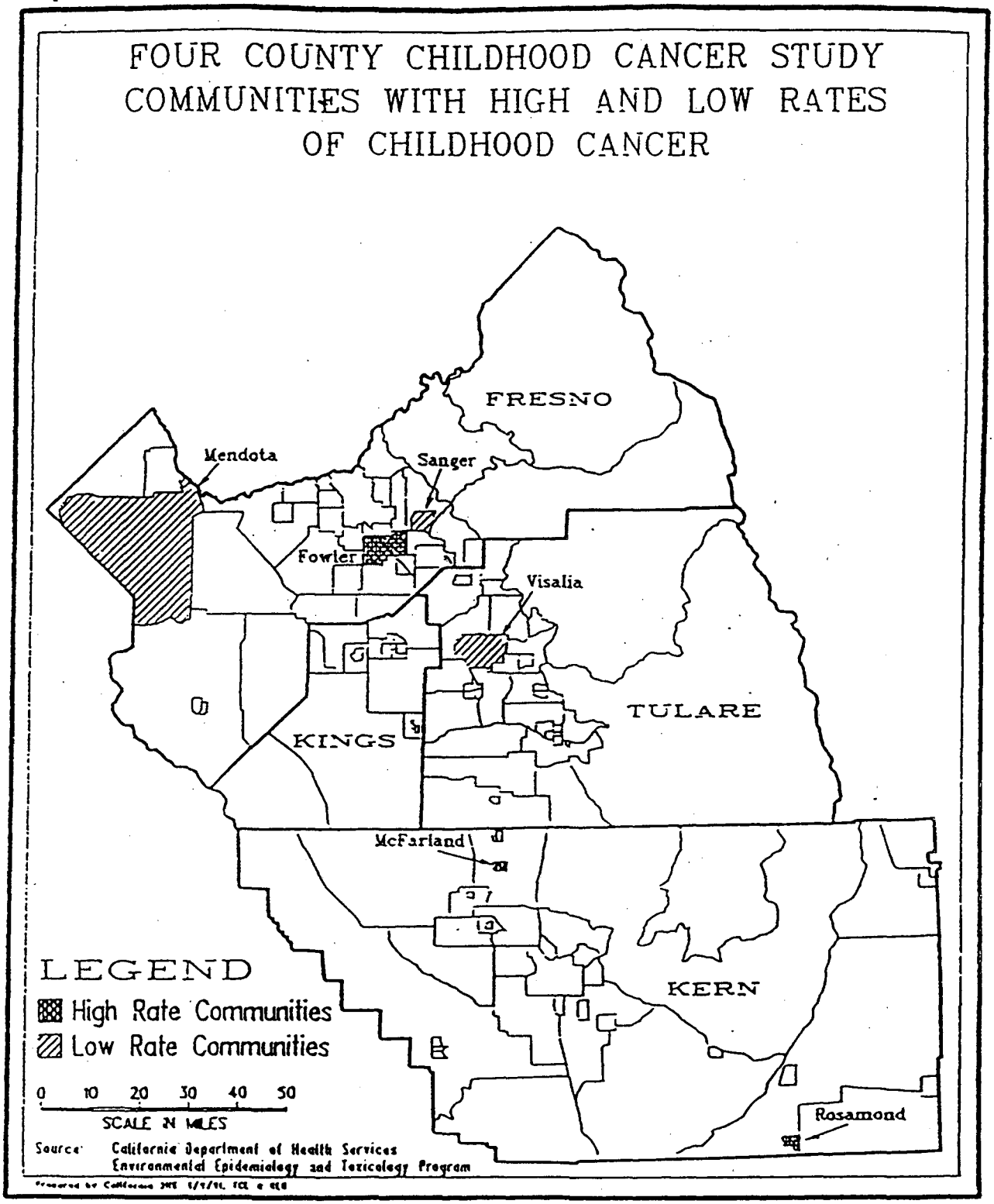



Figure 3.

Regarding the question of clustering among childhood cancer cases in the four-county area, the present report confirms the negative result of the earlier analysis. Unfortunately, the DHS population data, for the nine-year period 1980-88, were lost in a computer mishap and are no longer available. The population data used in the present study, which are from the 1980 Census, demonstrate the feasibility of the DEMP technique as an analytic tool. However, firm epidemiological conclusions cannot be drawn from the DEMP analysis until the analysis has been repeated with correct 1980-88 population data.

Regarding the negative findings of the present study, we note that the nearest neighbor technique is not the most sensitive technique that could have been used. A different metric, for example the distance of the 5th or 10th nearest neighbor instead of the nearest neighbor, might have uncovered effects not observed in the DHS analysis or the LBL analysis, provided such effects exist. To avoid the statistical implications of multiple testing we have purposely limited the present investigation to a single technique (nearest neighbor analysis) which was chosen before the analysis began.

## DATA PREPARATION

### Case locations

In April 1993, DHS provided to LBL a file containing the following for each of the 401 cases:

- unique case identifier
- year of diagnosis
- census tract code
- abbreviated name of community
- longitude
- latitude
- county code
- age category at diagnosis (either 0-4 or 5-14)
- gender
- race (white, Hispanic, or other)

In the present analysis all 401 cases were analyzed as a single data set, without regard to year of diagnosis, age category, gender or race. The census tract codes provided by DHS do not agree exactly with the 1980 census data and map files at LBL, and were not used. Only the longitude and latitude were used in the present analysis.

### Population data

The first DHS report [SATA90] describes the estimation of population at risk by age group, gender and race, for the period 1980-88. The input data used were 1980 Census data and intercensal population estimates from the California Department of Finance. Consistency checks were performed with the use of preliminary 1990 Census data. Following the completion of the DHS analysis, the file containing the 1980-88 population estimates was inadvertently erased, and cannot be easily replaced.

For the present analysis LBL used the 1980 Census population for children of ages 0-17, which is readily available in LBL SEEDIS (Socio-Economic Environmental Demographic Information System) [SEED94]. For a few very small tracts, the population of children 0-17 was estimated from the total population. The estimation process is described in [MERR93].

A correct analysis of the four-county data set would have used the 1980-88 population at risk for children 0-14. The considerable additional work necessary to obtain the correct population estimates is described in [MERR93] and will be performed later.

**Geographic map files**

The map data used in the present analysis are proprietary 1980 Census tract boundary files which were purchased from National Planning Data Corporation in 1985 and incorporated in LBL SEEDIS. The input map file is shown in Figure 4 along with the locations of the 401 cases. Figure 4 agrees closely with Figure 1, which was included in the second DHS report [REYN91]. Note, however, that Figure 4 includes census tract boundaries and Figure 1 does not. The additional geographic detail occurs primarily in Fresno and Bakersfield.

The map files were further processed for use in the DEMP analysis. The pre-DEMP map processing is described in [MERR94B] and includes the following steps:

- making the separate county map files match at the county boundaries (the heavy lines in Figure 4 indicate the locations that required special processing);
- removing unnecessary geographic detail for efficiency in the DEMP analysis;
- insertion of connection segments in order to represent doughnut-shaped tracts as a single polygon;
- decomposition of each census tract into triangles (the unique Delaunay triangulation was used);
- subdivision of each segment to convert the triangles into hexagons;
- conversion among various file formats, for DEMP processing and display of the resulting map files.

The resulting map, ready for DEMP analysis, is shown in Figure 5. Additionally, for the conventional rate comparison described in the next section, a point-in polygon routine was used to determine the census tract of each case.

Figure 4.

Four-county map from SEEDIS, with 401 cases.

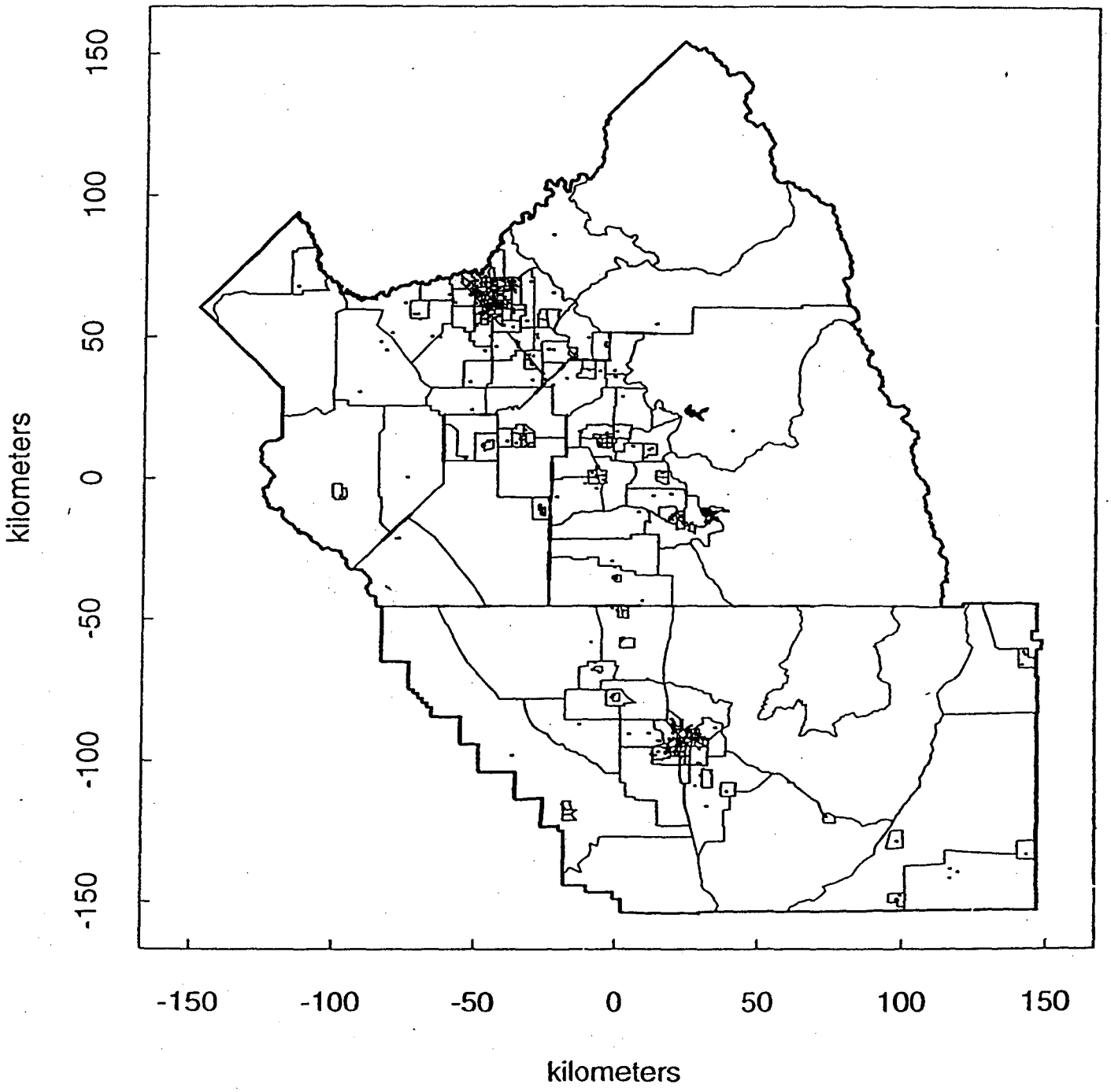
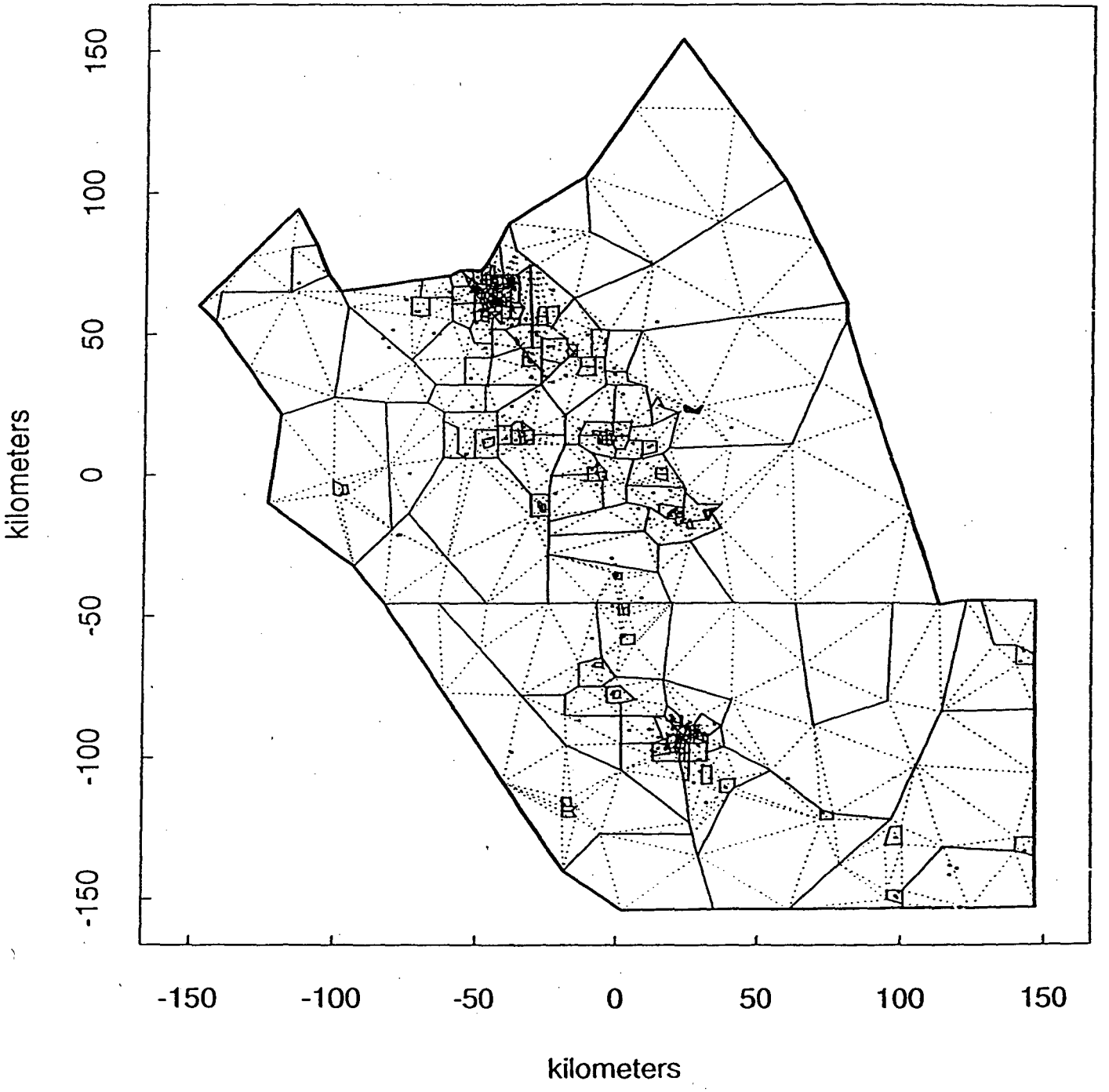


Figure 5.

Four-county map, filtered and triangulated, with 401 cases.



## CONVENTIONAL ANALYSIS: RATE RATIOS

A typical epidemiologic approach involves the comparison of the tract-specific rates to the overall rate. Tract-specific rates $p_i$ are calculated in the usual fashion $p_i = d_i/n_i$ (i.e., number of cases in a specific tract divided by the population-at-risk in that tract). In fact, these values are estimated probabilities but are often referred to as rates. If the population is assumed constant over time, as in the present analysis, $n_i$ is estimated from the population.

Equivalently, the number of cases $d_i$ observed in tract $i$ is compared with the number of cases $\hat{e}_i$ expected under the null hypothesis that rates are uniform. Specifically,

$$\hat{e}_i = \text{overall rate} \times n_i$$

In the present study, $n_i$ is the 1980 Census population of children of ages 0-17; in the entire four-county area there were 382,546 children of ages 0-17 and 401 cases were observed, so

$$\hat{e}_i = ( 401 / 382546 ) \times n_i$$

Values of $n_i$, $d_i$, and $\hat{e}_i$ are obtained for each census tract $i$. Under the hypothesis that cancer cases occur at random, the number of cases $d_i$ in each census tract has a Poisson distribution. For these conditions, a test statistic which is an excellent approximation to the exact Poisson distribution is [BRES87]

$$z_i = \sqrt{9D_i}\left( 1 - \frac{1}{9D_i} - \left( \frac{\hat{e}_i}{D_i} \right)^{1/3} \right)$$

where $D_i = d_i$ if $d_i$ exceeds $\hat{e}_i$ and $D_i = d_i+1$ otherwise. The value $z_i$ has an approximate standard normal distribution (mean = 0 and variance = 1) if no systematic pattern exists in the distribution of the cancer cases among the census tracts. The tract-specific values $z_i$ are displayed in Figure 6. (Two very small tracts with $d_i = 0$ and $\hat{e}_i < 0.02$ were excluded). If no clustering exists among the cancer cases, then 2.5%, or approximately seven of these values, should exceed 1.96 (the upper dotted line). In fact ten tracts exceed this value (p-value = 0.13). We conclude that the conventional analysis, like that performed by DHS, provides no significant evidence for non-uniformity of rates.

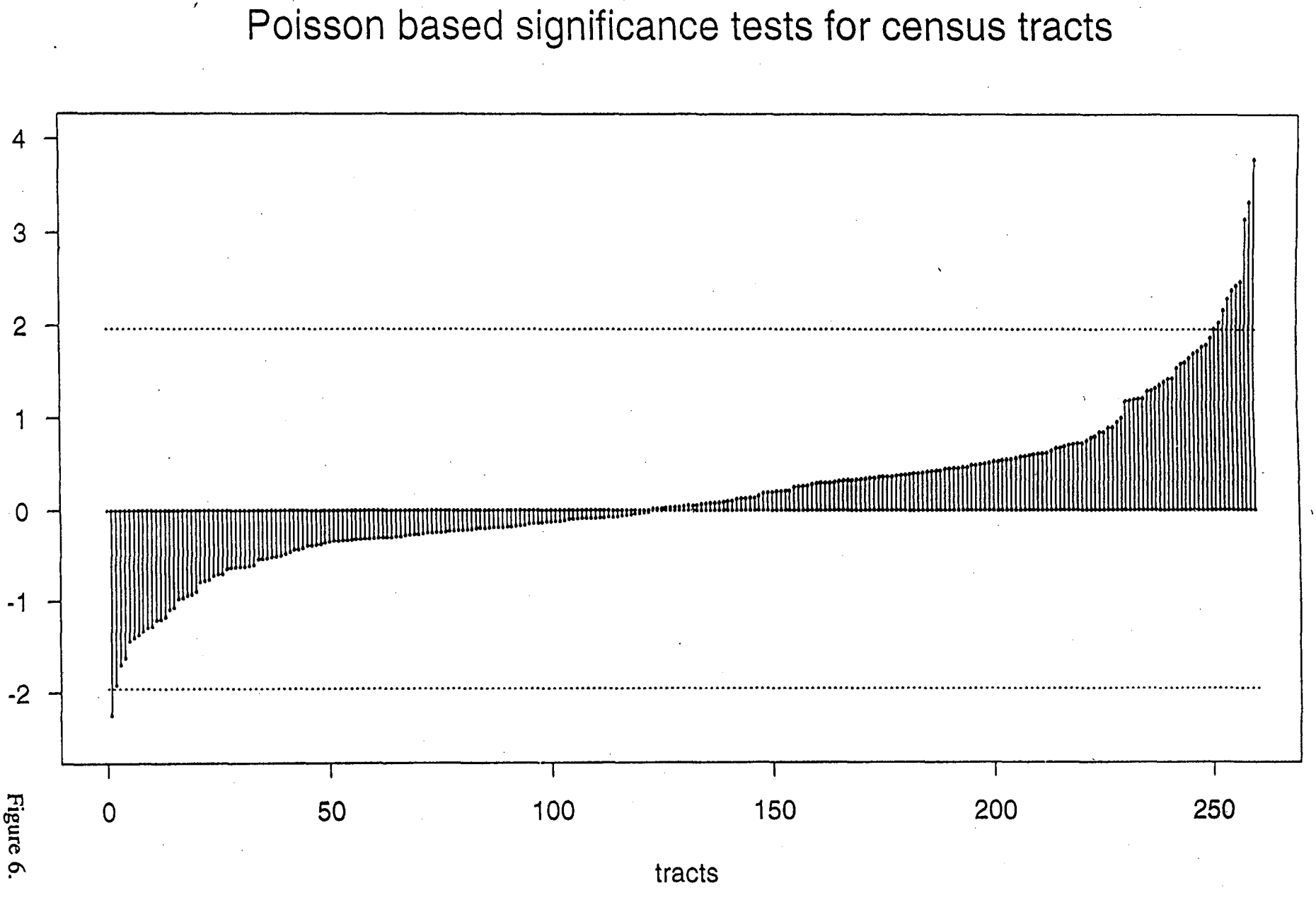

Poisson based significance tests for census tracts
4
3
2
1
0
-1
-2
0
50
100
150
200
250
tracts

Figure 6.

## DEMP ALGORITHM

Development of a practical DEMP algorithm has been a goal of the LBL PAREP project since 1985. The history of the LBL effort is summarized in Appendix D. At least a dozen computer algorithms have been developed elsewhere; most of these are suitable for graphic display purposes, but not for the requirements of statistical analysis. Earlier LBL algorithms addressed this problem but were too slow for problems of reasonable size. To our knowledge no other algorithm has been applied to a study area as complex and nonuniform as the one described here.

A theoretical breakthrough occurred with the 1993 publication of a new algorithm by two Russian authors [GUSE93]. The algorithm was independently implemented and extensively tested at LBL; that effort is described in a 130-page technical report [CLOS94].

We find that the Russian algorithm as published is inadequate for mixed urban-rural areas like the four-county study area, where population densities are extremely nonuniform. Problems can be avoided by introducing more points or taking more numerous smaller steps, but then computing time and memory requirements become excessive. An additional scaling factor described in [CLOS94] was found to be essential in the present analysis.

Also, the Russian algorithm provides no mechanism for detecting or correcting map errors (boundary intersections) produced by density equalization. This problem can also be avoided by using more points or proceeding in smaller steps, but at the expense of computing time.

Two separate equalization runs were completed. In each run the target areas of the 262 tracts were determined from the 1980 population of ages 0 through 17. The units which were density equalized are the 1212 triangles shown in Figure 5. Within each tract the target areas of the triangles were apportioned in the same ratio as the original areas. The areas of two small lakes, which are less than 0.3 percent of the land area, were allowed to float freely; their effect on the analysis is negligible.

- In the primary run "hex10" we took ten equal steps using 1212 hexagons, which were obtained by bisecting the segments of the triangle map in Figure 5. Forty map errors were introduced (and ignored) but these did not bias the statistical results, as will be shown. The primary run hex10 is described in Appendix B.

- In the secondary run "tri10" we took equal steps using triangles which were successively subdivided as necessary in order to prevent map errors. After the seventh step the triangles were converted to hexagons to permit a solution to be reached. Only seven map errors were introduced in the final steps. The secondary run tri10 is described in Appendix C. For reasons that are explained in Appendix C, the tri10 run introduced artificial clusters, so it was not used in the statistical analysis of the present report.

In summary, the successful hex10 run incorporates improvements worked out during ten years of development effort. The rejection of the tri10 run shows that spurious results can be detected and avoided. The practical feasibility of the DEMP approach has been demonstrated for a sizeable and complex data set.

The hex10 run required about 20 hours on a Sun SPARC 10 work station. Computing time increases approximately as the square of the number of regions in the map. If required for future applications, improved programming techniques can reduce computing time by a factor of 2 to 5; another factor of 10 or even 100 can be achieved by implementing the algorithm on a massively parallel computer.

## DENSITY EQUALIZED FOUR-COUNTY MAPS

Figure 7 shows the distribution of the 401 cases in the four-county study area. The distribution is identical to that in Figures 1 and 4. As expected, most of the cases are in Fresno and Bakersfield, where the population is concentrated. For reference, the county boundaries appear as faint dotted lines.

Figure 8 shows the distribution of the cases on the density equalized map. (The exterior points are artificial points plotted at random, which were used *only* in Table 5, page 39.) The map scale "kilometers" is not really true distance, but can be interpreted as "equivalent kilometers" since the equalized map is normalized to have the same area as the original map.

We note the presence of a few localized clusters in the density equalized map of Figure 8. At least partially, these are due to the fact that the population on the original map (Figure 7) is not uniformly distributed within individual census tracts; people live in houses which are on streets, and these are not randomly scattered throughout a tract. Given populations and map boundaries for 262 census tracts, the DEMP algorithm can only equalize densities *among* the 262 *different* tracts, not *within* a *single* tract. To equalize densities within a tract would require population data and map files with detail below the tract level, for example for block groups or blocks.

There also appears to be a general lack of cases in the northeast and southern regions of the study area, but further analysis is required to determine whether this nonuniformity is significant.

In Figures 9 and 10 we present the same maps, but this time with 401 artificial cases which were generated assuming equal risk. (The exterior points in Figure 10 are additional points plotted at random, which were used *only* in Table 5 and Appendix A.) Prior to density equalization a tract was randomly selected with probability proportional to its population, and a point randomly plotted within that tract; then the process was repeated 401 times. As expected, no clustering is observed in the equalized map of Figure 10.

In Figures 11 and 12 we present the same maps again, but this time with 401 locations which were randomly generated in the *same* tracts as the real cases. (The exterior points in Figure 12 are additional points plotted at random, which were used *only* in Table 5.) The intent is to remove the small clusters *within individual* census tracts, while retaining the true distribution of cases *among different* tracts. This analysis will be discussed more fully later.

Figure 7.

Actual locations of 401 real cases,
before density equalization.

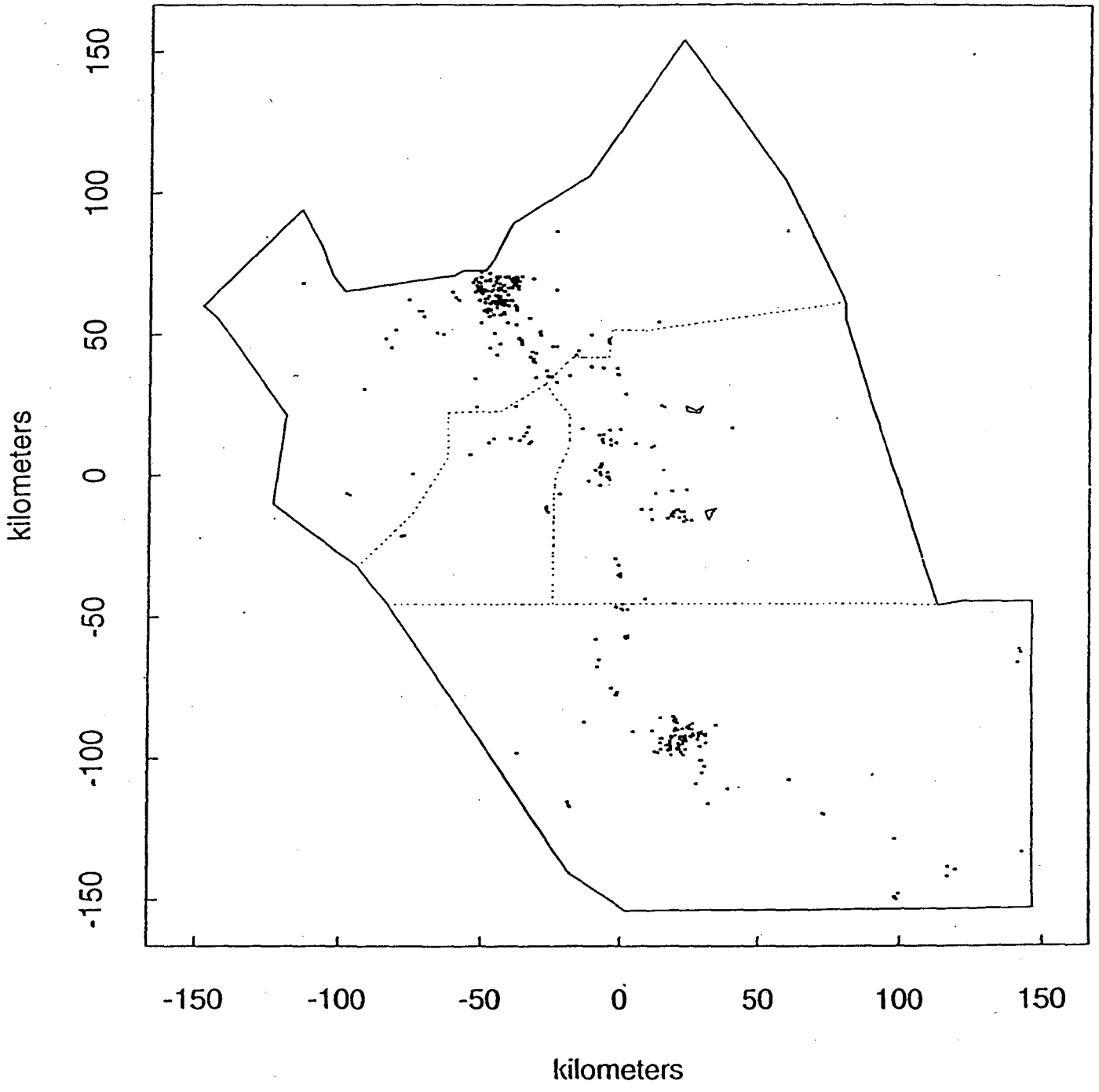


Figure 8.

Actual locations of 401 real cases, after density equalization. The external points are random artificial cases used *only* in Table 5 (page 39).

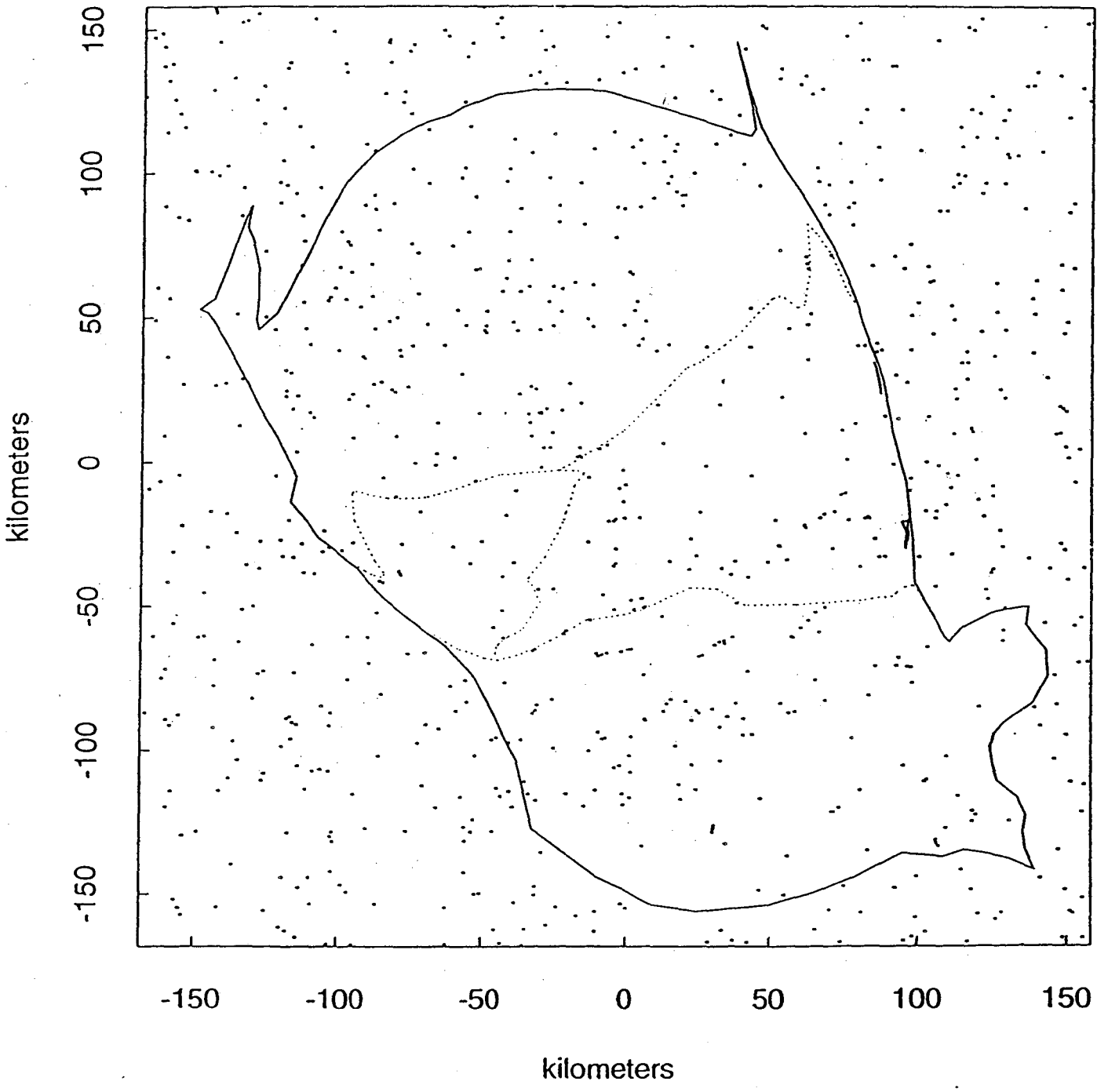


Figure 9.

Locations of 401 artificial cases assuming
equal risk, before density equalization.

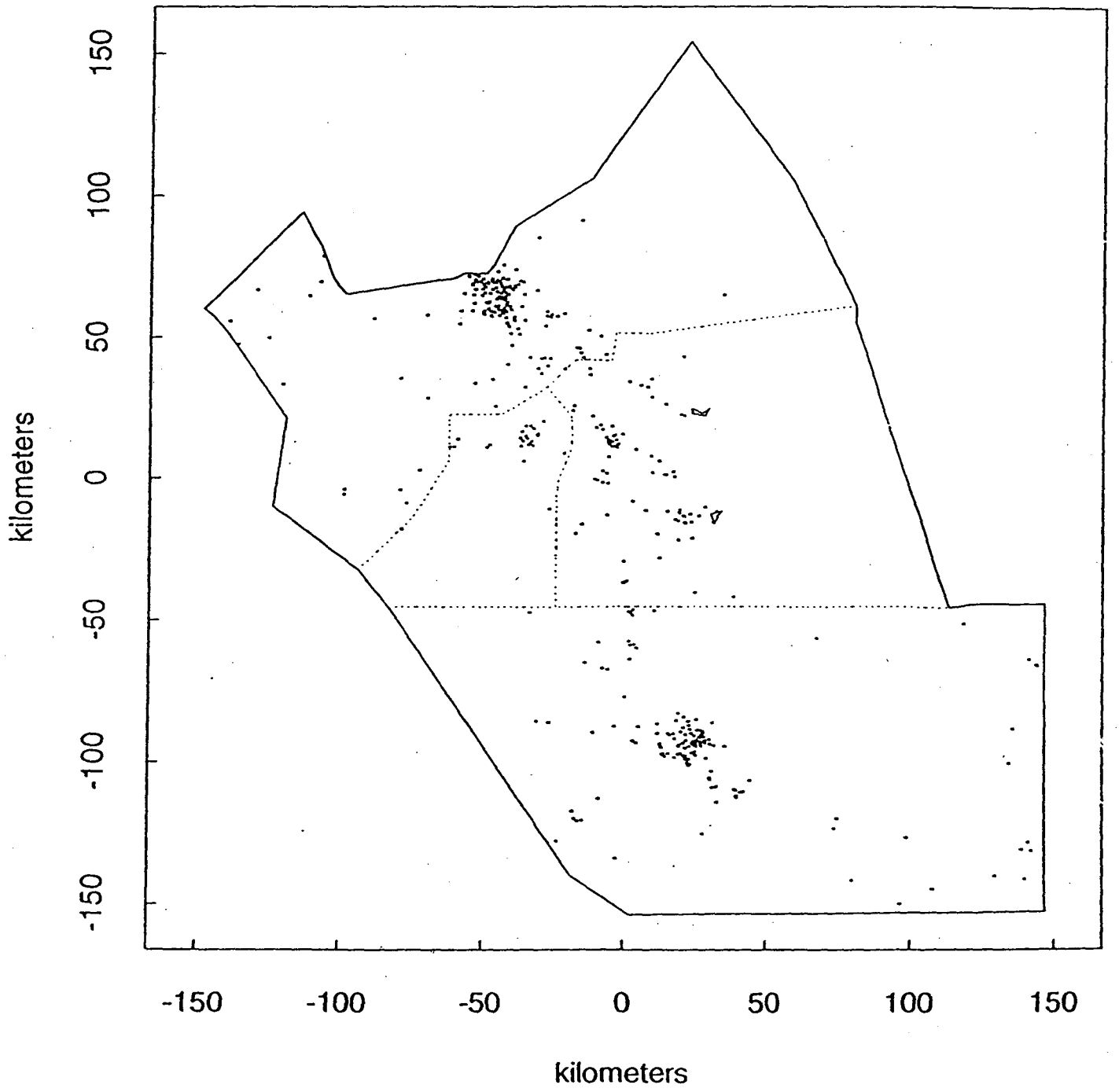


Figure 10.

Locations of 401 artificial cases assuming equal risk, after density equalization. The external points are additional random cases used *only* in Table 5 (page 39) and Appendix A.

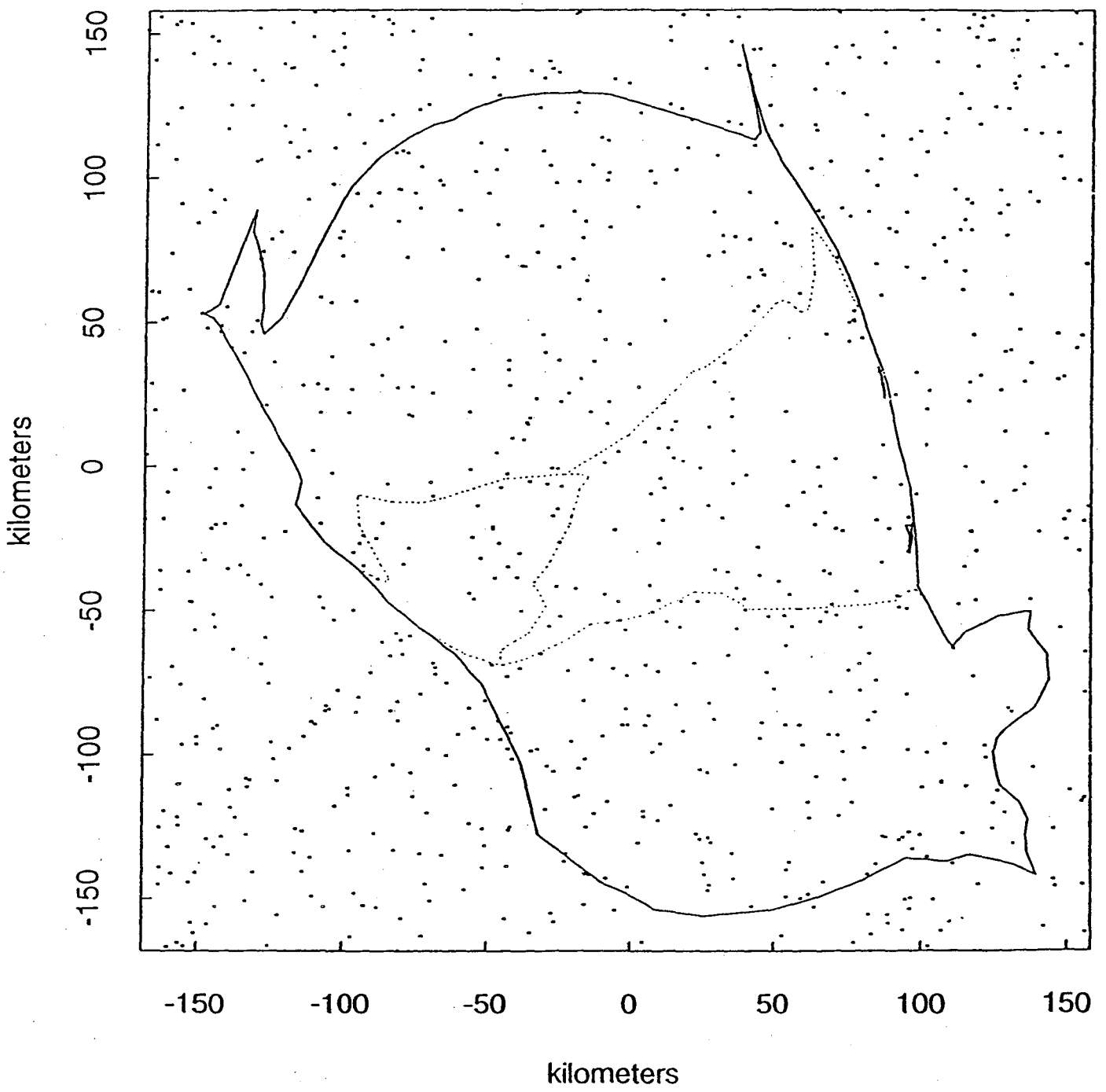


Figure 11.

401 real cases, each plotted at a random location
in its own tract, before density equalization.

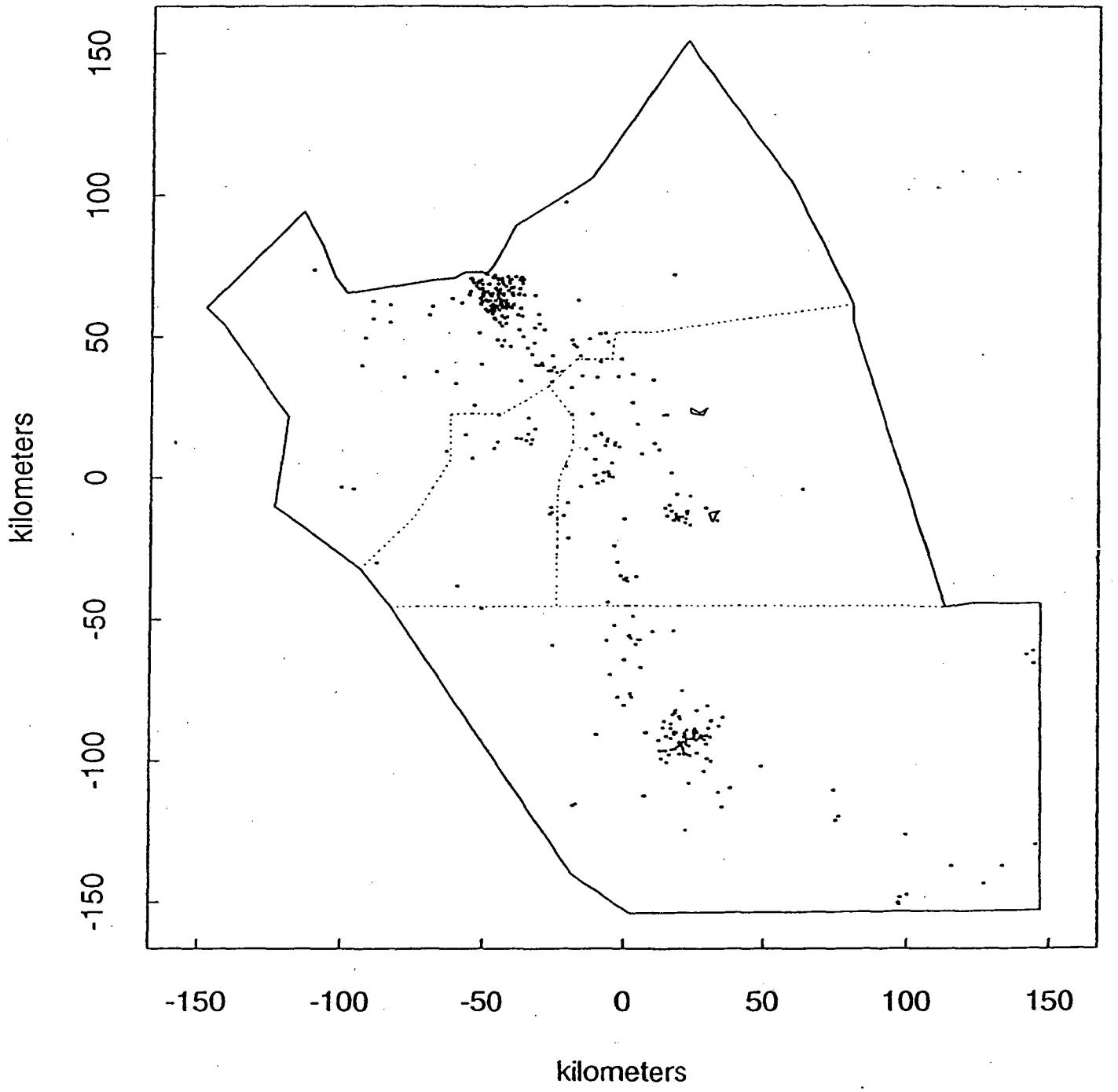
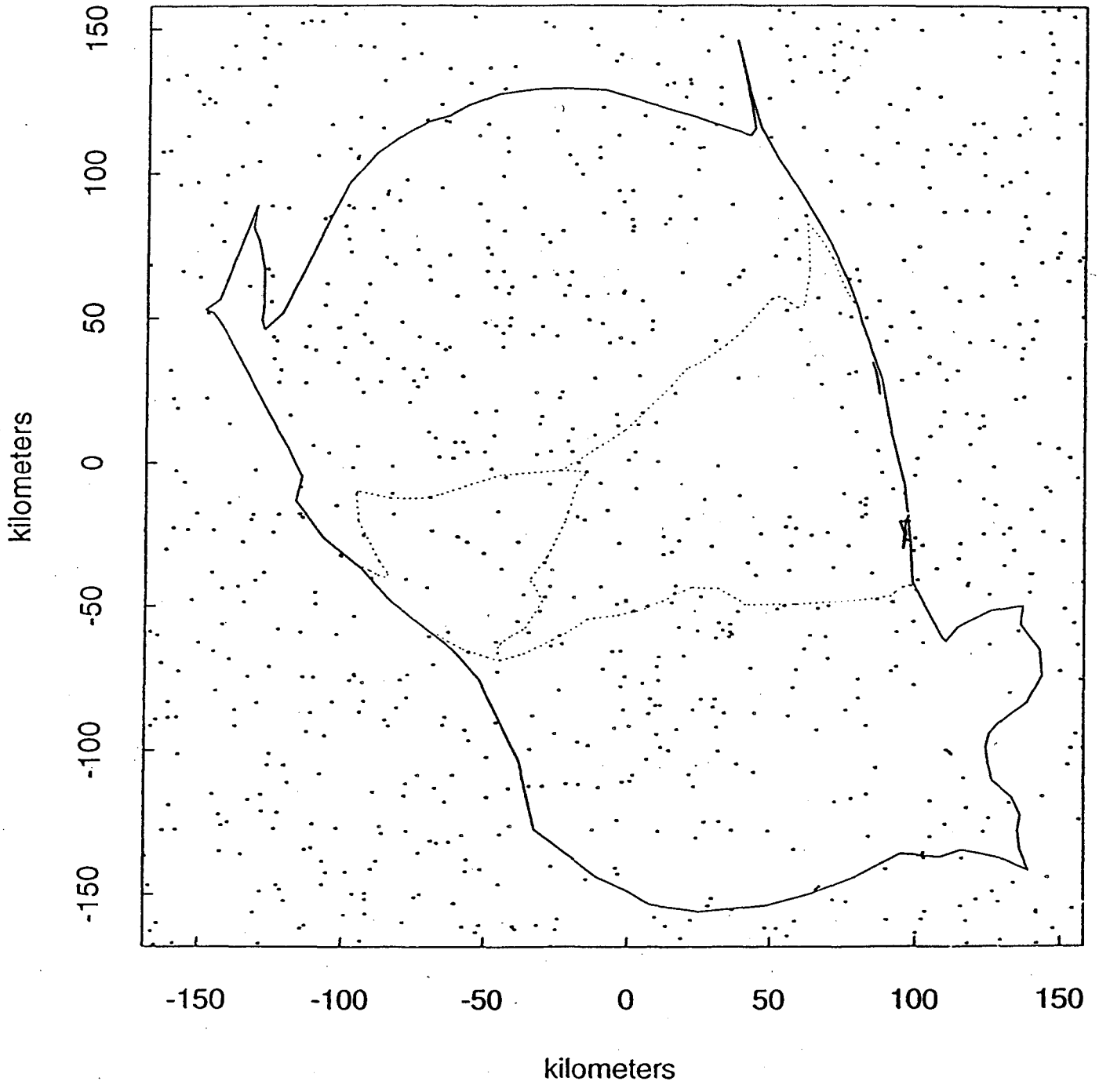


Figure 12.

401 real cases, each plotted at a random location in its own tract, after density equalization. The external points are additional random cases used *only* in Table 5 (page 39).



## General considerations

The most elementary approach to the analysis of spatial data involves dividing the area of interest into a series of subareas, and counting the occurrences of the phenomenon under study. Under specific conditions these counts can be evaluated by comparison with the Poisson distribution (as in the earlier section discussing rate ratios). Employing discrete counts is not as effective as directly analyzing a continuous variable, particularly for small numbers of observations. One method for investigating the question of spatial randomness, which utilizes the actual distance between points, is a nearest neighbor analysis.

A nearest neighbor is basically what the name implies. For $n$ observed points, distances to all other points under consideration are calculated. The nearest neighbor is the minimum distance among the $n$ - 1 measurements. The collection of these $n$ minimum distances constitutes a set of nearest neighbor data. The expected mean and the variance of the distribution of a set of nearest neighbor values can be derived under the conditions that the spatial distribution generating the data is random.

When interest lies in the distance from a specific observation to its nearest neighbor in any direction, the radius $r$ of a circle is an appropriate measure of distance. The density of points over a defined geographic region equals $n/A$, where $A$ is the total area under consideration and $n$ the total number of observed points.

The cumulative distribution function $F(r)$ is the probability that the nearest neighbor associated with a given point will occur at a distance less than $r$. The probability density $f(r) = dF(r)/dr$ is the derivative of this function and $f(r) \times dr$ is the infinitesimal probability that the nearest neighbor will occur at a distance between $r$ and $r + dr$. If the points generating the nearest neighbor distribution are distributed spatially at random, then

$$F(r) = e^{-\pi r^2 \frac{n}{A}}$$

$$f(r) = 2\pi r \frac{n}{A} e^{-\pi r^2 \frac{n}{A}}$$

(The theoretical functions $F(r)$ and $f(r)$ are included in Figures A-2 and A-1, respectively, of Appendix A.)

Knowledge of the probability function $F(r)$ allows the calculation of various summary statistics associated with nearest neighbor distances for a sample of randomly distributed points. For example, the expected median distance, expected mean distance, and the variance and standard error of the observed mean $\bar{r}$ are [SELV91]:

$$median\ (\ r\ ) = 0.470\ \sqrt{\frac{A}{n}}$$

$$mean\ (\ r\ ) = 0.500\ \sqrt{\frac{A}{n}}$$

$$variance\ (\ \bar{r}\ ) = 0.068\ \frac{A}{n^2}$$

$$\sigma = \sqrt{variance\ (\ \bar{r}\ )}\ = 0.262\ \frac{\sqrt{A}}{n}$$

The fact that the median and mean are approximately equal implies that the density function *f(r)* is nearly symmetric about its maximum value. This implies that the mean nearest neighbor distance $\bar{r}$ has approximately a normal distribution, which greatly simplifies the interpretation of the results.

**Specific results**

For the four-county data, $n = 401$ and $A = 51{,}500$ km$^2$, giving *mean(r)* = 5.66 km and $\sigma = 0.148$ km. Here, "km" means equivalent kilometers; namely, the map units of Figure 8. Given the model used to derive *F(r)*, a *z*-statistic

$$z = (\ \bar{r}\ - mean(r)\ ) / \sigma$$

provides an assessment of the difference between the observed and theoretical mean values. The value $\bar{r} = 4.93$ km is the observed mean nearest neighbor distance of the 401 cancer cases, from the density equalized map in Figure 8. The value *z* has an approximate standard normal distribution (mean = 0 and variance = 1) if no spatial pattern exists. Therefore,

$$z = (4.93 - 5.66) / .148\ = -\ 4.9 \text{ standard deviations}$$

The result is summarized in Table 1. The value of the test-statistic $z = -\ 4.9$ implies that it is highly unlikely ( *p*-value < 0.001) that the observed cancer cases are distributed at random over the four-county area.

**Table 1.** Summary statistics from the nearest neighbor analysis of the four-county data (equivalent kilometers)

| $\bar{r}$ (km) | *mean(r)* (km) | $\sigma$ (km) | *z*-statistic | *p*-value |
|---|---|---|---|---|
| 4.93 | 5.66 | .148 | - 4.9 | < 0.001 |

**Boundary bias**

The mathematical derivation of *F(r)* implies a study area without boundaries, which does not occur in a real application. Therefore, some bias is incurred by using any test derived from the theoretical expression *F(r)*. Table 2 summarizes results from data simulations which explore the impact of this "boundary" bias. In the simulations, artificial cases were randomly generated within the boundary of the density equalized map. A series of sample sizes $n$ = 5, 10, 20, 30, 50, 100, 200 and 400 shows a decrease in the expected nearest neighbor distance *mean(r)* and $\sigma$ as the sample size increases. In Table 2, *mean(r)* and $\sigma$ are the theoretical mean and standard error for sample size $n$; $\bar{r}$ and $S$ are the observed mean and standard error calculated from the simulations; bias = $\bar{r}$ - *mean(r)* is the bias introduced by the boundary effect.

**Table 2.** Simulation results from random artificial cases within the four-county density equalized map. 2000 trials were performed for each sample size $n$.

| n | mean(r) (km) | $\sigma$ (km) | $\bar{r}$ (km) | S (km) | bias (km) | bias/$\sigma$ |
|---|---|---|---|---|---|---|
| 5 | 50.76 | 11.87 | 63.24 | 16.64 | 12.48 | 1.052 |
| 10 | 35.89 | 5.93 | 41.90 | 7.67 | 6.01 | 1.013 |
| 20 | 25.38 | 2.97 | 28.26 | 3.66 | 2.88 | 0.971 |
| 30 | 20.72 | 1.98 | 22.64 | 2.39 | 1.92 | 0.970 |
| 50 | 16.05 | 1.19 | 17.15 | 1.34 | 1.10 | 0.928 |
| 100 | 11.35 | 0.59 | 11.91 | 0.66 | 0.56 | 0.950 |
| 200 | 8.03 | 0.32 | 8.31 | 0.32 | 0.29 | 0.973 |
| 400 | 5.67 | 0.14 | 5.82 | 0.16 | 0.15 | 0.982 |

The observed means $\bar{r}$ are systematically higher than the theoretical means *mean(r)*. This occurs because cases near the boundary of the study area have reduced probability of having close nearest neighbors, and so their nearest neighbor distances are biased upward. The absolute bias in column 6 decreases with increasing sample size $n$; however, the bias relative to the standard error (bias/$\sigma$, in column 7) is essentially constant, approximately one standard deviation. In Table 1, after correcting $\bar{r}$ downward by one standard deviation to compensate for boundary bias, the corrected $z$-statistic is about - 6 standard deviations.

The bias is sufficiently large that an alternative approach must be used to analyze the spatial distribution of cancer cases in Figure 8.

## Scaling of nearest neighbor distances

In the preceding analysis we have expressed nearest neighbor distances in equivalent kilometers; namely, the map units of the density equalized maps. This was done in order to show the behavior of the boundary bias as a function of sample size.

Beginning with the following section, nearest neighbor distances are dimensionless; that is, they are expressed in units which are equal to the square root of *A/n*. With this convention the expressions of the previous section become:

$$F(r) = e^{-\pi r^2}$$

$$f(r) = 2\pi r\, e^{-\pi r^2}$$

$$median\ (r) = 0.470$$

$$mean\ (r) = 0.500$$

$$variance\ (\bar{r}) = \frac{0.068}{n}$$

$$\sigma = \sqrt{variance\ (\bar{r})} = \frac{0.262}{\sqrt{n}}$$

**Table 3.** Summary statistics from the nearest neighbor analysis of the four-county data (dimensionless).

| $\bar{r}$ | *mean(r)* | $\sigma$ | *z-statistic* | *p-value* |
|---|---|---|---|---|
| 0.435 | 0.500 | 0.013 | - 4.9 | < 0.001 |

As in Table 1, with $\bar{r}$ corrected downward to compensate for the boundary bias, the corrected *z*-statistic is about - 6 standard deviations.

## Nonparametric analysis

Figure 10 shows the locations of random cases generated under the assumption of equal risk. That is, every individual was assumed to have the same probability of being diagnosed as a cancer case. The distributions of Figure 10 and Figure 8 can be directly compared to see whether these two spatial patterns differ only because of chance variation. The two distributions are equally affected by the boundary bias, so no further correction is necessary.

The observed distribution of 401 nearest neighbor distances from the actual cases in Figure 8 (scaled to be dimensionless) is shown in Figure 13. The mean of this distribution, as given in Table 3, is $\bar{r} = 0.435$. The observed distribution of 401 nearest neighbor distances from the random cases in Figure 10 (also scaled to be dimensionless) is shown in Figure 14. The mean of this distribution is $\bar{r} = 0.523$.

Figure 13.

Nearest neighbor distances of 401 real cases, after density equalization. From Figure 8, ignoring the external artificial cases.

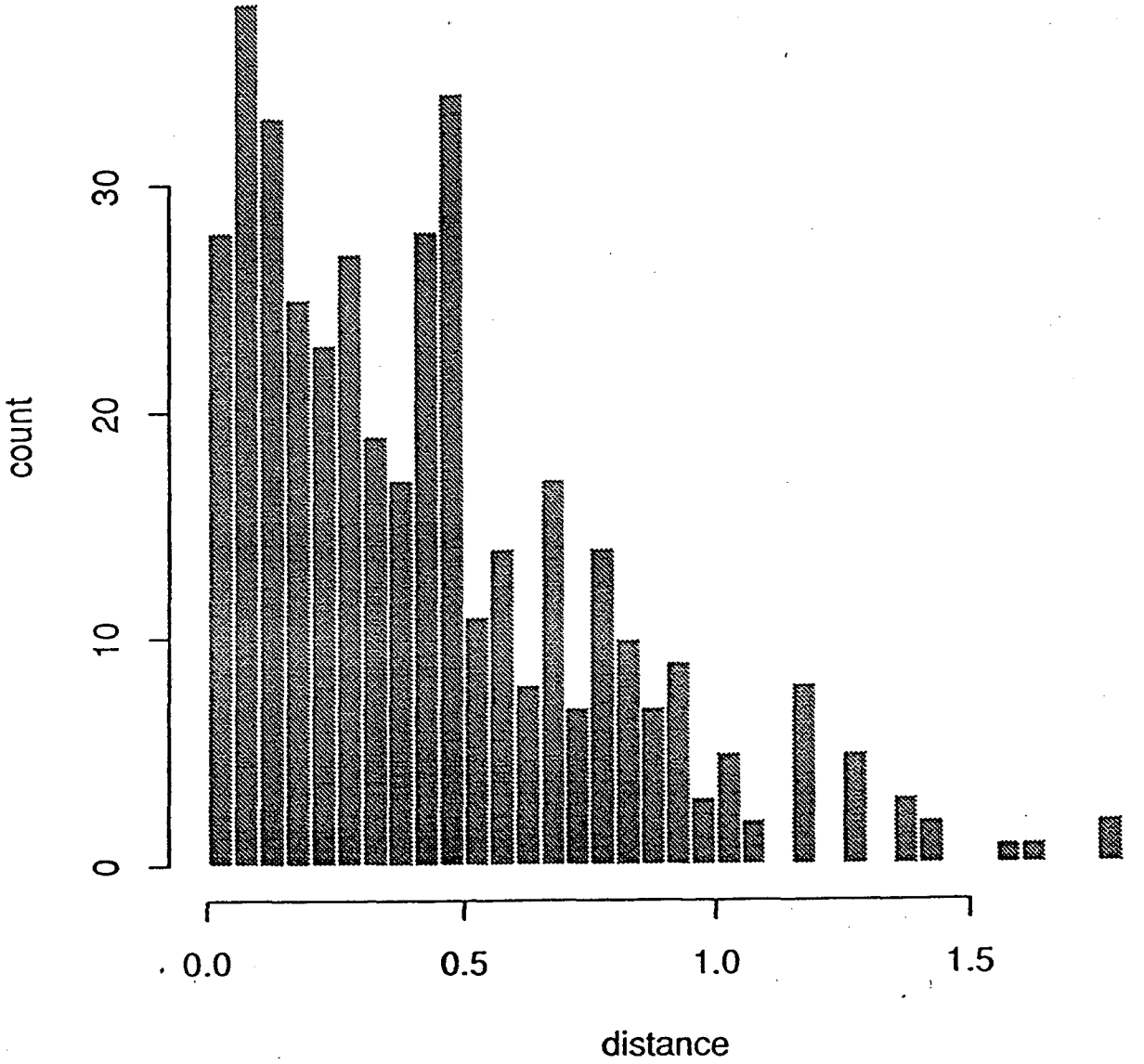
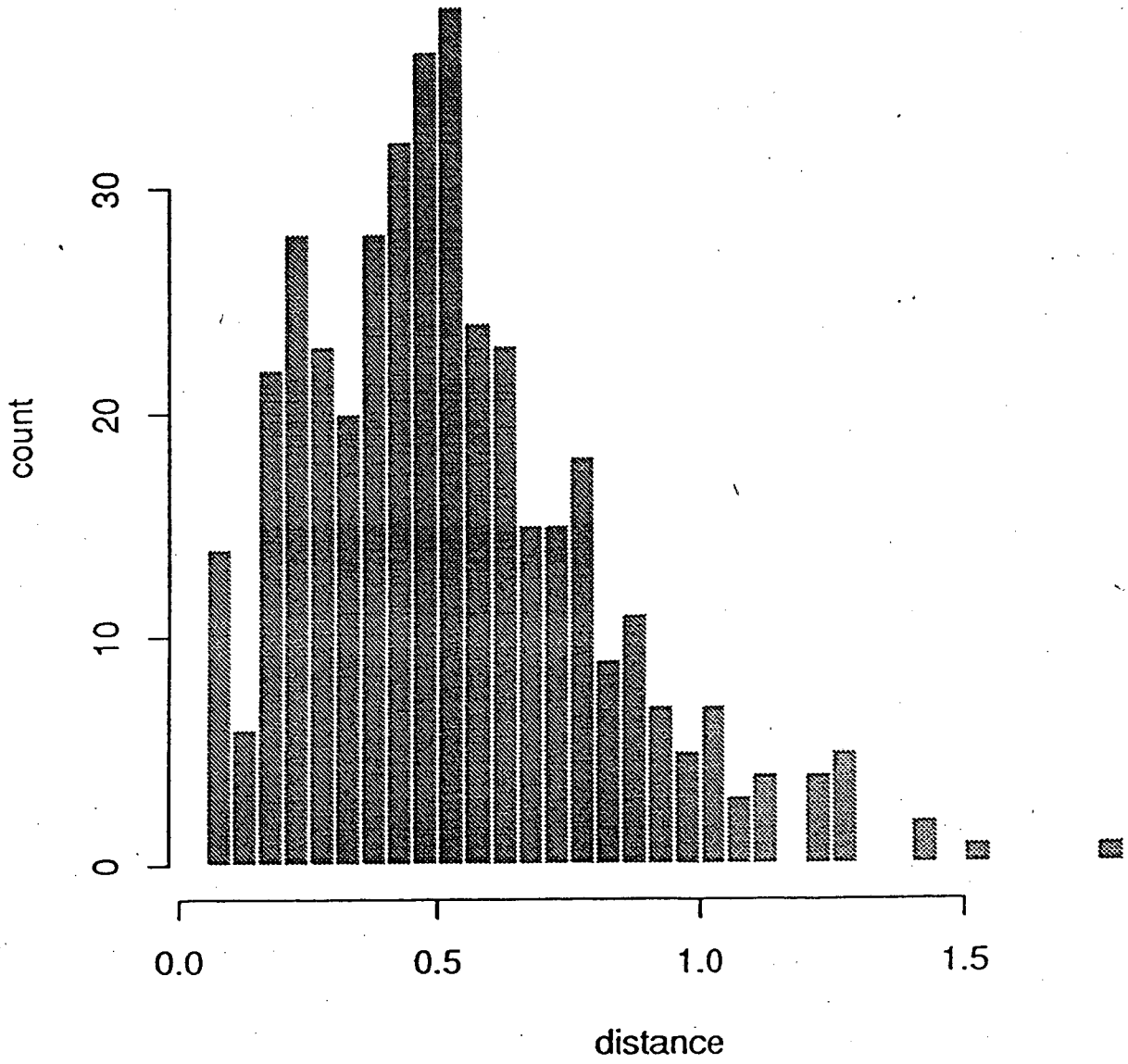


Figure 14.

Nearest neighbor distances of 401 artificial cases assuming equal risk, after density equalization. From Figure 10, ignoring the additional external artificial cases.



With the use of nonparametric smoothing, the distributions of Figure 13 and Figure 14 were combined in Figure 15 (the solid and dotted line respectively). These distributions are analogous to the theoretical density distribution *f(r)* but are non-parametric; that is, they do not depend on a statistical model.

The same data are presented in Figure 16, this time as cumulative probability distributions (the probability that a given case will have a nearest neighbor less than a certain value). These distributions are analogous to the cumulative probability distribution *F(r)*, but are also non-parametric.

Finally, the same data are presented in Figure 17, this time as a "quantile-quantile plot" or "QQ plot". Here the cumulative distribution of the real cases (y axis) is plotted against the cumulative distribution of the random cases (x axis). Agreement between the two distributions would have produced a QQ plot differing from the diagonal reference line only because of random variation.

All three figures (15, 16, and 17) demonstrate the same effect -- that the observed cases have an excess of small nearest neighbor distances relative to randomly generated cases. The same effect is at least partially responsible for the *z*-statistic of - 4.9 noted in Table 1 and Table 3.

The likelihood that a transformed map will detect spatial patterns is related to the size of the subareas used to transform the map. In the four-county map, the subareas are the 262 census tracts. On the other hand, the locations of cancer cases reflect exact longitude and latitude based on residential address. This degree of precision in the case data permits one to observe clustering that may have nothing to do with disease, and which the DEMP technique cannot remove due to limitations of the map files and population data.

Figure 15.

Estimated densities -- real cases and random cases.

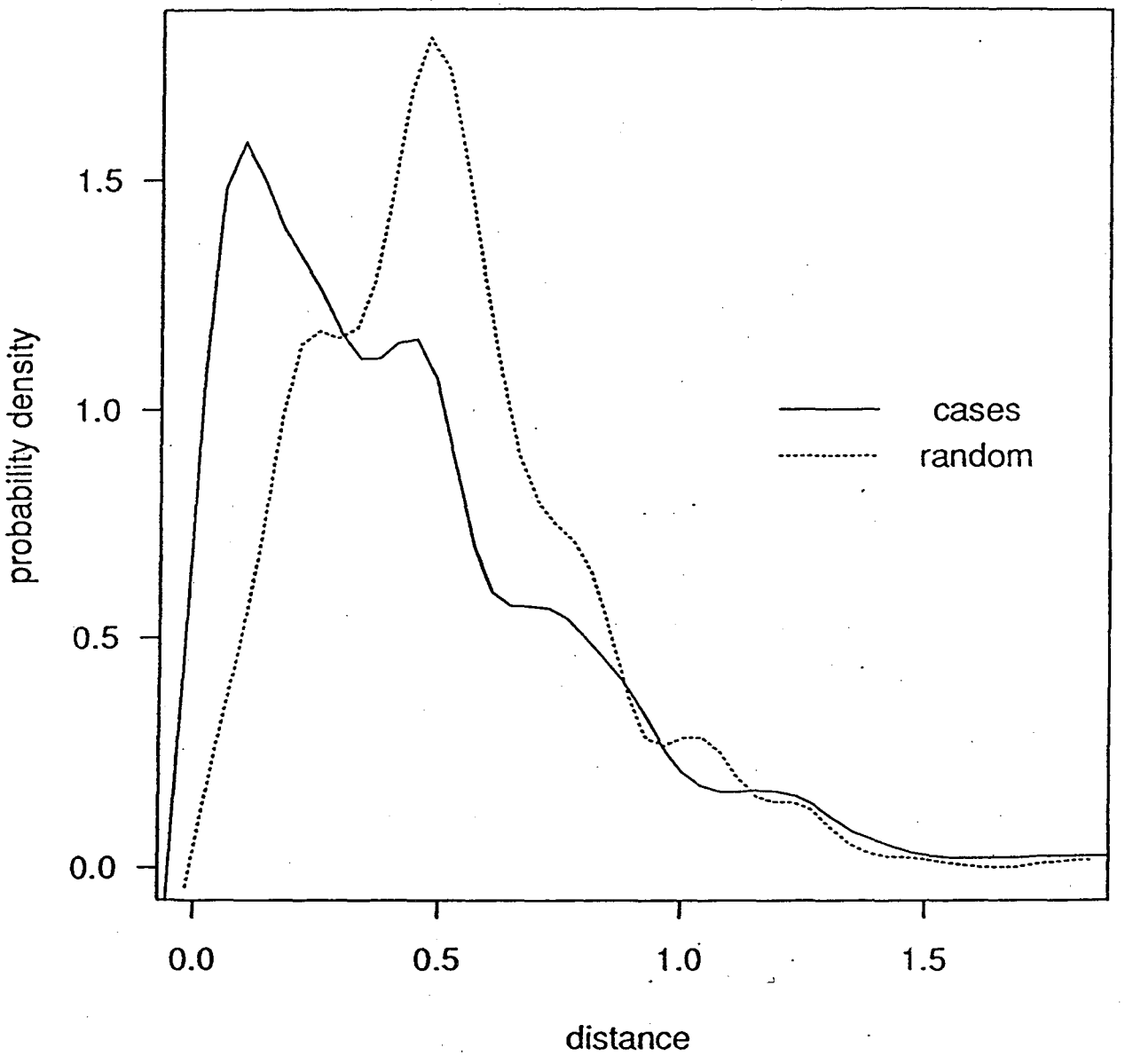


Figure 16.

Cumulative distributions -- real cases and random cases.

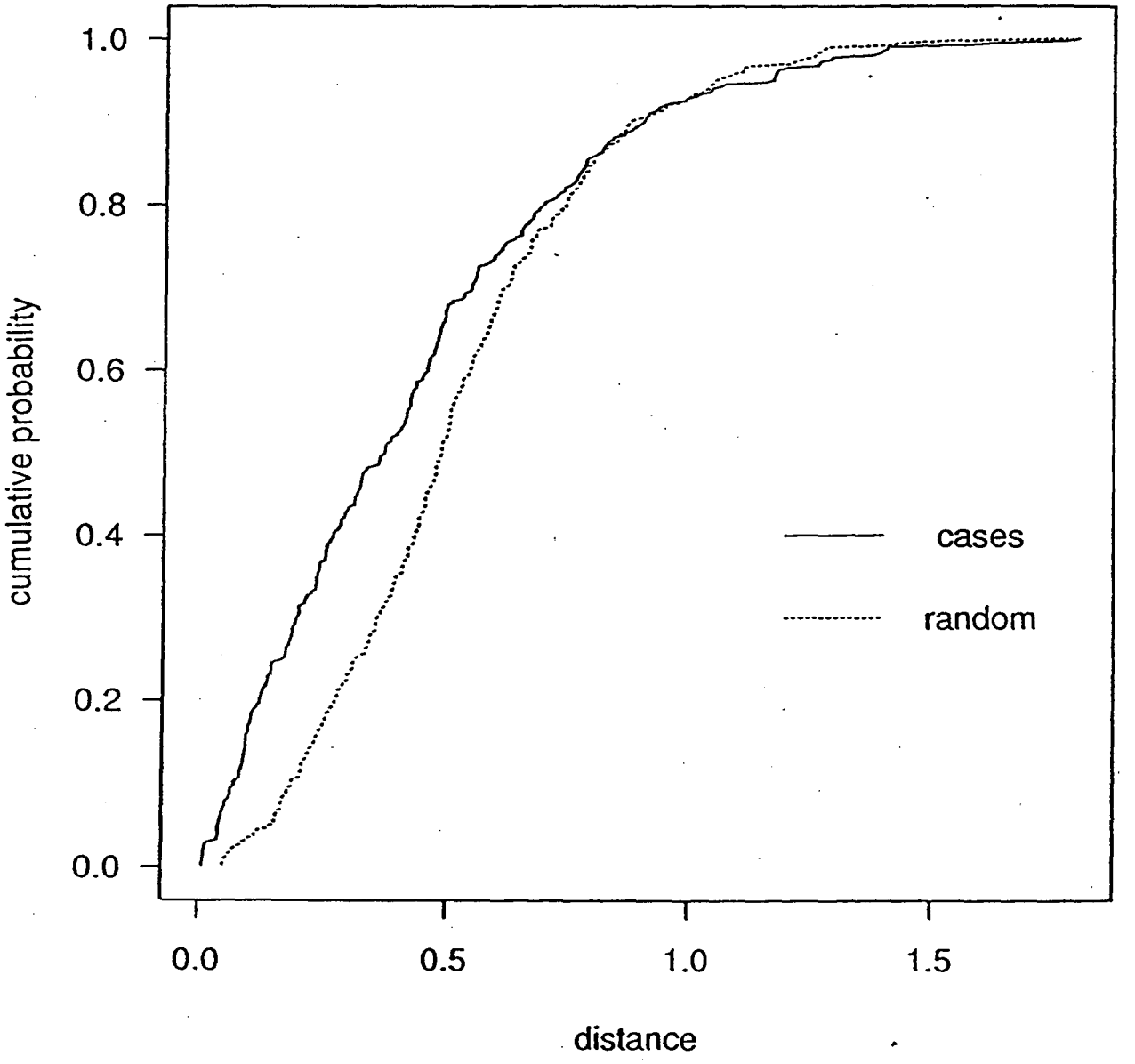
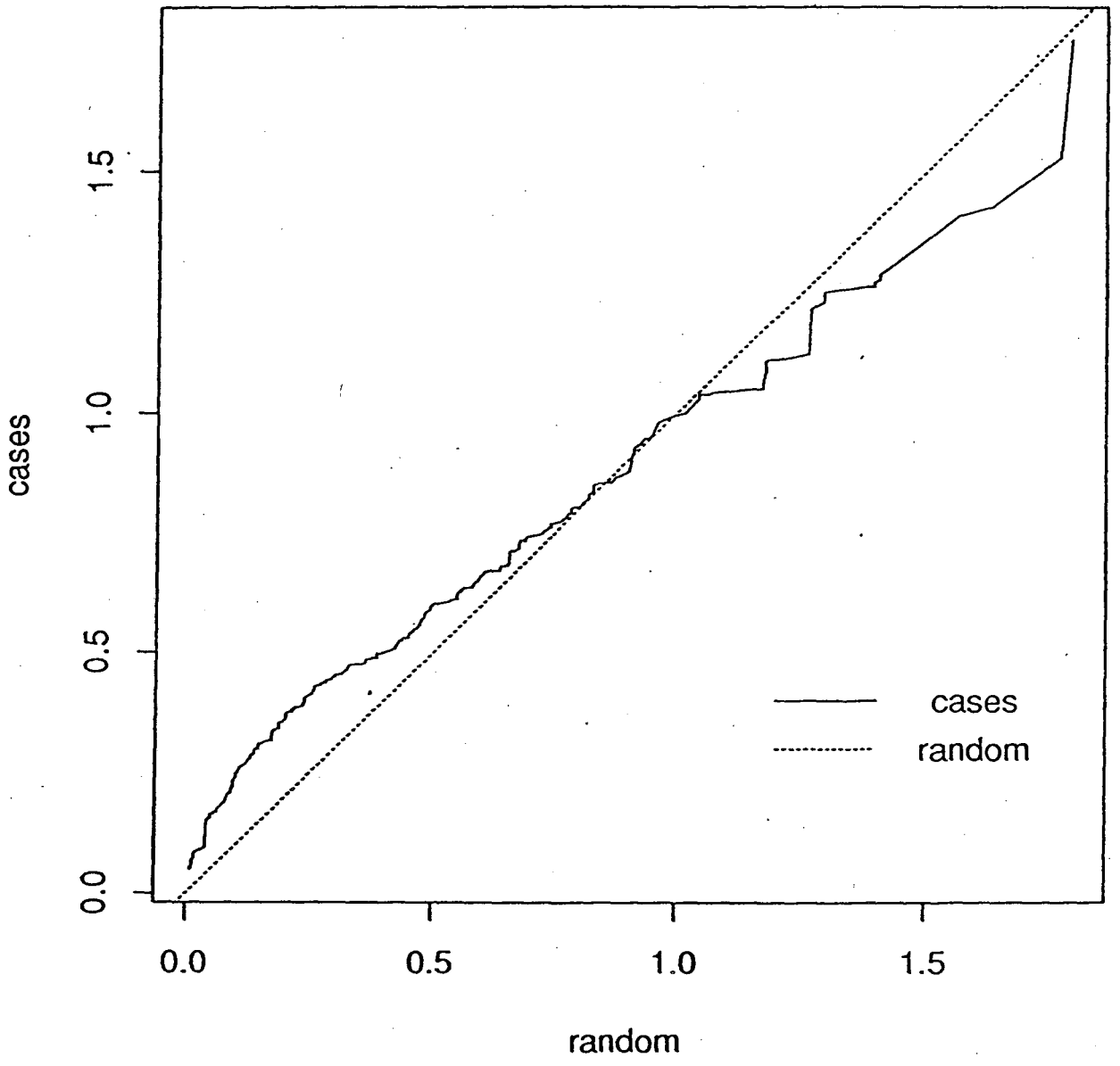


Figure 17.

QQ plot -- real cases versus random cases.



**Elimination of clusters within individual tracts**

In the present analysis we are interested in the variation of disease rates relative to a population density which is assumed constant. To remove the clustering of cases within individual tracts, each of the 401 cases was reassigned to a random location within its own tract. This process brings the case data into conformity with the assumption, implicit in the DEMP technique, that the observations are uniformly distributed within the subareas used to make the transformation. The case data, so adjusted and density equalized (Figure 12) were subjected to the same non-parametric analysis as the original case data in Figure 8. The results are presented in Figures 18, 19, 20, and 21 in exactly the same format as the original case data in Figures 13, 15, 16, and 17, respectively. In the revised plots, the differences between the adjusted case data and the purely random cases are seen to be greatly reduced.

Figure 18.

Nearest neighbor distances of 401 real cases, each plotted at a random location in its own tract, after density equalization. From Figure 12, ignoring the additional external random cases.

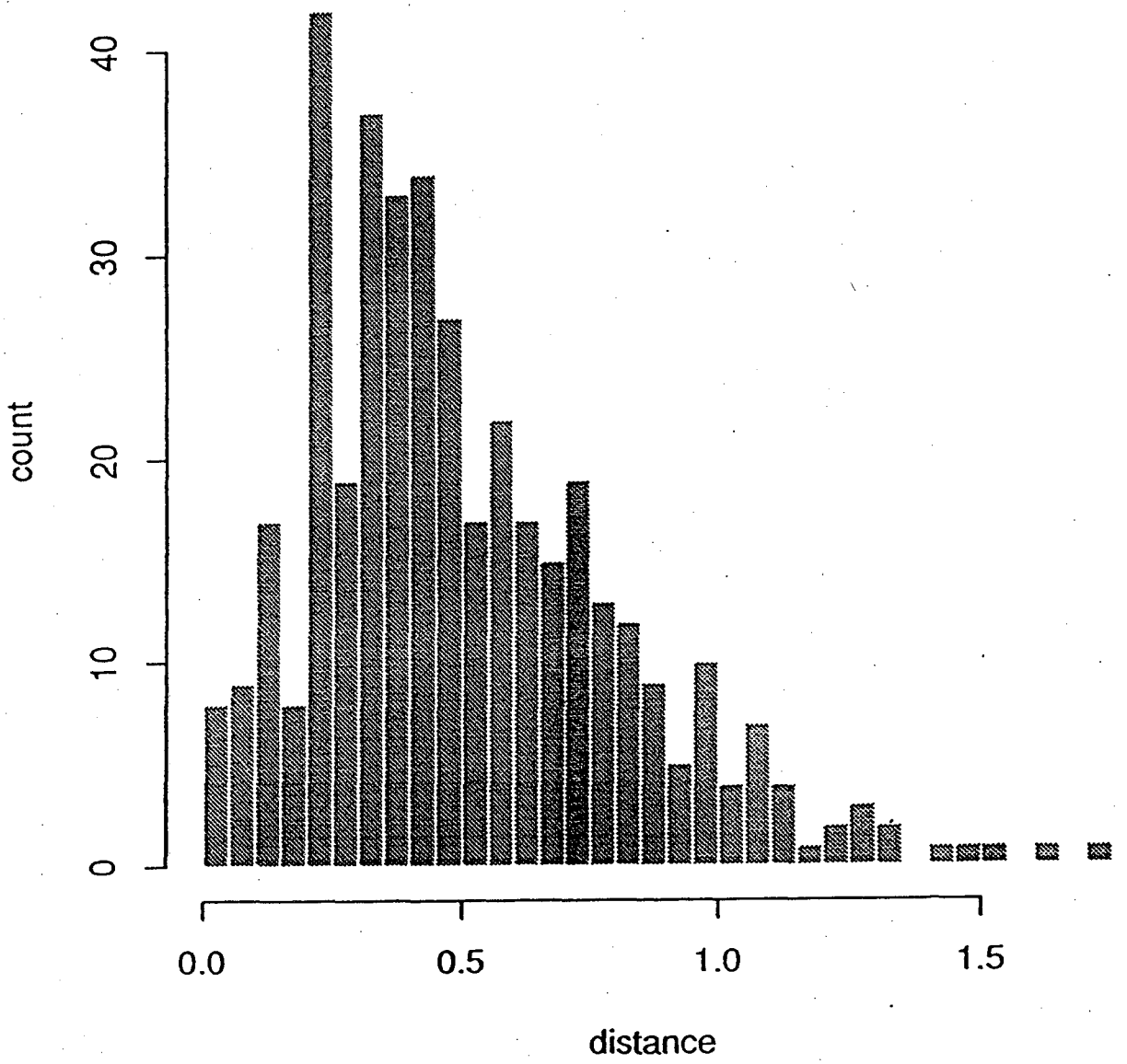


Figure 19.

Estimated densities -- cases at random location in tract, and random cases.

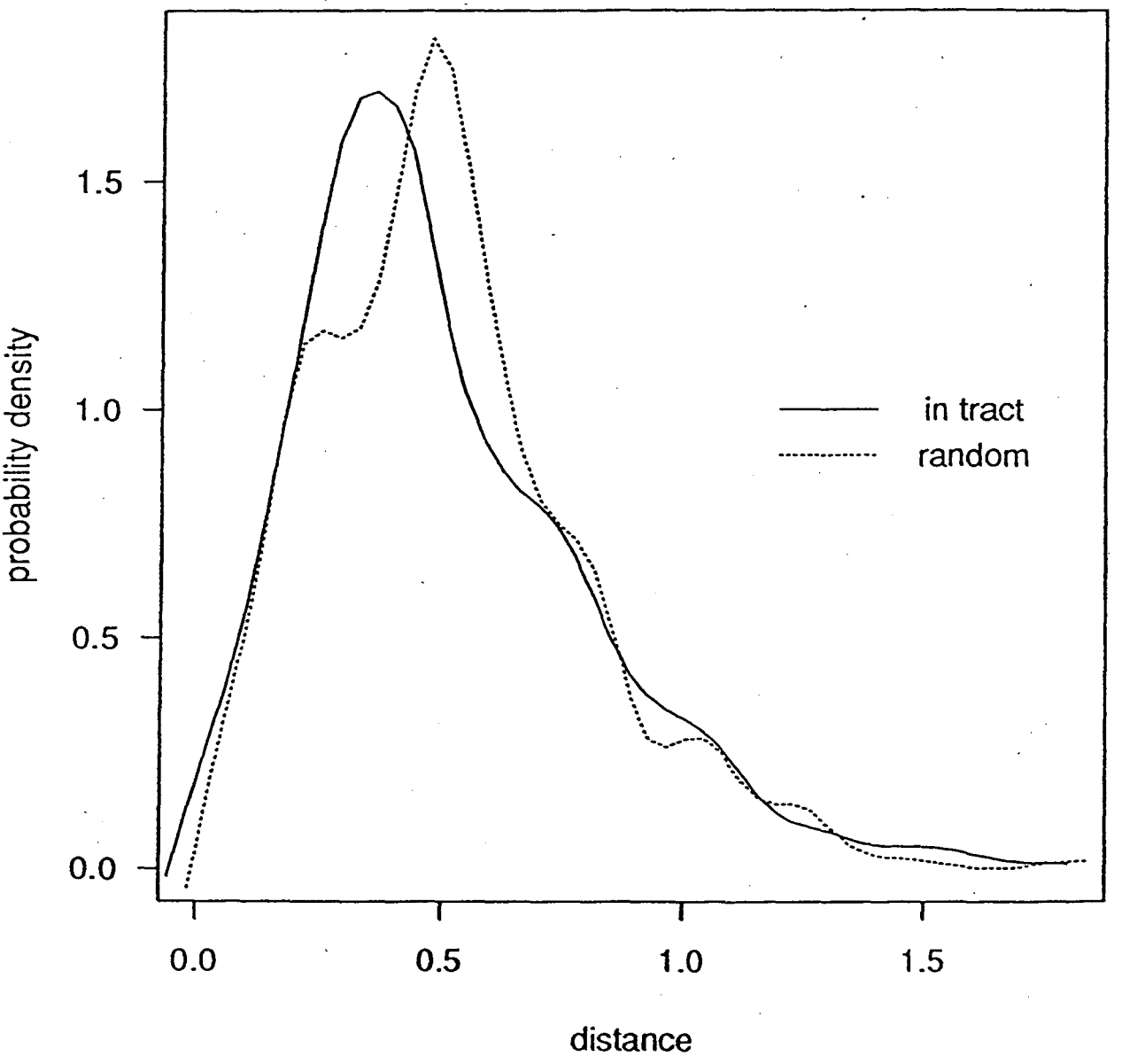


Figure 20.

Cumulative distributions -- cases at random location in tract, and random cases.

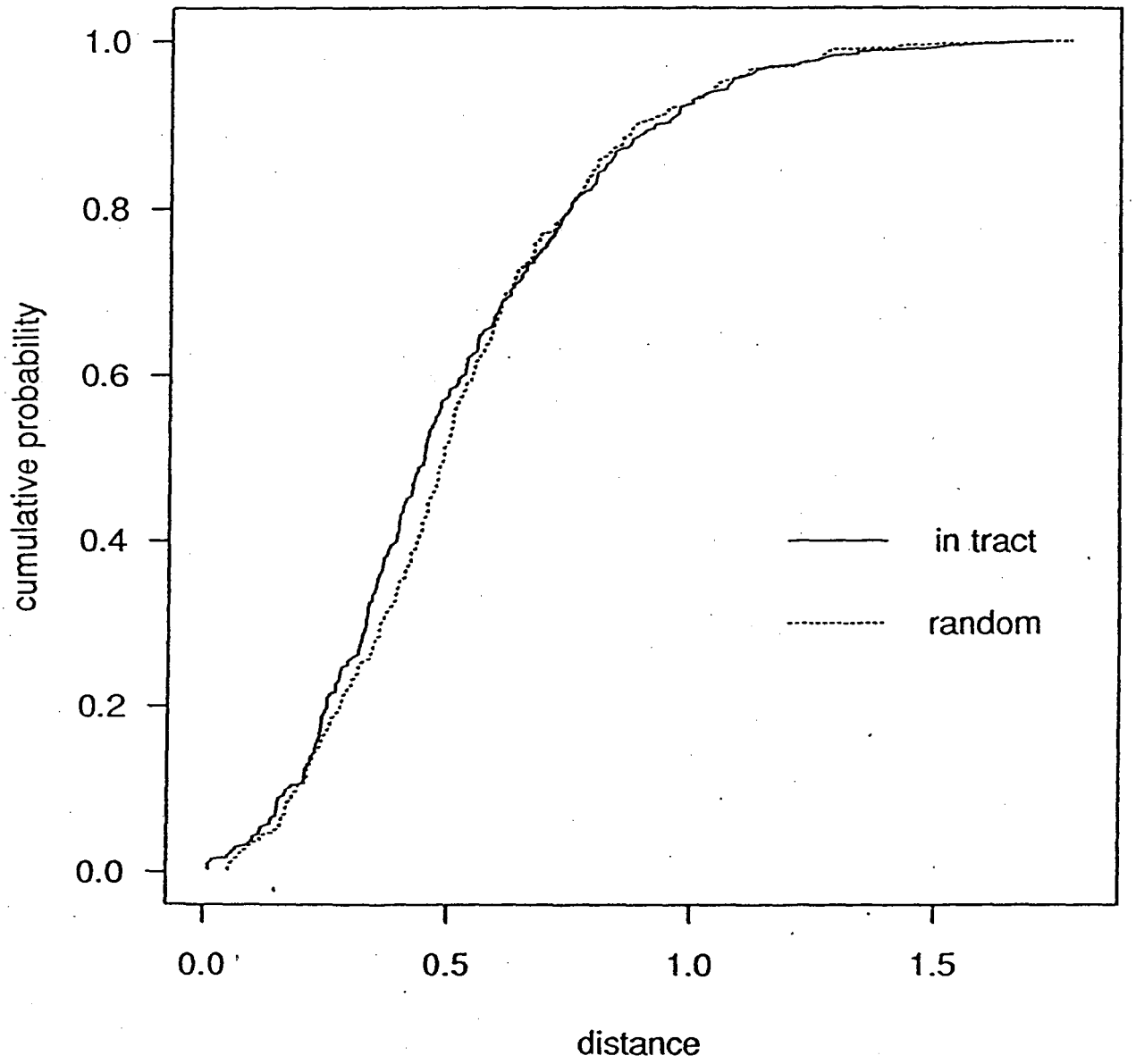
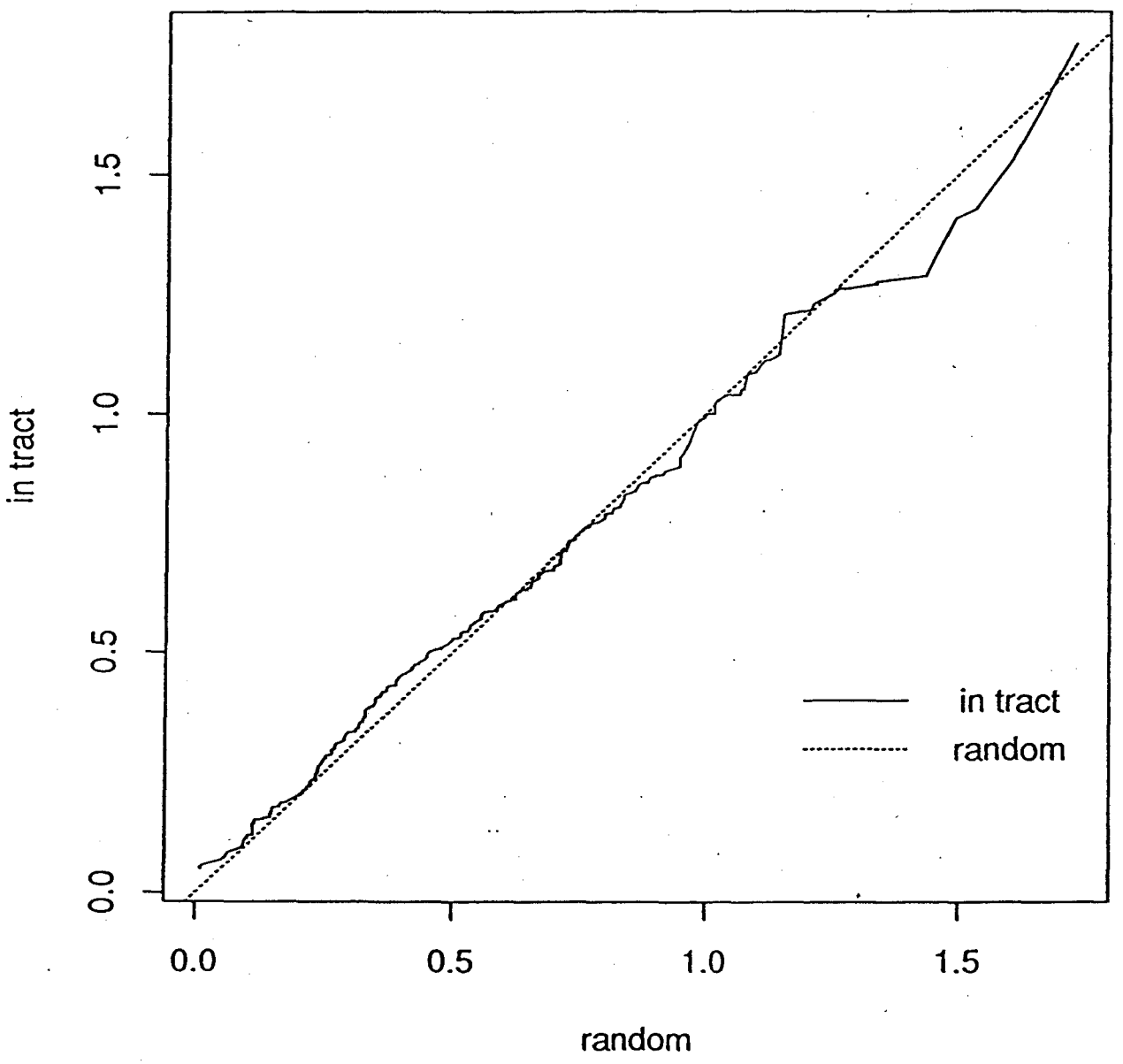


Figure 21.

QQ plot -- cases at random location
in tract, versus random cases.



**Summary of nearest neighbor distances, without external cases**

Table 4 summarizes measurements of $\bar{r}$, the mean nearest neighbor distance (dimensionless) from the density equalized maps in Figures 8, 10, and 12 respectively. The random external cases in those figures were not used. Standard errors $\hat{\sigma}$ were estimated from the jackknife technique [SELV91]

$$\hat{\sigma} = \frac{1}{n} \sum_{i=1}^{n} (n-1) \, ( \bar{r}_i - \bar{r} )^2$$

where $\bar{r}_i$ is the modified sample mean obtained by omitting the contribution of case $i$ to the sample mean.

The value of $\bar{r}$ from the actual cases in Figure 8 is .435, with an estimated standard error $\hat{\sigma}$ = .017. The random cases of Figure 10 (sample r1) yielded $\bar{r}$ = .523 and $\hat{\sigma}$ = 0.014. The data of Figure 12 (sample t1); namely, the cases reassigned to arbitrary locations within their own tract, yielded $\bar{r}$ = .507 and $\hat{\sigma}$ = .015.

For the single samples r1 and t1 in Figures 10 and 12 respectively:

$\bar{r}$ (cases, actual location) is lower than $\bar{r}$ (random samples)
by (.435-.523)/.014 = - 6.3 standard deviations;

$\bar{r}$ (cases, random in tract) is lower than $\bar{r}$ (random samples)
by (.507-.523)/.014 = - 1.1 standard deviation.

Twenty different random samples r1-r20 were generated exactly as the random cases in r1. The overall mean was .508. The variance among the 20 samples produced a standard error estimate $S$ = .014, in agreement with the jackknife estimate $\hat{\sigma}$ = .014. Twenty different random locations t1-t20 in the tract of each case were also generated, exactly as the random locations in t1. The overall mean was .496. The variance among the 20 samples produced a standard error estimate $S$ = .012. This value is smaller than the jackknife estimate $\hat{\sigma}$ = .015 since the 20 samples t1-t20 are not independent.

For the 20 samples r1-r20 and t1-t20:

$\bar{r}$ (cases, actual location) is lower than $\bar{r}$ (random samples)
by (.435-.508)/.014 = - 5.2 standard deviations;

$\bar{r}$ (cases, random in tract) is lower than $\bar{r}$ (random samples)
by (.496-.508)/.014 = - 0.9 standard deviation.

The boundary bias need not be considered here, since it affects all the samples equally.

**Summary of nearest neighbor distances, with external cases**

In Table 5 we again summarize measurements of $\bar{r}$, the mean nearest neighbor distance (dimensionless) from the density equalized maps in Figures 8, 10, and 12 respectively. This time, however, the random external cases in each figure were considered as nearest neighbor candidates, in assigning nearest neighbors to each of the 401 points inside the boundary. This is one method of correcting the boundary bias discussed earlier. Jackknife standard error estimates were calculated as in Table 4.

As expected, the results are similar to those in Table 4, except that all values of $\bar{r}$ have been systematically shifted downward, relative to those in Table 4, by about one standard deviation.

For the single samples r1 and t1 in Figures 10 and 12 respectively:

$\bar{r}$ (cases, actual location) is lower than $\bar{r}$ (random samples)
by (.425-.512)/.015 = - 5.8 standard deviations;

$\bar{r}$ (cases, random in tract) is lower than $\bar{r}$ (random samples)
by (.499-.512)/.015 = - 0.9 standard deviation.

For the 20 samples r1-r20 and t1-t20:

$\bar{r}$ (cases, actual location) is lower than $\bar{r}$ (random samples)
by (.426-.494)/.015 = - 4.5 standard deviations;

$\bar{r}$ (cases, random in tract) is lower than $\bar{r}$ (random samples)
by (.481-.494)/.015 = - 0.9 standard deviation.

The results from Tables 4 and 5 are consistent with each other, and with the results found earlier:

(1) The boundary effect biases measured values of $\bar{r}$ upward by about one standard deviation. The following estimates (2) and (3) are corrected for the boundary bias, or are unaffected by it.

(2) The observed cases have a measured value of $\bar{r}$ about five or six standard deviations lower than that expected under the null hypothesis of equal risk. This includes the effect of within-tract clustering that cannot be equalized with the available data.

(3) If each case is plotted at a random location in its own tract to eliminate the effect of within-tract clustering, the resulting value of $\bar{r}$ is only about one standard deviation lower than that expected under the null hypothesis. This residual effect is due entirely to clustering of cases in *different* tracts.

**Table 4. Mean nearest neighbor distance, no external cases**

$\bar{r}$ = mean nearest neighbor distance (dimensionless)
$\hat{\sigma}$ = standard error of $\bar{r}$, from jackknife method
$S$ = standard error of $\bar{r}$, from variance among 20 samples

| | $\bar{r}$ | $\hat{\sigma}$ | $S$ |
|---|---|---|---|
| Figure 8. | | | |
| 401 case locations | | | |
| actual data | .435 | .017 | |
| Figure 10. | | | |
| 401 random locations assuming equal risk | | | |
| sample r1 | .523 | .014 | |
| samples r1-r20 | .508 | | .014 |
| Figure 12. | | | |
| 401 random locations in same tract as case | | | |
| sample t1 | .507 | .015 | |
| 20 samples t1-t20 | .496 | | .012 |

**Table 5. Mean nearest neighbor distance, with external cases**

$\bar{r}$ = mean nearest neighbor distance (dimensionless)
$\hat{\sigma}$ = standard error of $\bar{r}$, from jackknife method
$S$ = standard error of $\bar{r}$, from variance among 20 samples

| | $\bar{r}$ | $\hat{\sigma}$ | $S$ |
|---|---|---|---|
| Figure 8. | | | |
| 401 case locations | | | |
| sample e1 | .425 | .016 | |
| samples e1-e20 | .426 | | .002 |
| Figure 10. | | | |
| 401 random locations assuming equal risk | | | |
| sample r1 | .512 | .013 | |
| samples r1-r20 | .494 | | .015 |
| Figure 12. | | | |
| 401 random locations in same tract as case | | | |
| sample t1 | .499 | .014 | |
| 20 samples t1-t20 | .481 | | .013 |

# CONCLUSIONS

The most important accomplishment described in this report was the successful density equalization of a complex and highly non-uniform map. For the first time, the practicality of the DEMP method for a substantial problem has been demonstrated on a computer of moderate size.

Possibilities exist for improvement. The four-county problem with 262 subareas required about 20 hours on a SPARC 10 work station. Improvement by a factor of 2 to 5 can be obtained through simple code optimization. An additional factor of 10 or even 100 can be achieved on a massively parallel computer.

Numerous errors and irregularities in the input map files were successfully eliminated, by automatic procedures which can be re-used to process map files from other geographic areas.

Cross-checks demonstrated that the density equalization, though not perfect, is sufficiently "clean" to permit unbiased analysis of the case locations on the density equalized map.

The utility of the DEMP map was demonstrated by applying one simple analytic method - nearest neighbor analysis - to the transformed case locations. This analysis is only one of many simple techniques available. Measurements of $\bar{r}$, the mean nearest neighbor distance, yielded the following results:

(1) The boundary effect biases measured values of $\bar{r}$ upward by about one standard deviation, relative to values expected under the null hypothesis of equal risk. The following estimates (2) and (3) are corrected for the boundary bias, or are unaffected by it.

(2) The observed cases have a measured value of $\bar{r}$ about five or six standard deviations lower than that expected under the null hypothesis of equal risk. This includes the effect of within-tract clustering that cannot be equalized with the available map files and population data.

(3) If each case is plotted at a random location in its own tract to eliminate the effect of within-tract clustering, the resulting value of $\bar{r}$ is only about one standard deviation lower than that expected under the null hypothesis of equal risk. This residual effect is due entirely to clustering of cases in *different* tracts. We conclude that the nearest neighbor analysis provides no evidence for clustering among different census tracts.

Regarding the epidemiological conclusions from the four-county data set, the negative findings of the earlier DHS report are basically confirmed. However, epidemiologic conclusions cannot be drawn at this time because the population data needed for a correct analysis are unavailable. In addition, stratification of the data by risk factors such as age group and race is required for a thorough epidemiologic investigation.

The DEMP technique is an innovative and powerful tool that is just now becoming practical for problems of reasonable size. It 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.

# APPENDIX A.

## RANDOM AND THEORETICAL DISTRIBUTIONS

As a check that the DEMP algorithm is not creating artificial clusters, we have verified that the density equalized random cases in Figure 10 have the correct nearest neighbor distribution. An excess of cases with small nearest neighbor distances could occur if the DEMP algorithm were not working properly.

Figure A-1 compares the observed random distribution (solid line) with the theoretical density function (dashed line). The dotted line is corrected for the boundary bias, by including (random) nearest neighbors outside the study area.

Figures A-2 and A-3 present the same data as a cumulative probability function and a QQ plot, respectively.

After correction for the boundary bias, there is no significant discrepancy between the theoretical and observed random distributions.

Figure A-1.

Estimated densities -- random cases,
corrected random cases, and theoretical.

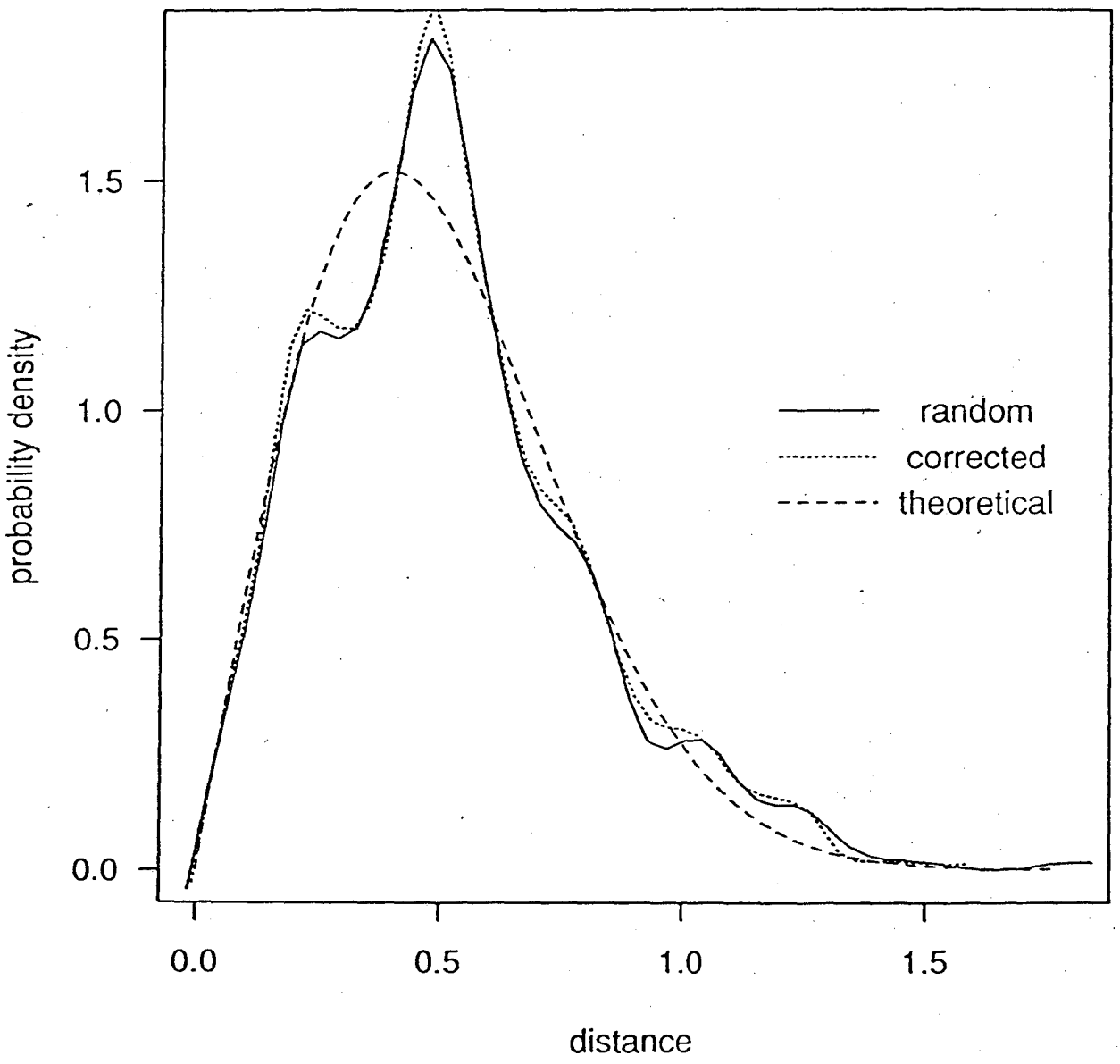


Figure A-2.

Cumulative distributions -- random cases,
corrected random cases, and theoretical.

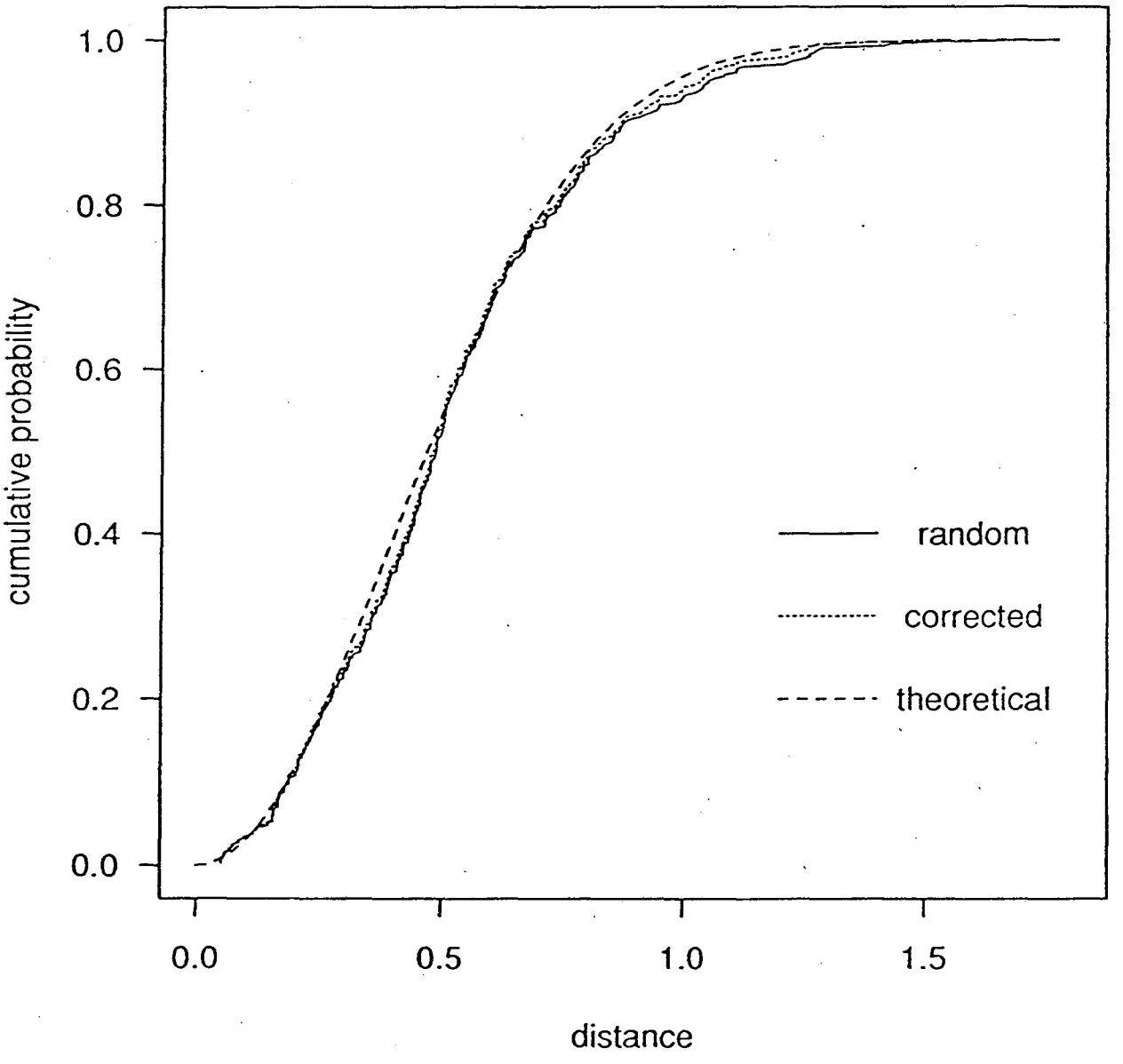
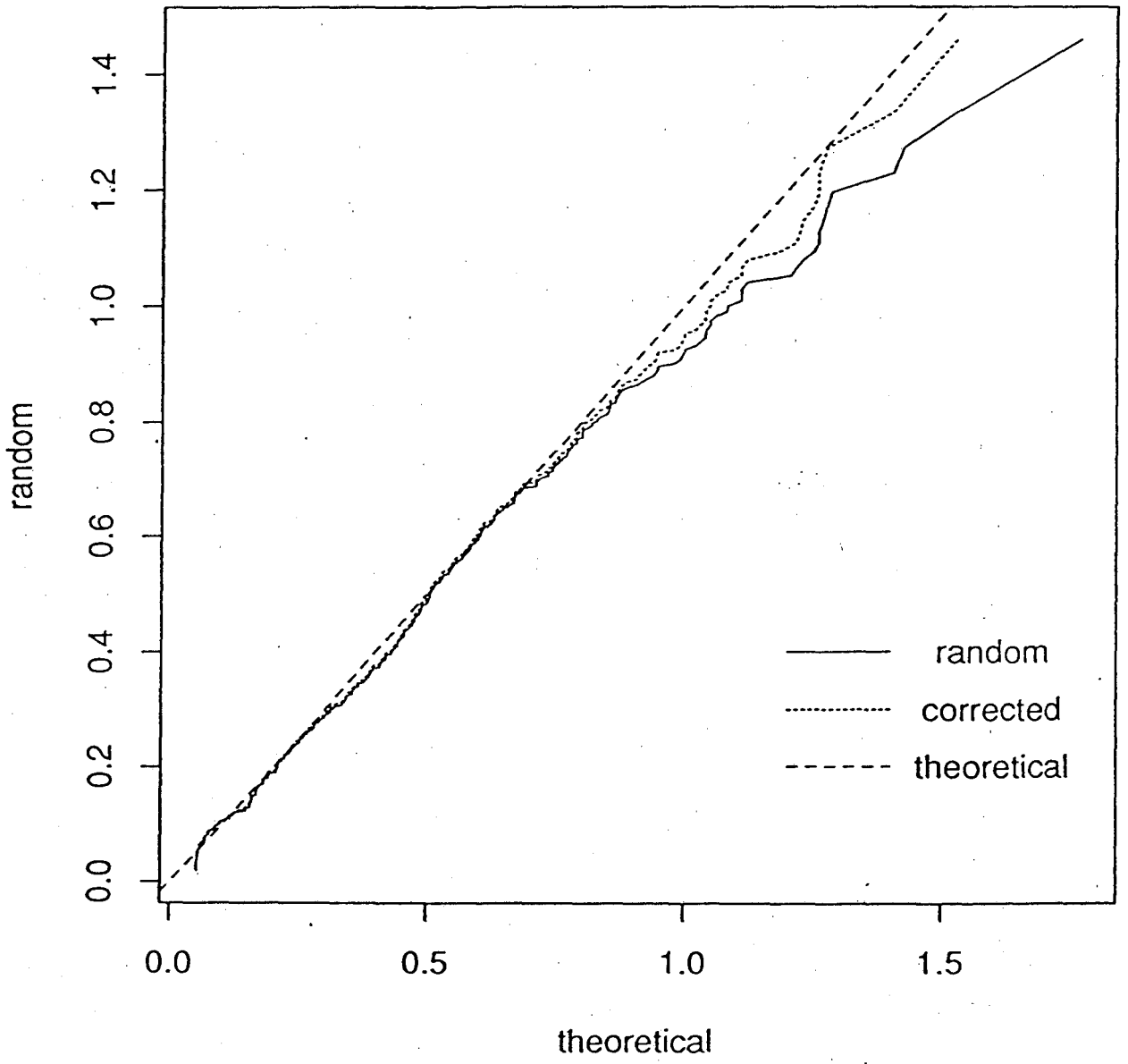


Figure A-3.

QQ plot -- random cases and corrected
random cases versus theoretical.



APPENDIX B.

## DESCRIPTION OF THE PRIMARY RUN HEX10

The primary run hex10 was used for all the statistical analysis in this report. Ten equal steps were taken. The population units were 1212 hexagons, obtained by bisecting the boundary segments of the 1212 triangles in Figure 5. The run included 1212 × 7 = 8484 boundary points and 16,441 non-boundary points:

| | |
|---|---|
| 401 | case locations |
| 8020 | 20 samples, random case locations |
| 8020 | 20 samples, random locations in tract |
| 16441 | total non-boundary points |

Table B-1 is the history of the hex10 run, including the polygon type (hexagon), step size $c_i$, computing time (in a Sun SPARC 10 work station), number of total polygons and negative-area polygons, and value of *hsum* after the step. For a technical description of the program, see [CLOS94]. The parameters *minangle* and *minseg*, not used in the hex10 run, are defined in Appendix C. The dimensionless quantity *hsum*, which is zero for a perfectly equalized map, is an area-weighted average of the squared relative difference between adjusted polygon areas and target polygon areas:

$$hsum = \frac{1}{atotal} \sum_{k=1}^{npoly} atarg_k \left( \frac{anow_k}{atarg_k} - 1 \right)^2$$

where *npoly* is the total number of polygons, $anow_k$ and $atarg_k$ are the present and target areas of polygon $k$, and *atotal* is the sum of all target areas.

**Table B-1. History of run hex10**

| step | poly type | $c_i$ | *min angle* | *min seg* | time hrs | poly tot | poly < 0 | *hsum* |
|---|---|---|---|---|---|---|---|---|
| 0 | hex | 0 | NA | NA | | 1212 | 0 | 17.04 |
| 1 | hex | 1/10 | NA | NA | 2.0 | 1212 | 0 | 12.10 |
| 2 | hex | 1/9 | NA | NA | 2.0 | 1212 | 0 | 8.30 |
| 3 | hex | 1/8 | NA | NA | 2.0 | 1212 | 1 | 5.47 |
| 4 | hex | 1/7 | NA | NA | 2.0 | 1212 | 3 | 3.43 |
| 5 | hex | 1/6 | NA | NA | 2.0 | 1212 | 3 | 2.03 |
| 6 | hex | 1/5 | NA | NA | 2.0 | 1212 | 6 | 1.11 |
| 7 | hex | 1/4 | NA | NA | 2.0 | 1212 | 10 | 0.552 |
| 8 | hex | 1/3 | NA | NA | 2.0 | 1212 | 22 | 0.667 |
| 9 | hex | 1/2 | NA | NA | 2.0 | 1212 | 23 | 0.134 |
| 10 | hex | 1/1 | NA | NA | 2.0 | 1212 | 40 | 0.142 |
| tot | | | | | 20.0 | | | |

Figure B-1 shows the present and target area of each hexagon after step 0 (the initial map of Figure 5). Hexagons to be expanded or reduced lie below or above a 45 degree line, respectively.

Figure B-2 shows the same data after step 5. Three hexagons have negative areas and no longer contribute to the mapping.

Figure B-3 shows the same data for the (approximately) density equalized map, after step 10. If the density equalization were perfect, all points would lie exactly on a 45 degree line. In Figure B-3, 40 hexagons have negative areas. In addition, an unknown number of positive-area hexagons may have boundaries that self-intersect.

Figures 5, B-4, and B-5 show the tract boundaries (solid) and hexagon boundaries (dotted) after step 0, 5 and 10 respectively. In Figure B-5 one can distinguish a few overlapping hexagon boundaries. With a little effort one can determine which areas correspond to each other on the three maps.

Figures B-6, B-7, and B-8 show the locations of 8020 random cases after step 0, 5, and 10 respectively. In Figure B-8 the uniform density of the transformed points shows that the DEMP algorithm is transforming points approximately correctly. Near the center of the map some minor distortion results from the overlap of a few hexagons. A few points fall slightly outside the map boundary due to insufficient detail in the boundary segments.

Figures B-3, B-5, and B-8 indicate that the density equalization is not perfect, but the problems are minor. In any event, the statistical comparisons among

(a) case locations
(b) random cases
(c) random locations within tract

are valid since (a), (b) and (c) were transformed identically; perfect density equalization is not essential.

Figure B-1.

Present areas versus target areas, initial map.

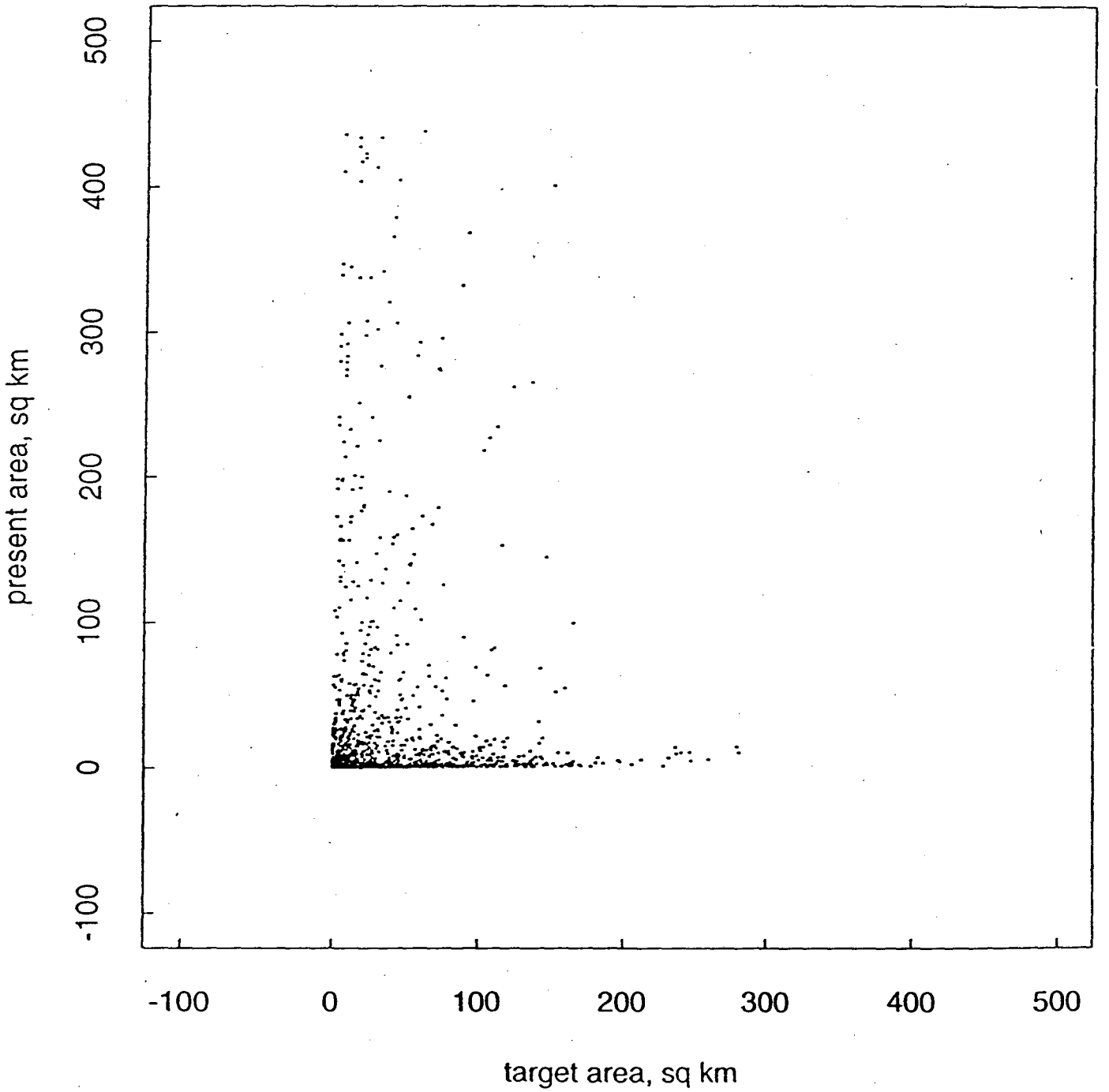


Figure B-2.

Present areas versus target areas,
run hex10, after step 5.

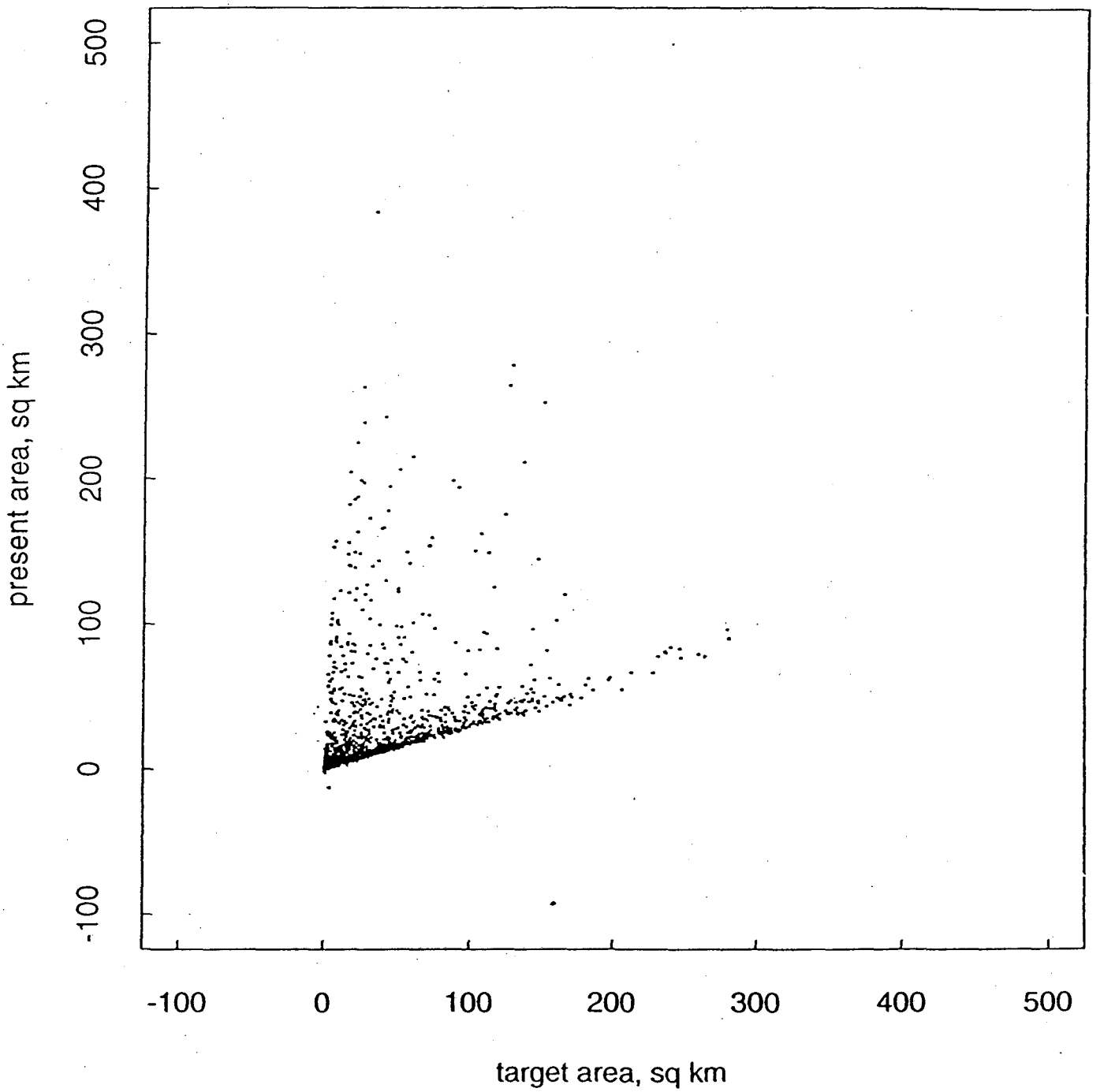


Figure B-3.

Present areas versus target areas,
run hex10, after step 10.

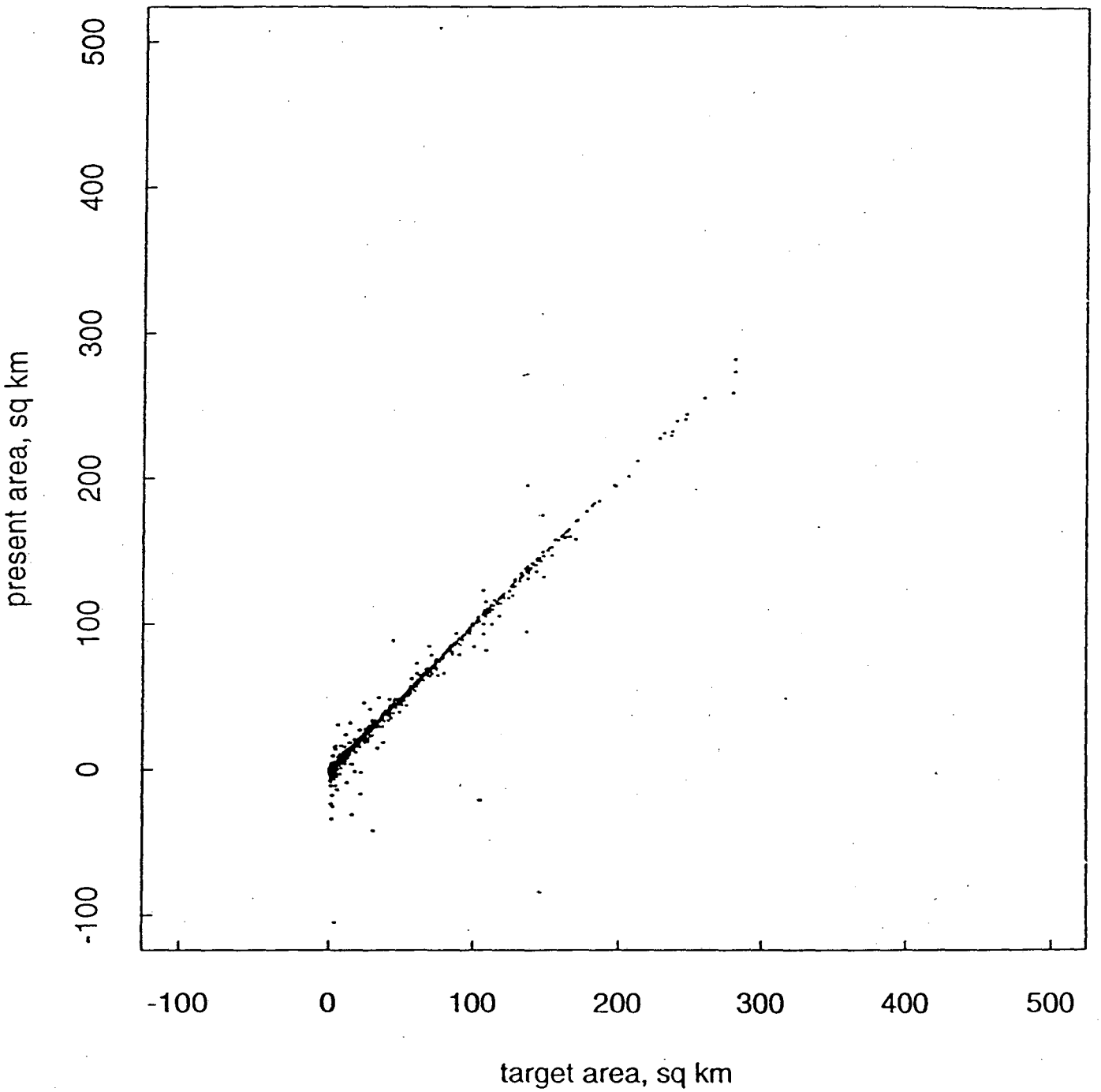


Figure B-4.

Tract boundaries, hexagon boundaries, and 401 cases;
run hex10, after step 5.

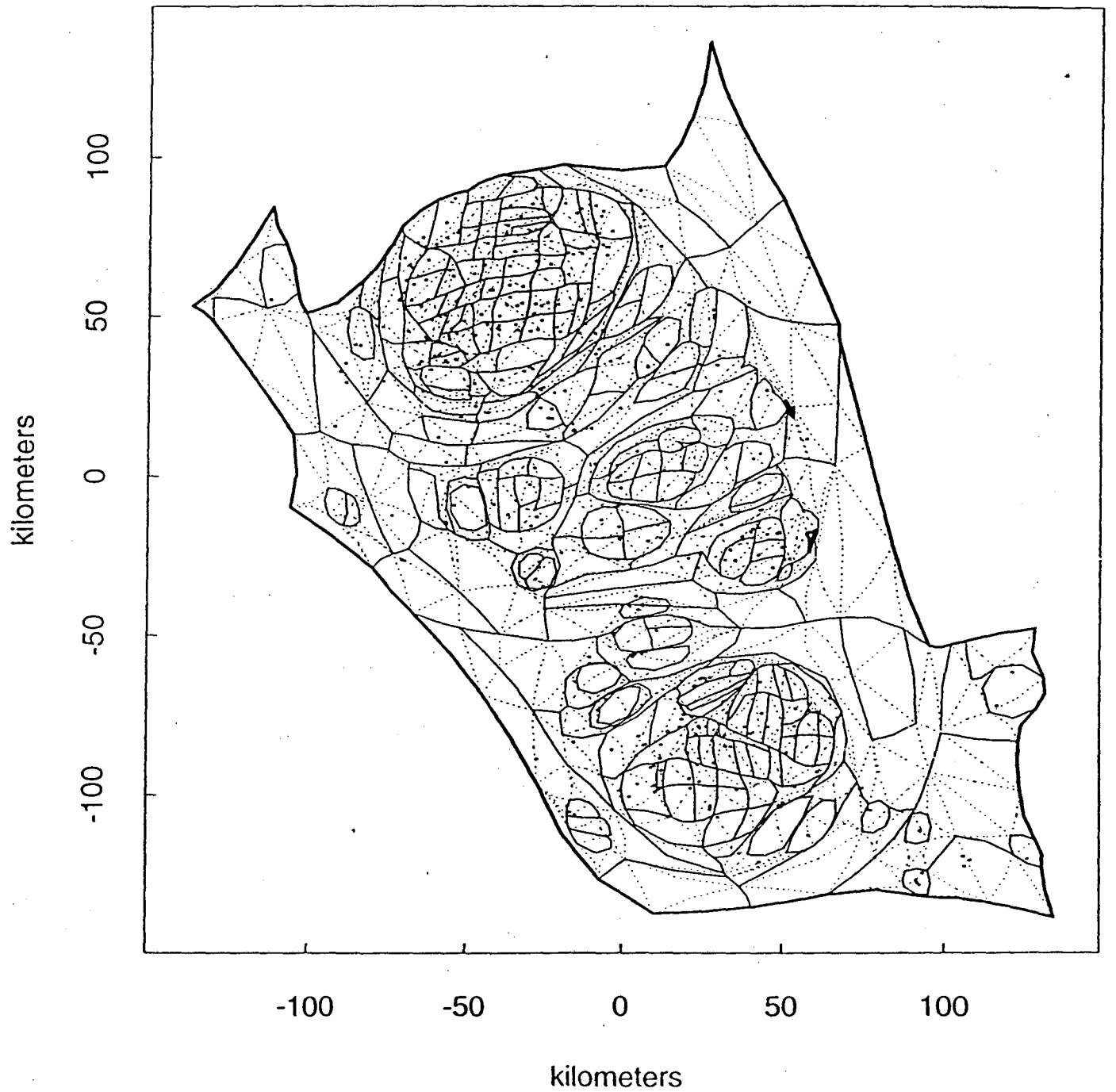


Figure B-5.

Tract boundaries, hexagon boundaries, and 401 cases;
run hex10, after step 10.

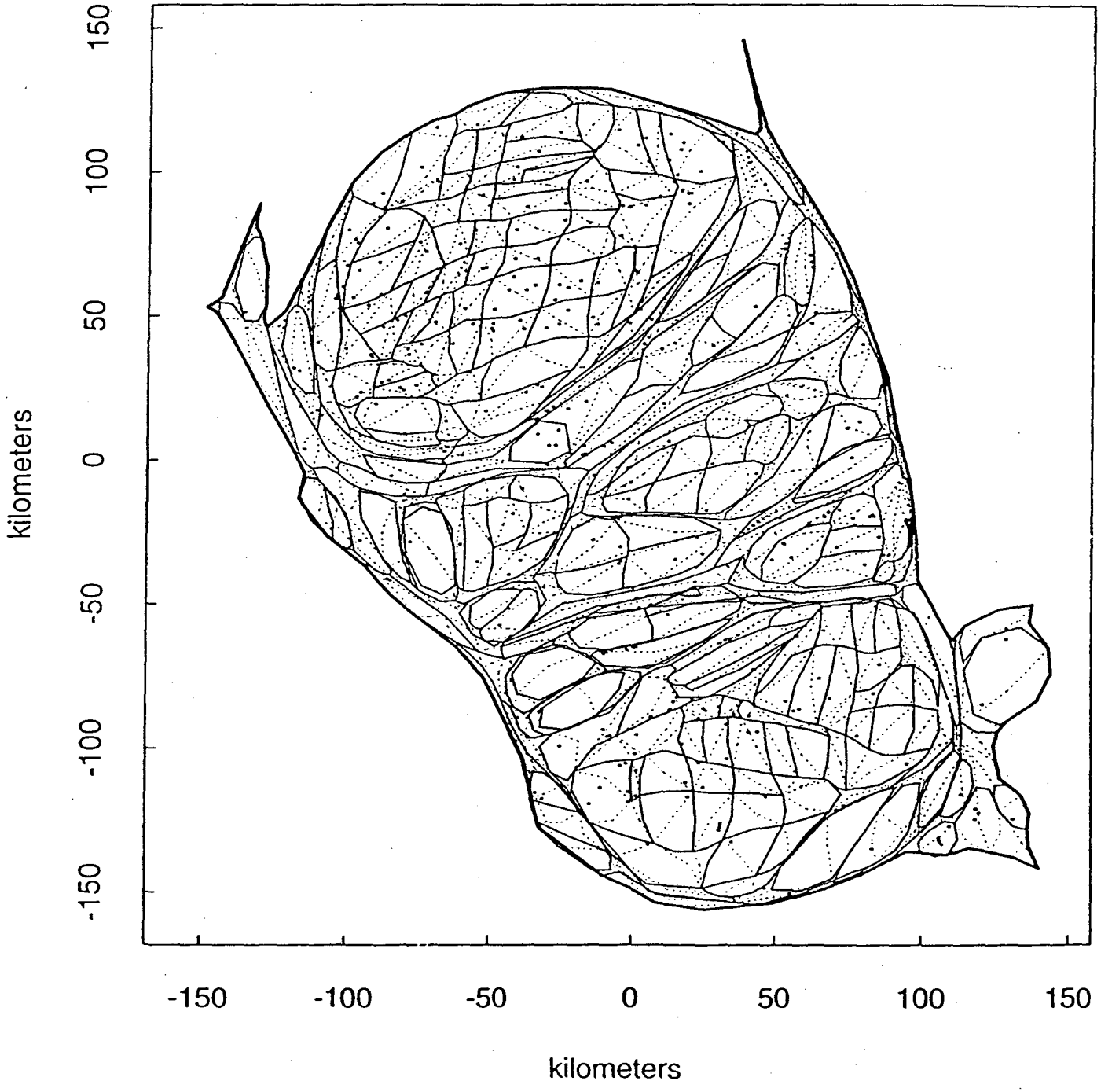


Figure B-6.

8020 random cases, initial map.

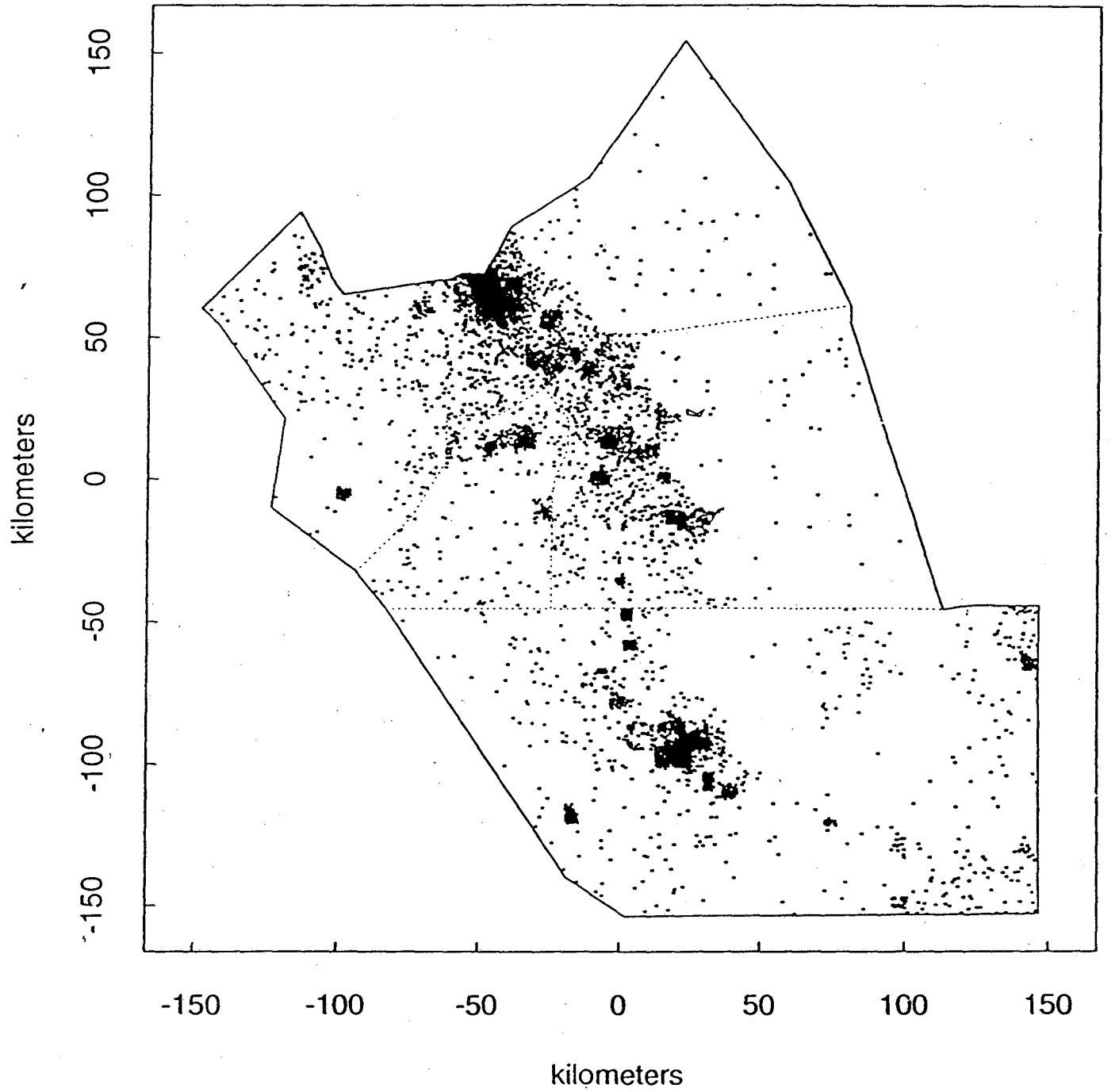


Figure B-7.

8020 random cases, run hex10, after step 5.

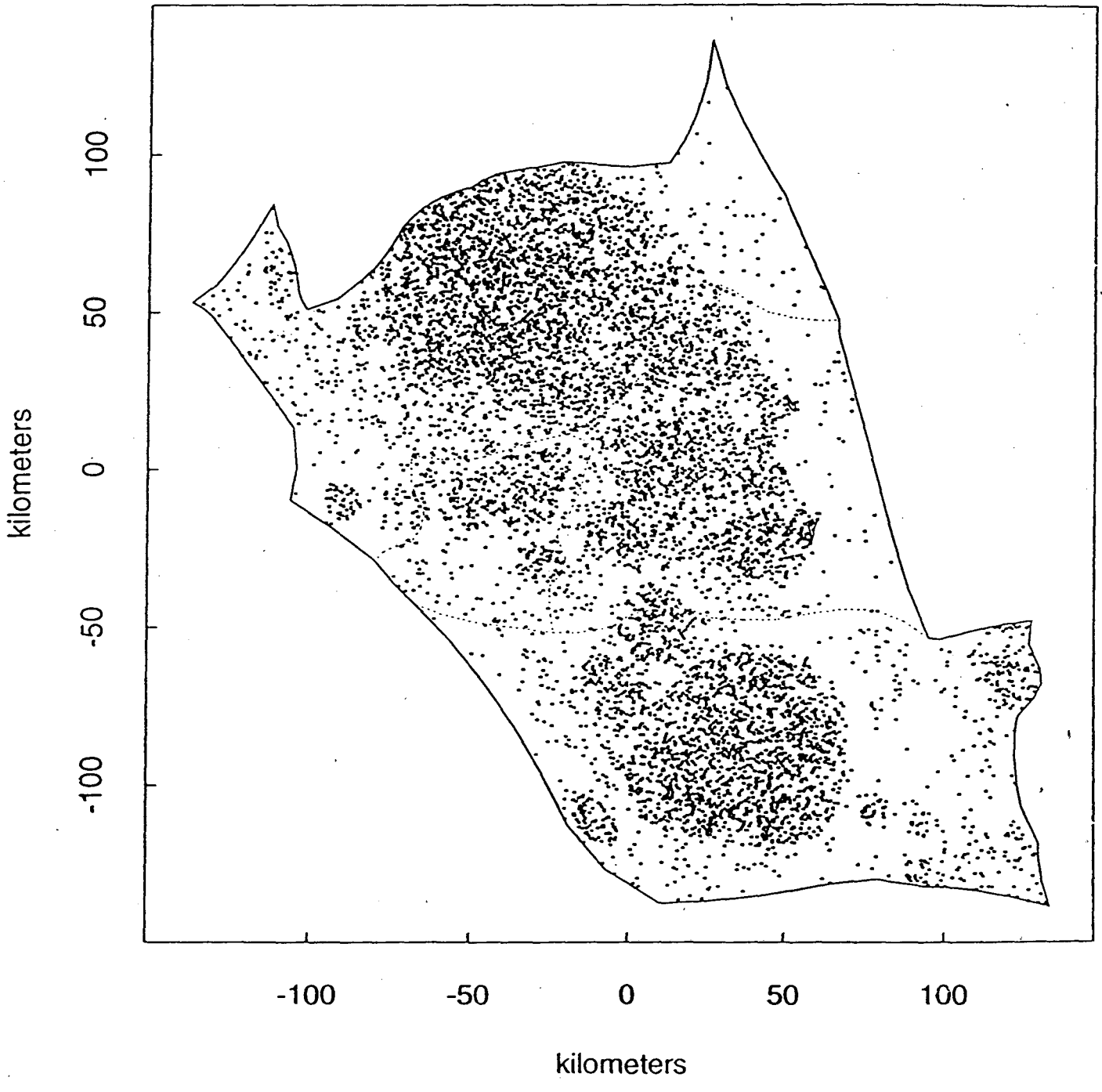
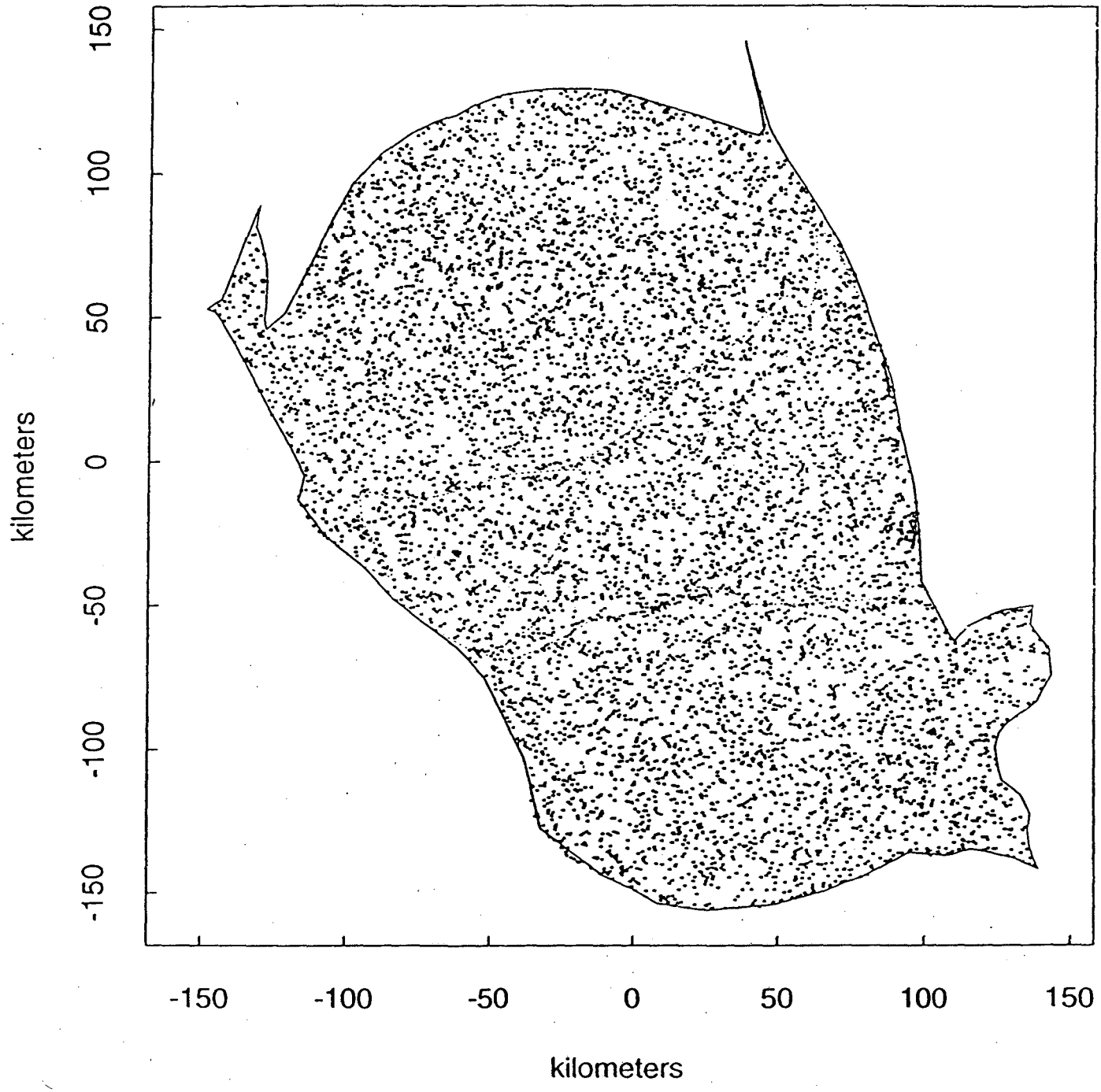


Figure B-8.

8020 random cases, run hex10, after step 10.



## APPENDIX C.

## DESCRIPTION OF THE SECONDARY RUN TRI10

The secondary run "tri10" was *not* used for statistical analysis in this report. The run included the same 16,441 non-boundary points as the primary run "hex10". In the first seven steps, the population units were triangles.

From previous runs (not presented here) we observed that illegal boundary crossings occurred only for triangles which had become highly oblique on the previous step; i.e. those having an internal vertex angle near 180 degrees. In the tri10 run, highly oblique triangles were subdivided after each step. We defined the "turning angle" of each triangle vertex as the complement of the internal vertex angle:

$$\text{turning angle} = (180 - \text{vertex angle})$$

Turning angles cannot exceed 180 degrees. Highly oblique triangles have turning angles near zero. A perfectly collinear triangle with zero area has a zero turning angle, and an inverted triangle with negative area has a negative turning angle.

Oblique triangles were defined as those having a turning angle less than *minangle* = 20 degrees. After each of the first seven steps, every oblique triangle was subdivided by dropping a perpendicular from the oblique vertex to the opposite boundary segment. The "complementary" triangle sharing the same bisected boundary segment was also subdivided. The populations and target areas of each bisected triangle were assigned to the two resulting triangles in the same ratio as their current areas. If a new oblique triangle was created, the process was repeated as many times as necessary. With triangle subdivision after each step, negative or zero turning angles (collinear or inverted triangles) did not occur in the first seven steps.

To avoid numeric underflow during the density equalization, it was necessary to avoid creating triangles having two vertices very close together. After triangle subdivision and before the next step, we removed from the map each pair of triangles sharing a segment whose length (after map scaling) was less than $minseg \times zero$, where $minseg = 10$ and $zero = 10^{-5}$. (See [CLOS94] for a discussion of map scaling and the constant *zero*.) The minimum segment length is equivalent to about 150 km × 10 × *zero*, or 15 meters. The population (and corresponding target area) associated with a discarded triangle, typically less than 0.01 person, was removed from the map and not reassigned to other triangles.

Triangle subdivision, followed by triangle removal, resulted in a net addition of triangles after each of the first seven steps. After seven triangle steps, *hsum* had declined from 17.04 to 0.87 and the number of triangles had increased from 1212 to 2064. There were no triangles with negative area.

Significant further improvement necessitated additional degrees of freedom, so at this point the triangles were converted to hexagons by bisecting each triangle boundary segment. In step 7a *minseg* was first increased from 10 to 20, reducing the number of triangles from 2064 to 2046; then in step 7b the 2046 triangles were converted to hexagons. The tri10 run was completed by taking five equal steps (steps 8-12) with the 2046 hexagons, with $c_i$ = 1/5,1/4,...1/1.

Table C-1 is the history of the tri10 run, including the step size $c_i$, *minangle*, *minseg*, total time including triangle subdivision and triangle removal, total polygons and negative-area polygons after the step, and the value of *hsum* after the step.

**Table C-1. History of run tri10**

| step | poly type | $c_i$ | *min angle* | *min seg* | time hrs | poly tot | poly < 0 | *hsum* |
|---|---|---|---|---|---|---|---|---|
| 0 | tri | | | | | 1212 | 0 | 17.04 |
| 1 | tri | 1/10 | 20 | 10 | 2.0 | 1246 | 0 | 12.25 |
| 2 | tri | 1/9 | 20 | 10 | 2.0 | 1292 | 0 | 8.63 |
| 3 | tri | 1/8 | 20 | 10 | 2.5 | 1366 | 0 | 5.94 |
| 4 | tri | 1/7 | 20 | 10 | 2.1 | 1489 | 0 | 3.90 |
| 5 | tri | 1/6 | 20 | 10 | 2.4 | 1629 | 0 | 2.47 |
| 6 | tri | 1/5 | 20 | 10 | 2.0 | 1840 | 0 | 1.49 |
| 7 | tri | 1/4 | 20 | 10 | 3.3 | 2064 | 0 | 0.87 |
| 7a | tri | 0 | 20 | 20 | 0.2 | 2046 | 0 | 0.87 |
| 7b | hex | 0 | NA | NA | 0.2 | 2046 | 0 | 0.87 |
| 8 | hex | 1/5 | NA | NA | 5.1 | 2046 | 0 | 0.51 |
| 9 | hex | 1/4 | NA | NA | 5.1 | 2046 | 0 | 0.27 |
| 10 | hex | 1/3 | NA | NA | 5.1 | 2046 | 0 | 0.12 |
| 11 | hex | 1/2 | NA | NA | 5.1 | 2046 | 1 | 0.034 |
| 12 | hex | 1/1 | NA | NA | 5.1 | 2046 | 7 | 0.0031 |
| tot | | | | | 42.2 | | | |

In Figure B-1 we showed the present area and target area of each hexagon after step 0 of run hex10 (the initial map of Figure 5). The plot is identical for the triangles of run tri10. Triangles to be expanded or reduced lie below or above a 45 degree line, respectively.

Figure C-1 shows the present area and target area of the 2064 triangles after step 7. No triangles have negative area.

Figure C-2 shows the same data for the density equalized map, after step 12. Density equalization is better than that of the hex10 run, shown in Figure B-3. A value of *hsum* = 0.0031 was obtained in the tri10 run, compared with *hsum* = 0.142 in the hex10 run. In Figure C-2, only seven hexagons have negative area. In addition, an unknown number of positive-area hexagons may have boundaries that self-intersect.

Figures 5, C-3, and C-4 show the tract boundaries (solid) and polygon boundaries (dotted) after step 0, 7 and 12 respectively. No overlapping polygon boundaries can be distinguished visually.

Figures B-6, C-5, and C-6 show the locations of 8020 random cases after step 0, 7, and 12 respectively. Figure C-6 from the tri10 run (to be compared with Figure B-8 of the hex10 run) shows significant non-uniformities. This can be understood by carefully comparing Figure 5 and Figure C-3. In triangles that are expanded, the non-linear RLInt transformation causes many non-boundary points (case locations) to be pushed outside the boundaries of the triangles to which they belong. In the final steps of the tri10 run the map is almost perfectly equalized, but the non-boundary points are not where they belong. The same problem can occur with hexagons but is much less severe; the problem would not occur at all if if each polygon had infinitely many points in its boundary.

For completeness, we also present Figures C-7, C-8, and C-9 from the tri10 run, which correspond to Figures 8, 10, and 12 from the hex10 run. But because the tri10 run created artificial clusters, data from the tri10 run were not used in the statistical analysis of this report.

A simple modification to the RLInt program can remove the problems with the tri10 run: in the triangle transformation steps, a simple linear transformation can be used to map the non-boundary points within each triangle; the usual non-linear transformation will be used in the final hexagon steps.

Figure C-1.

Present areas versus target areas,
run tri10, after step 7.

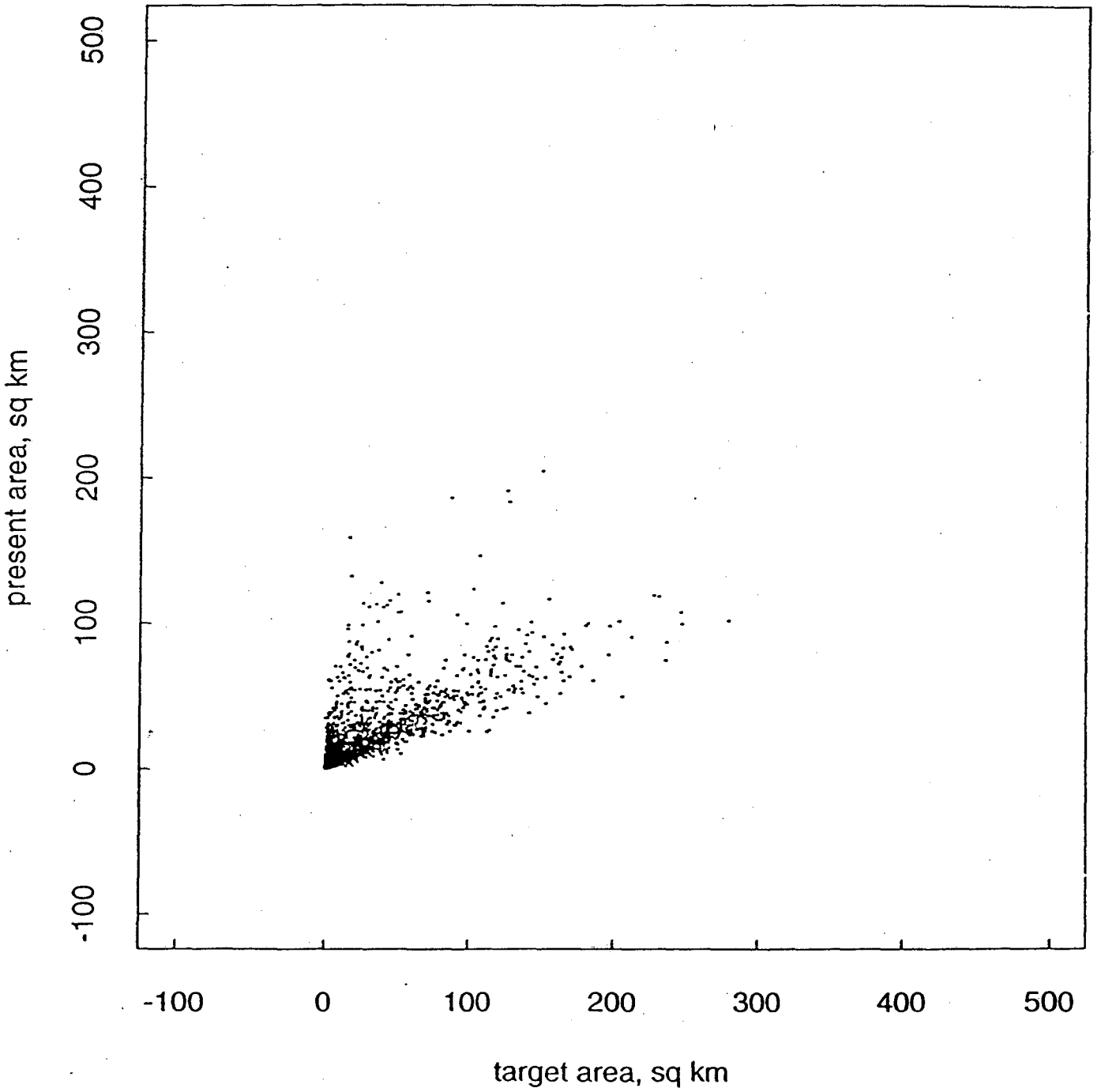


Figure C-2.

Present areas versus target areas,
run tri10, after step 12.

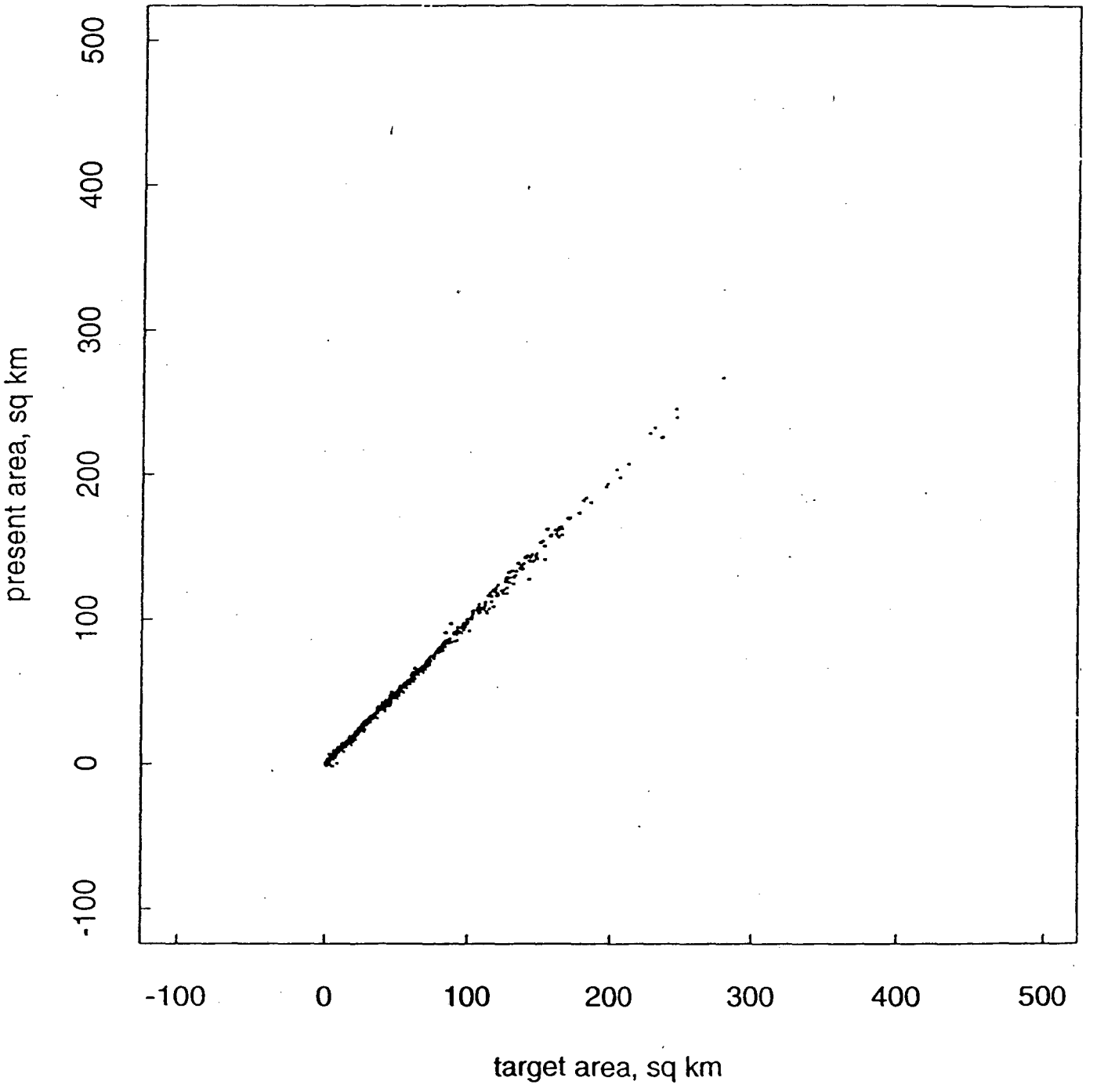


Figure C-3.

Tract boundaries, triangle boundaries, and 401 cases;
run tri10, after step 7.

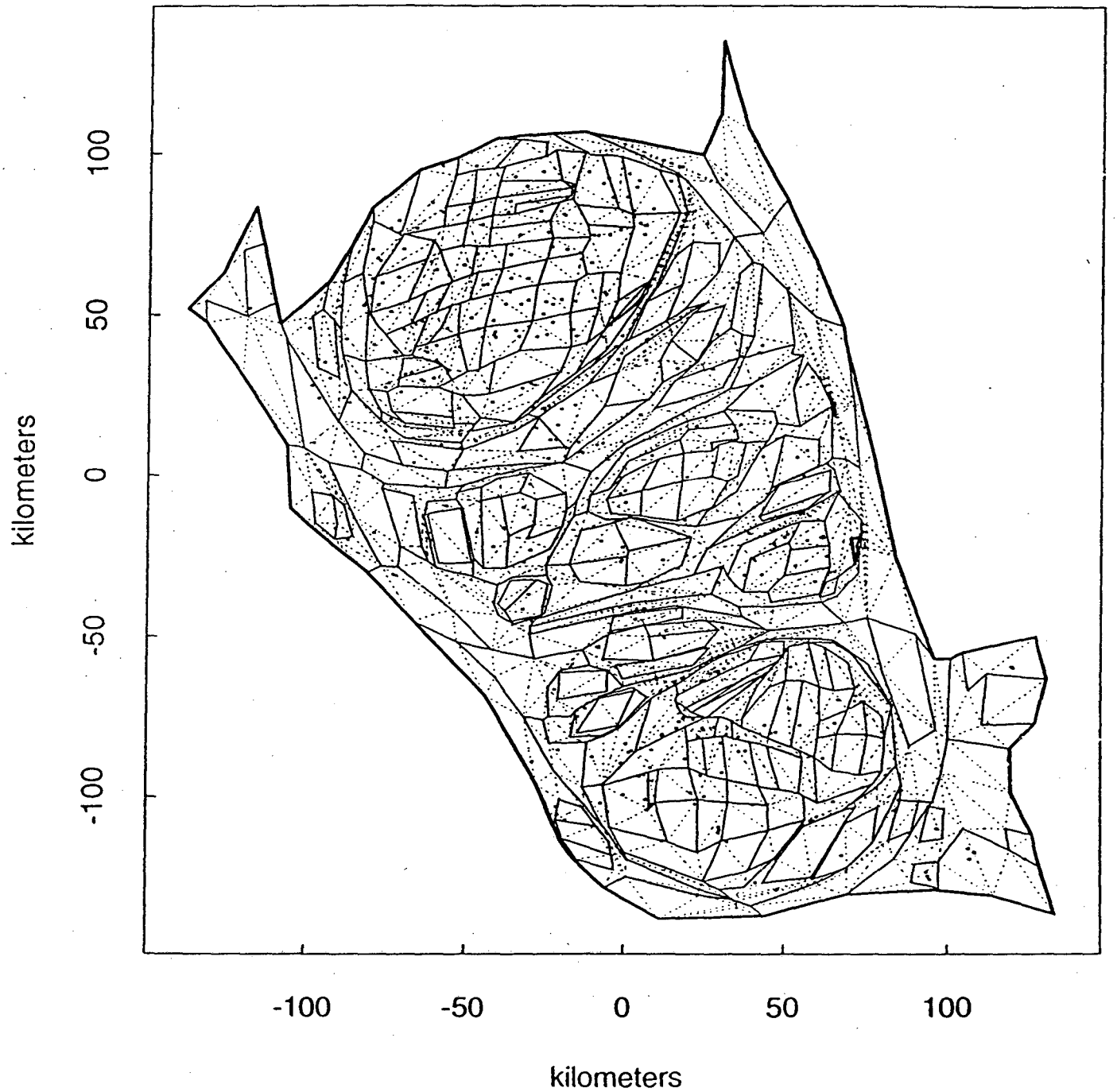


Figure C-4.

Tract boundaries, hexagon boundaries, and 401 cases;
run tri10, after step 12.

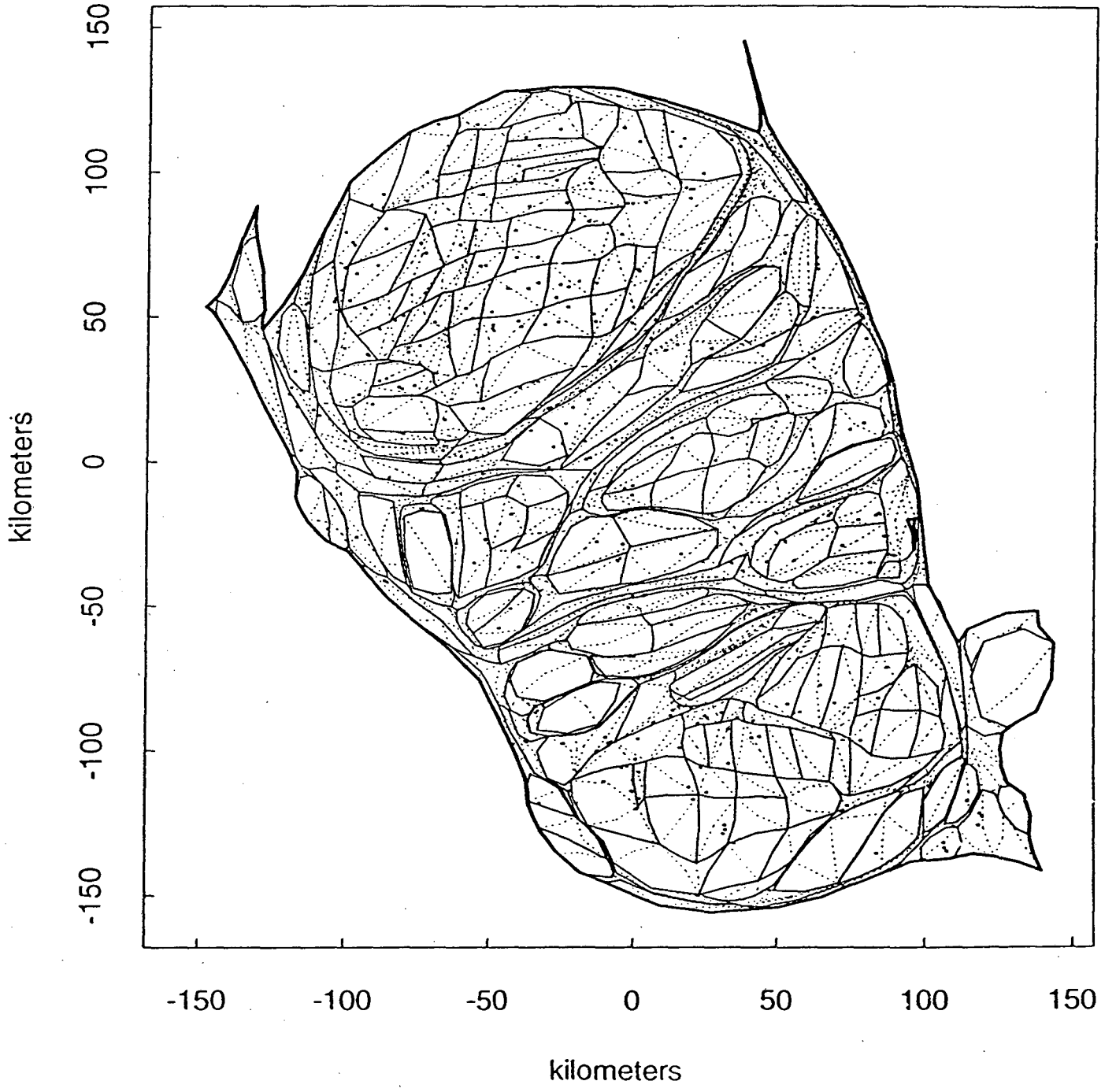


Figure C-5.

8020 random cases, run tri10, after step 7.

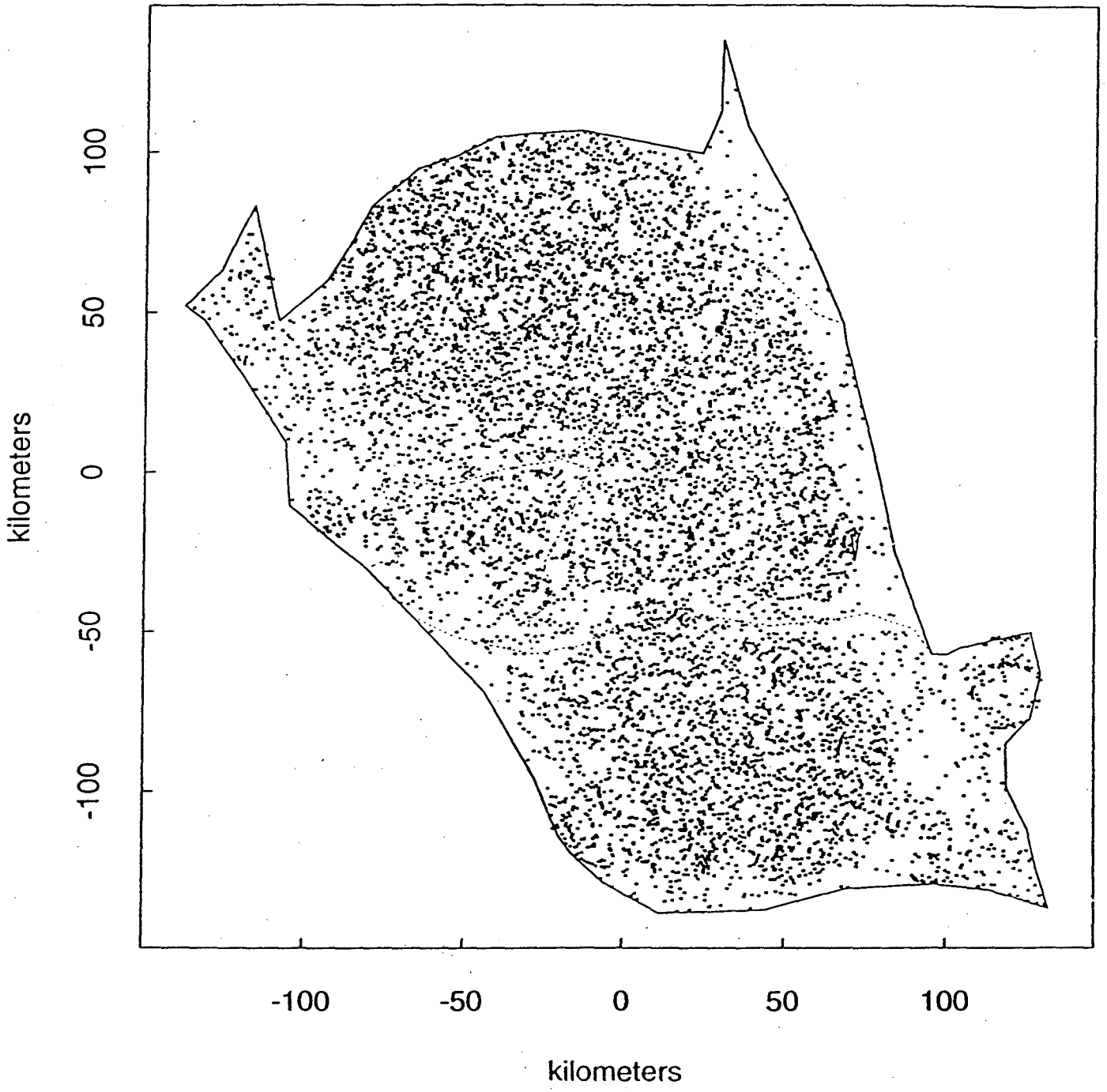


Figure C-6.

8020 random cases, run tri10, after step 12.

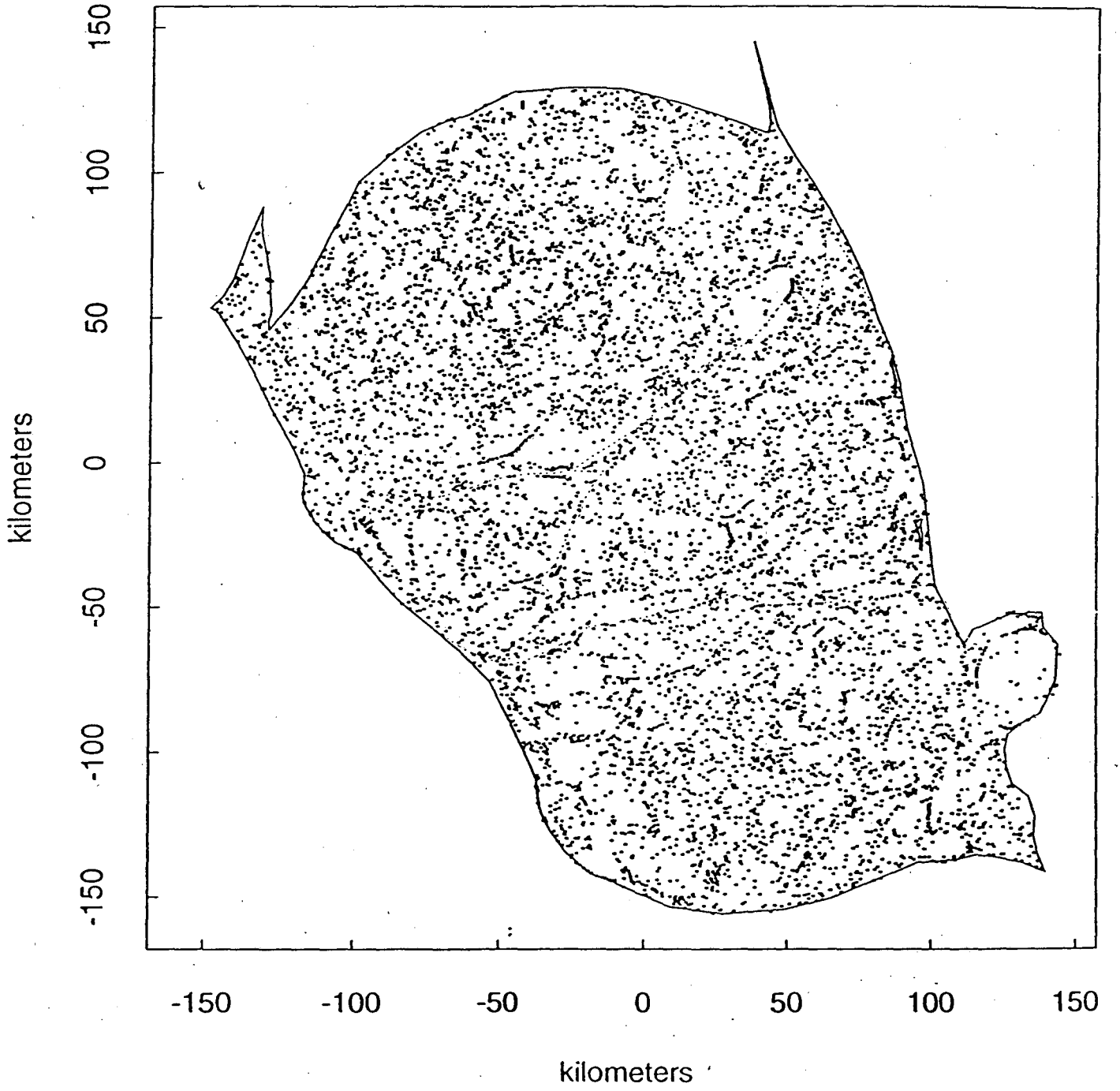


Figure C-7.

Actual locations of 401 real cases,
after density equalization, run tri10.
The external points are random artificial cases.

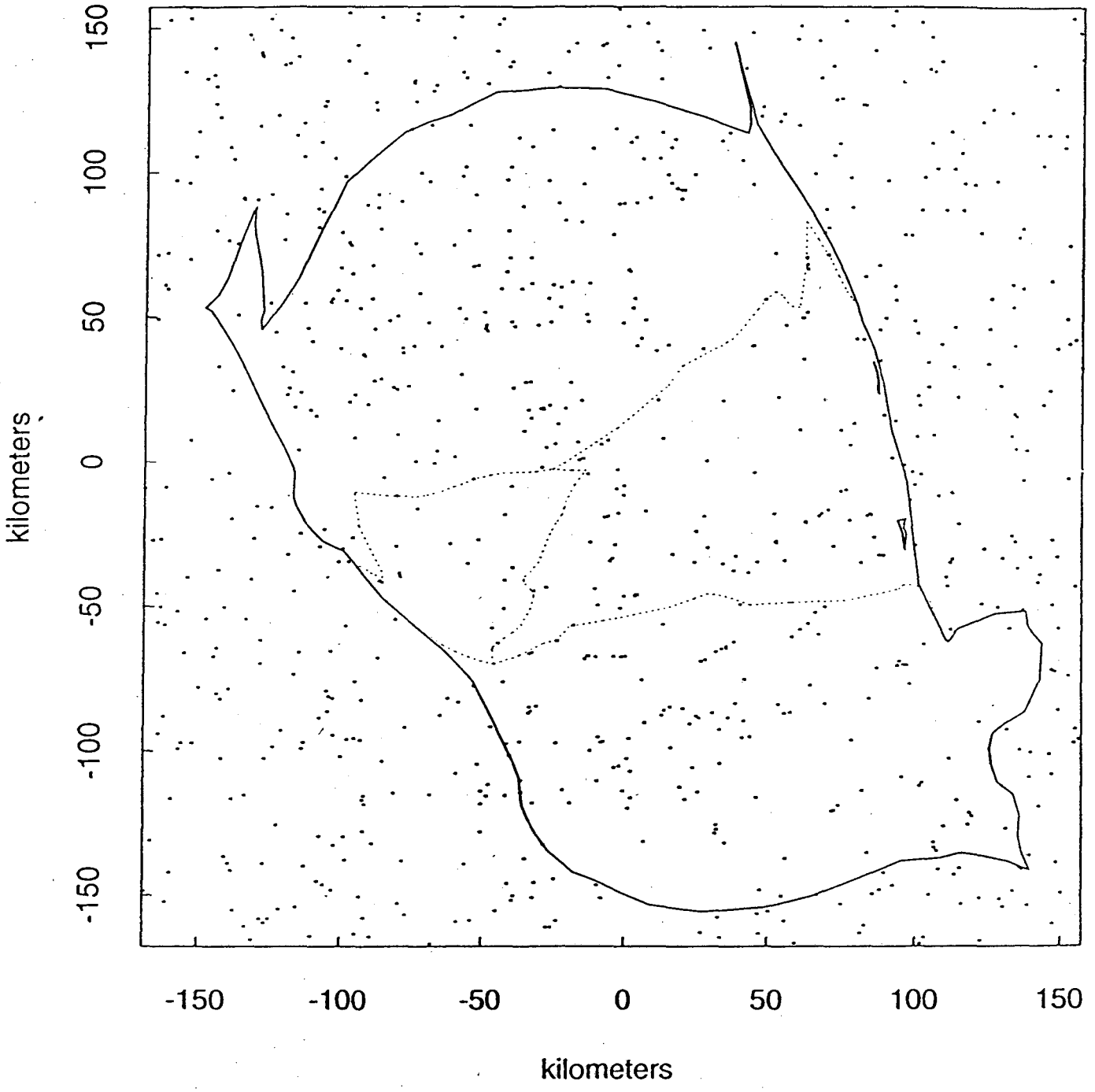


Figure C-8.

Locations of 401 artificial cases assuming equal risk, after density equalization, run tri10. The external points are additional random cases.

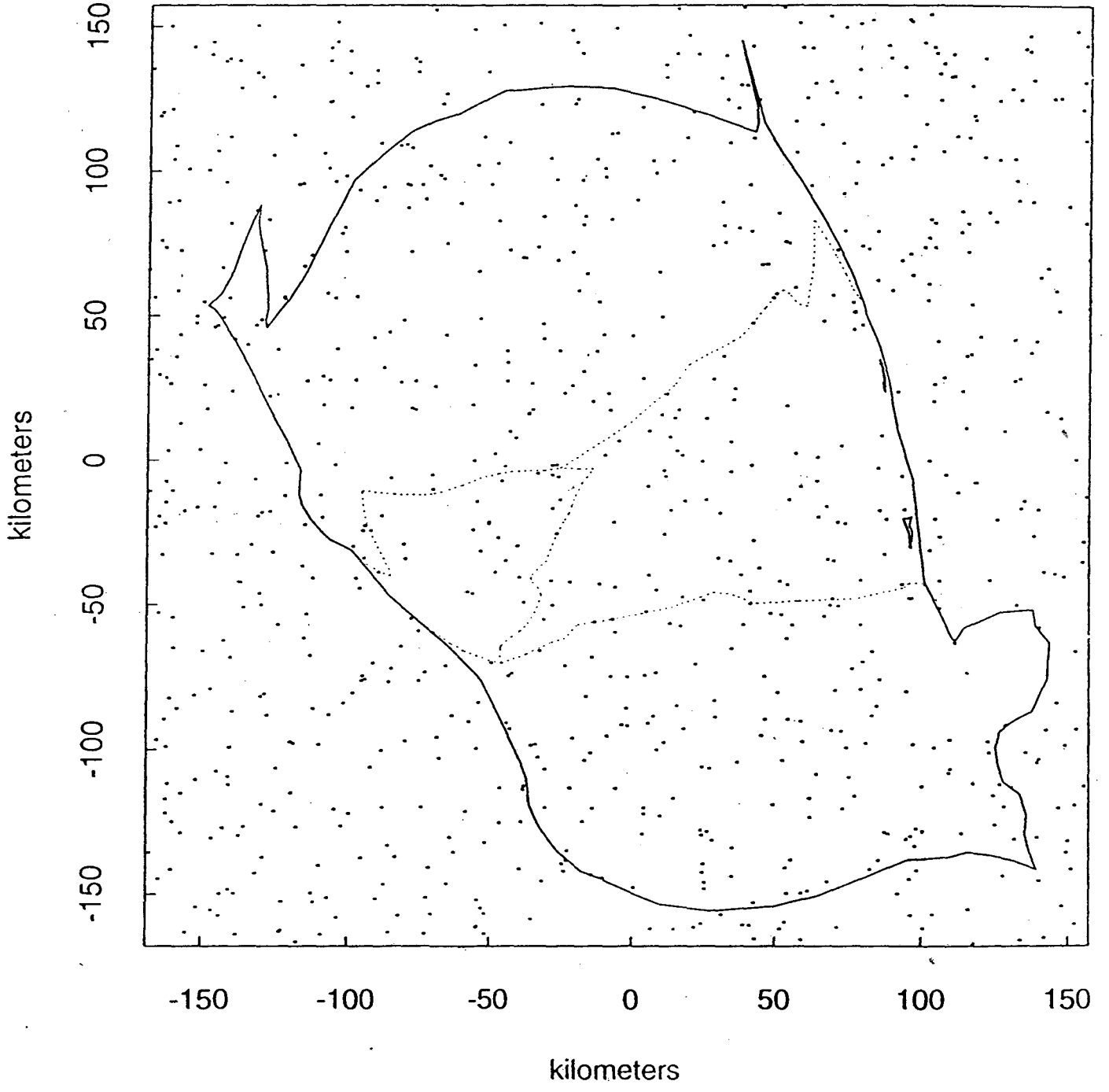
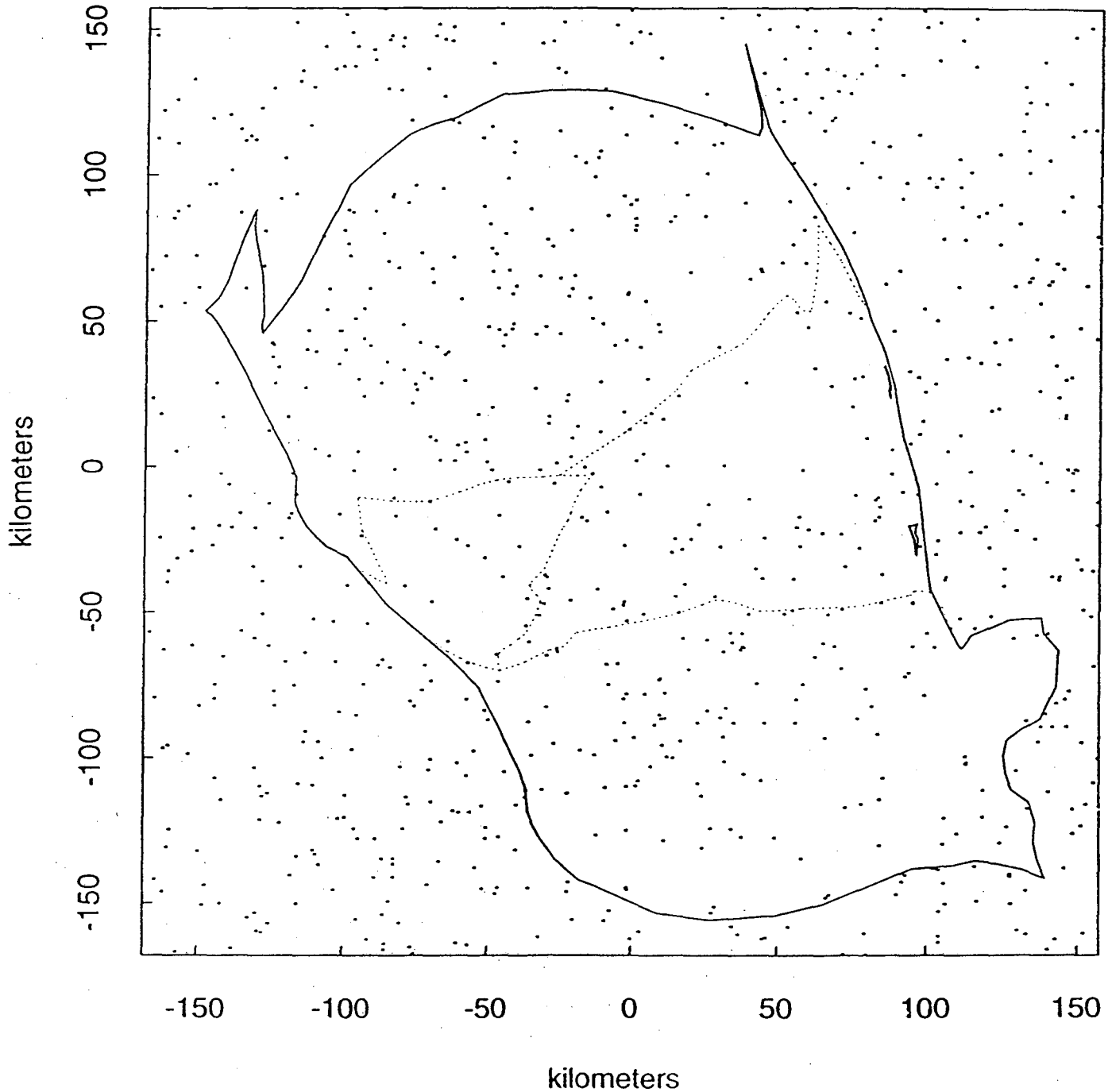


Figure C-9.

401 real cases, each plotted at a random location in
its own tract, after density equalization, run tri10.
The external points are additional random cases.



# APPENDIX D.

# HISTORY OF DEMP RESEARCH AT LBL

The purpose of the PAREP (Populations at Risk to Environmental Pollution) Project at Lawrence Berkeley Laboratory, an ongoing DOE project since 1978, is to develop resources (data, computing techniques, and biostatistical methodology) applicable to DOE's needs. Specifically, the PAREP project has developed techniques for statistically analyzing disease distributions in the vicinity of supposed environmental hazards. Such techniques can be applied to assess the health risks in populations residing near DOE installations, provided adequate small-area health data are available.

Since 1985, the research effort of the PAREP project has focused on the innovative approach of density equalizing map projections (DEMP), usually known as cartograms. Cartograms have long been used to display thematic data, and their value for analyzing public health data was recognized as early as the 1920's. Computer algorithms became available in the 1970's but so far have not been routinely used for the statistical analysis of disease distributions.

In a DEMP transformation, boundaries of geographic subareas (for example census tracts) are transformed so that population density is uniform over the entire transformed map. On the transformed map, the statistical analysis of the distribution of disease cases is simplified because the confounding effect of population density has been removed.

The unique contribution of the PAREP project has been the development of improved DEMP algorithms and statistical techniques for analyzing the resulting maps.

## 1988 LBL Algorithm

The first LBL algorithm, published in 1988 [SCHU88] employed a radial expansion or contraction relative to the centroid of each subarea in the map. The radial transformation changed the area but not the shape of the particular subarea in question, while changing the shape but not the area of all other subareas. The resulting map depended on the arbitrary order in which the subareas were transformed; in addition, it was possible for subarea boundaries to overlap after the transformation. Case locations were transformed along with subarea boundaries during the DEMP transformation.

### 1991 LBL Algorithm

A second LBL algorithm, completed in 1991 [MERR91] subdivided the map into triangles. As a function of all the coordinates of all the triangle vertices, we defined (1) a constraint function $H$ which vanishes only when each triangle reaches its desired target area, and (2) an objective function $G$ which measures overall distortion relative to the original map. (The function $H$ is equivalent to the function *hsum* defined in Appendix B of the present report.) A minimization program adjusted all the vertex coordinates so as to minimize $G$ subject to the constraint $H = 0$. The final solution defined a linear transformation for each triangle, which was applied to all the case locations within that particular triangle. With the 1991 LBL algorithm, solutions were uniquely defined and overlapping boundaries were avoided; however, the time of required for computation was prohibitive. To limit computation time, considerable geographic detail had to be sacrificed.

### 1993 Russian Algorithm

In 1993 a new algorithm was published by Gusein-Zade and Tikunov [GUSE93], in which the vector translation of each geographic map coordinate is calculated from the expansion or contraction of each infinitesimal area in the entire map. The translation due to a given subarea is calculated as a line integral around the boundary of that subarea. Convergence is achieved in a small number of iterations. Case locations are transformed along with subarea boundaries during the DEMP transformation.

### 1994 LBL Algorithm

In 1994 the Russian algorithm was independently implemented at LBL. A 130-page technical report [CLOS94] describes the implementation and extensive testing of the LBL implementation, which is known as *RLInt* (Russian Line Integral). New features in *RLInt* but not in [GUSE93] include the so-called "HH scaling factor," which was found to be necessary for equalizing highly non-uniform populations like that of the four-county area.

### 1995 LBL Algorithm

Additional *RLInt* program options not described in [CLOS94] were implemented and used in this report. The new options are described in Appendix E.

## APPENDIX E.

## 1995 LBL DEMP ALGORITHM - NEW PROGRAM OPTIONS

Since December 1994 the following program options, which are not described in [CLOS94], were added to the program *RLInt*. They are activated by specifying the following optional parameters in the file *RLInt.par*. Except for the *makebdy* option, they should be used only with triangle files (a file in which only the boundary polygon, region 1, has a number of points different from 3).

| | |
|---|---|
| *minangle* > 0 | (with *itstp* = 0 or 1) Subdivide oblique triangles which have turning angle less than *minangle* degrees; write the result to *RLInt.new.0000* or *RLInt.new.0001*. |
| *minseg* > 0 | (with *minangle* > 0, *iara* = 1, and *itstp* = 0 or 1) Eliminate triangles with a segment shorter than *minseg* × *zero* (*zero* is specified in the code as $10^{-5}$); write the result to *RLInt.fix.0000* or *RLInt.fix.0001*. |
| *makehex* = 1 | (with *itstp* = 0) Convert a triangle file to hexagons by bisecting every line segment; write the result to *RLInt.out.0000*. |
| *makebdy* = 1 | (with *itstp* = 0) Remake external boundary polygon (region 1). May be required if the *minangle* or *minseg* option has been previously used; write the result to *RLInt.out.0000*. |
| *nranpts* > 0 | (with *itstp* = 0, *nransamples* = 0) Generate *nranpts* points, randomly placed in the same tract as the case, for every non-boundary point (case) in the file *RLInt.dat*; include the result in *RLInt.out.0000*. |
| *nransamples* > 0 | (with *itstp* = 0, *nranpts* = 0) Generate *nransamples* points, randomly placed in the map, for every non-boundary point (case) in the file *RLInt.dat*; include the result in *RLInt.out.0000*. |

To generate *RLInt.new.0000* and *RLInt.fix.0000* for any option specified here, specify *minangle* = 1 and *minseg* = 1. Those values are usually small enough that no triangles will be subdivided or removed.

A sample file *RLInt.par* is provided below. Because the parameters are read by position, all lines should be included in the following order. The example shown eliminates triangles having a (scaled) segment length less than 20 × *zero*, about 30 meters in the four-county map. The resulting output file is written to *RLInt.fix.0000*.

```
HHH tri 0 steps, ci=1/0, 1 deg, minseg 1, bdy 0, hex 0, ranpts 0
10 iprint       0 no print, .gt. 0 is print out Print Flag
10 itable       table to RLInt.out.sum, RLInt.out.plot
0  maxit0       iteration max. Fixed point
0  maxit        iteration max. Transformation
0  itstp        Stop at exactly iteration itstp
1  iscale       0 Russian, 1 HHH Transformation scaling
1  iara 0       no scale, 1 scale Data Region scaling
1  icheck       0 no save, 1 save Result saving flag in core
0  ireset       0 float, no push - 1 reset iteration on Neg. Mag.
0  iciset       0 Russian, 1 ci = 1/2
1  idisk        0 no intermed disk output, .ne.0 output mod idisk
1  isum         0 no disk summary file, 1 write output
0  inow         no current iteration files, .gt.0 write output
16 nfdel        .lt.0 no file deletes, .gt.=0 delete old files
0  iprmpt       0 no interactive prompting,.ne.0 prompting
1  minangle     .gt.0 split tri with turnangle .lt. minangle
20 minseg       .gt.0 drop tri segs with length .lt. minseg.zero
0  makehex      1 to make hexagons from triangles
0  makebdy      1 to remake external boundary
0  nranpts      number of random pts to add in same tract
0  nransamples  random samples to add - null hypothesis
End of parameter data
```

## APPENDIX F.

## PROGRAMS AND DATA FILES

The program and data file locations listed here are subject to change. In each case, try first to obtain the file from the location listed here. If it is no longer there, obtain the current location from the current electronic version of the document you are reading.

The current electronic version of this document is in WWW URL:
`http://cedr.lbl.gov/pdocs/cdc9501/cdc9501.html`

If the electronic version of this document is no longer in that location, consult:
`http://cedr.lbl.gov/~merrill/index.html`

or send electronic mail to `dwmerrill@lbl.gov`.

### RLInt program

The RLInt program source code is publicly available, and is in the following locations. Please send electronic mail to `dwmerrill@lbl.gov` if you plan to use the code, so you can be informed of future modifications. You may request that the source code be mailed to your electronic address.

RLInt Fortran (f77) source code:
`parep2.lbl.gov:/CEDRCD/data1_new/merrill/Puff/Version5/RLInt.f`

Makefile for compiling and linking RLInt:
`parep2.lbl.gov:/CEDRCD/data1_new/merrill/Puff/Version5/Makefile`

sample csh program for running RLInt:
`parep2.lbl.gov:/CEDRCD/data1_new/merrill/Puff/Version5/RLInt.csh`

### Data from this analysis

Data for the 401 individual cases are confidential. To copy or use these data you must obtain permission from the California Department of Health Services. *The same applies to data in any derived files that could be used to identify individual subjects in the four-county study.* Other data files, such as the population files or map files, are locked but can be distributed upon special request.

For further information, send electronic mail to:

| | |
|---|---|
| Peggy Reynolds, DHS | (hw1.preynold@hw1.cahwnet.gov) |
| Raymond Neutra, DHS | (hw1.rneutra@hw1.cahwnet.gov) |
| Deane Merrill, LBL | (dwmerrill@lbl.gov) |

## REFERENCES

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


BRES87. Breslow NE and Day NE. 1987. *Statistical Methods in Cancer Research.* Oxford University Press.

CLOS94. Close ER, Merrill DW, and Holmes HH. 1994. 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.

GUSE93. Gusein-Zade SM and Tikunov VS. 1993. A New Technique for Constructing Continuous Cartograms; *Cartography and Geographic Information Systems*, Vol. 20, No. 3, 1993, 167-173.

MERR91. Merrill D, Selvin S and Mohr MS. 1991. Analyzing Geographic Clustered Response. Report LBL-30954, (44 pages), June 1991. Invited paper presented at 1991 Joint Statistical Meetings of the American Statistical Association, Atlanta GA, August 1991. WWW URL: http://cedr.lbl.gov/pdocs/asa91/asa91.txt.html. Summary version (6 pages) in proceedings, Section on Statistics and the Environment, American Statistical Association, pp. 96-101, published June 1992. WWW URL: http://cedr.lbl.gov/pdocs/asa91/short.txt.html.

MERR92. Merrill D, Selvin S and Mohr MS. 1992. Density Equalizing Map Projections: Techniques and Applications. Report LBL-32640, July 1992. Presented at Workshop on Statistics and Computing in Disease Clustering, Stony Brook NY, July 23-24, 1992. WWW URL: http://cedr.lbl.gov/stonybr/stonybr.txt.html.

MERR93. Merrill DW. 1993. Data required for prototype small-area analysis. Task Completion Report due 12/15/93. WWW URL: http://cedr.lbl.gov/mdocs/ftp/status/tr931215.asc.html.

MERR94A. Merrill DW. 1994. FY 1994 PAREP task completion report. WWW URL: http://cedr.lbl.gov/pdocs/fy94complete.html.

MERR94B. Merrill DW. 1994. Preparation of geographic map files for DEMP transformation. DOE task completion report due 1/15/94 (revised). WWW URL: http://cedr.lbl.gov/pdocs/tr940115/all.html.

MERR94C. Merrill D, Selvin S and Close ER. 1994. Use of density equalizing map projections (DEMP) in the analysis of a reported childhood cancer cluster in McFarland, California. Presented at the Second Conference on Statistics and Computing in Disease Clustering, Vancouver, B.C., Canada, July 21-22, 1994. WWW URL: http://cedr.lbl.gov/pdocs/vancouver/vancouver.html.

MERR95. (this report) Merrill D, Selvin S and Close ER. 1995. Use of density equalizing map projections (DEMP) in the analysis of childhood cancer in four California counties. Submitted to 1995 CDC/ATSDR Symposium on Statistical Methods: Small Area Statistics in Public Health: Design, Analysis, Graphic and Spatial Methods; Atlanta GA, January 25-26, 1995. WWW URL: http://cedr.lbl.gov/pdocs/cdc9501/cdc9501.html.

REYN91. Reynolds P, Satariano E, Smith D. 1991. The Four County Study of Childhood cancer incidence: Interim report II. Environmental Epidemiology and Toxicology Program, California Department of Health Services, October 1991.

SATA90. Satariano E, Reynolds P, Smith D, Goldman L. 1990. The Four County Study of childhood cancer incidence: Interim report I. Environmental Epidemiology and Toxicology Branch, California Department of Health Services. May 1990.

SCHU88. Schulman J, Selvin S and Merrill DW. 1988. Density Equalized Map Projections: A Method for Analyzing Clustering Around a Fixed Point. *Statistics in Medicine* 7:491-505.

SEED94. SEEDIS (Socio-Economic Environmental Demographic Information System). WWW URL: http://cedr.lbl.gov/mdocs/seedis/seedis.html.

SELV91. Selvin S. 1991. *Statistical Analysis of Epidemiologic Data.* Oxford University Press.