LBNL-41624

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

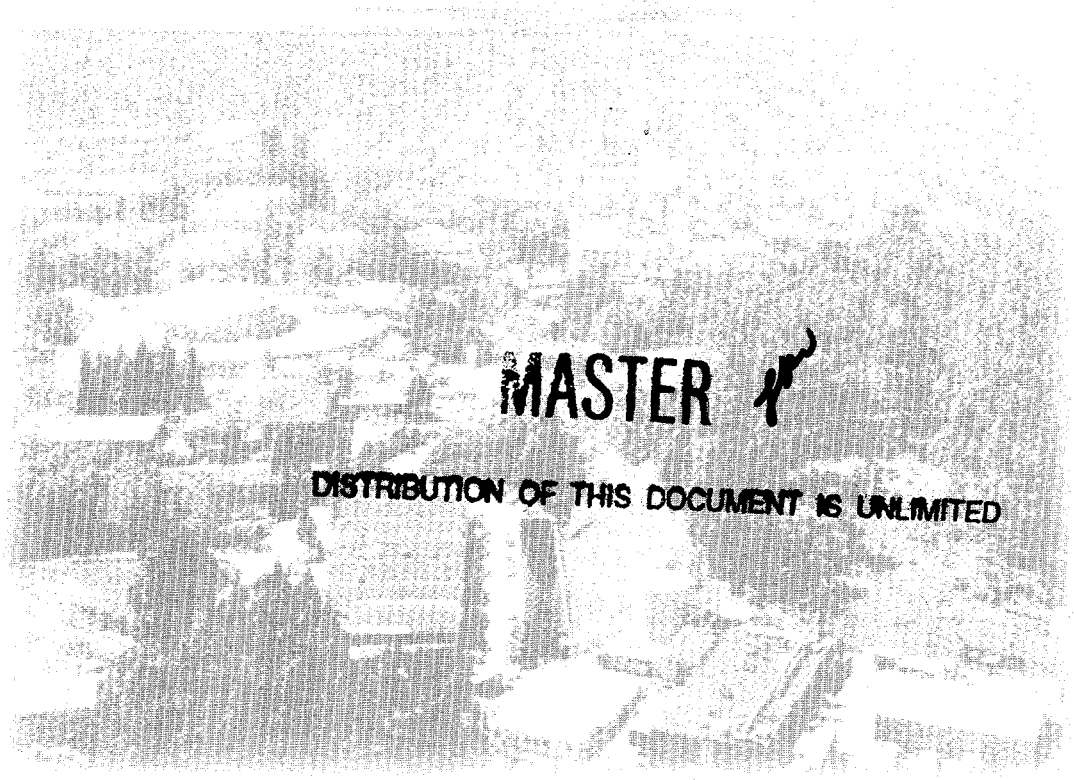
Density Equalizing Map Projections (Cartograms) in Public Health Applications

Deane W. Merrill

Information and Computing Sciences Division

May 1998

Dr.P.H. Thesis



DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

LBNL-41624
UC-405

Density Equalizing Map Projections (Cartograms) in Public Health Applications

Deane W. Merrill
(Dr.P.H. Thesis)

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

May 1998

Address all correspondence to Deane W. Merrill, Information and Computing Sciences Division, mail stop 50B-3238, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley CA 94720. Tel. (510) 486-5063. Fax: (510) 486-4004. Internet: dwmerrill@lbl.gov or merrill@crocker.com. Web home page: http://parep2.lbl.gov/~merrill or http://www.bearhaven.com.

The electronic version of this document is available at http://parep2.lbl.gov/~merrill/thesis/thesis or http://www.bearhaven.com/thesis or http://merrill.wwh.net/thesis. Future revisions will be incorporated in the electronic version.

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

Density Equalizing Map Projections (Cartograms) in Public Health Applications

by

Deane Whitney Merrill, Jr.

B.A. (Williams College) 1960
M.S. (University of California, Berkeley) 1962
Ph.D. (University of California, Berkeley) 1967

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Public Health

in

Public Health

in the

SCHOOL OF PUBLIC HEALTH

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Steve Selvin, Chair
Professor Warren Winkelstein, Jr.
Professor Kenneth Wachter

Spring 1998

ABSTRACT

Density Equalizing Map Projections (Cartograms) in Public Health Applications

by

Deane W. Merrill, Jr.

Doctor of Public Health in Public Health

University of California, Berkeley

Professor Steve Selvin, Chair

In studying geographic disease distributions, one normally compares rates among arbitrarily defined geographic subareas (e.g. census tracts), thereby sacrificing some of 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 risk is 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 calculated with validity.

The DEMP algorithm was applied to a data set previously analyzed with conventional techniques; namely, 401 childhood cancer cases in four counties of California. The distribution of cases on the transformed map was analyzed visually and statistically. To check the validity of the method, the identical analysis was

performed on 401 artificial cases randomly generated under the assumption of uniform risk. No statistically significant evidence for geographic non-uniformity of rates was found, in agreement with the original analysis performed by the California Department of Health Services (DHS).

Appendix A documents the electronic locations, of not only the data files used in this analysis, but of documents and data assembled during 30 years of related projects at Lawrence Berkeley National Laboratory (LBNL). These data are from SEEDIS (Socio-Economic Environmental Demographic Information System) and the PAREP (Populations at Risk to Environmental Pollution) project, and include comprehensive 1970 and 1980 U.S. Census data. Over 3200 tapes of historical government data, some of them unique and irreplaceable, have been archived and are documented online.

TABLE OF CONTENTS

TABLE OF CONTENTS (CONTINUED)

TABLE OF CONTENTS (CONTINUED)

TABLE OF CONTENTS (CONTINUED)

TABLE OF CONTENTS (CONTINUED)

LIST OF FIGURES

LIST OF FIGURES (CONTINUED)

LIST OF FIGURES (CONTINUED)

LIST OF TABLES

INTRODUCTION

A doctoral dissertation, even one for a professional degree such as Doctor of Public Health (Dr.P.H.), is intended to certify the candidate as ready to embark on a professional career. A specific study must be conducted from beginning to end, and a report produced that is of publication quality. This dissertation, which describes a re-analysis of the Four County Childhood Cancer data set by an innovative method, satisfies that formal requirement.

It seemed desirable and even imperative to expand the scope of this particular dissertation for several reasons:

(1) Previous analyses of the Four County Childhood Cancer set have already been published, by the California Department of Health Services (DHS), and by this author in preliminary form. The final results are completely negative, in agreement with the original DHS publication.

(2) This dissertation is unusual in that it is being written at the *end* of a 30-year professional career. As such, it is intended primarily for researchers wishing to make future use of the data and the techniques described here. The detailed Appendix A is the most valuable part of this report. It specifies the electronic locations of not only the figures and data of the Four County re-analysis, but of data, programs and documentation assembled during the entire course of the SEEDIS (Socio-Economic Environmental Demographic Information System) and PAREP (Populations at Risk to Environmental Pollution) projects. SEEDIS alone includes over 100 documented

databases and 10 gigabytes of data (100,000 data files with an average size of 100,000 characters).

In the 1970's and 1980's, LBNL (Lawrence Berkeley National Laboratory) assembled a unique archive of historical demographic and epidemiologic data, including most user tapes of the 1980 and 1970 U.S. Census. The author preserved the data through several physical migrations, and arranged for their recent move to a computer at the U.S. Bureau of the Census. Over 500 notebooks of paper documents are in the author's possession. Over 3200 9-track tapes, some unique and irreplaceable, are stored in a warehouse and cataloged online. By studying the documents and data in Appendix A, future researchers will be able to locate, obtain and use those data files. *(Public access is restricted, for those files containing proprietary or confidential data; file locations are provided for future use by authorized persons only.)*

(3) Cartograms, or Density Equalizing Map Projections (DEMP), have potential application beyond the scope of this dissertation, even beyond epidemiological applications. A technical discussion of the current DEMP© implementation appears elsewhere [CLOS94]. Later on, a new implementation is planned, which can find future use in commercial GIS applications.

(4) This dissertation, especially Appendix A, contains numerous references to public URL's (Uniform Resource Locators) in the World Wide Web. This is a practical necessity, for merely the electronic documents created in this project (not to mention the data files) would have required many hundred pages in paper form. The applicability of the World Wide Web as a supplementary publishing medium is not yet

formally recognized by the University of California ("Guidelines for submitting a Doctoral Dissertation," April 1994, UC Berkeley Graduate Division). The dynamic nature of the Web requires that documents referring to URL's *themselves* be dynamic Web documents. This catch-22 situation is circumvented by specifying where the electronic version of this dissertation is presently stored; namely, at http://parep2.lbl.gov/~merrill/thesis and http://www.bearhaven.com/thesis. The electronic version will be updated as necessary. Should this document itself be moved to a new location, a search with one of the Web search engines will reveal its new location.

(5) Steve Selvin has stated, "The dissertation is the one opportunity the student has to record for posterity anything he/she wants to include." In accordance with that advice, this dissertation contains a brief autobiography of the author's professional career, including some of the anecdotes that have made that career so enjoyable.

ACKNOWLEDGMENTS

Friends have asked me, why I chose to embark on a second Doctoral degree program. (Because of my work obligations, the degree took fifteen years to complete, and I am receiving it only after my retirement, which occurred in November 1997.) The answer, of course, is that I thoroughly enjoyed the Dr.P.H. program - for what I learned and for the colleagues I met along the way. In retirement I hope to further develop the DEMP and SEEDIS resources in collaboration with my junior colleagues. If my newly acquired Dr.P.H. degree bolsters our credibility in seeking new funding, that will be an added benefit.

Special thanks go to my colleague and friend Steve Selvin, who has worked with me and inspired me for twenty years, and who patiently encouraged my 15-year graduate career. Susan Sacks and Warren Winkelstein brought us all together, creating in 1976 what became the PAREP project (Populations at Risk to Environmental Pollution). I am truly grateful to Steve's assistant Bonnie Hutchings, who rescued me repeatedly from the embarrassment of administrative mishaps and forgotten deadlines. Thanks also to Dean Patricia Buffler for her helpful advice and consistent encouragement.

Programmatic support in the early years of PAREP development was provided through the efforts of Carl Quong and Donald M. Austin at Lawrence Berkeley Laboratory, and Robert Goldsmith in the Department of Energy (DOE). Harvard Holmes, Fred Gey, Chris Stuber, Valerie Gregg, Kathie Ragland, Mary White, Jack

Colford, Allan Konrad and many others provided tangible assistance in later years, when loss of DOE funding threatened permanent destruction of the PAREP resources. For now, SEEDIS, including unique copies of 1970 and 1980 Census data, are safely housed at the Census Bureau, but continued financial support will be required to prevent those data from being lost in the future.

Via the Internet, I have enjoyed a active and useful correspondence with Vladimir Tikunov and Sabir Gusein-Zade, the Russian authors of the DEMP algorithm that was implemented and modified at LBNL. Thanks to Dr. Eugene Shiryaev for nominating me to the International Eurasian Academy of Sciences. These colleagues in Russia, and dozens of others throughout Europe and North America and South America, continue to provide original ideas for the calculation and application of cartograms. Waldo Tobler, who is indisputably the "father of the computer cartogram," has provided guidance and many helpful discussions.

The author is 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. Elon Close implemented the Russian DEMP algorithm at LBNL. Michael Mohr wrote many of the programs used to manipulate map files. Christine Erdmann plans to use the DEMP technique in her own dissertation, and is working to integrate the program with ArcView© and S-Plus©. The author is grateful to the CEDR project for use of computer facilities to complete this dissertation.

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

CURRICULUM VITAE

Deane Merrill came from a background in experimental high energy nuclear physics. His Ph.D. (1967) was in the group of Nobel prizewinner Luis Alvarez at Lawrence Berkeley National Laboratory (LBNL).

In 1973 Deane changed direction and became a computer scientist, then a Census data expert. In the 1980's he helped develop the LBNL Socio-Economic Environmental Demographic Information System (SEEDIS), which became the largest integrated collection of historical data from the 1960's, 70's and 80's. The SEEDIS data now reside at the Census Bureau, where they will be integrated with future Census data systems.

In the 1990's Deane developed a CD-ROM information system and 1990 U.S. Census LOOKUP. The Census Bureau installed LOOKUP as its first Web-based information system for 1990 Census data. A modified version is installed at the University of California, Berkeley Library Web site.

For 20 years, Deane collaborated with Steve Selvin, UC Berkeley Professor of Biostatistics, in the Populations at Risk to Environmental Pollution (PAREP) project. The major focus of their research was the development and application of Density Equalizing Map Projections (DEMP).

In May 1998, Deane completed a second doctorate - a Dr.P.H. in Epidemiology at the UC Berkeley School of Public Health. Recently retired from LBNL, he will move to western Massachusetts where he and his wife will open a

small Bed and Breakfast. From his new location, he plans to maintain an active Internet presence and to further develop the PAREP and SEEDIS resources in collaboration with his former Berkeley colleagues.

e-mail: dwmerrill@lbl.gov or merrill@crocker.com

Web page: http://parep2.lbl.gov/~merrill or http://www.bearhaven.com/dwmerrill.html

HISTORY AND BACKGROUND

This dissertation satisfies one of the requirements of the Dr.P.H. degree; namely, the completion of a specific research project, which is the reanalysis of the Four County Childhood Cancer data set by the innovative method of Density Equalizing Map Projections (DEMP)©. Because this dissertation is being completed at the *end* of the author's professional career, this unique opportunity will be also be used to describe the motivation and history of the DEMP effort at Lawrence Berkeley National Laboratory (LBNL).

Appendix A, which for future researchers will be the most valuable part of this report, documents the electronic locations of documents and data files assembled by the SEEDIS (Socio-Economic Environmental Demographic Information System) and PAREP (Populations at Risk to Environmental Pollution) projects. The electronic version of this report, which is located at http://parep2.lbl.gov/~merrill/thesis and http://www.bearhaven.com/thesis, will be updated in the future as necessary.

Ernest Orlando Lawrence Berkeley National Laboratory (LBNL), in Berkeley, California, was founded in 1931 by its namesake Ernest O. Lawrence. Known for years as the "Rad Lab," LBNL is frequently been confused with its better known sister laboratory Lawrence Livermore National Laboratory (LLNL), a nuclear weapons laboratory fifty miles to the east. Both Lawrence Laboratories, and also Los Alamos National Laboratory in New Mexico, are owned by the Department of Energy (DOE) and operated for DOE by the University of California. Unlike Livermore and Los

Alamos, LBNL has performed no classified work since World War II. During the 1950's and 1960's LBNL was recognized as the pre-eminent nuclear research laboratory in the United States, and it acquired computing power that was unrivaled outside classified installations.

As a result of this computing power, LBNL was able to perform contract work for other government agencies, that they could not perform for themselves without similar resources. In the early 1970's, LBNL contracted with the Bureau of the Census and the Department of Labor (DOL) to produce the Urban Atlas report series, a large-format series of color maps displaying the socio-economic characteristics of metropolitan areas in the U.S. LBNL's contribution included the digitization of 1970 Census tract boundaries. LBNL also contracted with DOL's Employment and Training Administration to produce a comprehensive set of printed "manpower" reports from 1970 Census data. A third major project, with the U.S. Army Corps of Engineers, called for creation of an early on-line information system known as REAP.

To put the project effort into perspective, recall that in 1970 personal computer terminals did not exist, and all computer input was via punched cards. The mainframe computer could read the punched cards and print text, even performing the optional miracle of justifying the text to the right margin. Only the most expensive line printers could print lower case letters. In the punched cards, lower case letters had to be specified by "escape sequences" - awkward combinations of two or more characters.

Then came the 110 baud paper teletype - a huge advance because one could edit electronic images of the punched cards, and even submit programs directly from

one's office. This technological breakthrough became so popular that the LBNL computers quickly became swamped. During the day it was not uncommon to type one line of text, then go for a cup of coffee while waiting for the computer's response, so the next line could be typed. Until the mid-1970's, the key punch was still the most efficient means of creating and editing computer programs.

In the early 1970's, LBNL's Computing Science and Applied Mathematics (CSAM) Department, under Carl Quong, numbered close to 100 employees. Carl coined the term SEEDIS (Socio-Economic Demographic Information System) and integrated the separate projects into a *program* which could effectively unite the growing resources into a coherent system. Deane Merrill was hired in 1973 with the mandate to become the SEEDIS "data guru," the role that he filled until his retirement from LBNL in 1997.

The present Department of Energy (DOE) was the AEC (Atomic Energy Commission) in the 1950's and 1960's, and ERDA (Energy Research and Development Administration) in the 1970's. Deane and his supervisor Donald M. Austin (not the epidemiologist Donald F. Austin) were co-delegates to ERDA's IWGDE (Interlaboratory Working Group for Data Exchange). In 1974-76, IWGDE representatives from ERDA's major research laboratories (Argonne, Berkeley, Brookhaven, Livermore, Oak Ridge, Pacific Northwest, and Savannah River) developed an ANSI standard for data exchange, and pioneered in the exchange of electronic mail and small data files via the fledgling Internet.

The first SEEDIS system, completed in 1975, ran on Control Data Corporation computers. SEEDIS provided interactive dialup access to a dozen small data files at

the state and county level. The user could select the geographic areas and data elements desired, without recourse to printed code books. LBNL devised a self-documenting file format called "codata," which provided communication among SEEDIS modules. The selected data could be automatically routed to an interactive spreadsheet (the word did not yet exist) called CHART, and a mapping program called CARTE. CHART and CARTE, the first programs of their kind, provided graphic output on Tektronix storage tube devices. In 1975, using dedicated cross-country phone lines and the entire capacity of the LBNL computing center, SEEDIS was demonstrated live to an audience of ERDA officials in Washington DC.

In 1976, Warren Winkelstein and Susan Sacks of UC Berkeley's School of Public Health obtained funding from the Environmental Protection Agency (EPA) to construct a database of "Populations at Risk to Air Pollution (PARAP)," which would integrate county level data on air quality, population, and mortality. The project was adopted by ERDA/DOE in 1978 and renamed "Populations at Risk to Environmental Pollution (PAREP)." Extending the concepts in Kernighan and Plauger's *Software Tools,* Deane Merrill created a set of "codata tools" to manipulate and integrate data files in the codata format. Beginning in 1979, SEEDIS was re-implemented on a network of Digital VAX computers, using the "codata tools" as the underlying data exchange mechanism.

In 1980, the Department of Labor contracted with LBNL to produce publications from the 1980 Census, similar to the "manpower" reports of the 1970 Census. SEEDIS funding from the Department of Labor continued until 1985, and PAREP project funding from DOE until 1994. Related funding was received from the

Army Corps of Engineers, the Centers for Disease Control, and the Electric Power Research Institute. During the 1980's, data were continually added to SEEDIS, including all of the 1980 Census (Summary Tape Files1,2,3,4), county mortality, SEER (Surveillance, Epidemiology and End Results) cancer incidence by tract, 1980 census tract map files, and dozens of other databases.

In 1977, Susan Sacks introduced Deane Merrill to Steve Selvin, Professor of Biostatistics in UC Berkeley's School of Public Health. The resulting PAREP collaboration, which lasted until Deane's retirement in 1997, produced more than thirty publications and a dozen Master's and Doctoral dissertations. From the beginning, drawing upon LBNL's unique resources, the focus of the PAREP project has been the application of biostatistical techniques to summary "ecologic" data. The pitfalls of analyzing ecologic data as if they were unit record data are not appreciated by many epidemiologists; those pitfalls were discussed in several early papers.

DENSITY EQUALIZING MAP PROJECTIONS

The PAREP project faced a classic dilemma in comparing disease rates among different geographic areas or time periods. Rates are inadequate, because in an area with small population, even one case can produce a rate of epidemic proportions. On the other hand, a level of significance such as "two standard deviations above normal" hides the rate, which is the quantity of underlying epidemiologic significance. Furthermore, such a significance level is correlated with population size.

Another classic dilemma is the problem of representing geographic variability; if the subareas are chosen too small, stable rates cannot be calculated; too large, and geographic detail is lost. Grouping of subareas to achieve stable rates requires arbitrary decisions that can affect the conclusions of the analysis.

A different mapping approach was first used as early as 1798, when Seaman plotted the locations of yellow fever cases in New York [SEAM1798]. Physician John Snow used the same technique to investigate a cholera outbreak in London in 1849 [SNOW1849]. Snow observed a cluster of cases in the vicinity of the Broad Street pump, concluded that the well was contaminated, and took it upon himself to remove the handle of the pump. Implicit in his interpretation of the map was the underlying assumption that the population density was relatively uniform.

The same approach, but with corrections for varying population density, was first described in the 1920's. [KARS23, WALL26, GILL27]. (An earlier cartogram is that of Haack in 1903, but it was not used to analyze disease distributions. [HAAC03]

) Prior to plotting the cases, county boundaries were adjusted so as to give to each county an area proportional to its population. Then a visible cluster of cases could be correctly interpreted as an increased rate in the region of interest. The first two authors constructed their maps manually with paper and pencil, but Gill [GILL27] was more imaginative. He weighed out lumps of plasticene with weights proportional to the individual county populations; he then assembled a map of the counties of California, with the lumps of plasticene in their proper relative locations. Then, rolling the lumps to uniform thickness with a rolling pin, he automatically created a density equalized map without recourse to a computer. In the present analysis the same trick is performed with a computer, but the underlying principle is no different from Gill's. Furthermore, his method was much faster than the computerized method, both in implementation and execution!

In the literature, density equalized maps have been called population maps, cartograms, contiguous-area cartograms, or anamorphoses. The quest for a computerized cartogram algorithm was pioneered by Tobler, who described the problem mathematically in 1961 and had a working computer program by 1970 [TOBL61, TOBL70]. Tobler's program, though primitive by today's standards, was no small feat considering the limitations of computers in 1970. Since 1970 at least a dozen different authors have implemented new algorithms; the programs are rich with innovative and original techniques. Comparing the robustness and speed of different algorithms from the authors' written descriptions is almost impossible. At LBNL, several different cartogram programs were obtained for comparative evaluation, but the task was abandoned as impractical. None of the programs would compile on the

LBNL computer without considerable effort, and each program required map input data in its own particular format.

Tobler and the other cartogram implementors, lacking health data, have not been particularly interested in analyzing geographic disease distributions. The authors of the 1920's public health papers, lacking computers, were unable to analyze their results quantitatively. Since both data and a computer were available at LBNL it was resolved, in 1985, to write a cartogram program for analysis of disease distributions. It was assumed (incorrectly) that if so many people were writing cartogram programs, it must be an easy task. Creating a working program was not too difficult, but to produce a robust algorithm that could reliably process large maps quickly and cheaply was quite another matter. With only part-time effort available for the task, almost ten years were required!

The first LBNL algorithm, published in 1988 [SCHU88], employs a radial expansion or contraction relative to the centroid of each subarea in the map. The radial transformation changes the area but not the shape of the particular subarea in question, while changing the shape but not the area of all other subareas. A solution is reached in one iteration; however, the resulting map is rather distorted, and the nature of the distortion depends on the arbitrary order in which the subareas are transformed. In addition, it is quite common for some subarea boundaries to illegally intersect one another during the transformation; once this occurs, the algorithm becomes nonsensical, and a solution cannot be reached.

A second LBNL algorithm, completed in 1991 [MERR91A] subdivides the map into triangles. Two functions of the coordinates of all the triangle vertices are

explicitly calculated: (1) a function H which shrinks to zero when density equalization is complete, and (2) a function G which measures distortion relative to the original map. An external minimization program adjusts all the vertex coordinates, minimizing G subject to the constraint H = 0. With the 1991 LBNL algorithm, unique solutions are found and overlapping boundaries are avoided; however, computation time is excessive. For a map composed entirely of triangles, the problem is almost but not quite overconstrained; there are so few degrees of freedom that a solution is never reached in practice. An additional inconvenience with the 1991 LBNL algorithm is the the program's reliance on a numerical minimization routine; efficient routines for large problems are available only in proprietary software packages.

In 1993 a mathematical algorithm was described by Gusein-Zade and Tikunov [GUSE93], in which the correction to be applied to each point in the map is calculated explicitly from the required expansion or contraction of each infinitesimal area in the entire map. Convergence is achieved in a small finite number of iterations. Every polygon in the map generates a radial "push" or "pull" on the the rest of the map, depending on whether its present area is smaller or larger than the target area determined by its population. The magnitude of the radial "push" or "pull" decreases with distance, exactly as required to keep constant the areas of polygons that are being passively transformed; *i.e.* which already have the correct target area.

The breakthrough in [GUSE93] is the application of Stoke's Theorem, by which the calculation of area integrals is replaced by line integrals around the boundary of each subarea. The calculation is not difficult since each subarea is a polygon of constant population density, delineated by a finite number of line

segments. A detailed mathematical description of the algorithm is included in [CLOS94].

In 1994 the Russian algorithm was independently implemented at LBNL. Additional features were added [CLOS94, MERR95A], which were found to be necessary for equalizing highly non-uniform populations like that of the four-county area. In addition, proper map preparation prior to density equalization is essential for reducing the calculation time. The problem is not the small urban tracts which need to be inflated, but rather the large sparsely settled tracts in rural areas, which need to "stretch" around the urban areas while their area is being reduced to practically zero. There must be sufficient geographic detail to avoid illegal overlapping of polygon boundaries during the iteration process. On the other hand, excessive detail must be avoided, because calculation time increases as the square of the number of points in the map. Finding the right balance is something of an art, and how to automate the process is not obvious.

FOUR COUNTY CHILDHOOD CANCER STUDY

As a demonstration and a test of the DEMP technique, it was decided to apply the DEMP methodology to the re-analysis of a data set that had been previously analyzed by conventional techniques. The data used were provided through a collaborative agreement between Lawrence Berkeley National Laboratory (LBNL) and the California State Department of Health Services (DHS). The data were originally collected and analyzed by DHS to investigate a reported childhood cancer cluster in the community of McFarland, California. [SATA90, REYN91, REYN96]. The data consist of 401 childhood cancer cases occurring between 1980 and 1988 in four counties (Fresno, Kings, Kern and Tulare).

To facilitate comparison, the same data and selection criteria as in the original study were used, as far as possible. However, the methodology for estimating population is intentionally different from that used by DHS, for three reasons: (1) the detailed population estimates produced by DHS had been lost, so that detailed comparisons were impossible; (2) 1990 Census data were not yet available at the time of the DHS analysis; and (3) the DHS population estimates incorporated some data that are available only for California. The present analysis uses only decennial Census data that are available nationwide, so the same techniques can be analogously applied to different data sets in the future.

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 an Appndix. 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 were found to be slightly higher than those in rural non-agricultural areas. The urban non-agricultural rates were comparable to urban rates elsewhere in California. Rates in agricultural areas were not elevated.

The present analysis is concerned primarily with the second DHS report [REYN91], which was published in final form as [REYN96]. The second DHS report examined differences among specific geographic areas; specifically, among 101 communities in the four-county area. 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 uniform underlying rates. One community (McFarland) had an elevated rate outside the 99% confidence limit, almost exactly what would have been expected from chance alone.

Map 5.

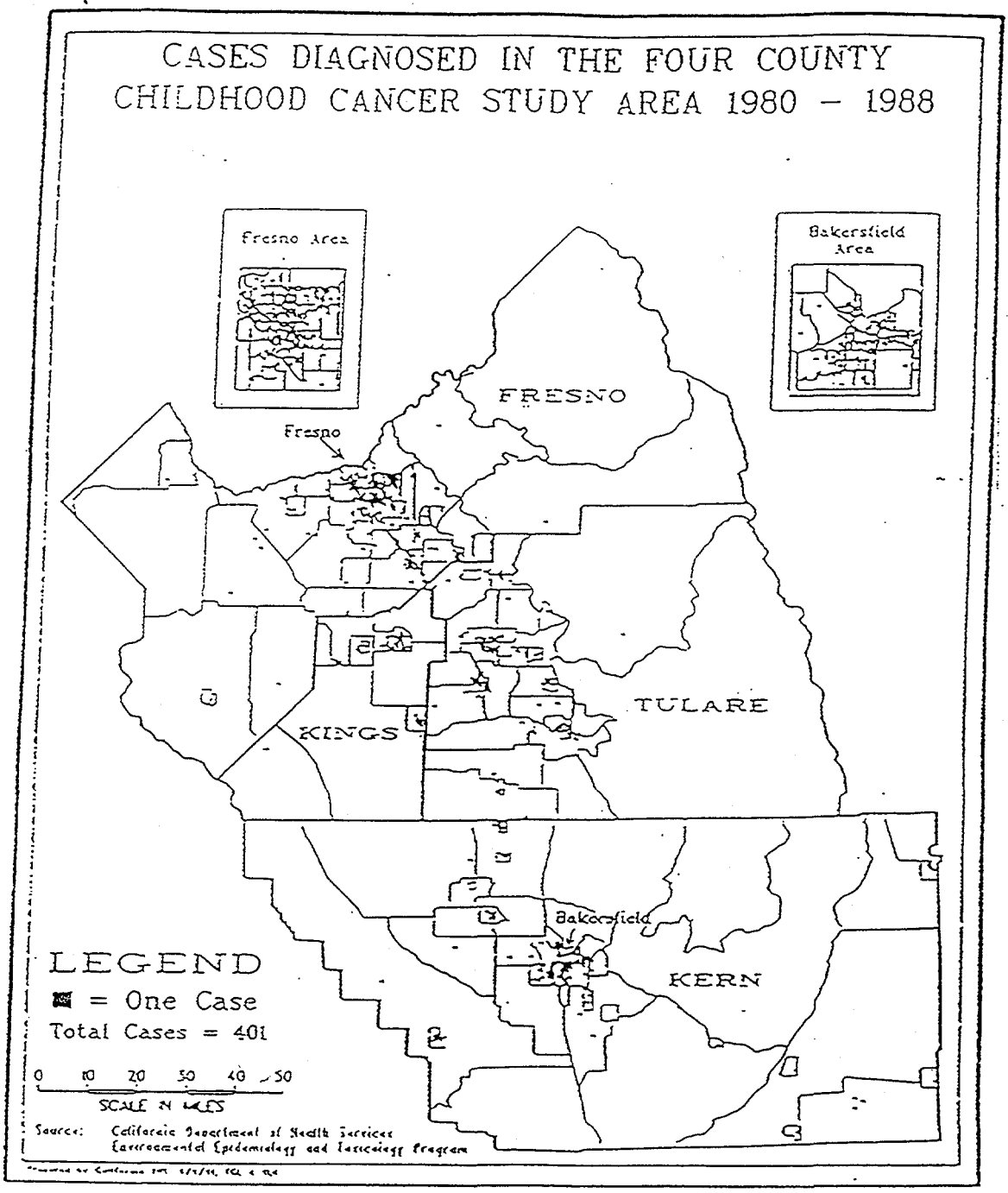


Figure 1. Cases Diagnosed in the Four County Childhood Cancer Study Area 1980-1988, from REYN91 (California Dept. of Health Services). The locations of the 401 childhood cancer cases, and the boundaries of the 101 communities used in the DHS analysis, are indicated.

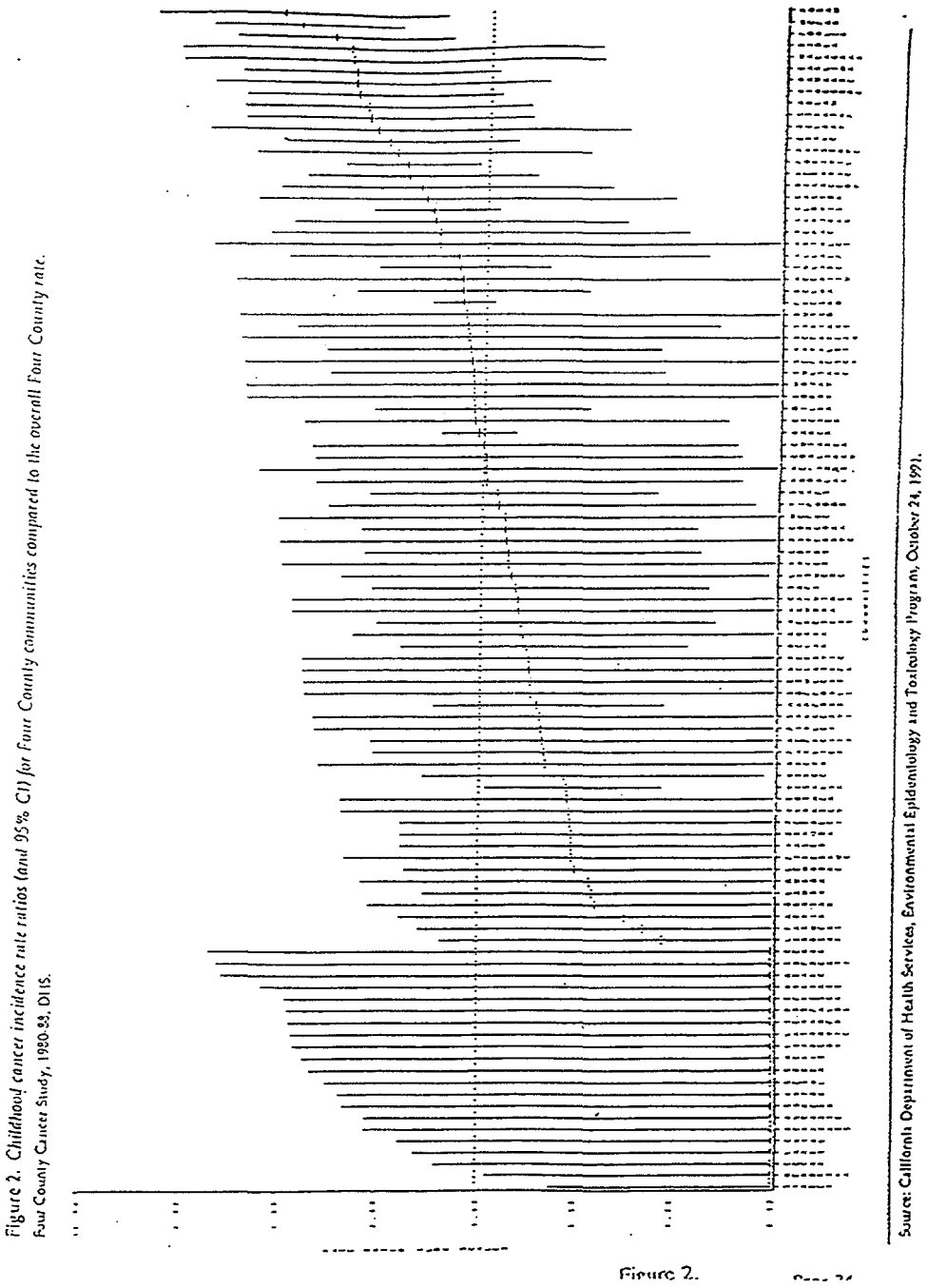


Figure 2. Childhood cancer incidence rate ratios (and 95% CL) for Four County communities compared to the overall Four County rate, from REYN91. At the 95% CL, three of the 101 communities have rates that are significantly low, and three have rates that are significantly high.

Map 6.

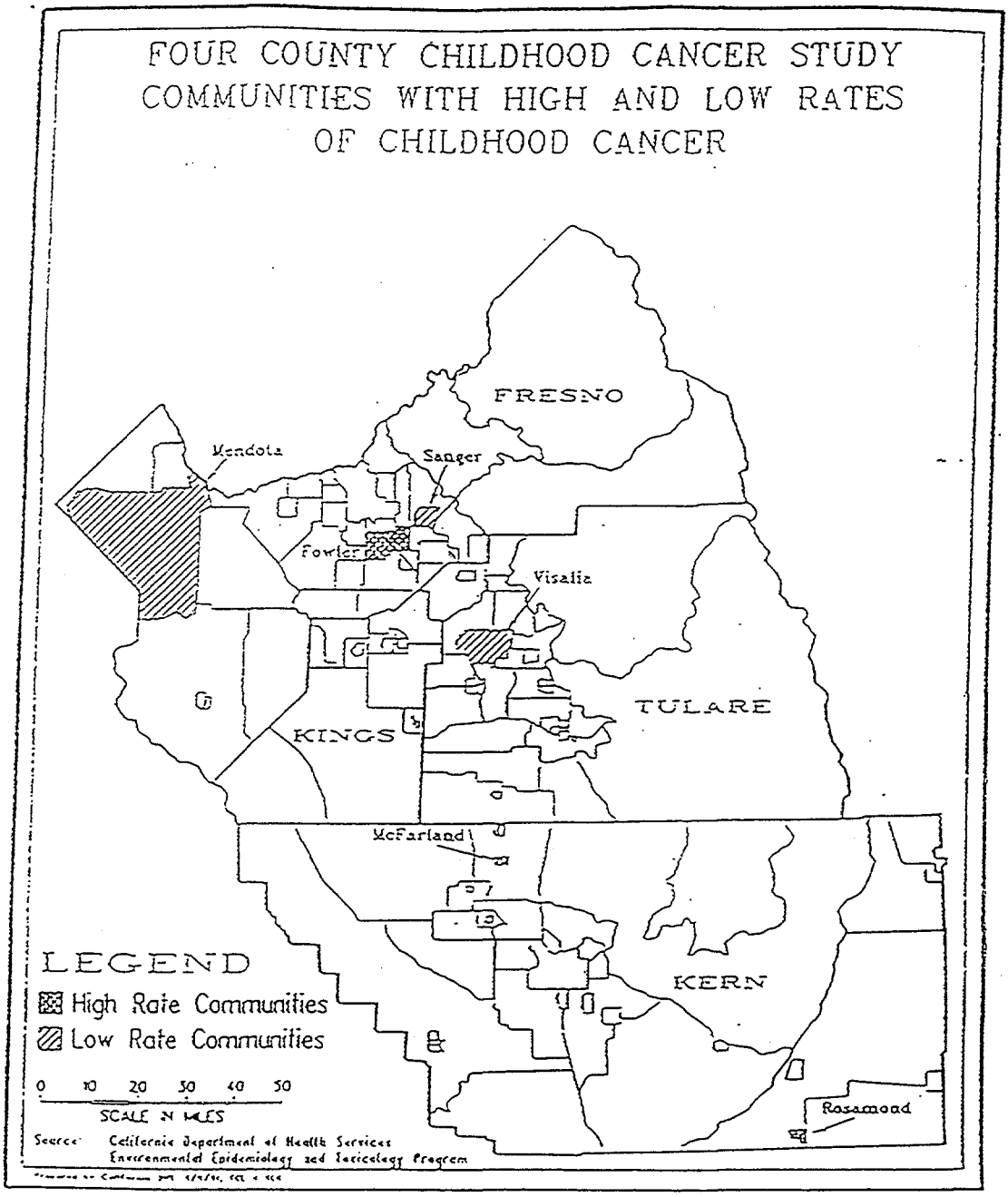


Figure 3. Four County Childhood Cancer Study communities with high and low rates of childhood cancer, from REYN91. The three communities with significantly low rates in Figure 2, and the three with significantly high rates, are indicated.

DATA ANALYSIS

In this section, preparation of the data files is described only briefly. Detailed documentation is provided in electronic files whose locations are given in Appendix A.

Case data

The study design was dictated by the characteristics of the case data. Each of the 401 childhood cancer had an associated census tract code, city code, latitude and longitude, race and ethnicity (white, hispanic, or other), year of incidence, age group (0-4 or 5-14), sex, and cancer site (leukemia, brain cancer, or other). The variables were distributed as shown in Table I.

Both in [REYN91] and in the present report, a single analysis was performed for the full data set of 401 cases. Also, each subsample listed in Table I was analyzed separately, to look for geographic effects that might be related to any of the five stratification variables (race/ethnicity, year, age, sex, cancer site). The results were consistently negative, and so the presentation focuses on the analysis of the complete data set.

In Table I, selection criteria that define subsets of the full data set are displayed in bold face type. The thirteen subsets listed in Table I are *not* statistically independent; however, the three race subsets (white, Hispanic, other) *are* independent from each other, as are the two time periods (1980-84 and 1985-88), etc.

Table I. Distribution of variables in case data.					
race and ethnicity	years	ages	sex	cancer site	number of cases
all	1980-88	0-14	both	all	401
white	1980-88	0-14	both	all	192
Hispanic	1980-88	0-14	both	all	166
other	1980-88	0-14	both	all	43
all	**1980-84**	0-14	both	all	209
all	**1985-88**	0-14	both	all	192
all	1980-88	**0-4**	both	all	211
all	1980-88	**5-14**	both	all	190
all	1980-88	0-14	**male**	all	226
all	1980-88	0-14	**female**	all	175
all	1980-88	0-14	both	**leukemia**	134
all	1980-88	0-14	both	**brain**	76
all	1980-88	0-14	both	**other**	191

Locations of the cases

In a preliminary analysis [MERR95] the exact latitude and longitude of each case were used, but this was not done in the present analysis for two reasons:

(1) The analysis is conducted at the census tract level. Plotting each case at its exact location within a tract leads to a statistical bias in the analysis of the final DEMP map if uniform rates are assumed, since population data below the tract level were not available in the present analysis.

(2) After removal of geographic detail (explained later) the LBNL map file differs from the one used by DHS. A case can be incorrectly assigned to the wrong tract if latitude and longitude are used to determine its tract. Although the latitude

and longitude provided by DHS were not used in the present analysis, it was verified that they yielded the correct tract assignments when compared with the LBNL pre-DEMP map of 1980 Census tracts.

The non-utilization of exact location within a tract has important benefits for the DEMP technique:

(1) It permits the analysis of data sets; e.g., SEER cancer incidence, or county mortality data, where only the geocode of residence is publicly available.

(2) The cases can be plotted on the density equalized map *after* the DEMP calculation has been completed. This means that analysts can use a DEMP map prepared in advance by a third party. The confidential case data, which are not needed for the DEMP calculation, remain in the hands of the analyst.

(3) On the final DEMP map where tract boundaries have been removed *(e.g.* Figures 19-32), one cannot identify the census tract of an individual cases, nor even count exactly the number of cases in a given tract. Such a plot conveys all the *significant* geographic information of the original map (e.g. Figure 1), without the risk of compromising confidentiality.

Units of analysis: Census tracts

In the DHS analysis, the units of analysis were the 101 communities which are shown in Figure 1. The communities are of vastly different size; for example, Fresno and Bakersfield each constitute a single community. The present analysis used Census tracts, which are approximately uniform in population, and which provide additional geographic detail in the densely populated areas. Because the case data

span the time period 1980-88, tract level population data from both the 1980 and 1990 Census were used.

Eestimates of population at risk, 1980-88

Next were obtained, separately, age/sex/race-specific 1980 population estimates for the 262 tracts of the 1980 Census, and age/sex/race-specific1990 population estimates for the 306 tracts of the 1990 Census. The 1980 and 1990 population estimates were aggregated to a consistent set of geographic units; namely, 259 modified 1980 Census tracts. Under the assumption that population change was linear in each tract between the two census dates (4/1/80 and 4/1/90), age/sex/race-specific estimates of population at risk were obtained for each of the 259 modified tracts, for each of the two time periods 1980-84 and 1985-88.

Preparation of geographic map files

There were 262 tracts in 1980 and 306 tracts in 1990 (Figures 4 and 5, respectively). SEEDIS contains a complete set of proprietary 1980 Census tract map files, which had been purchased from National Planning Data Corporation and installed in SEEDIS in 1986. Editing was required to repair topological errors, to remove small lakes, and to "sew" together the four separate county maps. The editing was partially automated with the use of routines described in Appendix A.

Maps of 1990 Census tract boundaries, for the four counties, were purchased from Geographic Data Technology, Inc. Some editing was required.

Unnecessary geographic detail was removed from the 1980 and 1990 geographic map files. The simplified maps are shown in Figures 6 and 7. Visual inspection, aided by routines written at LBNL, provided the proper correspondence between the five input data files: 1980 and 1990 map files, 1980 and 1990 Census data, and the 1980-88 cancer case data. The geographic units chosen for the analysis are 259 modified 1980 Census tracts, which are shown in Figure 8. These are identical to the 262 original 1980 Census tracts shown in Figure 6; except that five large 1980 tracts had to be aggregated into two larger "modifed tracts," in order to achieve correspondence with the 1990 Census tract definitions. Further map modifications were made: every polygon was subdivided by a Delaunay triangulation [BOOT87]; then every segment in the map was subdivided, thereby converting every triangle into a hexagon. This provided enough degrees of freedom that the DEMP calculation converged successfully.

Density equalized maps

With the use of the DEMP program described earlier [CLOS94], ten density equalized maps (Figures 9 through 18) were produced, one for each of the ten demographic subsets in Figure 1 (the total, three race/ethnic groups, two time periods, two age groups, and both sexes).

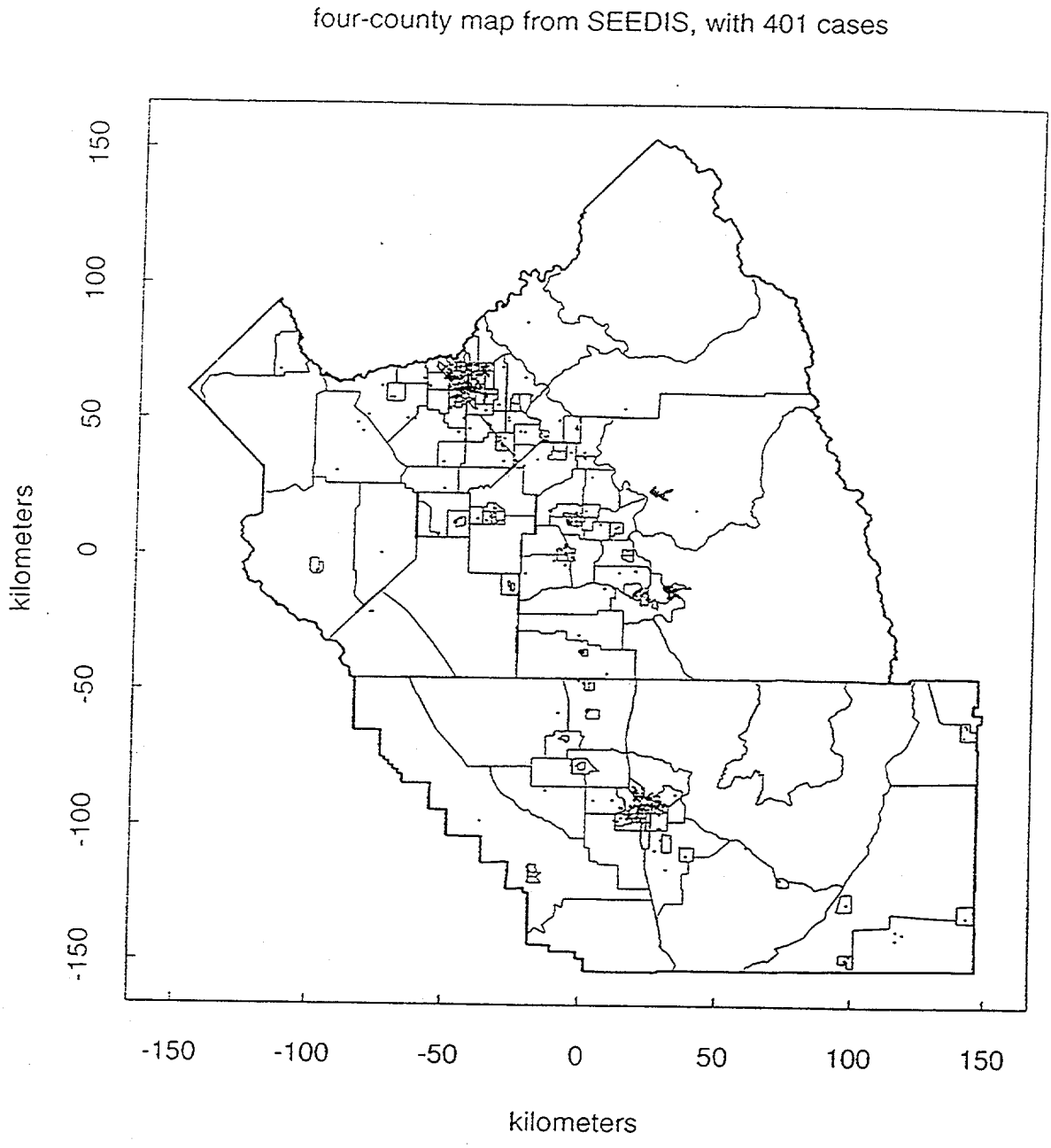


Figure 4. Four-county map from SEEDIS, with 401 cases. The case locations are the same as in Figure 1. The boundaries shown are those of the 262 tracts defined in the 1980 Census. The occasional darker segments along county boundaries indicate where the four separate county map files did not match exactly.

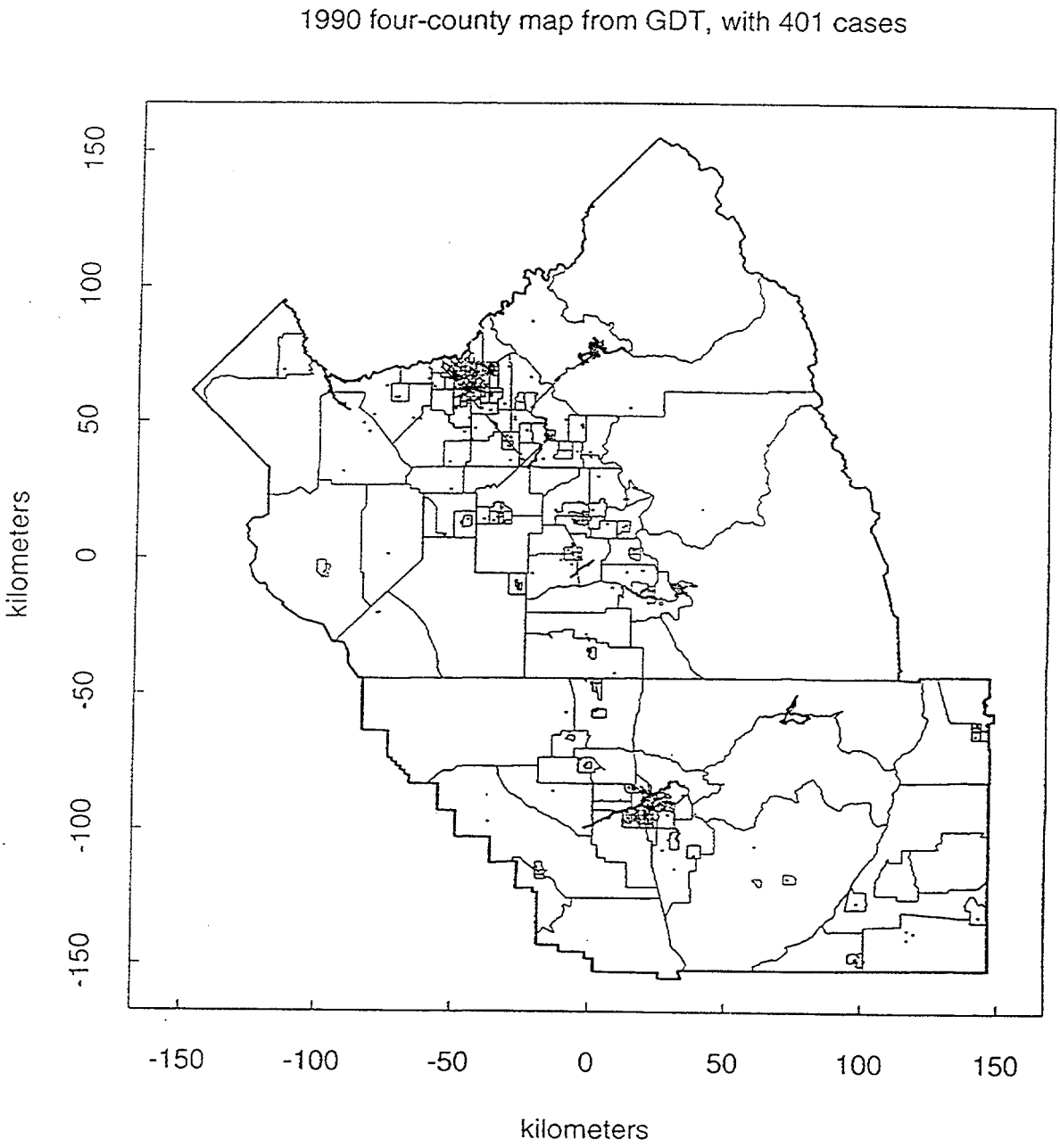


Figure 5. Four-county map from Geographic Data Technology, with 401 cases. The case locations are the same as in Figure 1. The boundaries shown are those of the 306 tracts defined in the 1990 Census.

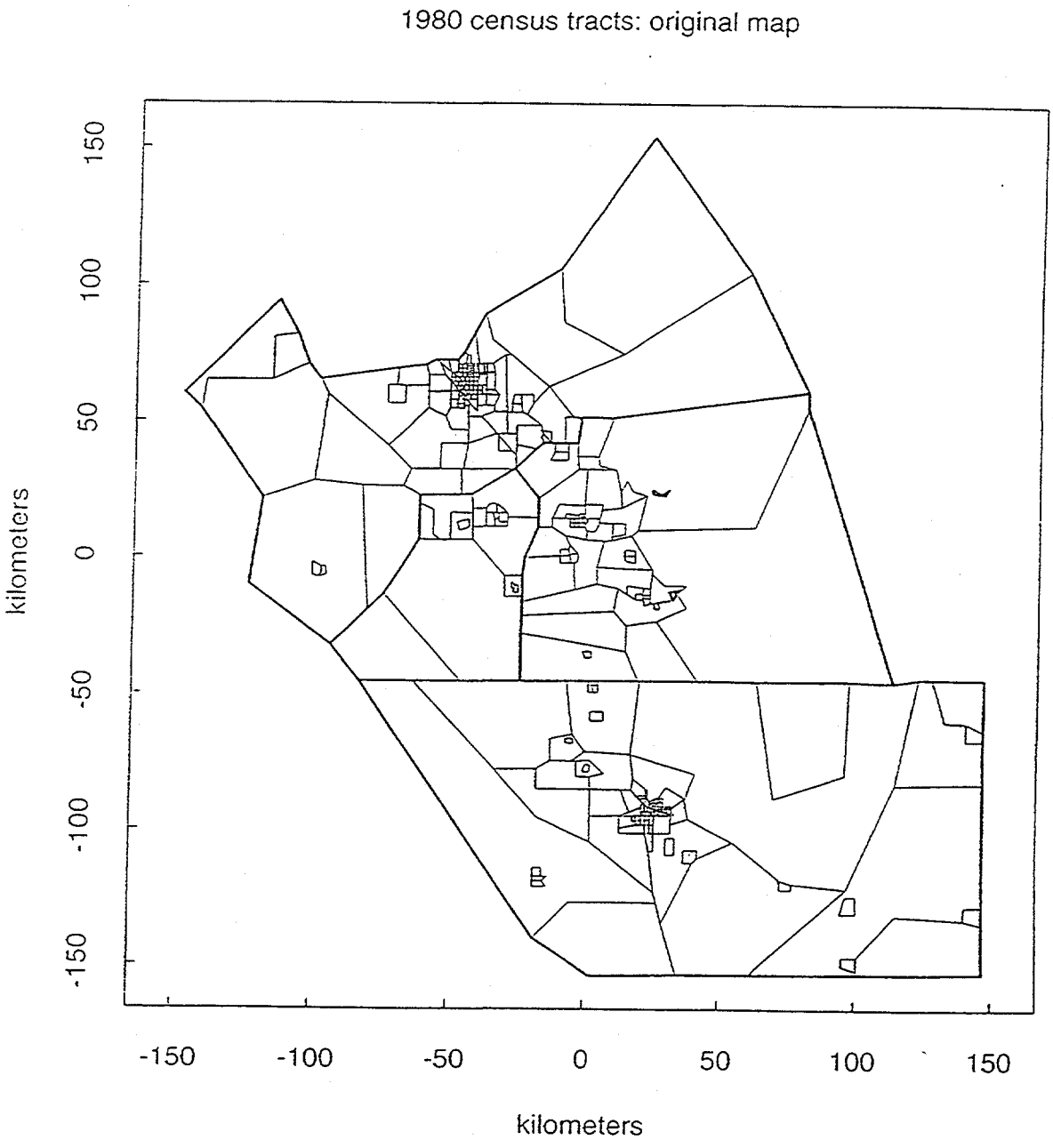


Figure 6. Boundaries of the 262 tracts defined in the 1980 Census, after removal of unneeded geographic detail. This map was automatically produced from the one in Figure 4, after some hand editing to remove map errors.

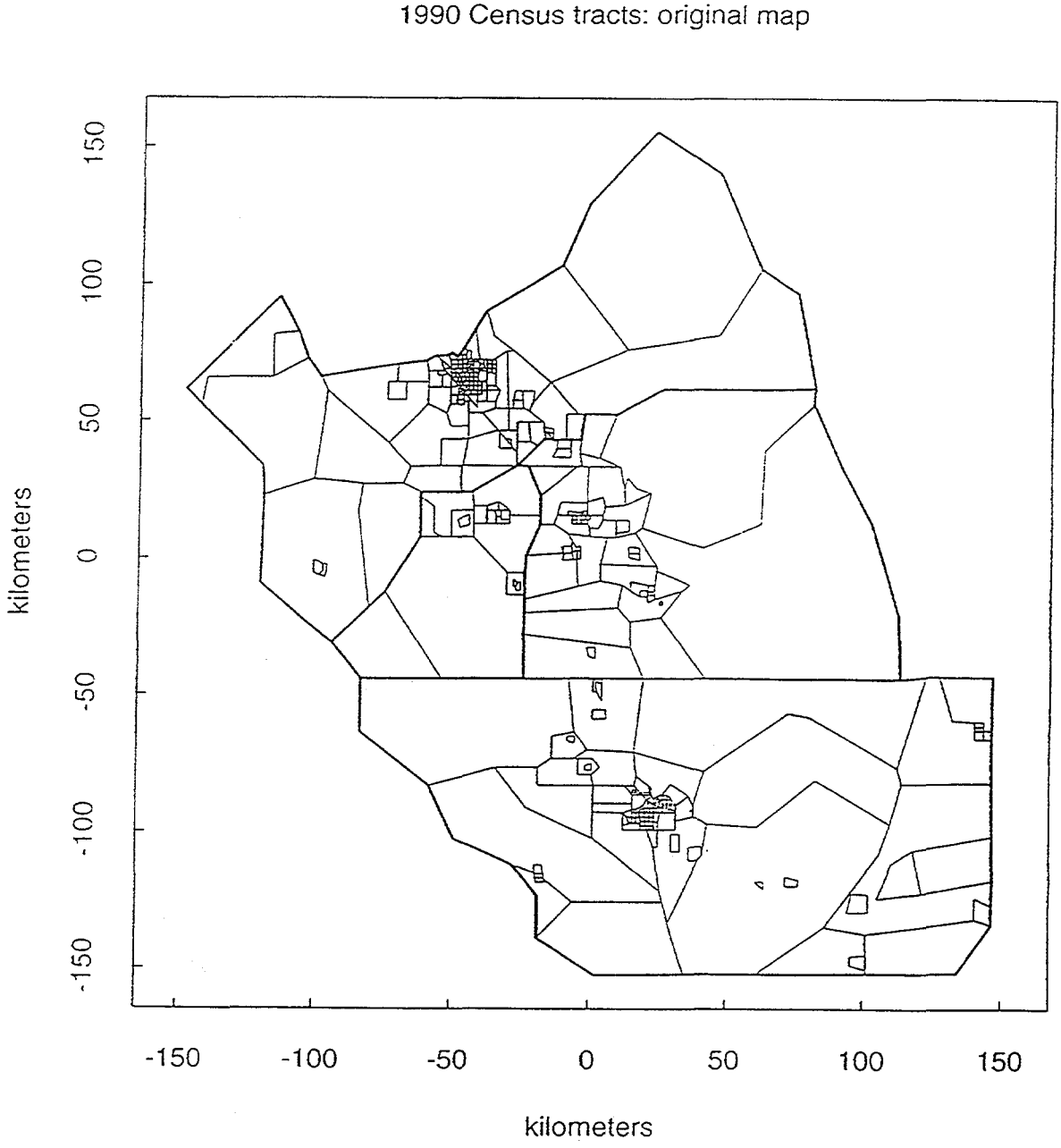


Figure 7. Boundaries of the 306 tracts defined in the 1990 Census, after removal of unneeded geographic detail. This map was automatically produced from the one in Figure 5. With only a few exceptions, the (306) 1990 tracts shown here nest within the (262) 1980 tracts shown in Figure 6, if minor boundary changes are ignored.

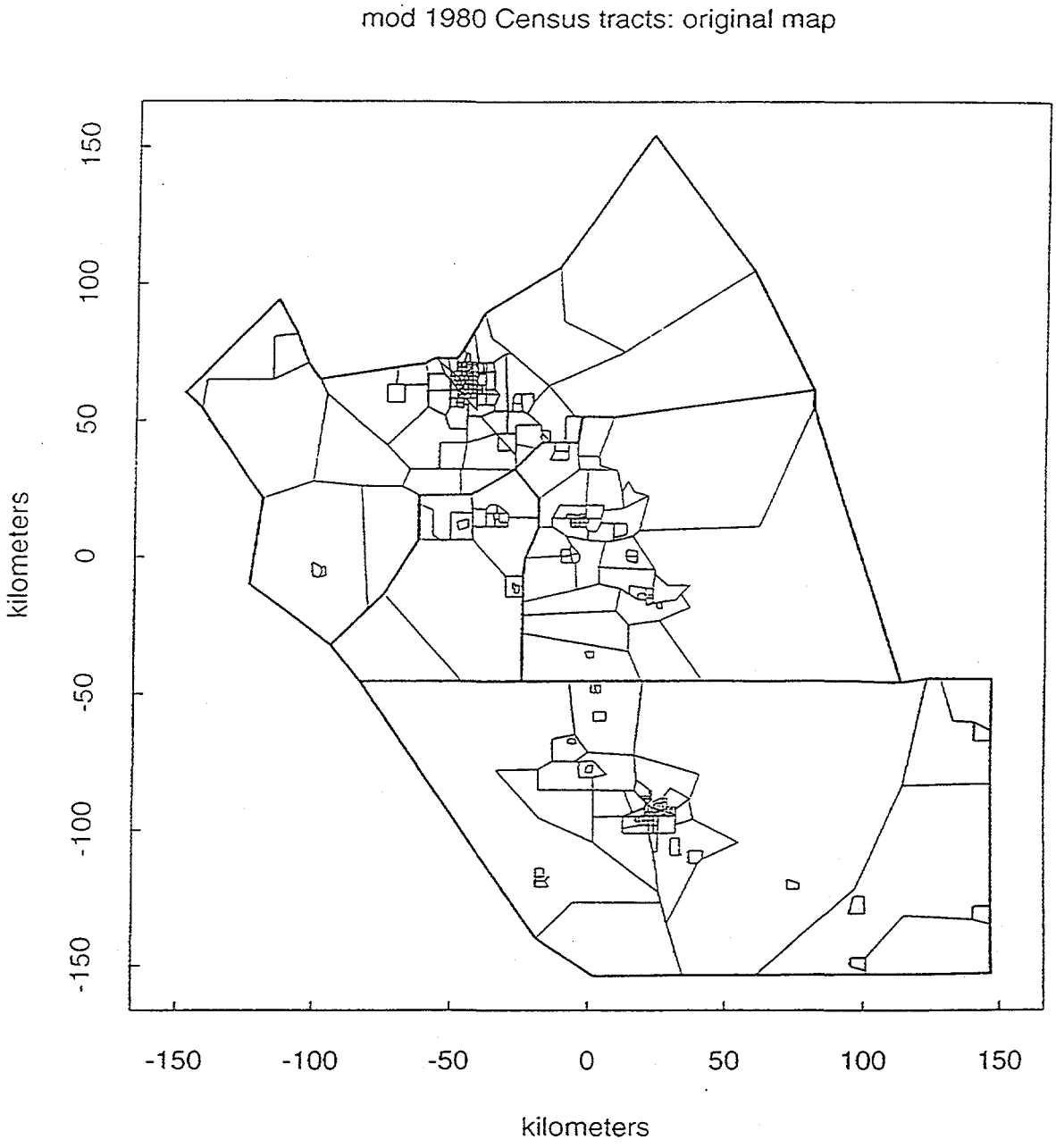


Figure 8. Boundaries of 259 modified 1980 Census tracts. These tracts are aggregates of *either* the (262) 1980 tracts in Figure 6, *or* the (306) 1990 tracts in Figure 7, if minor boundary changes are ignored. The 259 modified 1980 tracts were used for the remainder of the analysis in this report.

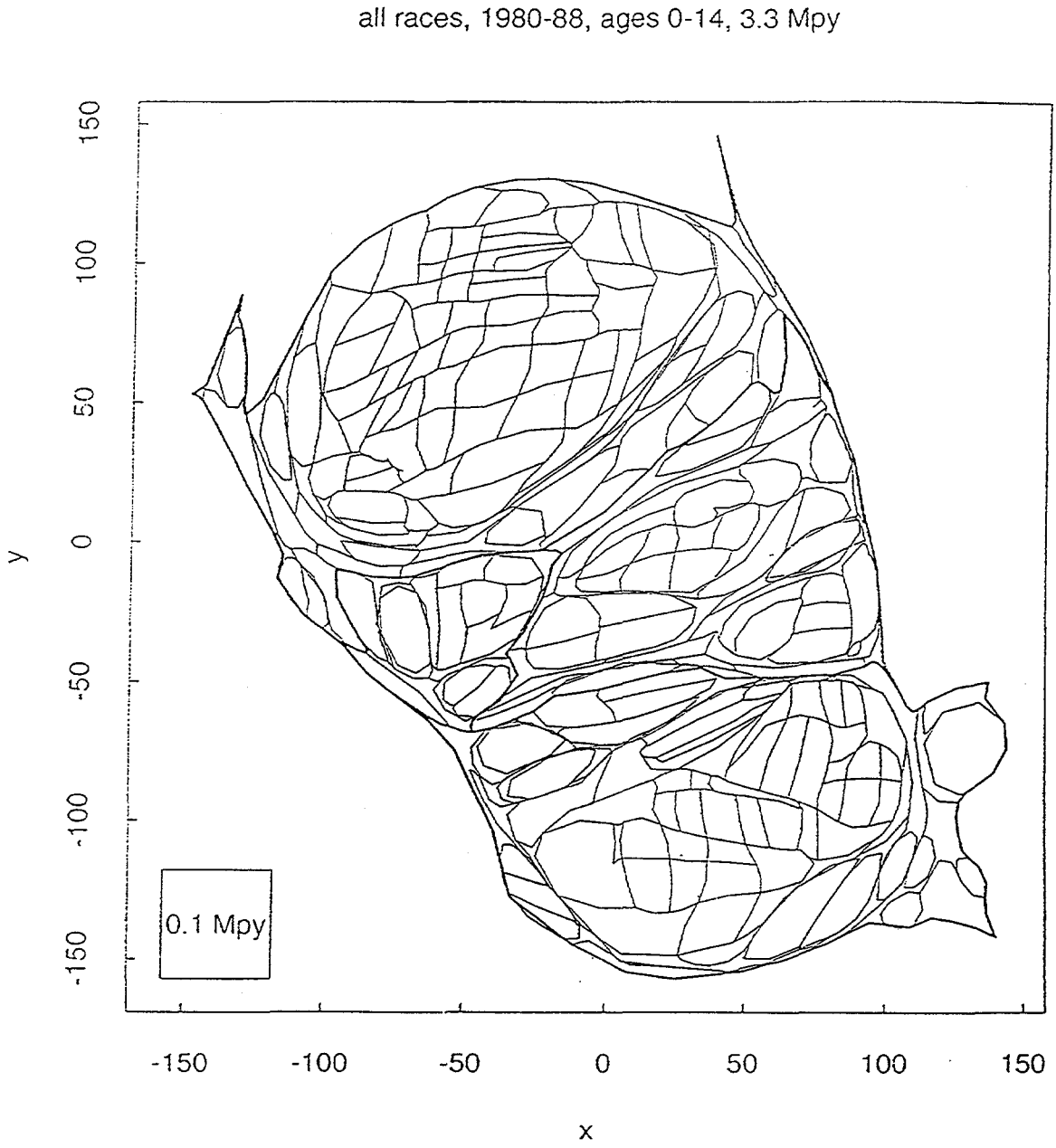


Figure 9. Density equalized map, all races, 1980-88, ages 0-14, both sexes, 3.3 million person-years at risk (Mpy). The square in the lower left corner shows the area corresponding to 0.1 Mpy. The subareas in the map are the 259 modified 1980 Census tract boundaries; the heavier lines in the map are county boundaries

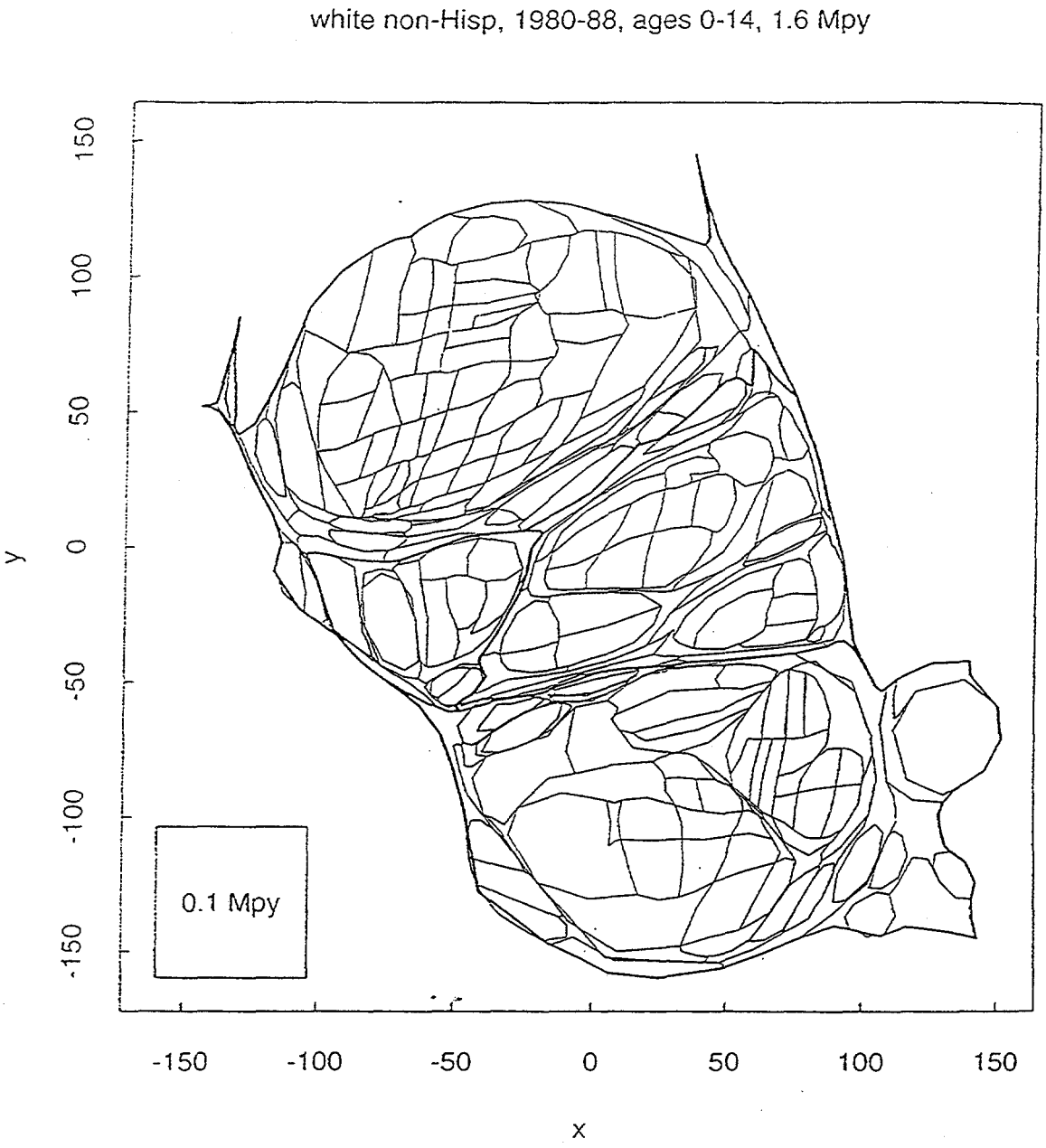


Figure 10. Same as Figure 9, for white non-Hispanics, 1980-88, ages 0-14, both sexes, 1.6 Mpy.

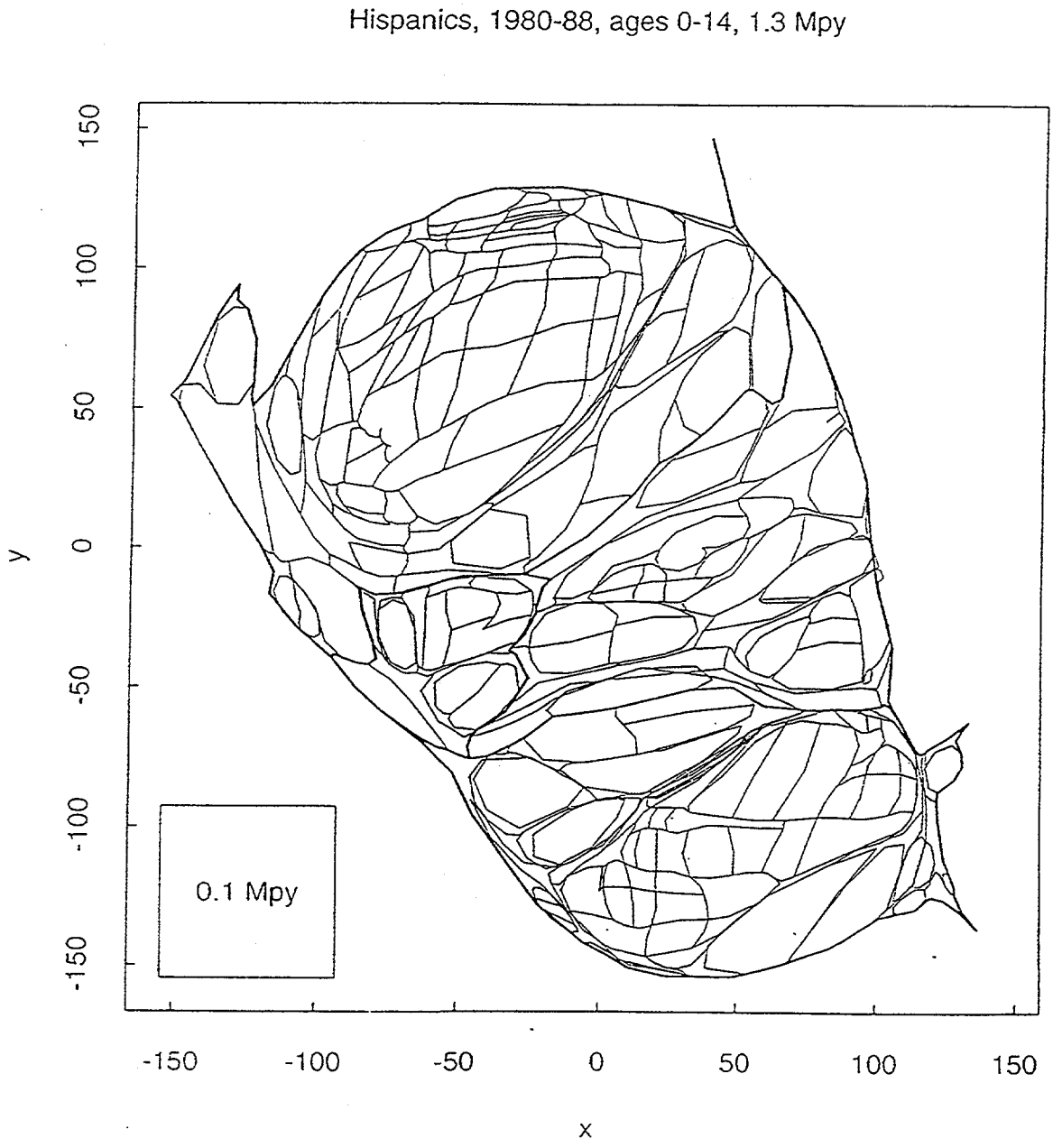


Figure 11. Same as Figure 9, for Hispanics, 1980-88, ages 0-14, both sexes, 1.3 Mpy. The differences between Figures 10 and 11 identify those subareas with different proportions of Hispanics.

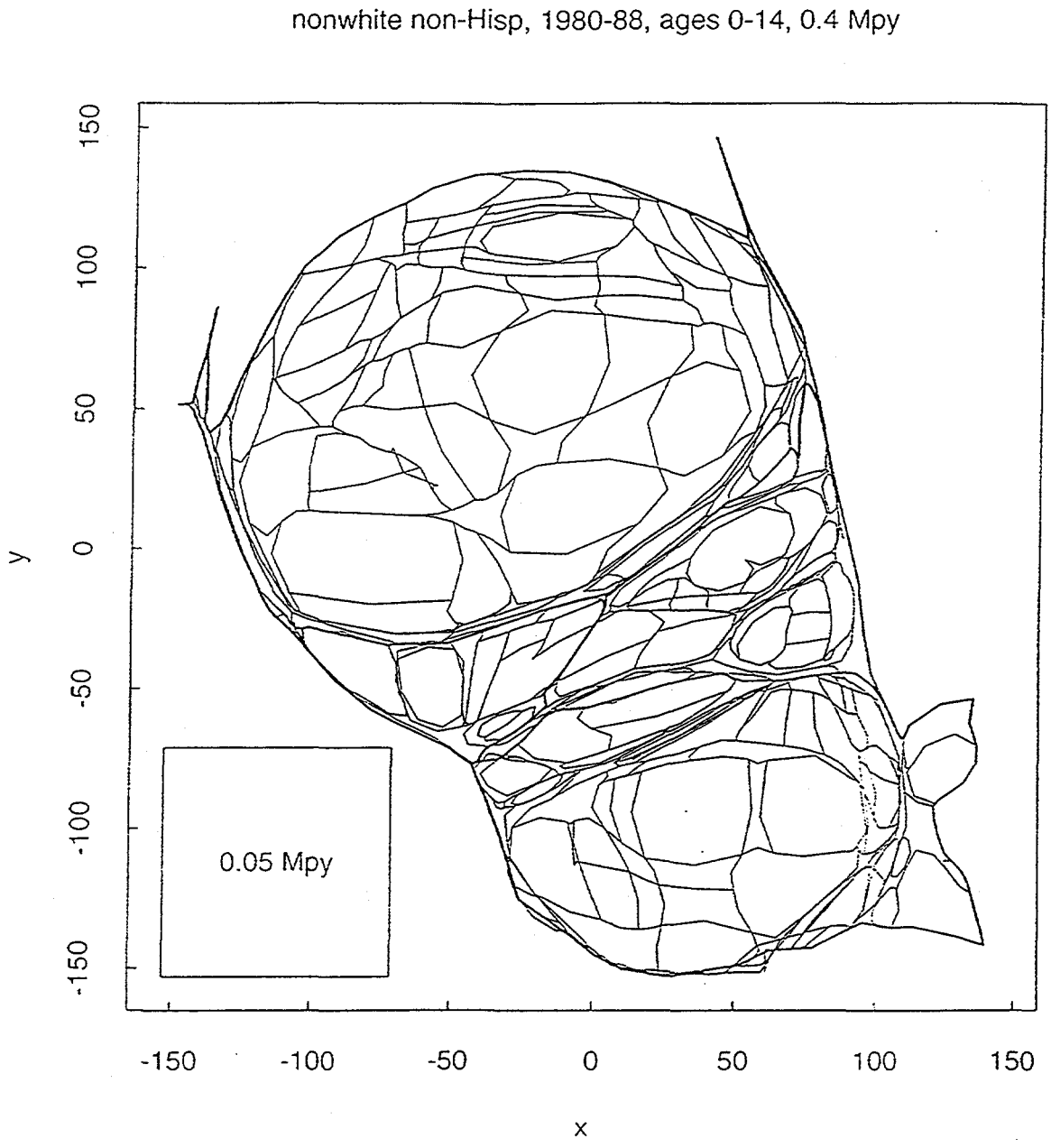


Figure 12. Same as Figure 9, for nonwhite non-Hispanics, 1980-88, ages 0-14, both sexes, 0.4 Mpy. The square in the lower left corner corresponds to 0.05 Mpy. Relative to whites and Hispanics, nonwhite non-Hispanics (mostly blacks) are more concentrated in the urban areas of Fresno and Bakersfield.

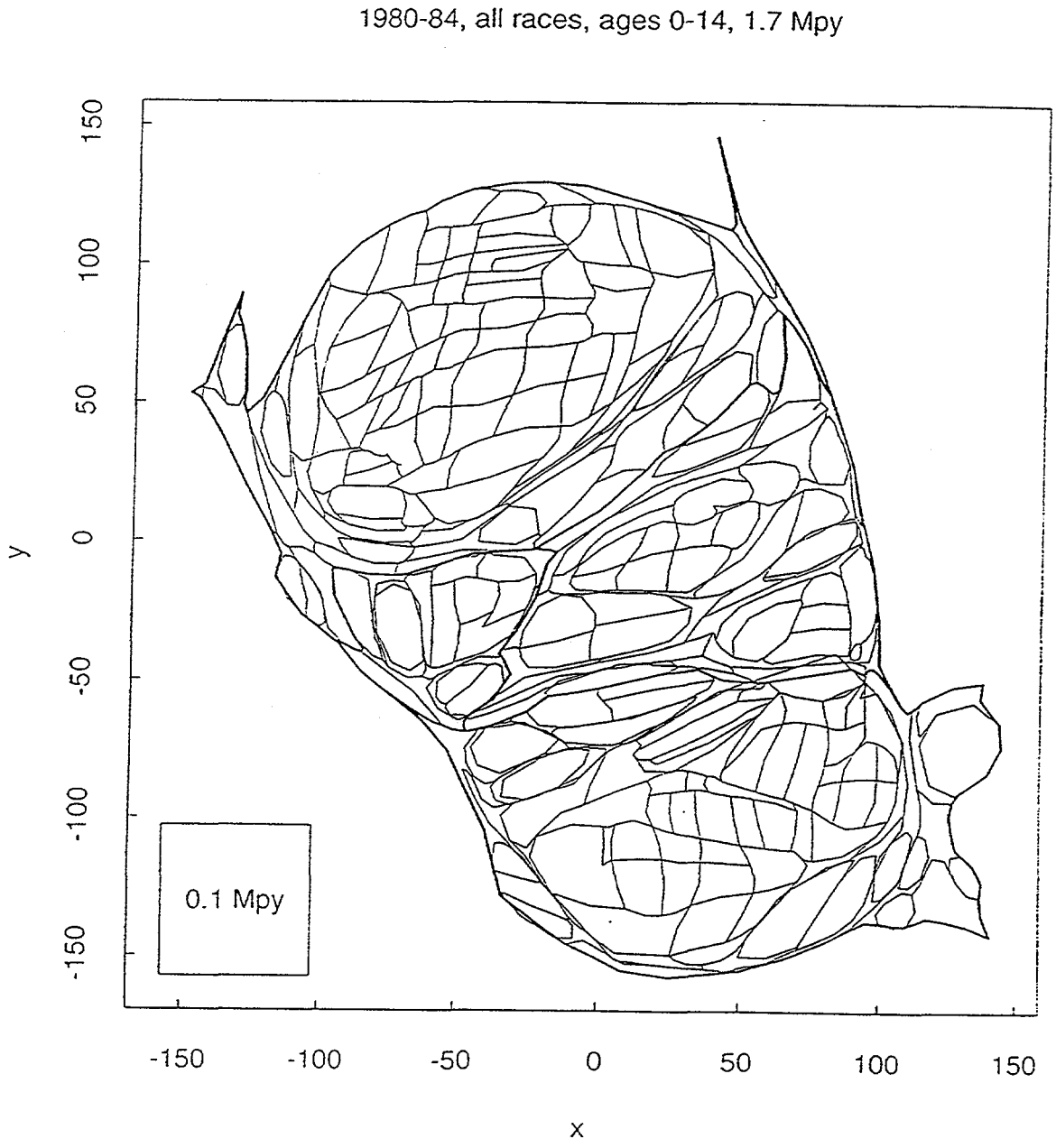


Figure 13. Same as Figure 9, for 1980-84, all races, ages 0-14, both sexes, 1.7 Mpy.

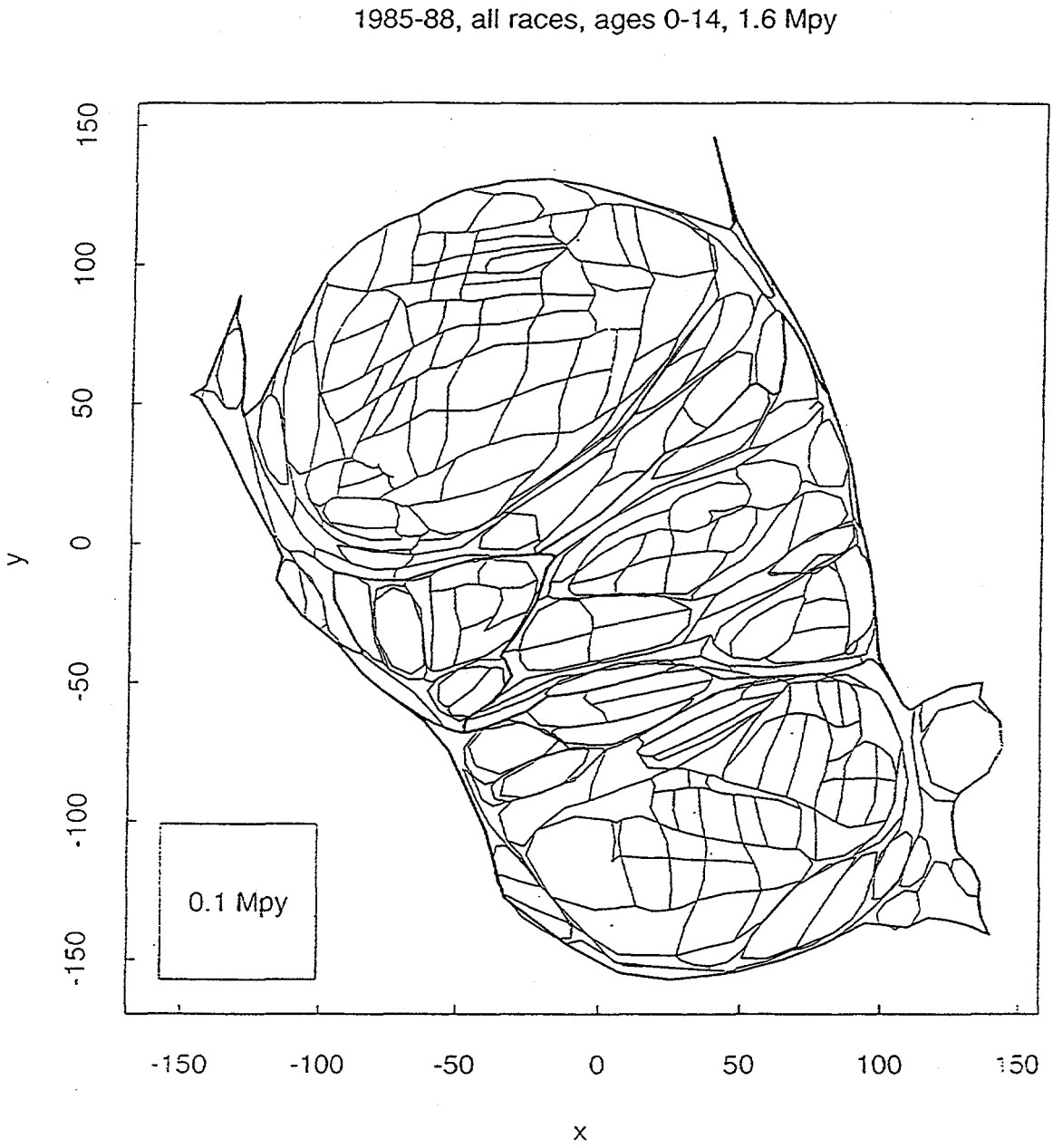


Figure 14. Same as Figure 9, for 1985-88, all races, ages 0-14, both sexes, 1.6 Mpy. No significant differences are observed between 1980-84 and 1985-88.

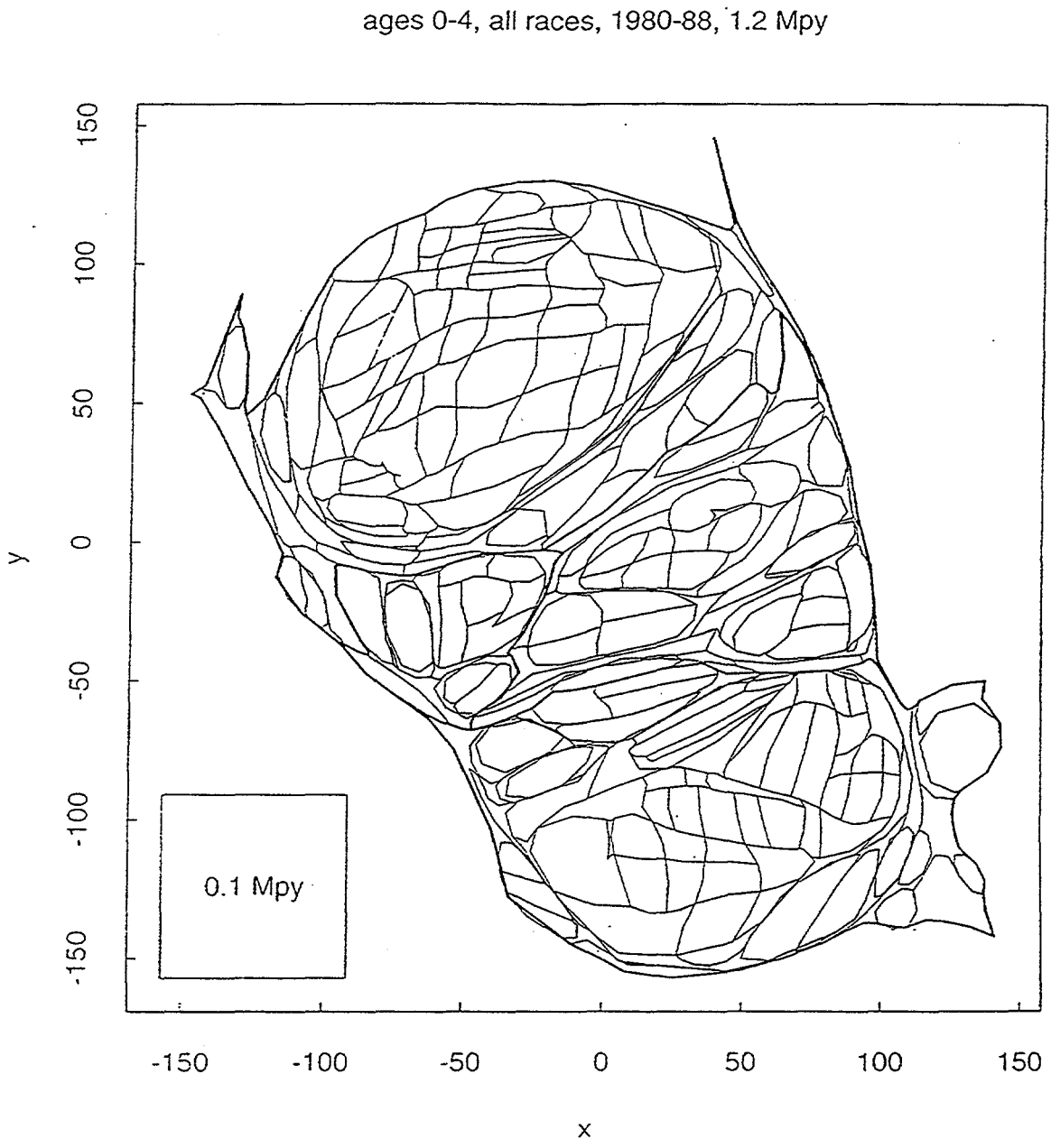


Figure 15. Same as Figure 9, for ages 0-4, all races, 1980-88, both sexes, 1.2 Mpy.

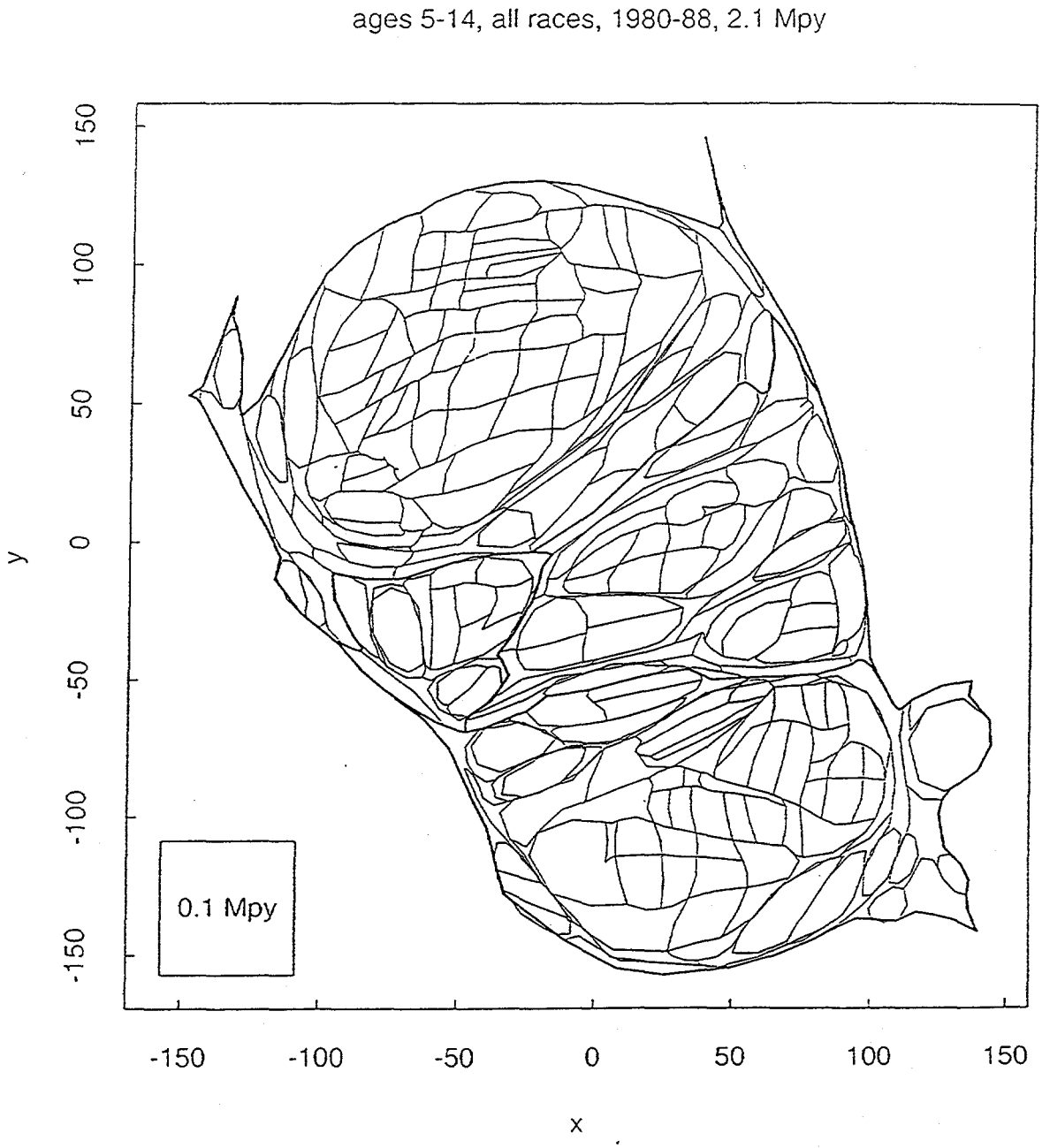


Figure 16. Same as Figure 9, for ages 5-14, all races, 1980-88, both sexes, 2.1 Mpy. No significant differences are observed between ages 0-4 and ages 5-14.

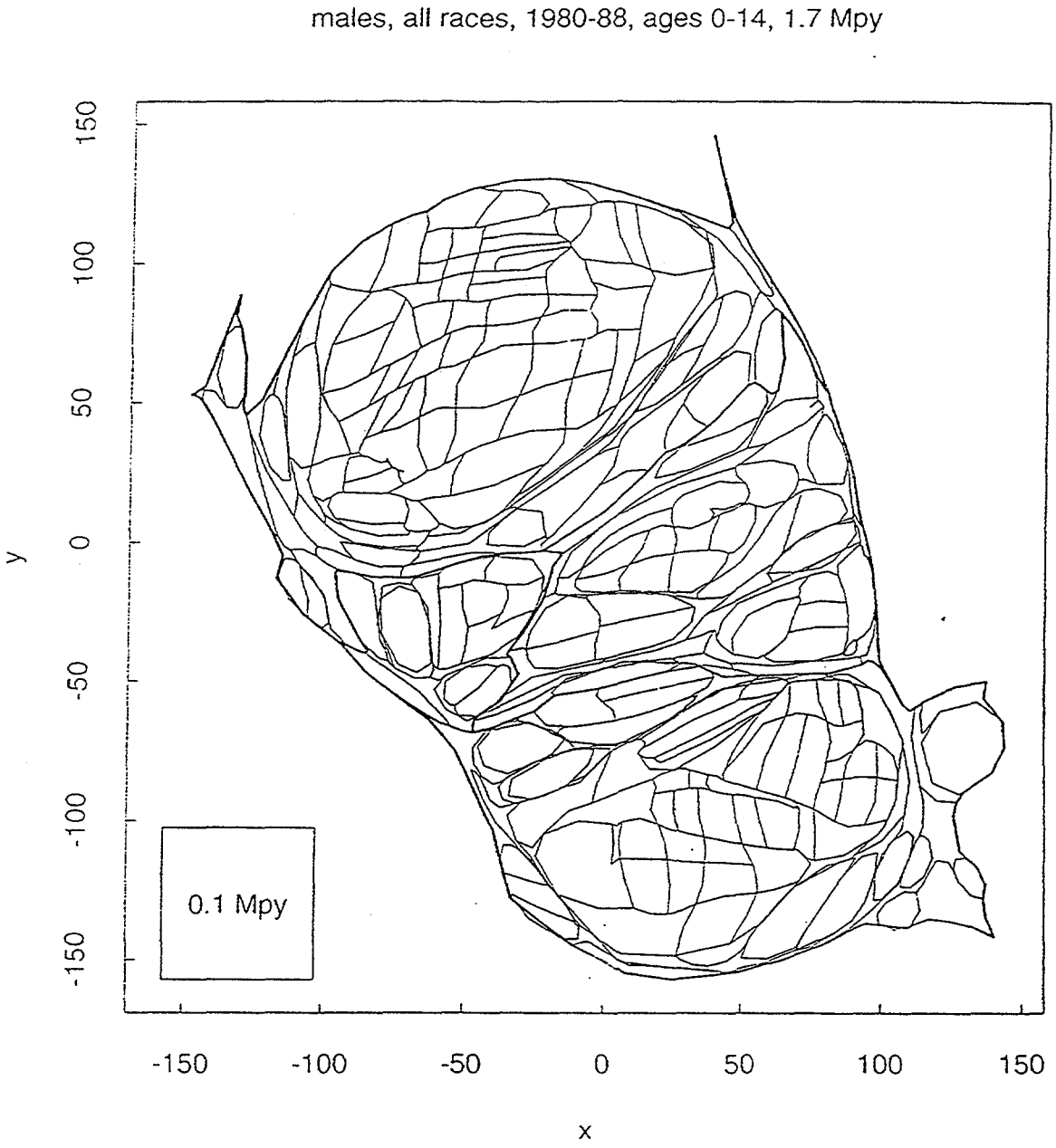


Figure 17. Same as Figure 9, for males, all races, 1980-88, ages 0-14, all races, 1980-88, 1.7 Mpy.

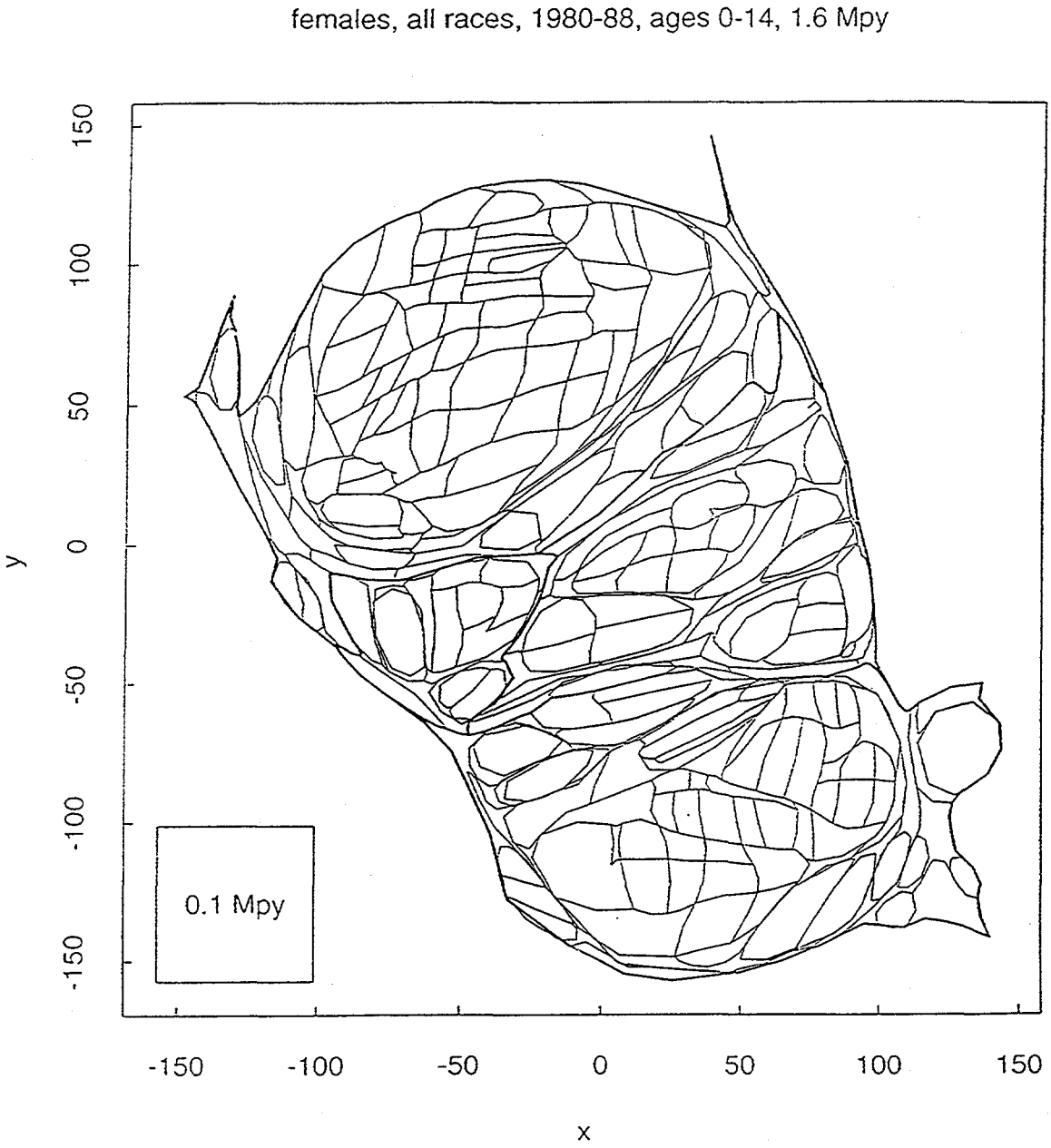


Figure 18. Same as Figure 9, for females, all races, 1980-88, ages 0-14, all races, 1980-88, 1.6 Mpy. No significant differences are observed between males and females.

Case distributions in the density equalized maps

The 401 cases were first plotted (in Figure 19) on the original map (from Figure 8). In the two upper insets of Figure 19, each case was plotted at two different random locations in the tract where it occurred. In the two lower insets, 401 artificial cases were similarly plotted, with the tract for each case chosen at random under the assumption that rates are everywhere equal.

Then, similar plots (Figures 20-29) were made from each of the ten density equalized maps in Figures 9 through 18. The "total" DEMP map (Figure 9) was additionally used for three subsets of the case data (leukemia, brain cancer, and other cancers), bringing the number of DEMP analyses to thirteen (Figures 20 through 32). The artificial cases in the *lower* two insets of each map are random *by construction;* any perceived clusters are due purely to random variation. Any perceived clusters among the real cases in the two *upper* insets must be significantly more noticeable than those in the lower insets, in order to be classified as non-random. *In none of the Figures 20 through 32 do we observe any significant patterns.*

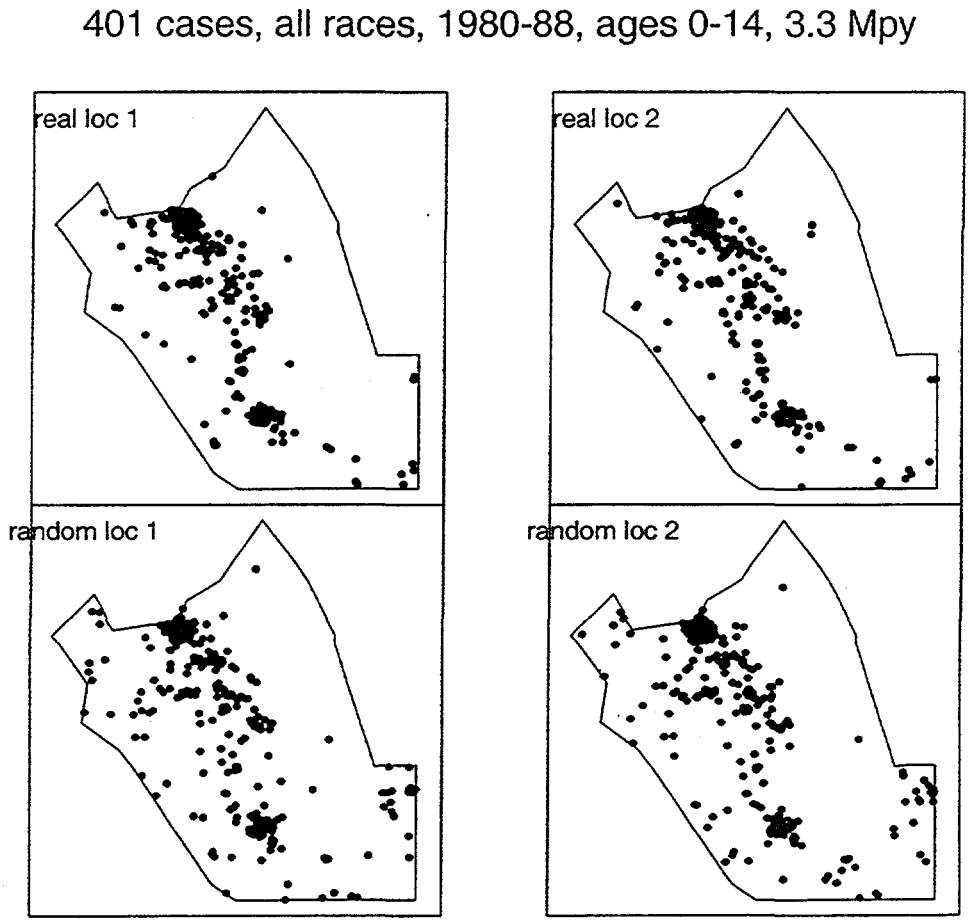


Figure 19. Real and random cases. 401 cases, original map, all races, 1980-88, ages 0-14, both sexes, 3.3 Mpy. In the two upper maps, each cases is plotted at two different random locations in the tract where it occurred. In the two lower maps, 401 artificial cases are similarly plotted, with the tract for each case chosen at random under the assumption that rates are everywhere equal.

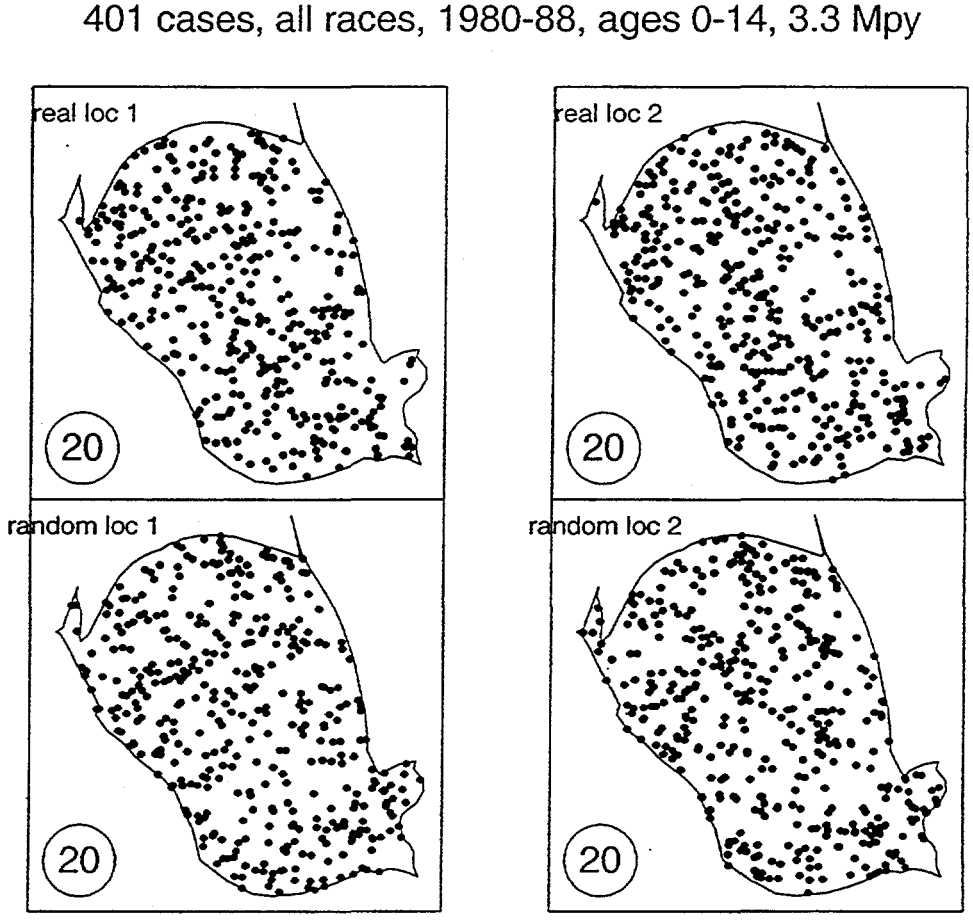


Figure 20. Same as Figure 19, on the density equalized map. In the two lower maps, the distribution of cases is random *by construction;* any apparent clusters are due to statistical variation. In the two upper maps, any apparent clusters are insignificant unless more extreme than the random fluctuations in the two lower maps. In all four maps, the circles indicate the size of an area within which 20 cases are expected.

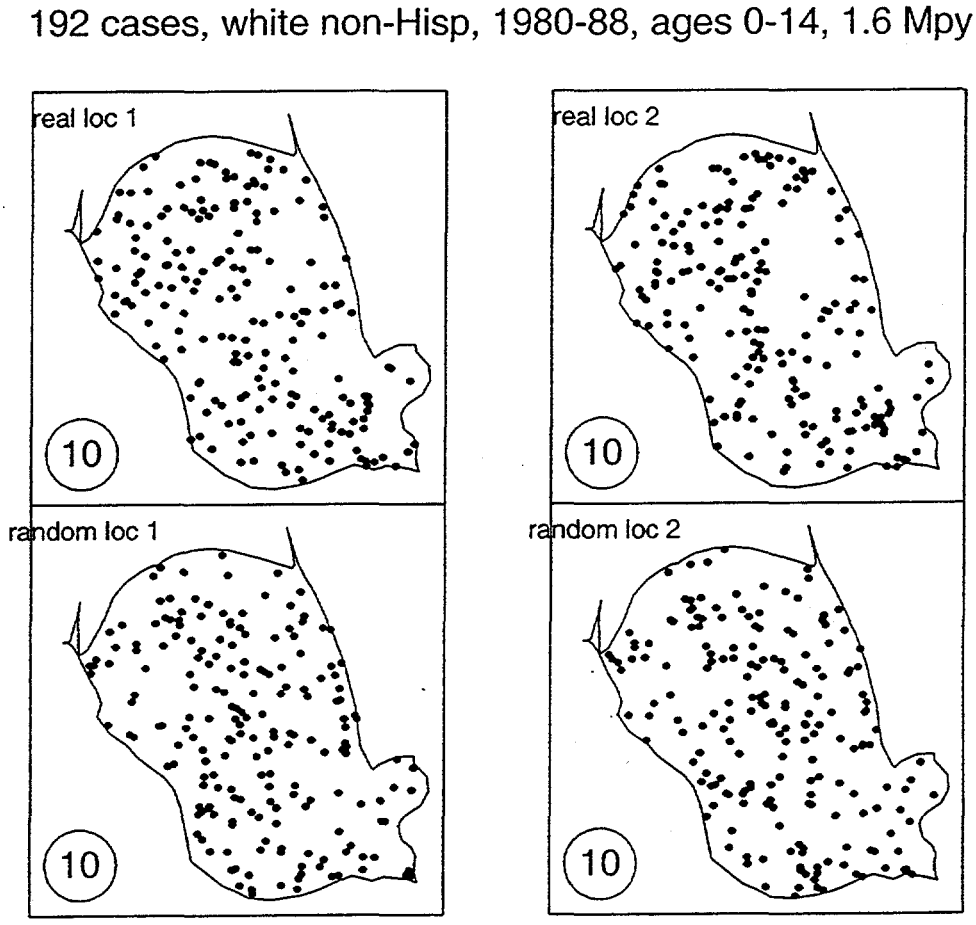


Figure 21. Same as Figure 20, for 192 white non-Hispanic cases, 1980-88, ages 0-14, both sexes, 1.6 Mpy. The density equalized map *and* the cases plotted pertain to white non-Hispanics only. The circles indicate the size of an area within which 10 cases are expected.

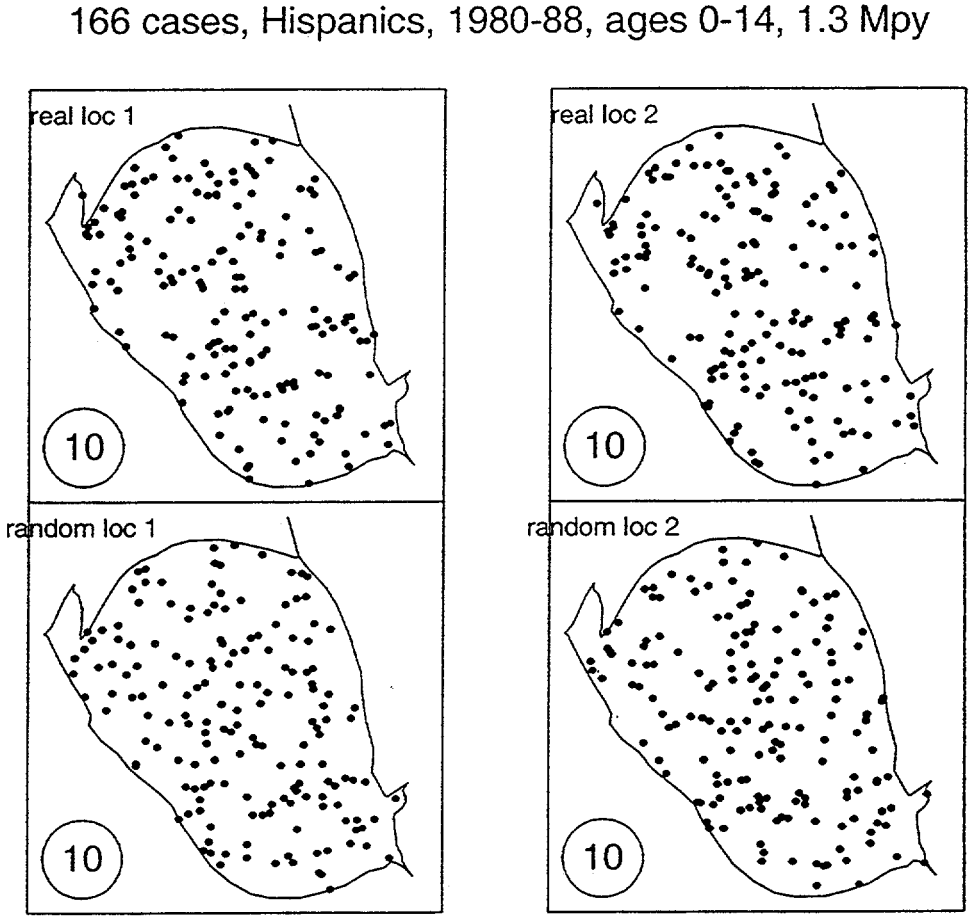


Figure 22. Same as Figure 20, for 166 Hispanic cases, 1980-88, ages 0-14, both sexes, 1.3 Mpy. The density equalized map *and* the cases plotted pertain to Hispanics only. The circles indicate the size of an area within which 10 cases are expected.

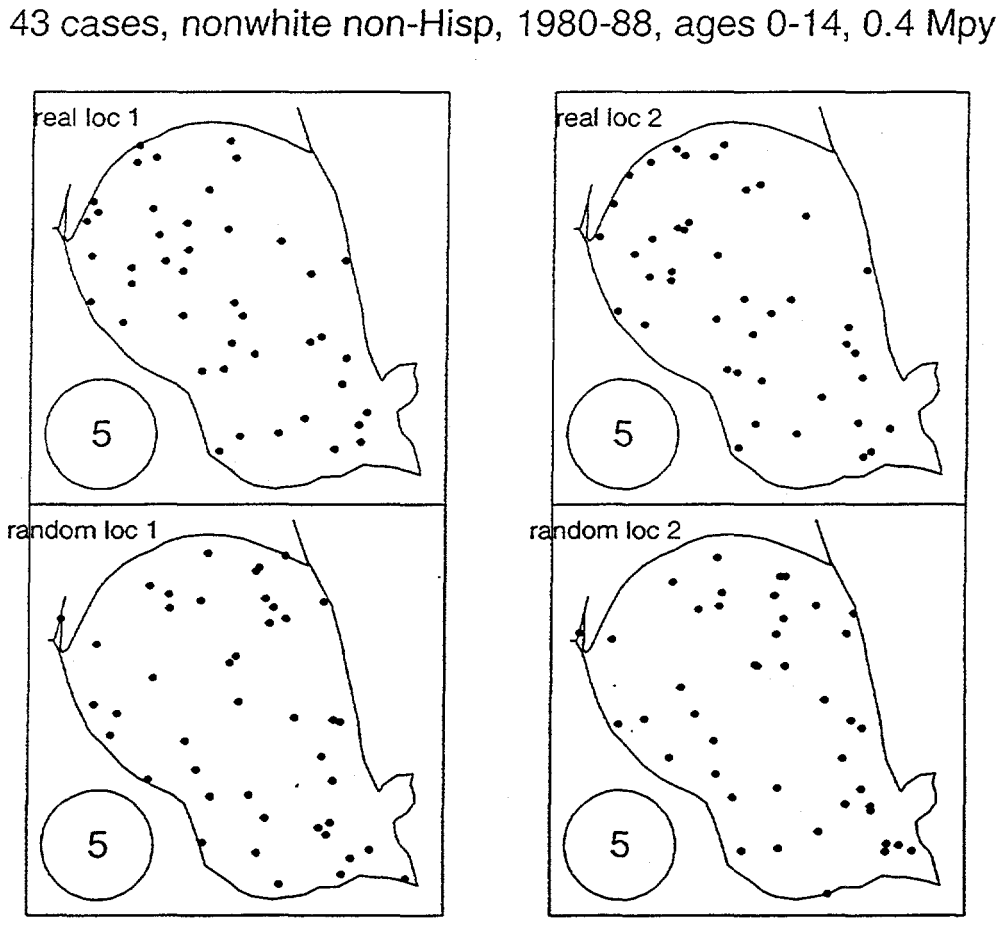


Figure 23. Same as Figure 20, for 43 nonwhite non-Hispanic cases, 1980-88, ages 0-14, both sexes, 0.4 Mpy. The density equalized map *and* the cases plotted pertain to nonwhite non-Hispanics only. The circles indicate the size of an area within which 5 cases are expected.

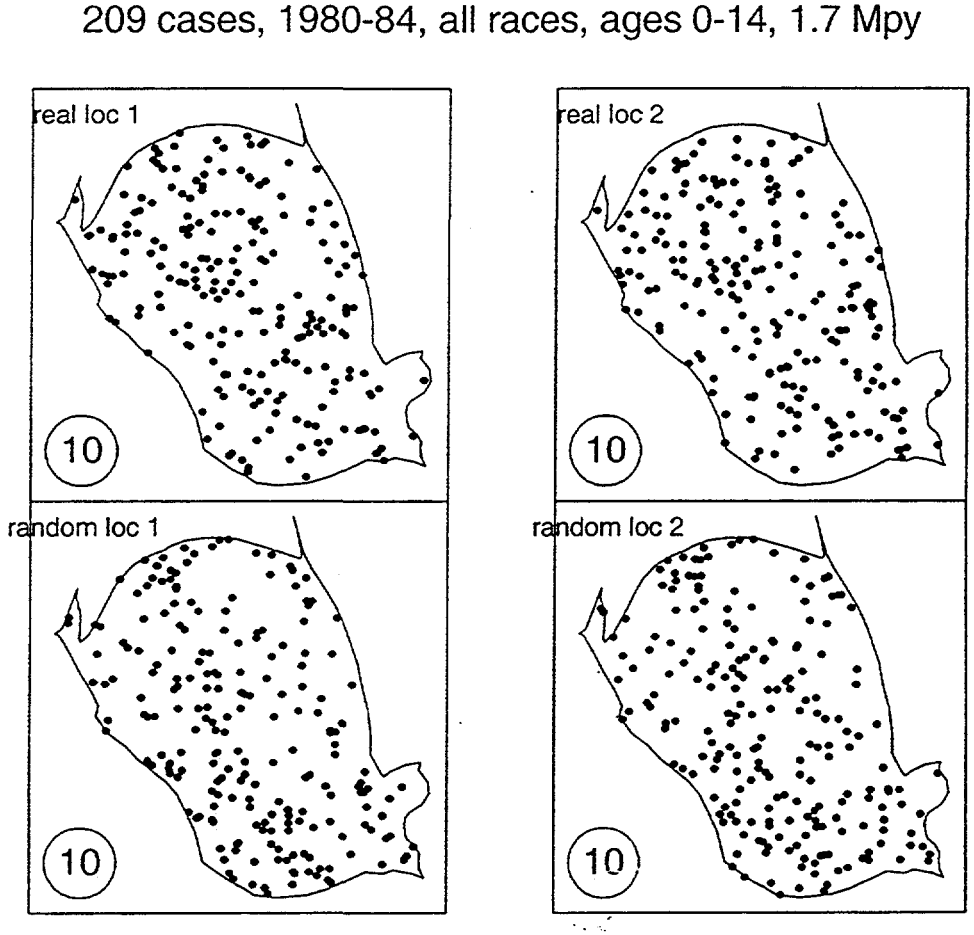


Figure 24. Same as Figure 20, for 209 cases, 1980-84, all races, ages 0-14, both sexes, 1.7 Mpy. The density equalized map *and* the cases plotted pertain to 1980-84 only. The circles indicate the size of an area within which 10 cases are expected.

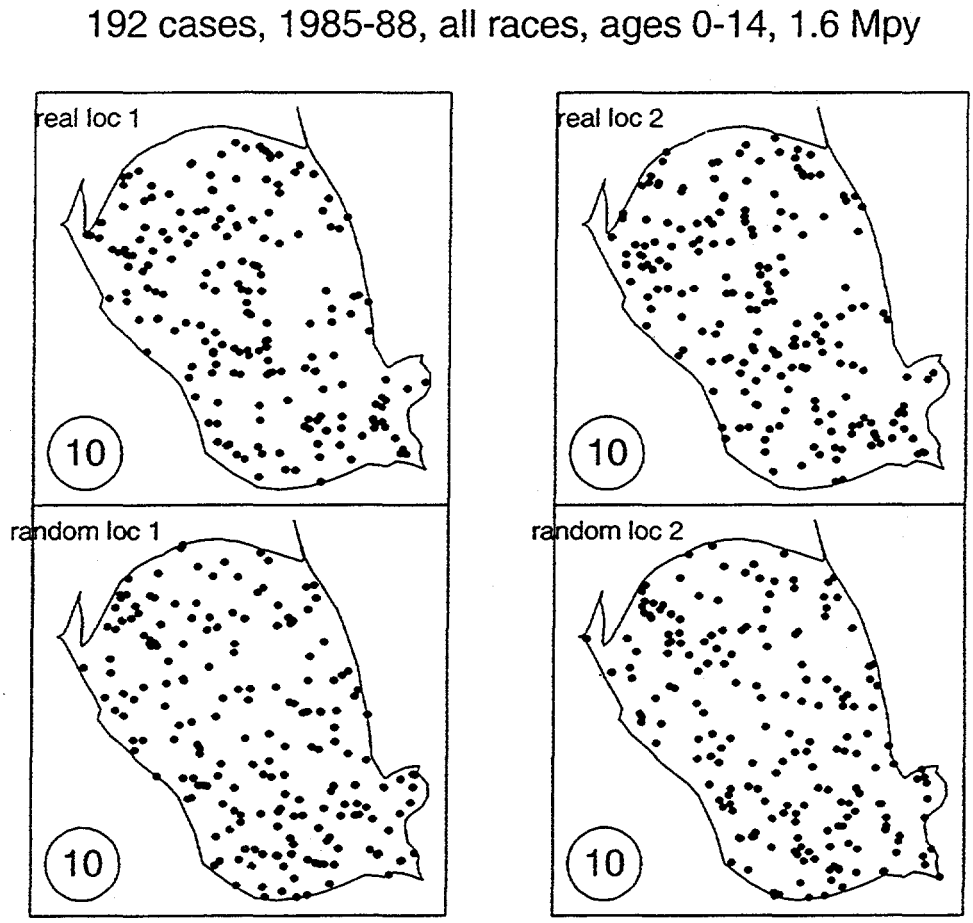


Figure 25. Same as Figure 20, for 192 cases, 1985-88, all races, ages 0-14, both sexes, 1.6 Mpy. The density equalized map *and* the cases plotted pertain to 1985-88 only. The circles indicate the size of an area within which 10 cases are expected.

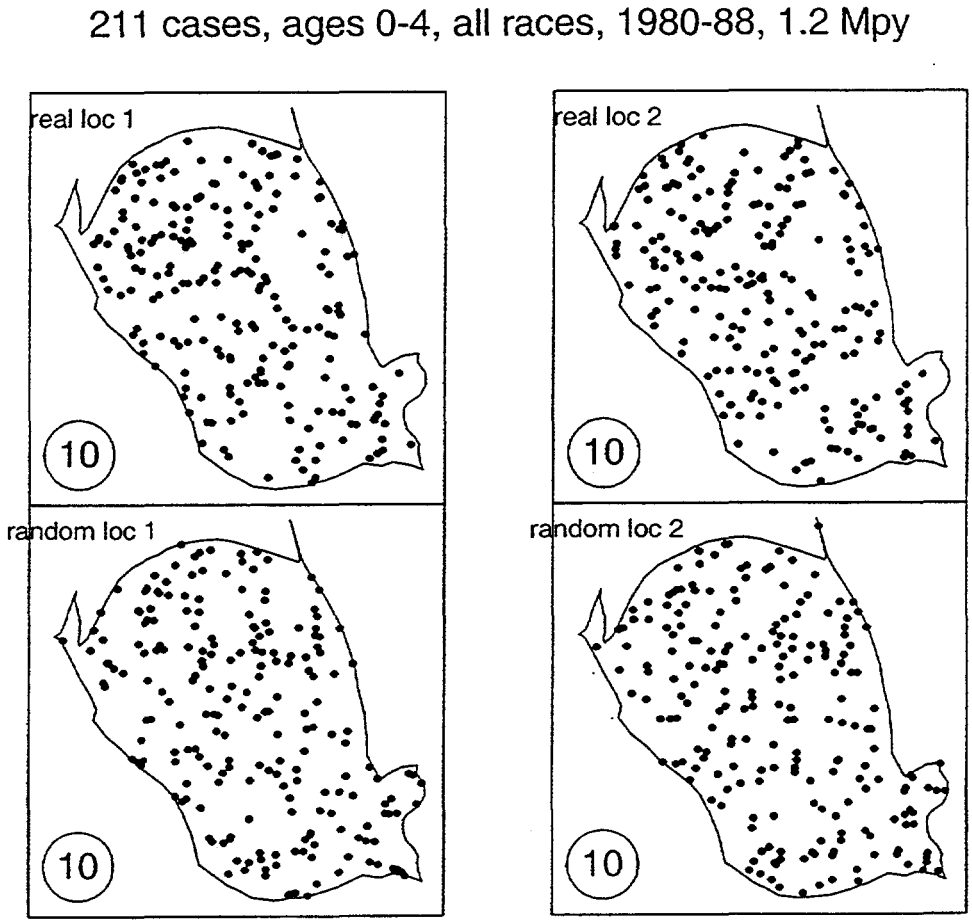


Figure 26. Same as Figure 20, for 211 cases, ages 0-4, all races, 1980-88, both sexes, 1.2 Mpy. The density equalized map *and* the cases plotted pertain to ages 0-4 only. The circles indicate the size of an area within which 10 cases are expected.

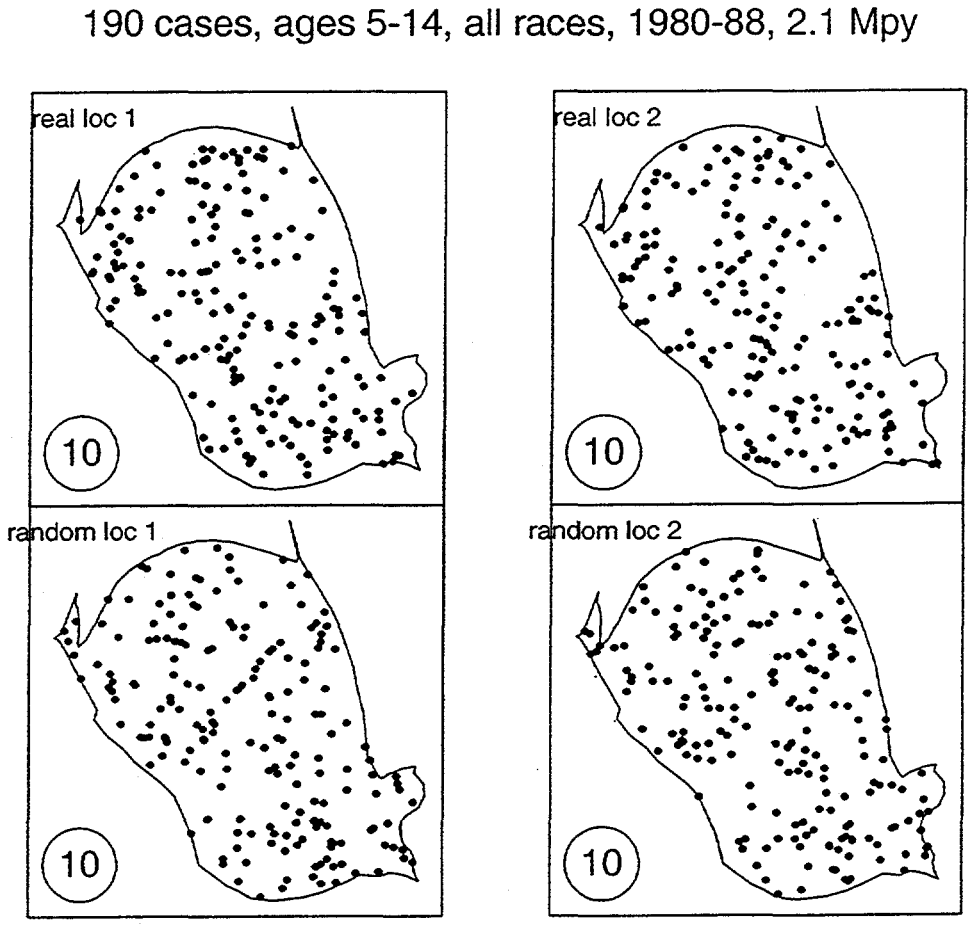


Figure 27. Same as Figure 20, for 190 cases, ages 5-14, all races, 1980-88, both sexes, 2.1 Mpy. The density equalized map *and* the cases plotted pertain to ages 5-14 only. The circles indicate the size of an area within which 10 cases are expected.

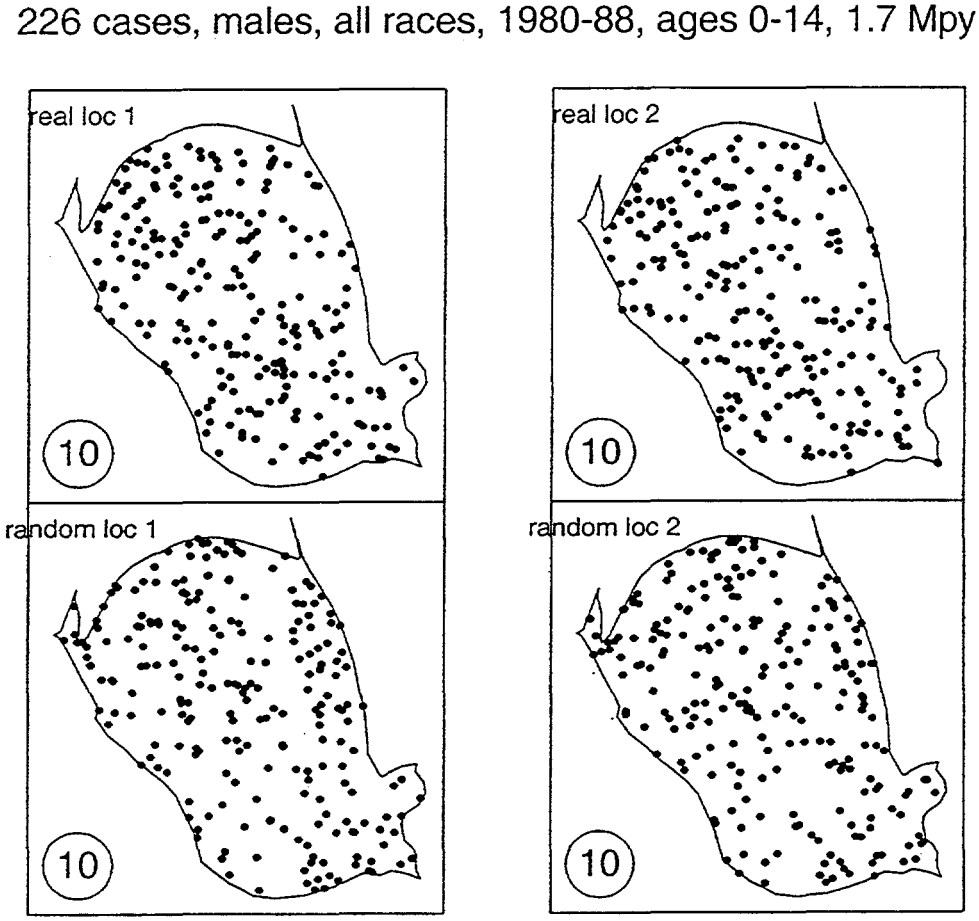


Figure 28. Same as Figure 20, for 226 cases, males, all races, 1980-88, ages 0-14, 1.7 Mpy. The density equalized map *and* the cases plotted pertain to males only. The circles indicate the size of an area within which 10 cases are expected.

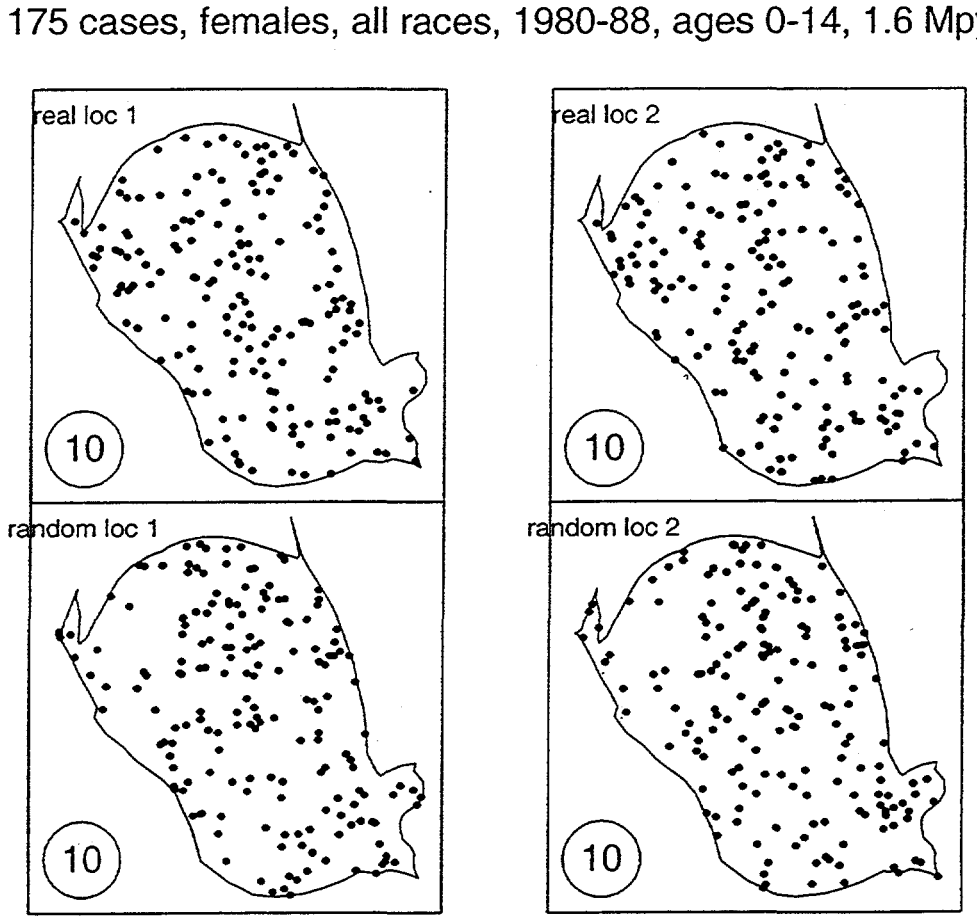


Figure 29. Same as Figure 20, for 175 cases, females, all races, 1980-88, ages 0-14, 1.6 Mpy. The density equalized map *and* the cases plotted pertain to females only. The circles indicate the size of an area within which 10 cases are expected.

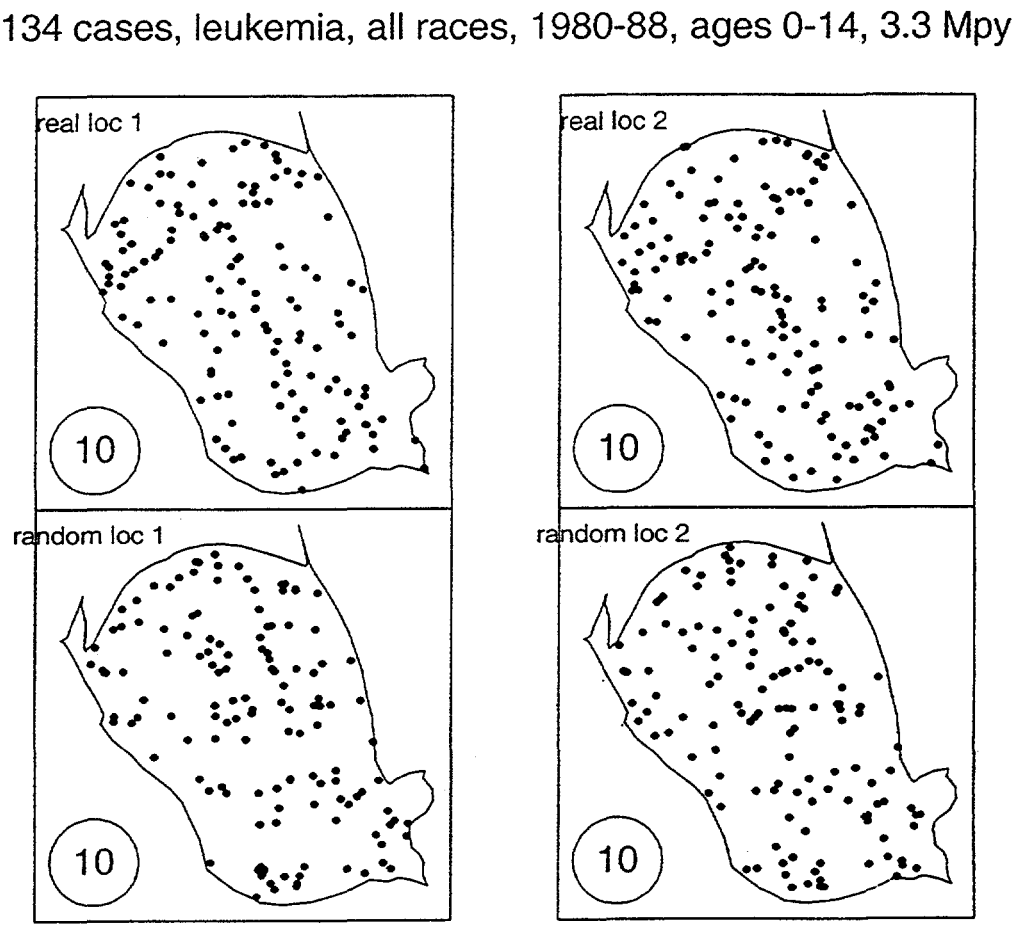


Figure 30. Same as Figure 20, for 134 leukemia cases, all races, 1980-88, ages 0-14, both sexes, 3.3 Mpy. The density equalized map pertains to all persons at risk; only leukemia cases are plotted. The circles indicate the size of an area within which 10 cases are expected.

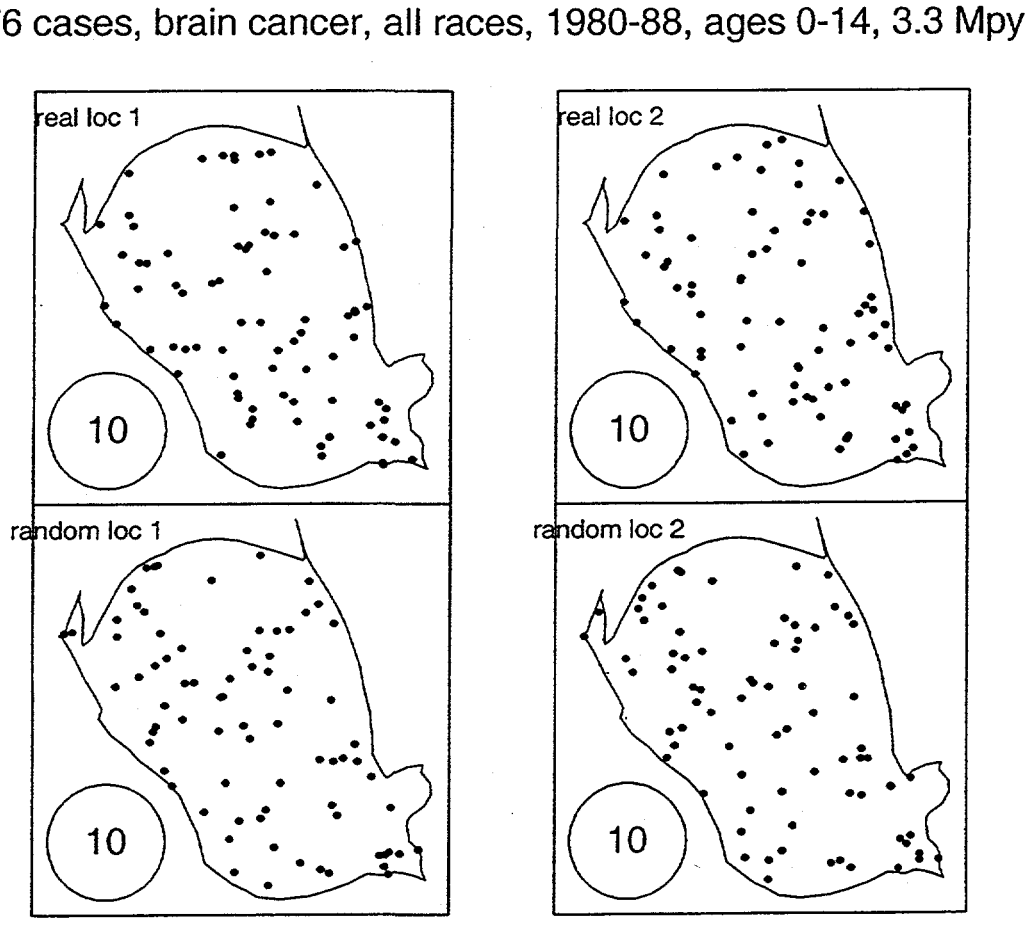


Figure 31. Same as Figure 20, for 76 brain cancer cases, all races, 1980-88, ages 0-14, both sexes, 3.3 Mpy. The density equalized map pertains to all persons at risk; only brain cancer cases are plotted. The circles indicate the size of an area within which 10 cases are expected.

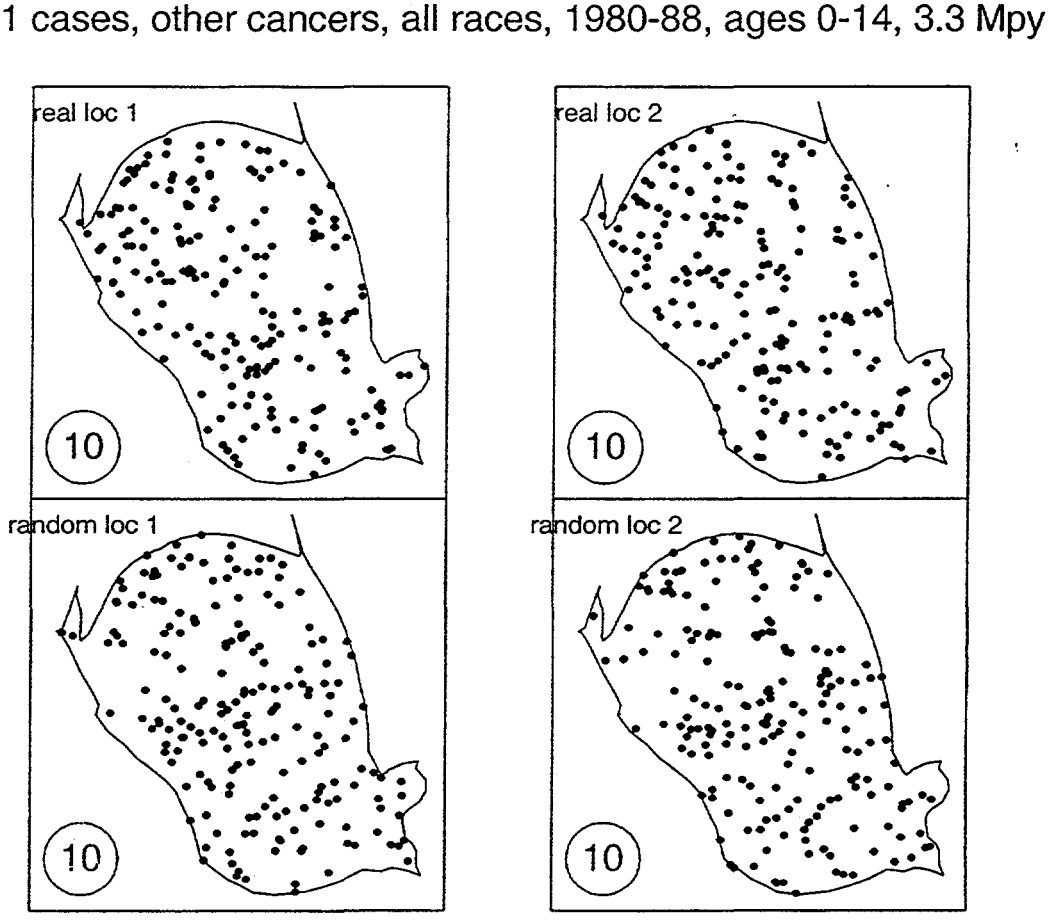


Figure 32. Same as Figure 20, for 191 cancer cases other than leukemia or brain cancer, all races, 1980-88, ages 0-14, both sexes, 3.3 Mpy. The density equalized map pertains to all persons at risk; only cancer cases other than leukemia or brain cancer are plotted. The circles indicate the size of an area within which 10 cases are expected.

Statistical analysis of RR (relative risk)

The case locations in Figures 20 through 32 were subjected to a statistical analysis, to quantitatively measure any differences that may exist between the real cases in the upper insets, and the random artificial cases in the lower insets. A quantity called RR (relative risk) was estimated by two methods at each point (x,y) in a regular sampling grid, over the entire area of each density equalized map.

The first method, which we denote NN, uses a kth nearest neighbor density estimator. From the number of cases and the area of the map one can calculate $A_{exp}(k) = kA_{tot} / N$, the area of a circle in which one *expects* to find k cases, under the null hypothesis of uniform risk. (Here A_{tot} is the area of the density equalized map and N is the number of cases.) At any point (x,y) in the map one can measure $A_{obs}(x,y,k)$, the area of a circle centered at (x,y) which *actually* contains k cases. (More precisely, the areas of the two circles which just pass through the kth nearest case and the (k+1)th nearest case are calculated, and $A_{obs}(x,y,k)$ is taken to be the average of those two areas.) Then at the point (x,y) the relative risk is $RR_{NN}(x,y,k) = A_{exp}(k) / A_{obs}(x,y,k)$.

The second method, which is denoted GK, uses a Gaussian kernel density estimator. Unlike the NN estimator, the GK estimator is a continuous function over the space (x,y). At every grid point (x,y) $RR_{GK}(x,y,k)$ is calculated as the sum over all cases j, of $C \exp -(d_j^2/d_0^2)$, where $C = 2/k$, d_j is the distance from (x,y) to case j, and $d_0^2 = k A_{tot} / (2 N pi)$,. The NN and GK estimators are approximately equal for values of k larger than about 20.

With both the NN and GK estimators, the parameter k specifies the desired spatial resolution. The number of points in the sampling grid is arbitrary, but the grid spacing should be small compared with the radius of the circle having area $A_{exp}(k)$. In order to avoid a bias near the boundary of the density equalized map, a uniform grid of artificial cases is laid down outside the map contour, with grid spacing such that RR(x,y,k) is exactly 1.0 outside the contour. Those external artificial cases are included when estimating RR(x,y,k) at internal points (x,y) near the boundary.

In Figures 33 through 36 the distribution of log RR is shown, for both the NN and GK estimators, and for k=10 and k=20. Figures 33 through 36 are for the full sample of 401 cases, shown in Figure 20. Plots corresponding to the 12 subsamples of Figures 21 through 32, not shown, are qualitatively similar.

For both the NN and GK estimators and for k=10 and k=20, RR has an approximately lognormal distribution. When this is true, the mean of RR is one, so the mean of log RR is zero. Furthermore, under the null hypothesis of uniform risk, the spatial distribution of cases is random, and the variance of the distribution of log RR can be predicted theoretically as k becomes large. The theoretical variance of log RR ("th var(log RR)" in Figures 33-36) is equal to 1/k and is independent of the arbitrary spacing of the grid points (x,y). The theoretical standard deviation error ("th sd(log RR)") is equal to sqrt(1/k). This theoretical s.d. error was used to plot the Gaussian curves that appear in Figures 33 through 36. If log RR has a normal distribution, then one expects about 31.7% ("th tail" in Figures 33-36) of the sampled grid points to have values of log RR outside the interval ±1 s.d.

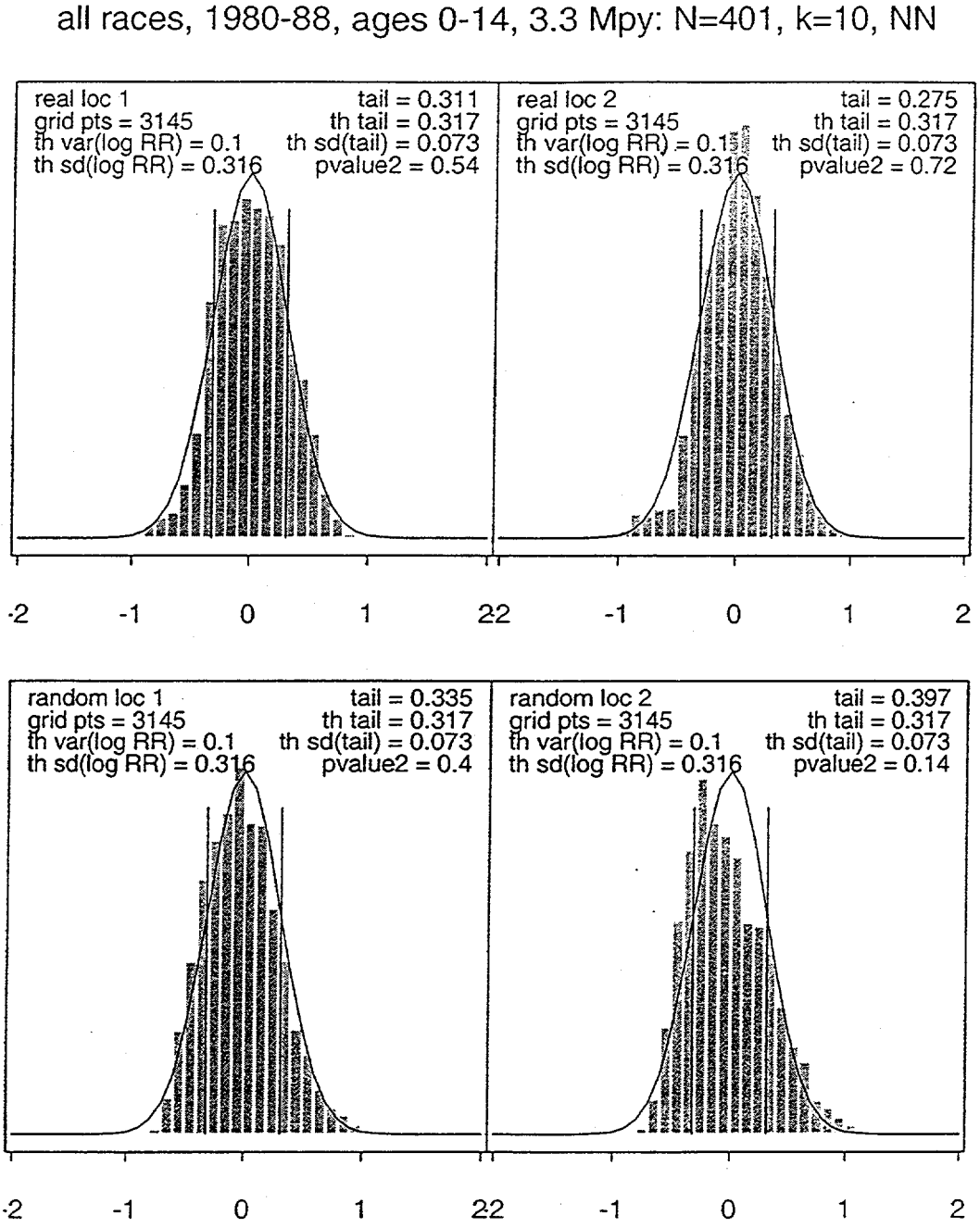


Figure 33. Distribution of log RR, for the four density equalized maps in Figure 20. The calculation is described in the text. A kth nearest neighbor (NN) formula, with k=10, was used. Clustering, if present, would cause the real cases (top) to have a broader distribution than the artificial cases (bottom). No significant differences were observed.

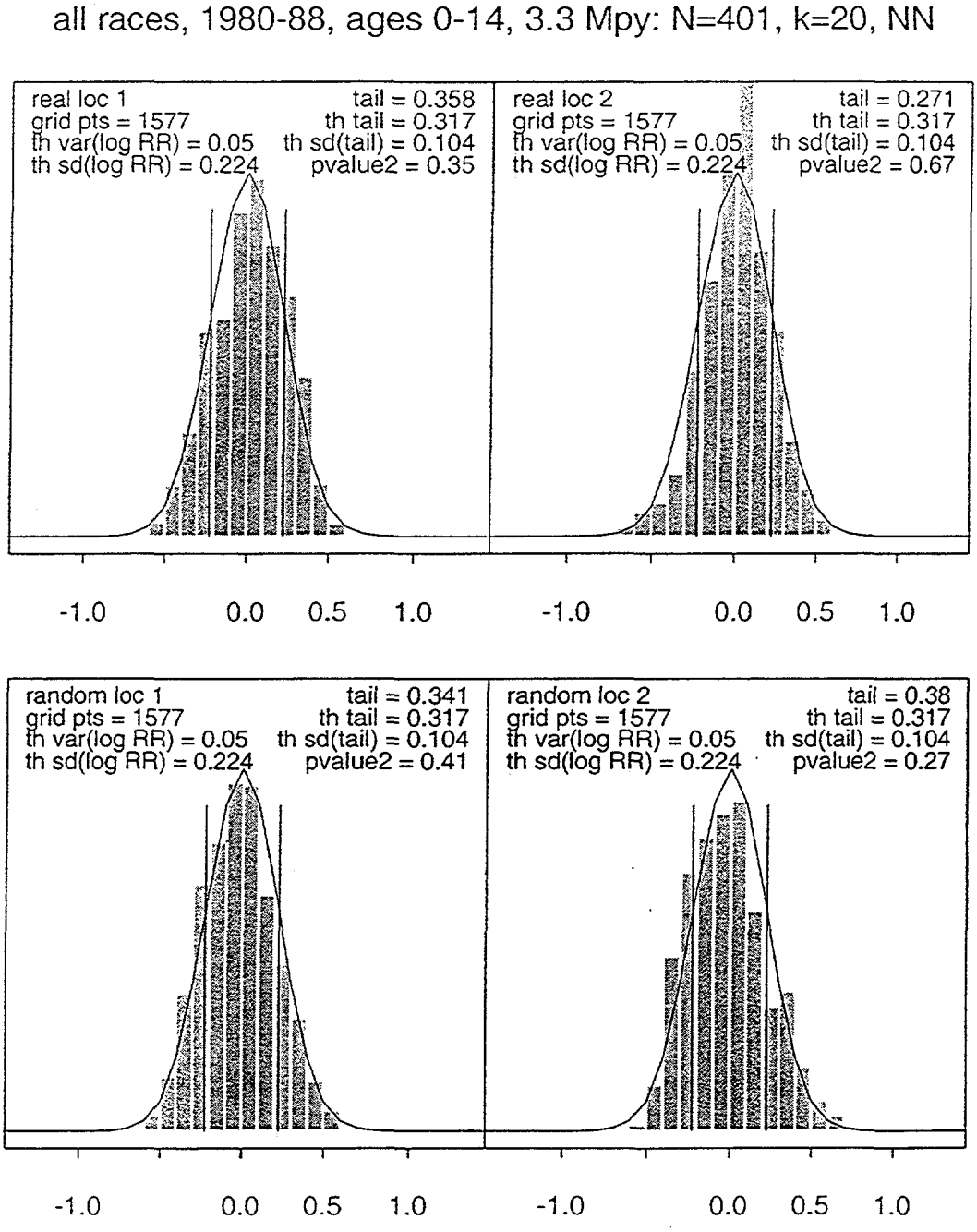


Figure 34. Same as Figure 33. A kth nearest neighbor (NN) formula, with k=20, was used.

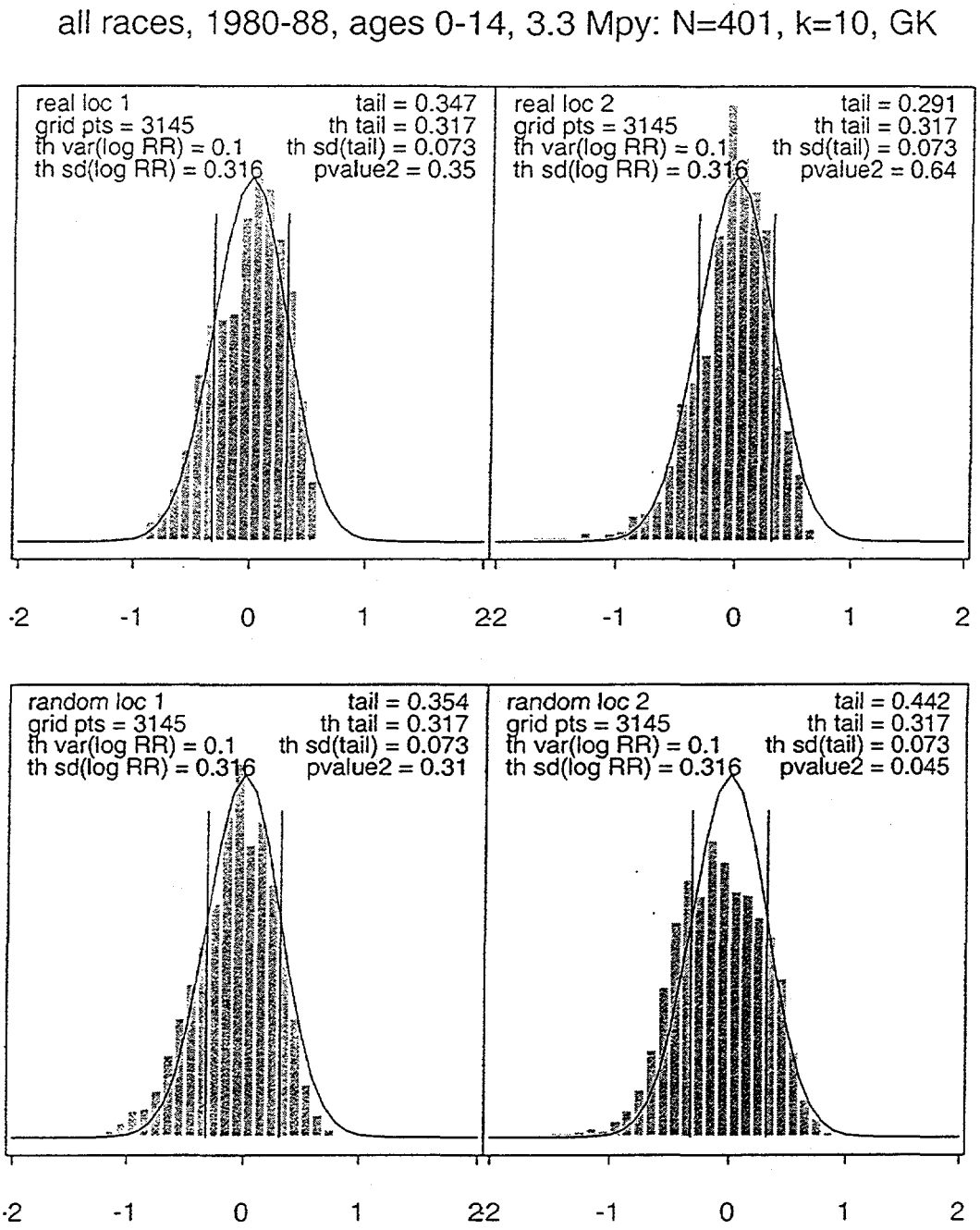


Figure 35. Same as Figure 33. A Gaussian kernel (GK) formula, with k=10, was used.

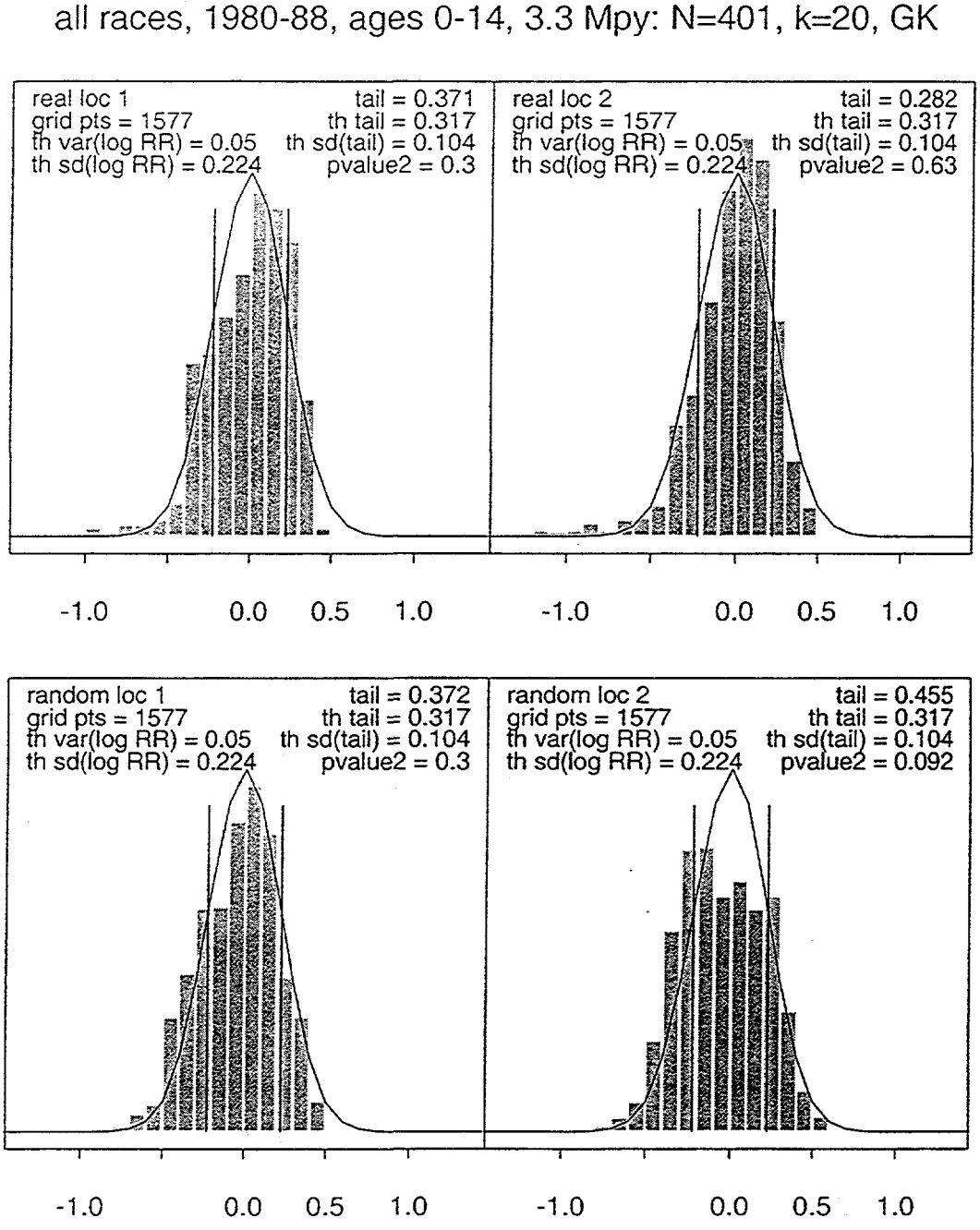


Figure 36. Same as Figure 33. A Gaussian kernel (GK) formula, with k=20, was used.

Now if the real cases were *not* randomly distributed, one would expect to see a broadening of the log RR distribution in the upper insets, with an excess of grid points having extreme values of log RR. This is because high values of log RR occur where cases are concentrated, and low values occur where cases are sparse. (One could also have a *narrowing* of the distribution, an *anticlustering,* if cases occurred in a pattern significantly *more* regular than random, for example on a uniform grid; but it is difficult to imagine that happening in the present situation.)

Analysis of T: real and random cases vs. theoretical

The quantity T ("tail" in Figures 33-36) is defined to be the *observed* fraction of grid points outside the ±1 s.d. interval. The theoretical variance of T is expected to be approximately T(1-T)k/N; the corresponding s.d. error ("th sd(tail)" in Figures 33-36) is used to estimate a p-value ("pvalue2"), which is the probability that a value of T at least as large as that observed, could have occurred by chance.

T and the corresponding p-value were calculated, for all 13 of the log RR plots (12 of them not shown), which were derived from Figures 20 through 32. The results are summarized in Table II, for the GK method and k=10. The two numeric values listed in each cell correspond to the two different random locations of each case within its tract. In assessing the values in Table II, remember that the 13 samples listed are not statistically independent! However, the three race/ethnicity groups *are* independent of each other, the two time periods *are* independent, etc.

Table II. Statistical analysis of T, for GK method with k=10 comparison of real and random cases, vs. theoretical distribution							
			T = fraction outside ±1 s.d.			p-value	
fig	sample	cases	expected	real cases	random cases	real cases	random cases
20	total	401	.32	.35, .29	.35, .44	.35, .64	.31, .05
21	white	192	.32	.34, .44	.31, .34	.41, .13	.51, .42
22	Hispanic	166	.32	.26, .33	.16, .22	.68, .47	.91, .79
23	other	43	.32	.23, .29	.24, .22	.66, .55	.64, .66
24	1980-84	209	.32	.27, .31	.42, .44	.68, .52	.15, .11
25	1985-88	192	.32	.35, .34	.33, .34	.39, .43	.46, .43
26	age 0-4	211	.32	.34, .35	.26, .24	.41, .39	.71, .78
27	age 5-14	190	.32	.34, .33	.27, .28	.43, .45	.67, .64
28	male	226	.32	.32, .30	.37, .33	.50, .59	.29, .44
29	female	175	.32	.32, .28	.39, .36	.48, .64	.27, .36
30	leukemia	134	.32	.33, .37	.36, .26	.46, .33	.36, .67
31	brain cancer	76	.32	.22, .22	.21, .25	.72, .72	.73, .66
32	other cancer	191	.32	.35, .27	.30, .35	.39, .67	.56, .39

The absence of p-values less than 0.05 in Table II shows that the observed distributions of RR, for both real and random cases, are consistent with a simple theoretical model: log RR is approximately normally distributed, with variance equal to 1/k.

The agreement with a theoretical model is gratifying, providing some confidence that a gross blunder has not been committed in the analysis. More important, however, is the comparison which follows, between the real cases and artificial random cases. For the analysis of the Four County data set, it is not important that log RR is normally distributed with variance 1/k, nor that the fraction of grid points outside ±1 s.d. is approximately 31.7%. What is important to discover

is whether the real cases and the artificial random cases exhibit the same behavior, when subjected to the same analysis.

Analysis of T_{AV}: real vs. random cases

In Table III are presented results from 13 analyses, analogous to those presented in Table II. As before, remember that the 13 analyses are not independent. In Table III each analysis is repeated with not *one* set of random cases (as in Figures 20 through 32 and Table II), but with *ten* sets of artificial random cases. Values of T_{AV} are presented, which are the average of the two values of T obtained by plotting each case (random or real) at two different locations within its tract.

Table III. Statistical analysis of T_{AV}, for GK method with k=10
comparison of real vs. random cases

		T_{AV} = fraction outside ±1 s.d.					std devs	p-value
sample	N	obs	exp	alloc err	plot err	tot err		
total	401	.340	.307	.032	.017	.037	+0.9	.18
white	192	.370	.291	.056	.028	.062	+1.3	.10
Hispanic	166	.321	.314	.054	.036	.065	+0.1	.46
other	43	.230	.207	.094	.056	.109	+0.2	.42
1980-84	209	.310	.311	.023	.035	.042	-0.1	.55
1985-88	192	.360	.261	.029	.038	.048	+2.0	.02
age 0-4	211	.350	.300	.022	.034	.041	+1.2	.11
age 5-14	190	.328	.296	.049	.033	.060	+0.5	.29
male	226	.352	.297	.048	.024	.053	+1.0	.15
female	175	.292	.287	.015	.031	.035	+0.2	.44
leukemia	134	.332	.286	.052	.046	.069	+0.7	.25
brain cancer	76	.188	.304	.067	.034	.075	-1.5	.94
other cancer	191	.291	.308	.039	.028	.047	-0.3	.63

The column "obs" contains the observed value of T_{AV} from the real cases; it is the same as the average of the two values of "real cases" in Table II, except that the values differ because different random plot locations were used in Tables II and III. The column "exp" contains the expected value of T_{AV}, which is estimated from the ten samples of artificial random cases. In each of the ten random samples, N cases were randomly allocated to the 259 tracts under the assumption of equal risk; then each case was plotted at two different random locations in its tract, exactly as for the real cases. *It is important to follow this randomization procedure exactly; see Appendix D for a fuller discussion.* The quantities "alloc err" and "plot err" are discussed in Appendix D.

The column "tot err" is the s.d. error of T_{AV}, estimated as the square root of the sample variance of the ten individual values of T_{AV}. The column "std devs", the number of standard deviations above or below the expected value, is equal to (obs-exp)/tot err; "p-value" is the probability that a value as large as "obs" can occur through chance alone.

For the combined sample of 401 cases, the ten values of T_{AV} from the random artificial cases have a sample mean of 0.307 and a s.d. error of 0.037. The observed value from the real cases is 0.340, which is 0.9 s.d. above the expected value. A value at least this large can occur through chance alone with a probability about 0.18.

One of the 13 samples analyzed has a p-value equal to 0.02. According to Bonferroni's multiple testing criterion, such a measurement would have to yield a p-value less than 0.05/13=0.004 in order to be considered statistically significant at the 95% confidence level.

The conclusion of the statistical analysis is that among the 401 cases of the Four County Childhood Cancer data set, there is no evidence for geographic clustering, beyond that expected from random variation alone.

Contour plots of relative risk

In Figures 37 and 38 are presented contour plots of RR calculated by the GK method. Plots calculated by the NN method, not shown, are similar, except that the contours are irregular due to spatial discontinuities in the NN function.

In Figure 37, which corresponds to the log RR distribution in Figure 35, a value k=10 was used. The contours RR=(0.73,1.37) i.e. log RR=±sqrt(1/k)=±0.31 correspond to ±1 s.d., the points in the Gaussian distribution marked with vertical lines in Figure 35. If non-uniformities in the RR distribution are due entirely to random variation, one expects about 31.7% of the measured RR values to lie outside the ±1 s.d. interval in Figure 35 or, in Figure 37, within the areas enclosed by the dashed and solid contours. All four insets in Figure 37 are consistent with purely random variation.

In Figure 38, which corresponds to the log RR distribution in Figure 36, a value k=20 was used. The contours RR=(0.80,1.25) i.e. log RR=±sqrt(1/k)=±0.22 correspond to ±1 s.d., the points in the Gaussian distribution marked with vertical lines in Figure 36. (Note that Figures 35 and 36 have a different horizontal scale.) If non-uniformities in the RR distribution are due entirely to random variation, one expects about 31.7% of the measured RR values to lie outside the ±1 s.d. interval in

Figure 36 or, in Figure 38, within the areas enclosed by the dashed and solid contours. All four insets in Figure 38 are consistent with purely random variation.

In Figures 39 and 40, the RR contours of Figures 37 and 38, respectively, are transformed back to the original geopolitical map. If the variation observed in Figures 37 and 38 had not been purely random, the contours drawn in Figures 39 and 40 would have indicated areas of high and low relative risk in the four county area. Due to the nonuniform population density, the spatial resolution of the contours is excellent in the urban areas around Fresno and Bakersfield, and extremely poor elsewhere.

In the absence of significant clustering among any of the 13 samples investigated, contour plots are presented only for the full sample of 401 cases, in Figures 37 through 40. Plots for the 12 subsamples, not shown, are qualitatively similar.

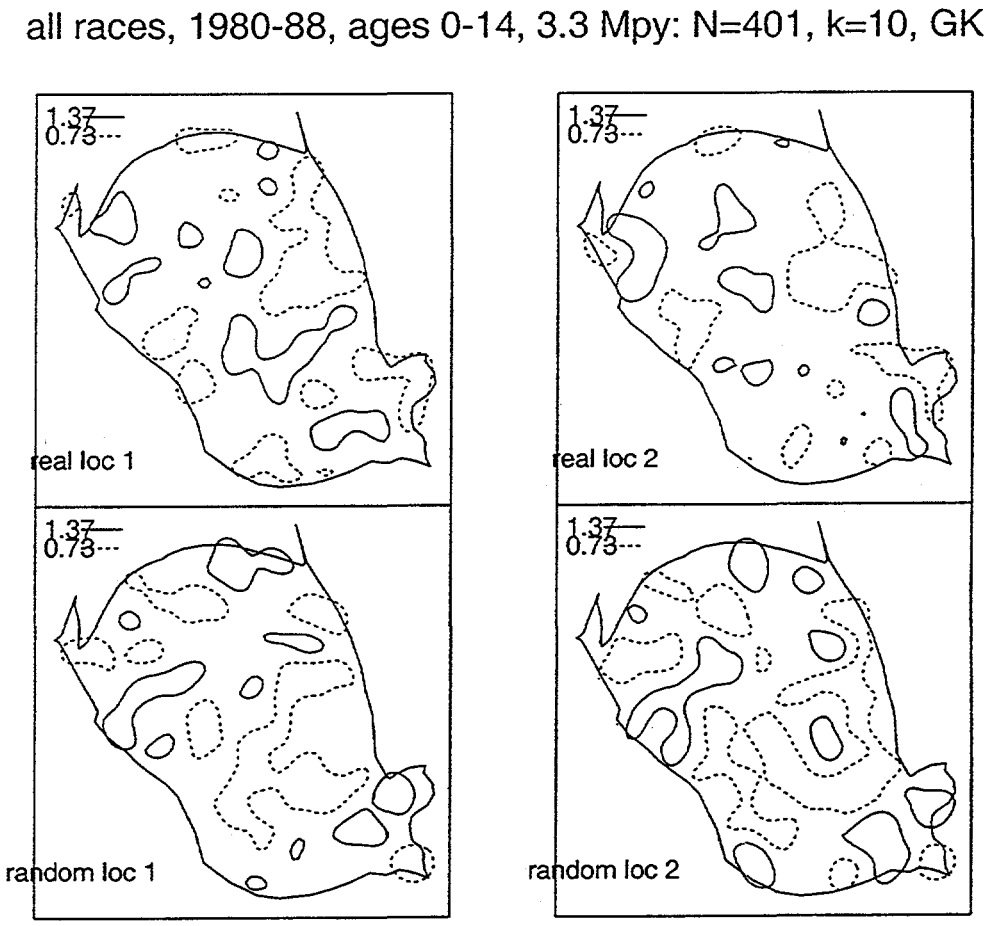


Figure 37. Contours of relative risk (RR), corresponding to Figure 20. A Gaussian kernel (GK) formula with k=10 was used. The contours RR=(0.73,1.37) i.e. log RR=(±0.31) correspond to ±1 s.d. in Figure 35.

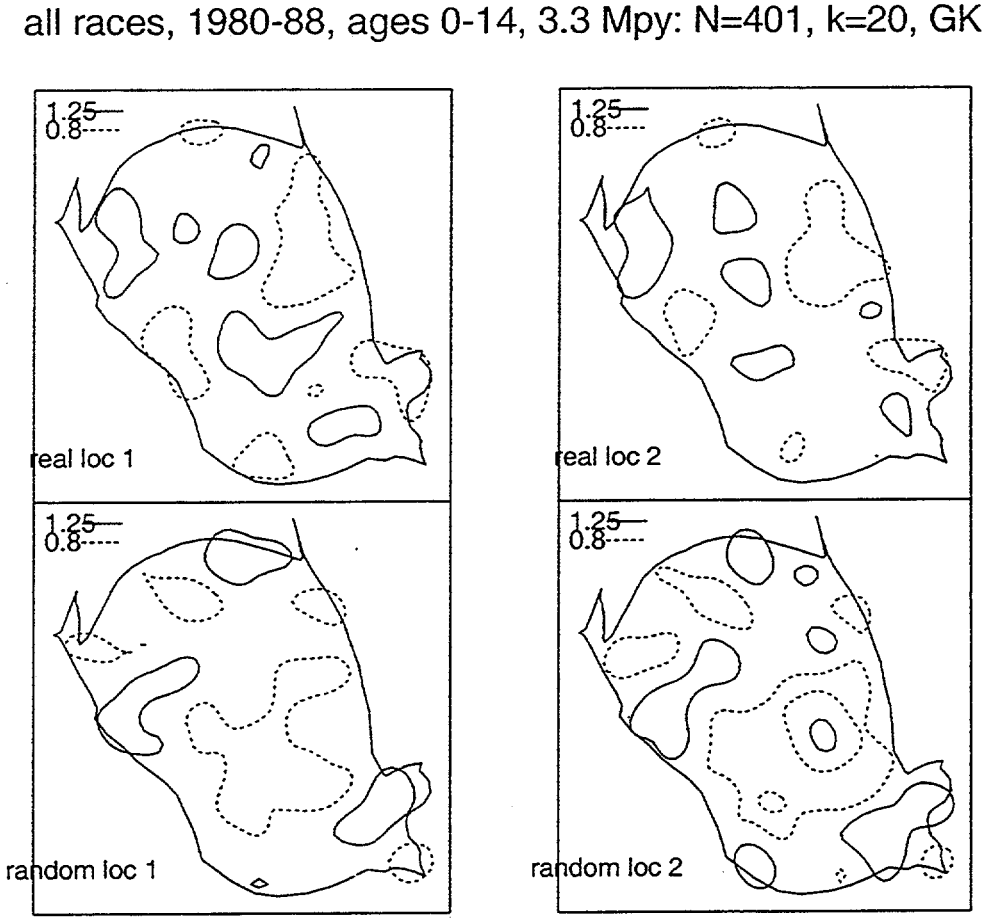


Figure 38. Contours of relative risk (RR), corresponding to Figure 20. A Gaussian kernel (GK) formula with k=20 was used. The contours RR=(0.80,1.25) i.e. log RR=(±0.22) correspond to ±1 s.d. in Figure 36.

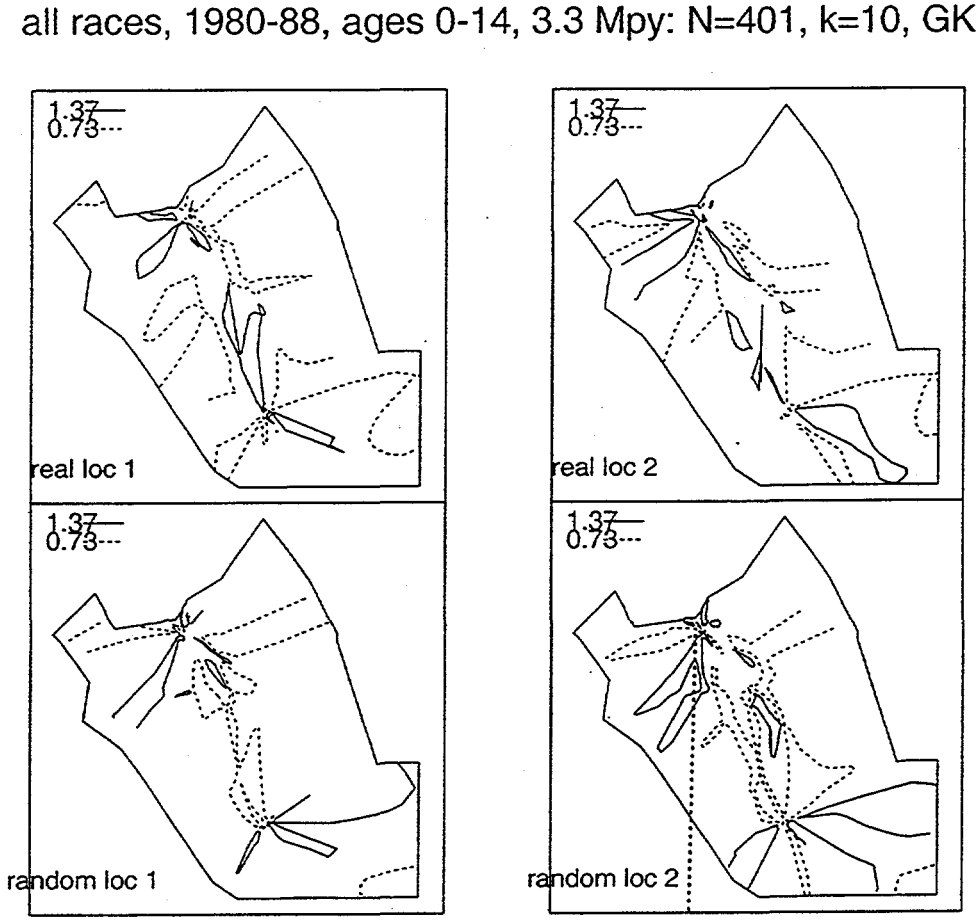


Figure 39. The RR contours of Figure 37, transformed back to the original geopolitical map. . A Gaussian kernel (GK) formula with k=10 was used. Due to the non-uniform population density of the four-county area, spatial resolution of the contours is very poor outside the urban areas of Fresno and Bakersfield.

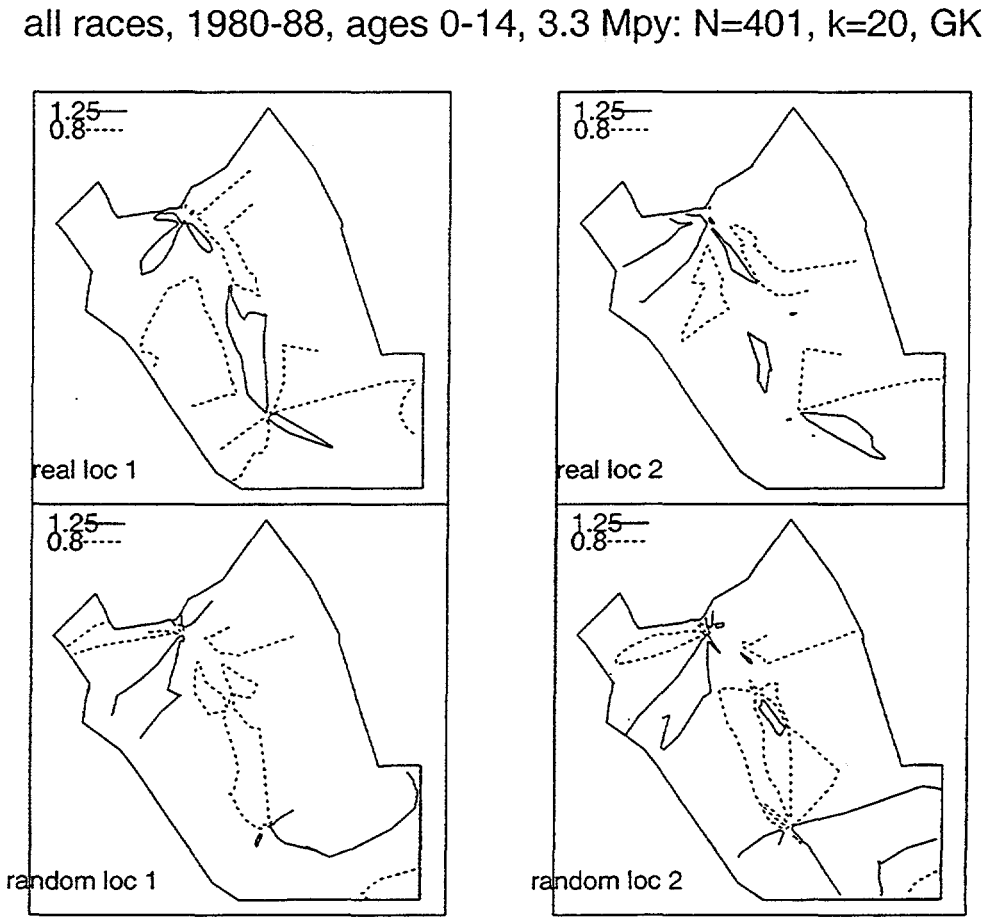


Figure 40. The RR contours of Figure 38, transformed back to the original geopolitical map. . A Gaussian kernel (GK) formula with k=20 was used.

Poisson-based test of real and random cases

An additional measurement was performed on both the real and random cases, which corresponds closely to the DHS test summarized in Figure 2. Like the DHS analysis, the following Poisson-based test contains no geographic information; it uses only the numbers of cases expected and observed in each tract, without regard to the tract locations.

A typical epidemiologic approach involves the comparison of the tract-specific rates to the overall rate. Tract-specific rates are $p_i = d_i / n_i$ (i.e., the 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. The number of cases d_i observed in tract i is compared with the number of cases e_i expected under the null hypothesis that rates are uniform. Specifically,

$$e_i = \text{overall rate} \times n_i$$

In the present calculation, n_i is the number of person-years in the period 1980-88, among children age 0-14 in the four-county area. In this cohort, 401 cases were observed, so

$$e_i = (401 / 3.3 \text{ Mpy}) \times n_i$$

Values of n_i, d_i and 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 a good approximation to the exact Poisson distribution is [BRES87]

$$z_i = \text{sqrt}(9 D_i) \times [1 - 1 / (9 D_i) - (e_i / D_i)^{1/3}]$$

where $D_i = d_i$ if d_i exceeds 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 Figures 41 and 42, for the real cases and the artificial random cases, respectively. (Two very small tracts with $e_i < 0.02$ 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, four tracts exceed this value among the real cases, and five among the artificial random cases.

It is puzzling that only 36 tracts are observed outside the interval (-1,+1) among the random cases, where one would have expected 32% of 257 tracts, or about 85. It appears that the statistic does not perform as well as advertised, at least not when d_i and e_i are small (their average value in this data set is about 1.5). But that is of no importance for the conclusions of this analysis: *since the real cases and random cases yield practically identical distributions of* z_i, *the Poisson-based test provides no evidence for non-uniformity of rates among the real cases.*

Also mysterious is the slight excess of very small values of z_i ($z_i < -1.5$) among the real cases relative to the random cases. This anomaly has nothing to do with the DEMP technique, which does not enter into the Poisson-based test. The most likely explanation is a slight misclassification bias, either in the case data or the Census data.

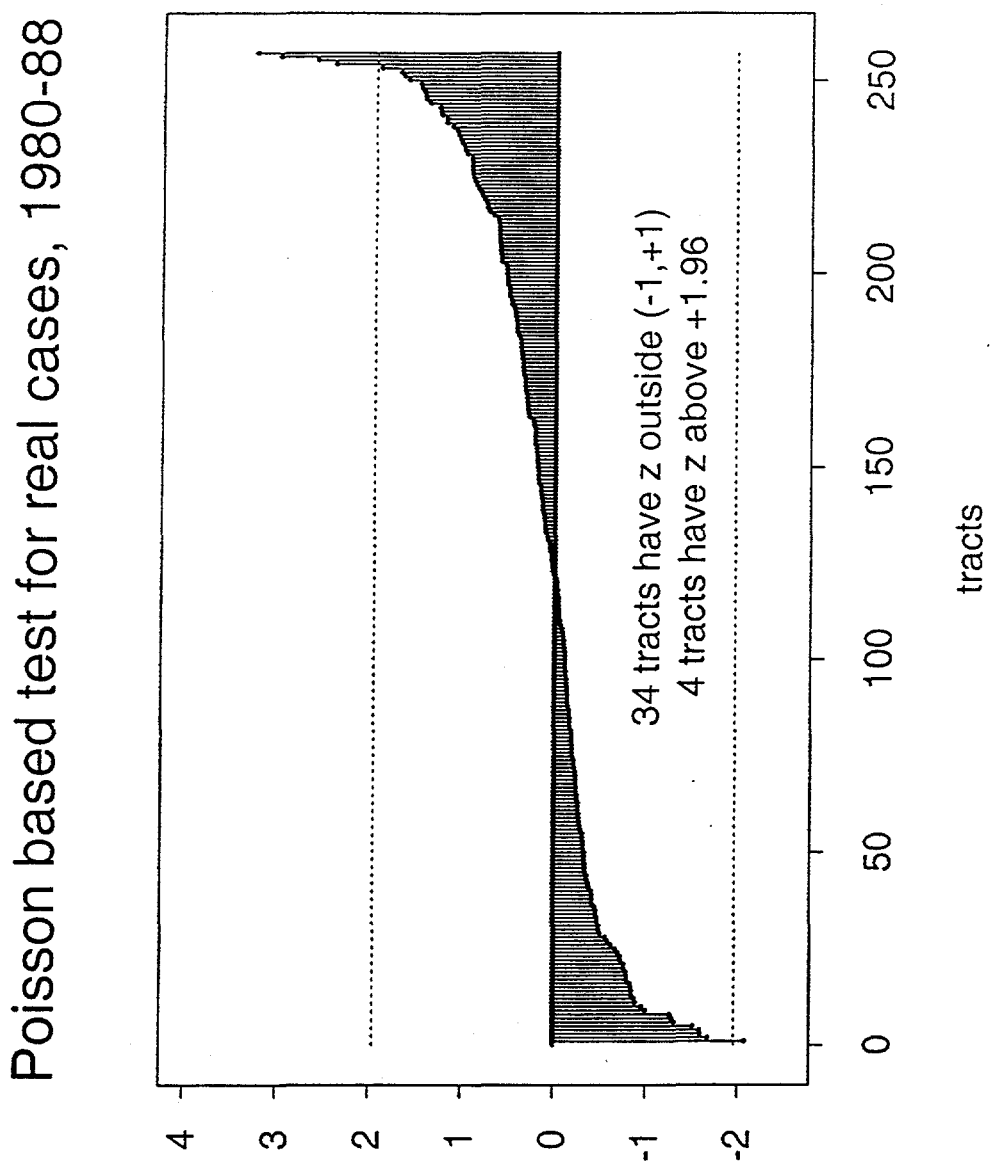


Figure 41. Distribution of the Poisson-based test statistic, calculated from the number of cases observed in each of the 259 tracts. Thirty-four tracts have z outside the interval (-1,+1); four tracts have z greater than 1.96. Figure 41 (real cases) should be compared with Figure 42 (artificial cases).

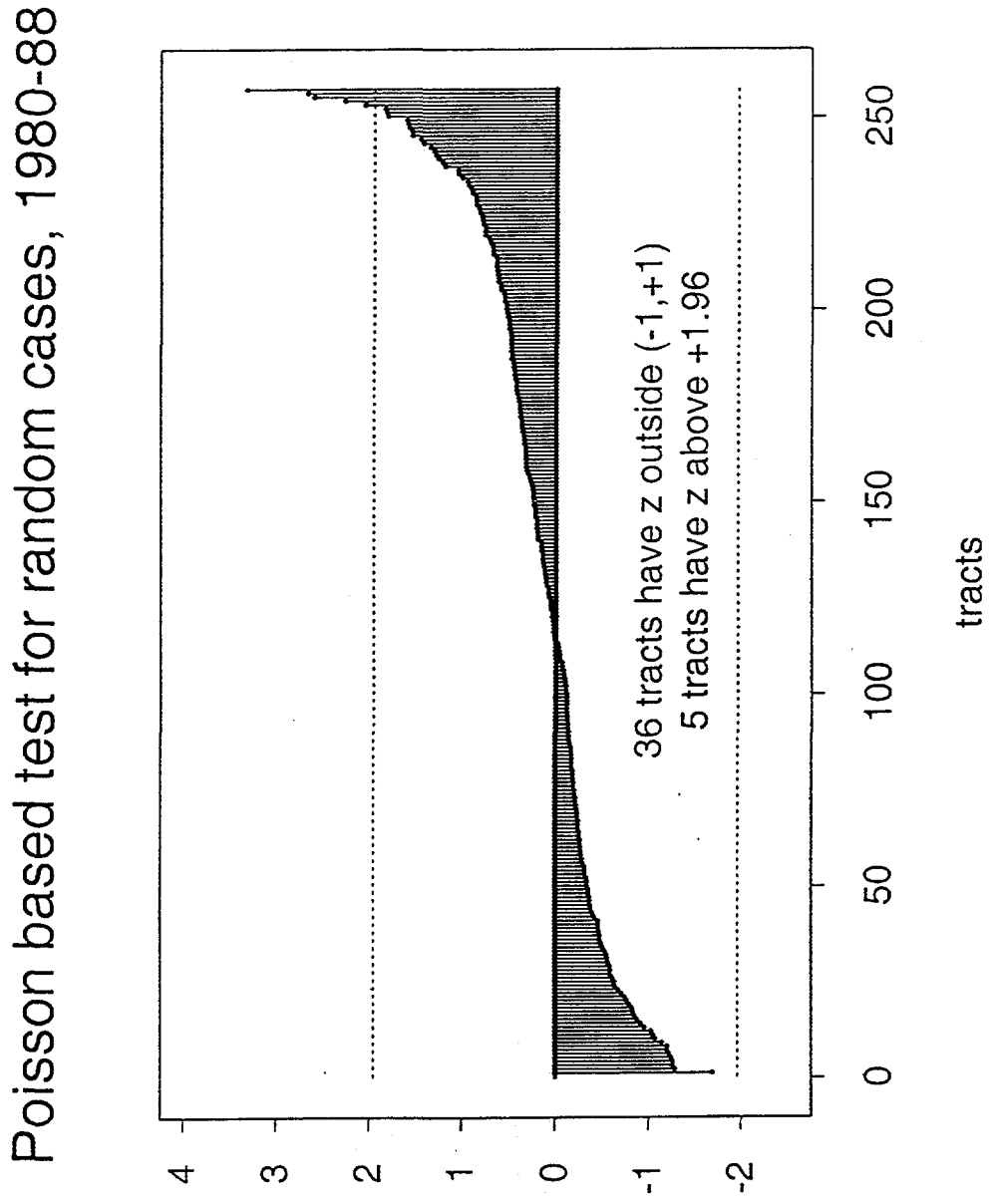


Figure 42. Distribution of the Poisson-based test statistic, calculated from the number of artificial random cases in each of the 259 tracts. Thirty-six tracts have z outside the interval (-1,+1); five tracts have z greater than 1.96. Figure 42 (artificial cases) should be compared with Figure 41 (real cases).

CONCLUSIONS

Proceeding from the specific to the general, conclusions are presented regarding: (1) the re-analysis of the Four County data set; and (2) the usefulness of the DEMP technique in general, and the future of "electronic epidemiology" on a grand scale.

Four County Childhood Cancer data set

The Four County Childhood Cancer data set previously analyzed by the California DHS [SATA90, REYN91, REYN96] has been re-analyzed, by the method of Density Equalizing Map Projections (DEMP). The data include all cancer cases incident among children 0-14 years of age, from 1980 through 1988, in the counties of Fresno, Kern, Kings, and Tulare. In agreement with [REYN91 and REYN96], our findings are consistent with the null hypothesis, that rates are geographically uniform over the four-county area.

In five separate subanalyses, the data were successively stratified by race/ethnicity (white, Hispanic, and other); time period (1980-84 and 1985-88), age (0-4 and 5-14), sex (male and female), and cancer site (leukemia, brain cancer, and other). Given the reduction in statistical power, it is not surprising that the same negative result pertains in all the subsamples. (The present finding concerns only geographic variation in the various subsamples; rates were *not* compared as a function of the stratifying variables, as was done in [SATA90].)

The geographic variation of rates was previously analyzed by DHS [REYN91, REYN96], and in preliminary investigations at LBNL [MERR95A, MERR95B, MERR96A, MERR96B]. Two of the preliminary LBNL reports, [MERR95B] and [MERR96A], described a kth nearest neighbor analysis that showed highly significant non-uniformity of rates. A subtle but important error was recently discovered, which invalidates the earlier LBNL conclusions; the error is explained in detail in Appendix D.

Three of the earlier LBNL analyses [MERR95A, MERR95B, MERR96A] used preliminary population estimates; namely, 1980 Census population for children 0-17, rather than 1980-88 population at risk for children 0-14. That deficiency was remedied in [MERR96B] and in the present report [MERR98]; the correction appears not to have altered the conclusions.

In [MERR96B] were presented scatter plots and contour plots similar to those in Figures 20, 37 and 38 of the present report. In [MERR96B] it was concluded that the rates were non-uniform, but that assessment was subjective, based only on visual inspection of the plots. The element that is new in the present analysis is the statistical analysis of the log RR distributions, summarized in Figures 33-36 and in Tables II and III. That analysis, and the Poisson-based statistic in Figures 41 and 42, are the *quantitative* measures of geographic variation in rates, and the results are negative.

An epidemiologic investigation can never *prove* the null hypothesis, it can only *reject* the null hypothesis, provided there is a real effect *and* the method is sufficiently sensitive to detect it. It is still an open question, as to whether there is any geographic

variation of rates in the four-county data set. One can only state that either (a) no such variations exist, or (b) the analysis was not sensitive enough to detect them.

Density Equalizing Map Projections

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 the distribution of cases should be random 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:

1. 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.
2. The DEMP technique avoids the calculation of unstable rates for small subareas having few cases.
3. The full geographic detail of the data can be used.
4. The DEMP analysis is appropriate, and even works best, in the analysis of rare diseases where the number of cases is small.
5. Systematic effects across broad regions of the map are easily detected, without the need for arbitrary grouping of subareas.
6. Rigorous, simple well-developed statistical techniques are available for analyzing the density equalized map.

7. 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.
8. If the null hypothesis is rejected, testing a different model can be simply performed by equalizing the map with respect to *expected cases*, rather than *population at risk*. The same method can be used to adjust for geographic variation of age, race, and other risk factors.

The present report illustrates some of the future potential of "electronic epidemiology" on a grand scale. Certainly, the need continues for traditional epidemiologic studies. For many investigations, there is no substitute for legwork and personal contact with one's study subjects. The data collected may be personal and confidential, not usable in more than one study. Computing needs can be met with a PC, perhaps even a hand calculator. That is the classic realm of epidemiology, but it is not the realm considered here.

The resources used in this analysis represent two decades of effort by dozens of individuals. The author alone spent a decade developing Census data, and another decade developing the technique of Density Equalizing Map Projections. Epidemiologic studies of this magnitude are not uncommon, but what is unique to this project is the residue of major resources that are *re-usable* in future projects. That residue includes 1970, 1980 and 1990 Census data and map files for the entire United States. With these data in hand, it will be possible in the future to automate all the steps in the present analysis, for any time period from 1970 through 1990.

This effort will have value not only for DEMP analyses, but for any future study in which age/sex/race/year-specific estimates of population at risk are required at the Census tract level.

An important new resource is the World Wide Web. Any institution or even an individual can, with minimal expense, obtain and integrate the data and tools for a particular task; and then, if one wishes, add one's own contribution to the growing store of public electronic resources.

For this growth to occur spontaneously, the tools (documents, data, and programs) must be generally useful, transportable, well-documented, inexpensive, public, non-proprietary, modular, and easy to use. Those are the goals that have guided this author, beginning with the integration and documentation of socio-economic, demographic, and health-related data in the mid-1970's. The product of that effort is the thousands of programs and data files listed in Appendix A, publicly documented in enough detail that they can be re-used without the author's assistance. *Those files which are proprietary or confidential are not publicly accessible, but can be readily used by authorized individuals.*

Most of the calculations described in Appendix A are straightforward but extremely tedious. This is especially true for the estimation of population at risk, the preparation of the pre-DEMP map files, and the post-DEMP statistical analysis. The author's general procedure is to document a series of small steps, and to provide public access to the intermediate data files at every step in the process. This permits the user or future implementor to proceed one small step at a time, checking the results of each step and inserting new procedures as required. The more complex

calculations are candidates for future Web implementation and integration, by an institution that has the resources to ensure continued maintenance and necessary user support in the future.

FUTURE DIRECTIONS

Any research project, by definition, opens up more new questions than it answers. In genealogy, for example, every "discovered" ancestor in a family tree leads one to two "undiscovered" ancestors; namely, the parents of the first. And so the search continues without end. There is *no possible* "completion" of a research project; there are only external constraints like budgets, retirements, and filing deadlines. In both areas mentioned above (the Four County data set, and DEMP-related analysis techniques in general), there is the potential to extend the present state of knowledge.

Four County Childhood Cancer data set

With the log RR statistical analysis described earlier, sensitivity is lost by plotting each case randomly in its tract of occurrence. (The random plotting procedure is a simple method of eliminating the bias due to within-tract clusters, which cannot be removed by the DEMP procedure.) Sensitivity can be improved by replacing the random plotting procedure with an exact area integral, as described in Appendix D. A closed-form calculation is computationally feasible, but the derivation is algebraically complex and was too ambitious for the present project. Appendix F, which provides the general formula for the area integral of a polynomial in a polygon, is a significant first step in that direction.

With that improvement, the log RR test described earlier will make *optimum* use of *all* the available data, and is guaranteed to be *at least* as sensitive as the DHS test or the Poisson-based test of Figures 41 and 42. This is true because in the limiting case k=0, the evaluation of RR involves no averaging over adjacent tracts, and is merely a geographically blind analysis of rates in the 259 tracts. If there exists *any* correlation in RR among adjacent tracts, there will be some optimum value of k, greater than zero, which will provide greater sensitivity than the k=0 test.

Additionally, the four-county analysis can be repeated with population data at a finer geographic level. The DHS analysis used 101 communities, and the present analysis used 259 Census tracts. All five of the required inputs (1980 and 1990 Census data, 1980 and 1990 map files, and the case data) are available for Census enumeration districts and block groups, which are generally one-fourth the size of Census tracts.

Density Equalizing Map Projections

In cooperation with former colleagues at the University of California, the author is applying for funding to apply the DEMP technique to additional data sets, possibly including HIV-related diseases in San Francisco, breast cancer in the San Francisco Bay area, and infections diseases in the agricultural San Joaquin valley of California. If obtained, that funding will permit further development of the DEMP algorithm, leading to a new implementation that can be used on a PC and integrated with commercial Geographic Information System (GIS) applications.

In cooperation with major statistical organizations (including the Bureau of the Census, and the Interuniversity Consortium for Political and Social Research) the author is applying for grant funding from the National Science Foundation and other sources. If obtained, that funding will permit preservation and integration of 1970-90 Census data in a user-friendly Web environment, which will greatly facilitate complex analyses like the one described in this report.

REFERENCES

(Some of the electronic addresses listed below are working documents in progress. If any of these have moved in the future, contact Deane Merrill at merrill@crocker.com for the current location.)

BOOT87. Boots B.N. Voronoi (Thiessen) Polygons. In series: Concepts and Techniques in Modern Geography (CATMOG), Geo Books, Norwich UK (1987).

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

CENS84. Intercensal Estimates of the Population of Counties by Age, Sex and Race: 1970-1980 Tape Technical Documentation. U.S. Bureau of the Census, Washington DC, 1984.

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. http://parep2.lbl.gov/pdocs/tr940401/all.html. #22 in DEMP Bibliography.

CRES91. Cressie, N. *Statistics for Spatial Data*, Wiley, New York; 1991.

GILL27. Population maps. AJPH 17:316-319, 1927.

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.

GUSE94. Gusein-Zade S.M. and Tikunov V.S. 1994. The Transformed Image: Current Status and Future Prospects, Mapping Sciences and Remote Sensing, 31, 1, 33-85. This monograph has three sections, which are listed in the DEMP Bibliography, below.

HAAC03. Haack H. and Wiechel H. 1903. Kartogramm zur Reichstagswahl. 2 Wahlkarten d. Deutschen Reiches in alter u. neuer Darstellung mit polit.-statist. Begleitworte u. kartogr. Gotha. Cited in [GUSE94].

HOWE70. Howe G.M. 1970. Some Recent Developments in Disease Mapping. Royal Soc. Hlth. J. 90, 16-20.

KARS23. Karl G. Karsten, *Charts and Graphs,* Ch. LII. Published in 1923. Cited in GILL27.

MERR91A. 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. http://parep2.lbl.gov/pdocs/asa91/asa91.txt.html. #14 in DEMP Bibliography. Summary version (6 pages) in proceedings, Section on Statistics and the Environment, American Statistical Association, pp. 96-101, published June 1992. http://parep2.lbl.gov/pdocs/asa91/short.txt.html. #15 in DEMP Bibliography.

MERR91B. Merrill D. 1991. Density Equalizing Map Projections (Cartograms) in Public Health Applications (draft), internal communication, pp. 49-55.

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. http://parep2.lbl.gov/mdocs/stonybr/stonybr.txt.html (text only). #16 in DEMP Bibliography.

MERR93. Merrill DW. 1993. Data required for prototype small-area analysis. Task Completion Report due 12/15/93. Internal documentation. http://parep2.lbl.gov/~merrill/docs/ftp/status/tr931215.asc.html.

MERR94A. Merrill DW. 1994. FY 1994 PAREP task completion report. Internal documentation. http://parep2.lbl.gov/~merrill/docs/parep/fy94complete.html.

MERR94B. Merrill DW. 1994. Preparation of geographic map files for DEMP transformation. DOE task completion report due 1/15/94 (revised). Internal documentation. http://parep2.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. http://parep2.lbl.gov/pdocs/vancouver/vancouver.html. #18 in DEMP Bibliography.

MERR95A. Merrill D, Selvin S, Close ER and Holmes HH. 1995. Use of density equalizing map projections (DEMP) in the analysis of childhood cancer in four California counties. Report LBL-36630, January 1995. http://parep2.lbl.gov/pdocs/cdc9501/lbl36630.html. #19 in DEMP Bibliography.

MERR95B. Merrill D, Selvin S, Close ER and Holmes HH. 1995. Use of density equalizing map projections (DEMP) in the analysis of childhood cancer in four

California counties. Graphics presented at 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. http://parep2.lbl.gov/pdocs/cdc9501/25 graphics.html. #20 in DEMP Bibliography.

MERR96A. Merrill DW, Selvin S, Close ER and Holmes HH. 1996. Use of density equalizing map projections (DEMP) in the analysis of childhood cancer in four California counties. Statistics in Medicine, Vol. 15, 1837-1848 (1996). #21 in DEMP Bibliography.

MERR96B. Merrill DW and Selvin S. 1996. . Use of density equalizing map projections (DEMP) in the analysis of childhood cancer in four California counties. Report LBL-36630 Rev. 3. Presented at National Cancer Institute GIS Advisory Meeting, November 25, 1996. #23 in DEMP Bibliography.

MERR98. Merrill DW. 1998. (this report) Density Equalizing Map Projections (Cartograms) in Public Health Applications. Dr.P.H. Dissertation, University of California, Berkeley School of Public Health, May 1998. Lawrence Berkeley National Laboratory Report LBNL-41624. http://parep2.lbl.gov/~merrill/thesis/thesis.html. #26 in DEMP Bibliography.

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.

REYN96. Reynolds P, Smith DF, Satariano E, Nelson DO, Goldman LR, Neutra RR. 1996. The Four County Study of Childhood Cancer: Clusters in Context. Statistics in Medicine, Vol. 15, 683-697 (1996).

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. #6 in DEMP Bibliography.

SEAM1798. Seaman V. 1798. An Inquiry into the Cause of the Prevalence of the Yellow Fever in New York, The Medical Repository, 1, 315-332. Cited in STEV65.

SEED94. SEEDIS (Socio-Economic Environmental Demographic Information System). http://parep2.lbl.gov/mdocs/seedis/seedis.html.

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

SNOW1849. Snow J. 1849. On the mode of communication of cholera. 1st edition, London: John Churchill. The first edition had no map. The famous map appeared in the 2nd edition, London, 1855. (The map faces p.45).

STEV65. Stevenson L.G. 1965. Putting Disease on the Map: the Early Use of Spot Maps in the Study of Yellow Fever, J.Hist. Med., 20, 226-261.

TOBL61. Waldo R. Tobler, Map Transformations of Geographic Space. Ph.D. Thesis, University of Washington, 1961.

TOBL70. Cartograms of Irregularly Shaped Areas, copyright 1970 by Waldo R. Tobler. Private communication, 1987.

WALL26. Wallace, J.W. Population map for health officers. AJPH 16:1023, 1926.

DEMP BIBLIOGRAPHY

The following is a bibliography of published and unpublished reports on computer implementations of Density Equalizing Map Projections (DEMP). It includes all documents of which the author has knowledge. Additions and corrections to this list will be appreciated.

For LBNL publications, the list is complete. URL locations are subject to change; for current locations consult http://parep2.lbl.gov/mdocs/demp/pubs.html. Some of the URL's cited below are text-only documents without figures. For paper copies of any of the LBNL publications, please send your e-mail address and your postal mailing address to merrill@crocker.com. Specify exactly which documents are desired.

LBNL algorithm #1: radial expansion at polygon centroids

1. Merrill, D.; Cartograms May Reveal Patterns in Cancer Incidence. In Currents (Lawrence Berkeley Laboratory newspaper), September 14, 1984. 1 page.

2. Shaw, G.; A Comparison of Techniques Used for the Detection of Spatial and Temporal-Spatial Disease Clustering; Dr.Ph. dissertation, Dept of Epidemiology, UC Berkeley; completed June 1986. 126 pages.

3. Schulman, J.; The Statistical Analysis of Density Equalized Map Projections; Ph.D. dissertation, Department of Biostatistics, UC Berkeley; Report LBL-22446; November 1986. 229 pages.

4. Selvin, S., Merrill, D., Schulman, J., Shaw, G., Benson, W. and Mohr, M.; Illustrations of a Density-Equalizing Map Projection Technique; Lawrence Berkeley Laboratory report LBL-23189; April 1987. 40 pages.

5. Selvin, S., Shaw, G., Schulman, J., and Merrill, D.W.; Spatial Distribution of Disease: Three Case Studies. Presented at Annual Meeting of the Society for Epidemiological Research, Chapel Hill NC, June 1985. Revised version: LBL-21162; Journal of the National Cancer Institute, 1987 Sep, 79(3):417-23. 7 pages.

6. Schulman J; Selvin S; Merrill DW. Density equalized map projections: a method for analysing clustering around a fixed point. Report LBL-24834. Statistics in Medicine, 1988 Apr, 7(4):491-505. 15 pages.

7. Selvin S, Merrill D, Schulman J, Sacks S, Bedell L, and Wong L; Transformations of Maps to Investigate Clusters of Disease; LBL-18550, October 1984; presented at the Annual Meeting of the American Public Health Association, Anaheim CA, November 1984. 33 pages.

8. Selvin S, Merrill D, Schulman J, Sacks S, Bedell L, and Wong L; Transformations of Maps to Investigate Clusters of Disease; LBL-18550(rev), June 1987; Social Science and Medicine, 1988, 26(2):215-21. 7 pages.

9. Merrill, D.; Density-Equalizing Map Transformations in Environmental Health Applications; in Proceedings of: Piecing the Puzzle Together: a Conference on Integrating Data for Decisionmaking; Washington DC; May 27-29, 1987; pp 122-125. Published by National Governors' Association, Washington, DC, 1988. 4 pages.

10. Shaw GM, Selvin S, Swan SH, Merrill D, and Schulman J; An Examination of Three Disease Clustering Methodologies; report LBL-23380, April 1987; International Journal of Epidemiology, 1988 Dec, 17(4):913-9. 7 pages.

11. Merrill, D.; Density Equalizing Map Projections (Cartograms) in Public Health Applications; presented at AutoCarto conference, American Congress of Surveyors and Mappers, Baltimore MD, April 2-7, 1989. 33 pages.

12. Schulman J., Selvin S., Shaw G. and Merrill D.; Detection of Excess Disease Near an Exposure Point: A Case Study; LBL-27628, July 1989; Archives of Environmental Health, 1990 May-Jun, 45(3):168-74. 7 pages.

LBNL algorithm #2: constrained minimization

13. Merrill D, Selvin S and Mohr MS. Density Equalizing Map Projections: A New Algorithm. Lawrence Berkeley Laboratory Report LBL-31984, February 1992. To be published in conference proceedings of Distancia '92: International Meeting on Distance Analysis, Rennes, France, June 22-26, 1992. 4 pages.

14. Merrill DW, Selvin S and Mohr MS; Analyzing geographic clustered response; report LBL-30954, June 1991; invited paper presented at 1991 Joint Statistical Meetings of the American Statistical Association, Atlanta GA, August 1991. WWW URL (text only): http://parep2.lbl.gov/pdocs/asa91/asa91.txt.html. 44 pages.

15. Merrill DW, Selvin S and Mohr MS; Analyzing geographic clustered response. Summary version in proceedings, Section on Statistics and the Environment, American Statistical Association, pp. 96-101, published June 1992. WWW URL (text only): http://parep2.lbl.gov/pdocs/asa91/short.txt.html. 6 pages.

16. Merrill D, Selvin S and Mohr MS. Density Equalizing Map Projections: Techniques and Applications. Lawrence Berkeley Laboratory report LBL-32640, July 1992. Presented at Workshop on Statistics and Computing in Disease Clustering, Stony Brook NY, 23-24 July 1992. WWW URL (text only): http://parep2.lbl.gov/mdocs/stonybr/stonybr.txt.html. 33 pages.

17. Selvin S, Schulman J, and Merrill D; Interpoint squared distance as a measure of spatial clustering; Report LBL-30430, February 1991; Social Science and Medicine, 1993 Apr, 36(8):1011-6. 14 pages.

LBNL algorithm #3: Russian line integral (RLInt)

18. Merrill, D., Selvin, S., and Close, E.R. 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. Presented at California Department of Health Services, 9/29/94. Draft project completion report submitted to Department of Energy on 7/27/94. WWW URL: http://parep2.lbl.gov/pdocs/vancouver/vancouver.html. 28 pages.

19. Merrill DW, Selvin S, Close ER and Holmes HH. 1995. Use of Density Equalizing Map Projections (DEMP) in the Analysis of Childhood Cancer in Four California Counties. Report LBL-36630, January 1995 (74 pages). Task completion report submitted to Department of Energy on 1/3/95. WWW URL: http://parep2.lbl.gov/pdocs/cdc9501/lbl36630.html. 74 pages.

20. Merrill DW, Selvin S, Close ER and Holmes HH. 1995. Use of Density Equalizing Map Projections (DEMP) in the Analysis of Childhood Cancer in Four California Counties. Graphics presented at the 1995 CDC/ATSDR Symposium on Statistical Methods: Small Area Statistics in Public Health: Design, Analysis, Graphic and Spatial Methods; January 25-26, 1995; Atlanta, Georgia. WWW URL: http://parep2.lbl.gov/pdocs/cdc9501/25graphics.html. 25 pages.

21. Merrill DW, Selvin S, Close ER and Holmes HH. 1995. Use of Density Equalizing Map Projections (DEMP) in the Analysis of Childhood Cancer in Four California Counties. Statistics in Medicine, Vol. 15, 1837-1848 (1966). 16 pages. For copyright notice and reprint information see WWW URL: http://parep2.lbl.gov/pdocs/cdc9510/reprint.html. Report LBL-36630 Rev.2, October 1995 (24 pages). Revised version of LBL-36630 Rev., April 1995.

22. Close ER, Merrill DW and Holmes HH. 1995. Implementation of a New Algorithm for Density Equalizing Map Projections (DEMP). Report LBL-35738, July 1995. WWW URL: http://parep2.lbl.gov/pdocs/tr940401/all.html. 134 pages.

23. Merrill DW and Selvin S. 1996. Use of Density Equalizing Map Projections (DEMP) in the Analysis of Childhood Cancer in Four California Counties. Report LBL-36630 Rev.3, November 1996. Presented at National Cancer Institute GIS Advisory Meeting, November 25, 1996. http://parep2.lbl.gov/pdocs/nci9611/nci9611.html. 12 figures with captions.

24. Merrill DW and Selvin S. 1997. Analyzing Spatial Patterns in Health Outcome Surveillance Data. Presented at 1997 CDC and ATSDR Symposium on Statistical Methods: Statistical Bases for Public Health Decision Making: From Exploration to Modeling, January 28-30, 1997, Atlanta, Georgia. http://parep2.lbl.gov/pdocs/cdc9701/cdc9701.html. 12 figures with captions.

25. Selvin S, Merrill DW, White MC, Ragland K, and Erdmann C. Breatst Cancer: Maps of Two Bay Area Counties. Journal of the American Public Health Association (in press, February 1998).

26. (this report) Merrill DW. 1998. Density Equalizing Map Projections (Cartograms) in Public Health Applications. Dissertation for Dr.P.H. in Epidemiology, University of California, Berkeley, School of Public Health. Lawrence Berkeley National Laboratory Report LBNL-41624, May 1998. http://parep2.lbl.gov/~merrill/thesis or http://www.bearhaven.com/thesis or http://merrill.wwh.net/thesis. About 220 pages.

Non-LBNL documents:

*Documents marked * are in the author's possession. UCB Information in parentheses following a citation indicates location in the UC Berkeley Library.*

* Allard R. 1997. Epidemiological Surveillance Bureau, Montreal General Hospital, Montreal, Canada. Private communication.

Angel S. and Hyman G. 1972. Transformations and Geographic Theory, Geographical Analysis, IV, 4, 350-367.

Angel S. and Hyman G. 1976. *Urban Fields*, Pion, London, 178 pp.

Bunge W. 1966. *Theoretical Geography*, 2nd edition, Gleerup, Lund.

* Cauvin C., Schneider C., and Cherrier G. 1989. Cartographic Transformations and the Piezopleth Maps Method., *The Cartographic Journal,* 26, 96-104.

Cerny J. 1972. Ontogenetic and Phylogenetic Perspectives on Cartograms, *The Monadnock*, 46 (June): 47-52.

Chrisman, N.R. 1974. The Impact of Data Structure on Geographic Information Processing. *Proceedings of AUTOCARTO I*, Amer. Congress on Surveying and Mapping, Bethesda, MD, pp. 165-177.

* Chrisman, N.R. 1988. Cartogram Computer Program. Private communication. September 1988.

Cook G.D. 1977. The Presentation of Two Algorithms for the Construction of Value-by-area Cartograms. M.Sc. thesis, University of Washington, Seattle, WA.

Cuff D.J., Pawling, J.W. and Blair E.T. 1984. Nested value-by-area cartograms for symbolizing land-use and other proportions, *Cartographica*, 21, 4, pp. 1-8.

* Dean AG. 1976. Population-Based Spot Maps: An Epidemiologic Technique. AJPH, 66, 10, 988-989.

Dent B. 1972. A Note on the Importance of Shape in Cartogram Communication, *Journal of Geography*, 71: 393-401.

* Dougenik J.A., Chrisman N.R., and Niemeyer D.R. 1985. An Algorithm to Construct Continuous Area Cartograms, *Professional Geographer*, 37, 1, 75-81.

Eastman J., NelsonW. and Shields G. 1981. Production Considerations in Isodensity Mapping, *Cartographica*, 18, 1: 24-30.

Elsasser H. 1970. Nichtflächenproportionale Kartogrammartige Darstellungen der Schweiz, *Geographica Helvetica*, 25, 2: 78-82.

* Forster F. 1966. Use of a Demographic Base Map for the Presentation of Areal Data in Epidemiology. *Br. J. Prev. Soc. Med.*, 20, 165-171. (UCB: PH unclassed).

Fronczak E.J. 1971. Plot System Description, Vol. 11 of *The Michigan Terminal System*, 3rd ed., Univ. of Michigan Computing Center, Ann Arbor, 170 pp.

Gilliard Q. 1979. Places in the News - Use of Cartograms in Introductory Geography Courses, *Journal of Geography*, 78, N3, 114-115.

* Gillihan AF. 1927. Population Maps. *AJPH*, 17, 316-319. (UCB: PH RA421.A4).

Griffin T.L. 1980. Cartographic Transformation of the Thematic Map Base, *Cartography*, 11, 3 (March): 163-174.

Griffin, T.L.C. 1983. Recognition of Areal Units on Topological Cartograms. *American Cartographer*, 10, N1, 17-29.

Gusein-Zade S.M. and Tikunov V.S. 1990. Numerical Methods of Compilation of Anamorphotic Cartographic Images, *Geodezy i Kartografia*, 1, 38-44 (in Russian).

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

* Guseyn-Zade S.M. and Tikunov V.S. 1994. The Transformed Image: Current Status and Future Prospects, *Mapping Sciences and Remote Sensing*, 31, 1, 33-85. This monograph has three sections, which are the next three papers listed.

* Guseyn-Zade S.M. and Tikunov V.S. 1994. Compilation of Linear Transformed Images, *Mapping Sciences and Remote Sensing*, 31, 1, 34-48. Part 1 of _____, The Transformed Image, above. This document contains a bibliography of 27 references.

* Guseyn-Zade S.M. and Tikunov V.S. 1994. Analog Methods in the Compilation of Areal Transformed Images, *Mapping Sciences and Remote Sensing*, 31, 1, 49-65. Part 2 of _____, The Transformed Image, above. This document contains a bibliography of 42 references.

* Guseyn-Zade S.M. and Tikunov V.S. 1994. Numerical Methods in the Compilation of Transformed Images, *Mapping Sciences and Remote Sensing*, 31, 1, 66-85. Part 3 of _____, The Transformed Image, above. This document contains a bibliography of 14 references.

Haack H. and Wiechel H. 1903. Kartogramm zur Reichstagswahl. 2 Wahlkarten d. Deutschen Reiches in alter u. neuer Darstellung mit polit.-statist. Begleitworte u. kartogr. Gotha. Cited in Gusein-Zade S.M. and Tikunov V.S. 1994. Analog Methods in the Compilation of Areal Transformed Images, above.

Härö A. 1968. Area Cartogram of the SMSA Population of the United States, *Annals*, AAG, 58, 3 (Sept): 452-460.

Harris C. and McDowell G. 1955. Distorted Maps - A Teaching Device, *Journal of Geography*, 55: 286-289.

* Howe G.M. 1970. Some Recent Development in Disease Mapping, *Royal Soc. Hlth. J.* 90, 16-20 (RA421.R69).

Hunter J. and Young J. 1968. A Technique for the Construction of Quantitative Cartograms by Physical Accretion Models, *Professional Geographer*, 20, 6: 402-407.

* Jackel C.B. 1997. Using ArcView to Create Contiguous and Noncontiguous Area Cartograms, *Cartography and Geographic Information Systems*, 24, 2, 101-109.

* Kadmon N. and Shlomi E. 1978. A Polyfocal Projection for Statistical Surfaces, *The Cartographic Journal*, 15, 1 (June): 36-41.

* Kadmon N. 1982. Cartograms and Topology. *Cartographica*, 19, 1-17.

Kelly J.I. and Neville R.J.W. 1985. A Population Cartogram of New Zealand, *New Zealand Journal of Geography*, 79, 7-11.

Kelly J. 1987. Constructing an area-value cartogram for New Zealand's population. *New Zealand Cartographic Journal*, 17, 1, pp. 3-10.

Kosinski L.A. 1984. The Roots of Population Geography, in *Geography and Population*, Pergamon, ed. Clarke J.I.

Lacko L. 1967. The Form and the Contents of Economic Maps, *Tijdschrift voor Econ. en Soc. Geografie*, Nov/Dec: 324-330.

* Levison M.E. and Haddon W. 1965. The Area Adjusted Map: An Epidemiological Device, *Public Health Reports*, 80, 1 (Jan): 55-59.

Monmonier M. 1977. Maps, Distortions, and Meanings, AAG Resource Paper 75-4, Washington, D.C.

Monmonier M. 1977. Nonlinear Reprojection to Reduce the Congestion of Symbols on Thematic Maps, *The Canadian Cartographer*, 14, 1 (June): 35-47.

Olson J. 1976. Noncontiguous Area Cartograms, *The Professional Geographer*, 28, 4: 371-380.

* Raisz E. 1934. The Rectangular Statistical Cartogram, *Geographical Review*, 2: 292-296.

Raisz E. 1936. Rectangular Statistical Cartograms of the World, *Journal of Geography*, 35, 1: 8-10.

Raisz E. 1938. *General Cartography*. McGraw Hill, New York.

* Rase W. 1992. Kartographische Anamorphosen, in *Kartographische Nachrichten*, vol. 1 or 2; and private communication.

* Ripley BD. 1981. *Spatial Statistics*. John Wiley & Sons, Inc., New York.

Ruston G. 1971. Map Transformations of Point Patterns: Central Place Patterns in Areas of Variable Population Density, *Papers and Proceedings*, Regional Science Assn, 28: 111-129.

Seaman V. 1798. An Inquiry into the Cause of the Prevalence of the Yellow Fever in New York, *The Medical Repository*, 1, 315-32. Cited by Stevenson L.G. (1965).

Sen A. 1970. Uniformizing Mappings, unpublished manuscript, Toronto.

Sen, A. 1975. A Theorem Related to Cartograms, *American Mathematical Monthly*, 82, 4: 382-385.

Sen A. 1976. On a Class of Map Transformations, *Geographical Analysis*, 8, 1: 23-37.

Skoda L. and Robertson J. 1972. Isodemographic Map of Canada, Geographical Paper No. 50, Department of Environment (En 36-506/50), Ottawa.

Slaby D., Casady R.J., Malin H.J., and Coakley J.F. 1979. Obstacles to Accurate and Valid Geographic Assessment of Vital Event Data. In *Proceedings of the*

International Symposium on Cartography and Computing: Applications in Health and Environment, Vol. 1, Nov. 1979, ed. Robert Aangeenbrug.

Snow J. 1849. On the Mode of Communication of Cholera. 1st. ed, London, J. Churchill. (This had no map.) The famous map appeard in the 2nd edition, London, 1855. (The map faces p.45).

Stevenson L.G. 1965. Putting Disease on the Map: the Early Use of Spot Maps in the Study of Yellow Fever, *J.Hist. Med.*, 20, 226-261. Cited by Howe G.M. (1970).

* Sutherland I.N. 1962. Representations of National, Regional, and Local Statistics. *Brit. J. Prev. Soc. Med.*, 16, 30-39.

Tikunov V.S. 1988. Anamorphated Cartographic Images: Historical Outline and Construction Techniques, *Cartography* (Australia), 17, 1, 1-8.

* Tobler W. 1961. Map Transformations of Geographic Space, Ph.D. Thesis, University of Washington, Seattle (U. Microfilm #61-4011).

* Tobler W. 1963. Geographical Area and Map Projections, *Geographical Review*, 53: 59-78.

* Tobler W. 1967. Computer Cartograms. Unpublished manuscript, Ann Arbor, about 20 pp.

* Tobler W. 1970. Cartograms of Irregularly Shaped Areas, Computer program in BASIC. Private communication, 1987.

* Tobler W. 1973. A Continuous Transformation Useful for Districting, *Annals*, New York Academy of Science, 119: 215-220.

* Tobler W. 1973. Cartogram Programs, Ann Arbor, Cartographic Laboratory Report, 110 pp.

Tobler W. 1974. Cartogram Programs for Geographic data given by Latitude/Longitude Quadrilaterals, by Irregularly Shaped Polygonal Areas, and by a Mathematical Equation, Cartographic Laboratory Report No. 3, Ann Arbor, 110 pp.

* Tobler W. 1974. Cartogram Programs. Program listings and documentation, upublished. About 120 pp.

* Tobler W. 1979. Cartograms and Cartosplines, pp. 53-58 in NCHS: *Proceedings of the 1976 Workshop on Automated Cartography and Epidemiology*, DHEW, Publ. No. PHS-79-1254. US Govt. Printing Office, Washington DC (UCB: PH XD80 5719).

* Tobler W. 1979. A Selection of References on Cartograms of the "varivalent" type. Unpublished manuscript, 2 pp.

* Tobler W. 1982. CART1 and CART2, Cartogram Programs in FORTRAN. Unpublished manuscript.

* Tobler W. 1984. Interactive Construction of Contiguous Cartograms (abstract), for presentation at NCGA Conference on Computer Graphics, Anaheim, CA, May 1984.

* Tobler W. 1986. Pseudo-Cartograms, *The American Cartographer*, 13, 1, 43-50.

* Tobler W. 1987. What is a Cartogram? Unpublished manuscript. 2pp.

* Tyroler H. and Smith H.L. 1966. IV. Epidemiology and Planning for the North Carolina Regional Medical Program. *AJPH*, 58, 6, 1058-1067.

* Wallace J.W. 1926. Population Map for Health Officers. *AJPH*, 16, 1023 (UCB: PH RA421.A4).

* Wesseling C. 1991. Faculty of Geographical Sciences, Utrecht University (Netherlands), private communication, June 1991.

APPENDIX A. ELECTRONIC FILE LOCATIONS

Abbreviations are explained at the end of each table. The file locations listed below are subject to change. For current file locations, check the current version of http://parep2.lbl.gov/~merrill/thesis or http://www.bearhaven.com/thesis, or contact Deane Merrill at dwmerrill@lbl.gov or merrill@crocker.com.
Proprietary and confidential files are locked to prevent public access.

A.1. FIGURES

Figures:

1980 Census tracts: Maps from California DHS:

described in:	ID	title	location	date	pubs	format
REYN91	af1	Cases diagnosed in the four county childhood cancer study area 1980-1988	$dhsfigs/fig1.ps	1/20/95	MERR95A fig.1; MERR98 fig.1	ps
REYN91	af2	Childhood cancer incidence rate ratios (and 95% Cl) for Four County communities compared to the overall Four County rate	$dhsfigs/fig2.ps	1/20/95	MERR95A fig.2; MERR98 fig.2	ps
REYN91	af3	Four County Childhood Cancer Study communities with high and low rates of childhood cancer	$dhsfigs/fig3.ps	1/20/95	MERR95A fig. 3; MERR98 fig.3	ps

$dhsfigs = http://parep2.lbl.gov/pdocs/cdc9501/dhsfigs = parep2.lbl.gov /data9/old/parep2/merrill/docs/parep/cdc9501/dhsfigs.
described in: where this figure is documented.
ID = identifier of this figure.
ps = Adobe PostScript.

Figures:

1980 Census Tracts: Poisson based significance test:

ID	title	location	date	pubs	format
af3a	Poisson based test, 1980 Census tracts, real cases	$spatial/sum.ps or $spatial2/sum.ps	1/18/95	MERR95A fig. 6	ps
af3b	Poisson based test, 1980 Census tracts, random cases	$spatial/sum.ps or $spatial2/sum.ps	1/18/95		ps

$spatial = parep2.lbl.gov/~merrill/selvin/parep/spatial
$spatial2 = parep2.lbl.gov/data9/old/parep2/merrill/docs/parep/cdc9501/25graphics = http://parep2.lbl.gov/pdocs/cdc9501/25graphics

Figures:

1980 Census Tracts, pre-DEMP Maps: Census tract boundaries:

described	ID	title	location	date	pubs	format
ad3	af4old	Four-county map obtained from SEEDIS (1980 Census tracts)	$figs/fig2.ps, gif	7/6/94		ps, gif
ad3	af4	four-county map from SEEDIS, with 401 cases (1980 Census tracts)	$myfigs/fig2.ps, gif	10/18/94	MERR95A fig.4; MERR98 fig.4	ps, gif
ad5	af5	Four-county map repaired by hand (1980 Census tracts)	$myfigs/fig2b.ps, gif	10/18/94		ps, gif
ad6	af6old	Four-county map with errors removed (1980 Census tracts)	$figs/fig3.ps, gif	7/6/94		ps, gif
ad6	af6	Four-county map with errors removed (1980 Census tracts)	$myfigs/fig3.ps, gif	10/18/94		ps, gif
ad7	af7	Reduction of Map Complexity (1980 Census tracts)	$figs/fig4.ps, gif	7/6/94		ps, gif
ad8	af8	Reduced four-county map (20 percent) (1980 Census tracts)	$figs/fig5.ps, gif	7/6/94		ps, gif
ad9	af9	Four-county map (20 percent), simply connected (1980 Census tracts)	$myfigs/fig24.ps, gif	10/18/94		ps, gif
ad10	af10	Four-county map, divided (40 km) (1980 Census tracts)	$myfigs/fig24b.ps, gif	10/18/94		ps, gif
ad11	af11	four-county map, filtered and triangulated, with 401 cases (1980 Census tracts)	$myfigs/fig25.ps, gif	12/28/94	MERR95A fig. 5	ps, gif
ad12	af12old	Four-county map, hexagons (1980 Census tracts)	$myfigs/fig25.ps, gif	11/15/94		ps, gif
ad31	af31old	Four-county map, age 0-17, iteration 0 of 10 (with tract and triangle boundaries and actual case locations)	$figs/fig31.ps, gif	7/18/94	MERR94 fig.1A	ps, gif
ad31	af31	401 case locations, initial map (with county boundaries and actual case locations)	$myfigs/fig31.ps, gif	11/15/94	MERR95A fig.7	ps, gif
cgi	af12	1980 census tracts: original map (reduced, hexagons)	$4county/orig_hex.ps, gif	10/8/96	MERR98 fig.6	ps, gif

$4county = http://parep2.lbl.gov/~merrill/maps/tr940115/4county = parep2.lbl.gov /export/home/u0/cedrdv/merrill/public_html/maps/tr940115/4county.
$figs = http://parep2.lbl.gov/pdocs/tr940115/figs = parep2.lbl.gov /data9/old/parep2/merrill/docs/parep/tr940115/figs

$myfigs = http://parep2.lbl.gov/~merrill/figs = parep2.lbl.gov /export/home/u0/cedrdv/merrill/public_html/figs.
af4old, etc = documents listed in Appendix A: Electronic Documentation
described: where this figure is documented.
gif = Graphic Interchange Format
ID = identifier of this figure.
ps = Adobe PostScript.

Figures:

1980 Census tracts: pre-DEMP maps: initial tract areas versus target areas:

described in:	ID	title	location	date	pubs	format
ad52	af52old	present areas versus target areas, initial map	$figs/fig52.ps, gif (not saved)	7/18/94	MERR94 fig.1B	ps, gif
ad52	af52	present areas versus target areas, initial map	$myfigs/fig52.ps, gif	11/15/94	MERR95A fig.B-1	ps,gif

$figs = http://parep2.lbl.gov/pdocs/tr940115/figs = parep2.lbl.gov /data9/old/parep2/merrill/docs/parep/tr940115/figs
$myfigs = http://parep2.lbl.gov/~merrill/figs = parep2.lbl.gov /export/home/u0/cedrdv/merrill/public_html/figs.
af52old, etc = documents listed in Appendix A.2: Electronic file locations: Documentation
described in: where this figure is documented.
gif = Graphic Interchange Format
ID = identifier of this figure.
ps = Adobe PostScript.

Figures:

1980 Census tracts: partially equalized maps:

ID	title	location	date	format
af41	4-county map, age 0-17, iteration 1 of 1	$figs/fig41.ps, gif	7/18/94	ps, gif
af43	4-county map, age 0-17, iteration 1 of 10	$figs/fig43.ps, gif	7/18/94	ps, gif
af44	4-county map, age 0-17, iteration 9 of 10	$figs/fig44.ps, gif	7/18/94	ps, gif
af45	4-county map, age 0-17, iteration 5 of 10	$figs/fig45.ps, gif	7/18/94	ps, gif

$figs = http://parep2.lbl.gov/pdocs/tr940115/figs = parep2.lbl.gov /data9/old/parep2/merrill/docs/parep/tr940115/figs
af 41, etc. = figures listed in Appendix A.1: Electronic file locations: Figures.
gif = Graphic Interchange Format
ID = identifier of this figure.
ps = Adobe PostScript.

Figures:

1980 Census tracts: density equalized maps:

described in:	ID	title	location	date	pubs	format
ad42	af42old	Four-county map, age 0-17, iteration 10 of 10 (with tract and triangle boundaries and case locations)	$figs/fig42.ps, gif	7/18/94	MERR94 Fig.3A	ps, gif
ad42	af42	401 case locations, run hex10, step 10 of 10 (with county boundaries and case locations, and random cases outside boundary)	$myfigs/fig42.ps, gif	11/15/94	MERR95 A Fig.8	ps, gif
ad50	af50old	Four-county map, age 0-17, iteration 10 of 10 (hexagons, age 0-17, total pop)	$figs/fig50.ps, gif	7/18/94	MERR94 Fig.3B	ps, gif
ad50	af50	present areas versus target areas, run hex10, step 10 (hexagons, age 0-17, total pop)	$myfigs/fig50.ps, gif	11/15/94	MERR95 A Fig.B-3	ps.gif

$figs = http://parep2.lbl.gov/pdocs/tr940115/figs = parep2.lbl.gov /data9/old/parep2/merrill/docs/parep/tr940115/figs.
$myfigs = http://parep2.lbl.gov/~merrill/figs = parep2.lbl.gov /export/home/u0/cedrdv/merrill/public_html/figs.
ad42 etc = documents listed in Appendix A: Electronic Documentation.
described in: where this figure is documented.
gif = Graphic Interchange Format
ID = identifier of this figure.
ps = Adobe PostScript.

Figures:

1990 Census tracts, pre-DEMP maps:

described in:	ID	title	location	date	pubs	format
ad15	af15	1990 four-county map from GDT, with 401 cases	$myfigs/fig106.ps, gif	4/25/95	MERR98 fig.5	ps, gif
ad16	af16	Reduced 1990 map (20 percent, 40 km)	$myfigs/fig107.ps, gif	4/26/95		ps, gif
ad17	af17	Land-cleaned 1990 map (0.5 sq km)	$myfigs/fig108.ps, gif	4/26/95		ps, gif
ad18	af18	1990 map, lakes removed (40 sq km)	$myfigs/fig109.ps, gif	4/27/95		ps, gif
ad19	af19	1990 map, tracts reassembled (20 percent, 40 km)	$myfigs/fig110.ps, gif	4/28/95		ps, gif
ad20	af20	1990 map, simply connected	$myfigs/fig111.ps, gif	4/28/95		ps, gif
ad21	af21	1990 map, triangulated	$myfigs/fig112.ps, gif	4/28/95		ps, gif
ad22	af22old	1990 map, hexagons	$myfigs/fig112.ps, gif	4/28/95		ps, gif
cgi	af22	1990 census tracts: original map (reduced)	$4county90/orig_hex.ps, gif	10/8/96	MERR98 fig.7	ps, gif

$4county90 = http://parep2.lbl.gov/~merrill/maps/tr940115/4county90 = parep2.lbl.gov /export/home/u0/cedrdv/merrill/public_html/maps/tr940115/4county90.
$myfigs = http://parep2.lbl.gov/~merrill/figs = parep2.lbl.gov /export/home/u0/cedrdv/merrill/public_html/figs.
ad15-df22 = documents listed in Appendix A.2: Electronic File Locations: Electronic Documentation.
cgi = Common Gateway Interface program located at http://parep2.lbl.gov/cgi-bin/merrilltest4.
described in: where this figure is documented.
$figs = http://parep2.lbl.gov/pdocs/tr940115/figs = parep2.lbl.gov /data9/old/parep2/merrill/docs/parep/tr940115/figs.
gif = Graphic Interchange Format
ID = identifier of this figure.
ps = Adobe PostScript.

Figures:

Modified 1980 Census Tracts: Poisson based significance test:

ID	title	location	date	pubs	format
af22a	Poisson based test, modified 1980 Census tracts, real cases	$spatial/sum8090.ps	3/20/98	MERR98 fig. 41	ps
af22b	Poisson based test, modified 1980 Census tracts, random cases	$spatial/sum8090.ps	3/20/98	MERR98 fig. 42	ps

$spatial = parep2.lbl.gov/~merrill/selvin/parep/spatial
ID = indentifier of this figure.
ps = Adobe PostScript.
pubs = where this figure is published.

Figures:

Modified 1980 Census tracts: pre-DEMP maps: no cases:

described in:	ID	title	location	date	pubs	format
cgi	af23	mod 1980 Census tracts: original map (reduced, hexagons)	$4county8090/orig_hex.ps, gif	10/10/96	MERR96 fig.1; MERR98 fig.8	ps, gif

$4county8090 = http://parep2.lbl.gov/~merrill/maps/tr940115/4county8090 = parep2.lbl.gov /export/home/u0/cedrdv/merrill/public_html/maps/tr940115/4county8090.
cgi = Common Gateway Interface program located at http://parep2.lbl.gov/cgi-bin/merrilltest4.
described in: where this figure is documented.
gif = Graphic Interchange Format
ID = identifier of this figure.
ps = Adobe PostScript format

Figures:

Modified 1980 Census tracts: pre-DEMP maps: real and random cases:

descr	ID	race	years	ages	sex	site	cases	location	date	pubs	format
cgi	af24	all races	1980-88.	0-14	m,f	all	401	$4c/four_orig_total.ps, gif	5/8/97	MERR98 fig.19	ps, gif
cgi	af25	**white non-Hisp**	1980-88	0-14	m,f	all	192	$4c/four_orig_anglo.ps, gif			ps, gif
cgi	af26	**Hispanics**	1980-88	0-14	m,f	all	166	$4c/four_orig_hisp.ps, gif	5/5/97		ps, gif
cgi	af27	**non-white non-Hisp**	1980-88	0-14	m,f	all	43	$4c/four_orig_nwnh.ps, gif			ps, gif
cgi	af28	all races	**1980-84**	0-14	m,f	all	209	$4c/four_orig_py8084.ps, gif			ps, gif
cgi	af29	all races	**1985-88**	0-14	m,f	all	190	$4c/four_orig_py8588.ps, gif			ps, gif
cgi	af30	all races	1980-88	**0-4**	m,f	all	211	$4c/four_orig_age0004.ps, gif			ps, gif
cgi	af31	all races	1980-88	**5-14**	m,f	all	190	$4c/four_orig_age0514.ps, gif			ps, gif
cgi	af32	all races	1980-88	0-14	**m**	all	226	$4c/four_orig_male.ps, gif			ps, gif
cgi	af33	all races	1980-88	0-14	**f**	all	175	$4c/four_orig_female.ps, gif			ps, gif
cgi	af34	all races	1980-88	0-14	m,f	**leuk**	134	$4c/four_orig_leuk.ps, gif			ps, gif
cgi	af35	all races	1980-88	0-14	m,f	**brain**	76	$4c/four_orig_brain.ps, gif			ps, gif
cgi	af36	all races	1980-88	0-14	m,f	**other**	191	$4c/four_orig_other.ps, gif			ps, gif

$4c = http://parep2.lbl.gov/~merrill/maps/tr940115/4county8090 = parep2.lbl.gov/export/home/u0/cedrdv/merrill/public_html/maps/tr940115/4county8090.
cgi = Common Gateway Interface program located at http://parep2.lbl.gov/cgi-bin/merrilltest4.
descr = where this figure is described.
gif = Graphic Interchange Format
leuk = leukemia.
ID = identifier of this figure.
ps = Adobe PostScript format
pubs = where this figure is published (including MERR98, this report).
race = race and ethnicity.
sex = gender.
site = cancer site.

Figures:

Modified 1980 Census tracts: density equalized maps: no cases:

descr	ID	race	years	ages	sex	Mpy	location	date	pubs	format
cgi	af24	all races	1980-88	0-14	m,f	3.3	$4c/hex_total.ps, gif	10/22/96	MERR96 fig.2; MERR98 fig.9	ps, gif
cgi	af25	**white non-Hisp**	1980-88	0-14	m,f	1.6	$4c/hex_anglo.ps, gif	1/23/97	MERR96 fig.3; MERR98 fig.10	ps, gif
cgi	af26	**Hispanics**	1980-88	0-14	m,f	1.3	$4c/hex_hisp.ps, gif	10/23/96	MERR96 fig.4; MERR98 fig.11	ps, gif
cgi	af27	**non-white non-Hisp**	1980-88	0-14	m,f	0.4	$4c/hex_nwnh.ps, gif	1/23/97	MERR96 fig.5; MERR98 fig.12	ps, gif
cgi	af28	all races	**1980-84**	0-14	m,f	1.7	$4c/hex_py8084.ps, gif	10/25/96	MERR98 fig.13	ps, gif
cgi	af29	all races	**1985-88**	0-14	m,f	1.6	$4c/hex_py8588.ps, gif	10/24/96	MERR98 fig.14	ps, gif
cgi	af30	all races	1980-88	**0-4**	m,f	1.2	$4c/hex_age0004.ps, gif	10/24/96	MERR98 fig.15	ps, gif
cgi	af31	all races	1980-88	**5-14**	m,f	2.1	$4c/hex_age0514.ps, gif	10/23/96	MERR98 fig.16	ps, gif
cgi	af32	all races	1980-88	0-14	**m**	1.7	$4c/hex_male.ps, gif	10/23/96	MERR98 fig.17	ps, gif
cgi	af33	all races	1980-88	0-14	**f**	1.6	$4c/hex_female.ps, gif	10/23/96	MERR98 fig.18	ps, gif

$4c = http://parep2.lbl.gov/~merrill/maps/tr940115/4county8090 = parep2.lbl.gov/export/home/u0/cedrdv/merrill/public_html/maps/tr940115/4county8090.
cgi = Common Gateway Interface program located at http://parep2.lbl.gov/cgi-bin/merrilltest4.
descr = where this figure is described.
gif = Graphic Interchange Format
ID = identifier of this figure.
Mpy = million person-years at risk.
ps = Adobe PostScript format
pubs = where this figure is published (including MERR98, this report).
race = race and ethnicity.
sex = gender.

Figures:

Modified 1980 Census tracts: density equalized maps: real and random cases:

descr	ID	race	years	ages	sex	site	cases	location	date	pubs	format
cgi	af37	all races	1980-88	0-14	m,f	all	401	$4c/four_demp_total.ps, gif	5/8/97	MERR96 fig.9; MERR98 fig.20	ps, gif
cgi	af38	**white non-Hisp**	1980-88	0-14	m,f	all	192	$4c/four_demp_anglo.ps, gif	10/31/96	MERR96 fig.10; MERR98 fig.21	ps, gif
cgi	af39	**Hispanics**	1980-88	0-14	m,f	all	166	$4c/four_demp_hisp.ps, gif	11/1/96	MERR98 fig.22	ps, gif
cgi	af40	**non-white non-Hisp**	1980-88	0-14	m,f	all	43	$4c/four_demp_nwnh.ps, gif	11/1/96	MERR98 fig.23	ps, gif
cgi	af41	all races	**1980-84**	0-14	m,f	all	209	$4c/four_demp_py8084.ps, gif	11/1/96	MERR98 fig.24	ps, gif
cgi	af42	all races	**1985-88**	0-14	m,f	all	190	$4c/four_demp_py8588.ps, gif	11/1/96	MERR98 fig.25	ps, gif
cgi	af43	all races	1980-88	**0-4**	m,f	all	211	$4c/four_demp_age0004.ps, gif	11/1/96	MERR98 fig.26	ps, gif
cgi	af44	all races	1980-88	**5-14**	m,f	all	190	$4c/four_demp_age0514.ps, gif	11/1/96	MERR98 fig.27	ps, gif
cgi	af45	all races	1980-88	0-14	**m**	all	226	$4c/four_demp_male.ps, gif	11/1/96	MERR98 fig.28	ps, gif
cgi	af46	all races	1980-88	0-14	**f**	all	175	$4c/four_demp_female.ps, gif	11/1/96	MERR98 fig.29	ps, gif
cgi	af47	all races	1980-88	0-14	m,f	**leuk**	134	$4c/four_demp_leuk.ps, gif	10/31/96	MERR98 fig.30	ps, gif
cgi	af48	all races	1980-88	0-14	m,f	**brain**	76	$4c/four_demp_brain.ps, gif	11/1/96	MERR98 fig.31	ps, gif
cgi	af49	all races	1980-88	0-14	m,f	**other**	191	$4c/four_demp_other.ps, gif	11/1/96	MERR98 fig.32	ps, gif

$4c = http://parep2.lbl.gov/~merrill/maps/tr940115/4county8090 = parep2.lbl.gov/export/home/u0/cedrdv/merrill/public_html/maps/tr940115/4county8090.
cgi = Common Gateway Interface program located at http://parep2.lbl.gov/cgi-bin/merrilltest4.
descr = where this figure is described.
gif = Graphic Interchange Format
leuk = leukemia.
ID = identifier of this figure.
ps = Adobe PostScript format
pubs = where this figure is published (including MERR98, this report).
race = race and ethnicity.
site = cancer site.

Figures:

Modified 1980 Census tracts: density equalized maps: adjusted tract areas versus target areas:

descr	ID	race	years	ages	sex	Mpy	hsum	location	date	pubs	format
cgi	cf1	all races	1980-88	0-14	m,f	3.3	0.0148	$4c/geoarea_total.ps, gif	10/24/96	MERR98: fig.C-1	ps, gif
cgi	cf2	**white non-Hisp**	1980-88	0-14	m,f	1.6	0.0273	$4c/geoarea_anglo.ps, gif	10/25/96		ps, gif
cgi	cf3	**Hisp**	1980-88	0-14	m,f	1.3	0.0658	$4c/geoarea_hisp.ps, gif	10/24/96		ps, gif
cgi	cf4	**non-white non-Hisp**	1980-88	0-14	m,f	0.4	0.333	$4c/geoarea_nwnh.ps, gif	10/24/96		ps, gif
cgi	cf5	all races	**1980-84**	0-14	m,f	1.7	0.166	$4c/geoarea_py8084.ps, gif	10/25/96		ps, gif
cgi	cf6	all races	**1985-88**	0-14	m,f	1.6	0.0461	$4c/geoarea_py8588.ps, gif	10/24/96		ps, gif
cgi	cf7	all races	1980-88	**0-4**	m,f	1.2	0.0269	$4c/geoarea_age0004.ps, gif	10/24/96		ps, gif
cgi	cf8	all races	1980-88	**5-14**	m,f	2.1	0.0084	$4c/geoarea_age0514.ps, gif	10/24/96		ps, gif
cgi	cf9	all races	1980-88	0-14	**m**	1.7	0.0124	$4c/geoarea_male.ps, gif	10/24/96		ps, gif
cgi	cf10	all races	1980-88	0-14	**f**	1.6	0.0136	$4c/geoarea_female.ps, gif	10/24/96		ps, gif

$4c = http://parep2.lbl.gov/~merrill/maps/tr940115/4county8090 = parep2.lbl.gov /export/home/u0/cedrdv/merrill/public_html/maps/tr940115/4county8090.
cf1 etc = figures in Appendix C: Checking Density Equalization.
cgi = Common Gateway Interface program located at http://parep2.lbl.gov/cgi-bin/merrilltest4.
descr = where this figure is described.
gif = Graphic Interchange Format
hsum = measure of degree of density equalization (average over 259 tracts, of squared relative error in tract area).
ID = identifier of this figure.
Mpy = million person-years at risk.
ps = Adobe PostScript.
pubs = where this figure is published (including MERR98, this report).
race = race and ethnicity.
sex = gender.

Figures:

Modified 1980 Census tracts: 8020 random cases:

descr	ID	title	location	date	pubs	format
cgi	cf11	8020 random cases, original map	$4c/big_orig.ps, gif	10/28/96	MERR96 fig.7; MERR98 fig.C-2	ps, gif
cgi	cf12	8020 random cases, density-equalized map	$4c/big_demp.ps, gif	10/28/97	MERR96 fig.8; MERR98 fig.C-3	ps, gif

$4c = http://parep2.lbl.gov/~merrill/maps/tr940115/4county8090 = parep2.lbl.gov /export/home/u0/cedrdv/merrill/public_html/maps/tr940115/4county8090.
cf11 etc = figures in Appendix C: Checking Density Equalization
cgi = Common Gateway Interface program located at http://parep2.lbl.gov/cgi-bin/merrilltest4.
descr = where this figure is described.
gif = Graphic Interchange Format
ID = identifier of this figure.
ps = Adobe PostScript.
pubs = where this figure is published (including MERR98, this report).

Figures:

Modified 1980 Census tracts: log of relative risk:

all races, 1980-88, ages 0-14, both sexes, all sites (401 cases, 3.3 Mpy)

descr	ID	k	method	location	date	pubs	format
cgi	af50	10	NN	$4c/varnn_10_total.ps, gif	5/8/97	MERR98 fig.33	ps, gif
cgi	af51	20	NN	$4c/varnn_20_total.ps, gif	5/8/97	MERR98 fig.34	ps, gif
cgi	af52	10	GK	$4c/vargk_10_total.ps, gif	5/8/97	MERR98 fig.35	ps, gif
cgi	af53	20	GK	$4c/vargk_20_total.ps, gif	5/8/97	MERR98 fig.36	ps, gif

$4c = http://parep2.lbl.gov/~merrill/maps/tr940115/4county8090 = parep2.lbl.gov /export/home/u0/cedrdv/merrill/public_html/maps/tr940115/4county8090.
cgi = Common Gateway Interface program located at http://parep2.lbl.gov/cgi-bin/merrilltest4.
descr = where this figure is described.
gif = Graphic Interchange Format
GK = Gaussian Kernel Method.
ID = identifier of this figure.
k = scaling parameter for relative risk analysis = 10 or 20. Equivalent to number of cases expected within the sampling area. Lower k provides better spatial resolution but less statistical power.
method = method used for relative risk analysis: NN or GK.
NN = nearest neighbor method.
ps = Adobe PostScript.
pubs = where this figure is published (including MERR98, this report).

Figures:

Modified 1980 Census tracts: contour maps of relative risk:

all races, 1980-88, ages 0-14, both sexes, all sites (401 cases, 3.3 Mpy), GK method:

descr	ID	k	projection	location	date	pubs	format
cgi	af54	10	DEMP	$4c/contour_10_total.ps, gif	5/7/97	MERR98 fig.37	ps, gif
cgi	af55	20	DEMP	$4c/coutour_20_total.ps, gif	5/7/97	MERR98 fig.38	ps, gif
cgi	af56	10	orig	$4c/origcontour_10_total.ps, gif	5/7/97	MERR98 fig.39	ps, gif
cgi	af57	20	orig	$4c/origcoutour_20_total.ps, gif	5/7/97	MERR98 fig.40	ps, gif

$4c = http://parep2.lbl.gov/~merrill/maps/tr940115/4county8090 = parep2.lbl.gov /export/home/u0/cedrdv/merrill/public_html/maps/tr940115/4county8090.
cgi = Common Gateway Interface program located at http://parep2.lbl.gov/cgi-bin/merrilltest4.
descr = where this figure is described.
gif = Graphic Interchange Format
ID = identifier of this figure.
k = scaling parameter for relative risk analysis = 10 or 20. Equivalent to number of cases expected within the sampling area. Lower k provides better spatial resolution but less statistical power.
Mpy = million person-years at risk
NN = nearest neighbor method.
projection = map projection: DEMP or original.
ps = Adobe PostScript.
pubs = where this figure is published (including MERR98, this report).

Figures:

Modified 1980 Census tracts: fraction of log RR in tail:

all races, 1980-88, ages 0-14, both sexes, all sites (401 cases, 3.3 Mpy), GK method, k=10:

ID	artificial cases	location	date	pubs	format
af58	uniform	$splus4\uniform.ps, gif	5/7/97	MERR98 fig.D-1	ps, gif
af59	random	$splus4\random.ps.gif	5/7/97	MERR98 fig.D-2	ps, gif

$splus4 = csr6.lbl.gov: c:\Program Files\splus4\users\merrill
artificial cases = method by which artificial cases were generated (see Appendix D).
gif = Graphic Interchange Format
GK = Gaussian Kernel method.
ID = identifier of this figure.
ps = Adobe PostScript.
pubs = where this figure is published (including MERR98, this report).

A.2. ELECTRONIC DOCUMENTATION

Electronic Documentation:

Map files: 1980 Census tracts: source maps:

ID	describes:	title	location	date	format
dm2	m2a-m7b	1980 Census Geographic Base Maps (installation in SEEDIS)	$seedict:tract80.hel	1/14/92	hel

$seedict = seedis.census.gov::sy$seedis: [seedis.seedict]
hel = format used by SEEDIS "page" program. To view tract80.hel, log into seedis.census.gov and type "page tract80".
m2-m7: map files listed in Appendix A.3: Electronic Documentation: Data Files: Map files.
SEEDIS = LBL Socio-Economic Environmental Demographic Information System. VMS program installed at seedis.census.gov::sy$seedis:[seedis]seedis.com.

Electronic Documentation:

Map files: 1980 Census tracts: pre-DEMP maps:

ID	describes:	title	location	date	format
ad1	ad2, ad13	Preparation of Geographic Map Files for DEMP Transformation: Figures. Included in MERR94B.	$tr94/97figs.html	11/5/95	html
ad2	ad3, ad...	Four California Counties: 1980 Census Tracts	$tr94/97figs4county.html	11/3/95	html
ad3	ad4-ad12	1980 Census Tracts: Pre-DEMP maps.	$tr94/97figs4county_predemp.html	4/25/95	html
ad4	af4old, af4	1980 Census Tracts: Pre-DEMP maps. Original map, from SEEDIS.	$tr94/972fig2.html	4/25/95	html
ad5	af5	1980 Census Tracts: Pre-DEMP maps. Four-county map repaired by hand.	$tr94/972bfig2b.html	10/18/94	html
ad6	af6old, af6	1980 Census Tracts: Pre-DEMP maps. Four-county map with errors removed	$tr94/973fig3.html	10/18/94	html
ad7	af7	1980 Census Tracts: Pre-DEMP maps. Reduction of map complexity.	$tr94/974fig4.html	5/11/95	html
ad8	af8	1980 Census Tracts: Pre-DEMP maps. Reduced four-county map (20 percent).	$tr94/975fig5.html	10/18/94	html
ad9	af9	1980 Census Tracts: Pre-DEMP maps. Four-county map, simply connected.	$tr94/9724fig24.html	9/8/95	html
ad10	af10	1980 Census Tracts: Pre-DEMP maps. Four-county map, divided (40 km).	$tr94/9724bfig24b.html	11/21/94	html
ad11	af11	1980 Census Tracts: Pre-DEMP maps. Four-county map, triangles, 40 km.	$tr94/9725fig25.html	12/21/94	html
ad12	af12	1980 Census Tracts: Pre-DEMP maps. Four-county map, hexagons.	$tr94/9726fig26.html	11/15/94	html
ad31	af31old, af31	Four-county map with case locations, age 0-17, hexagons, 1980 total pop, step 0 of 10	$tr94/9731fig31.html	11/15/94	html
ad52	af52old, af52	Present and target areas, age 0-17, hexagons, 1980 total pop, step 0 of 10	$tr94/9752fig52.html	5/11/95	html

$tr94 = http://parep2.lbl.gov/pdocs/tr940115 = parep2.lbl.gov /data9/old/parep2/merrill/docs/parep/tr940115.
ad2, etc. = electronic documents listed in Appendix A.2: Electronic File Locations: Electronic Documentation.
af4, etc. = figures in electronic format, listed in Appendix A1: Electronic File Locations: Figures.
describes: = ID(s) described by this document.
html = HyperText Markup Language.
ID = identifier of this document. All the documents ad1-ad12 are included in MERR94B.
SEEDIS = LBL Socio-Economic Environmental Demographic Information System. See http://parep2.lbl.gov/mdocs/seedis/seedis.html.

Electronic Documentation:

Map files: 1980 Census tracts: density equalized maps:

ID	describes:	title	location	date	format
ad42	af42old, af42	Four-county map, age 0-17, iteration 10 of 10 (hexagons, 0-17, total 1980 pop)	$tr94/9742fig42.html	11/15/94	html
ad50	af50old, af50	Present and target areas, age 0-17, hexagons, 1980 total pop, step 10 of 10	$tr94/9750fig50.html	2/25/98	html

$tr94 = http://parep2.lbl.gov/pdocs/tr940115 = parep2.lbl.gov /data9/old/parep2/merrill/docs/parep/tr940115.
af42, etc. = figures in electronic format, listed in Appendix A.1: Electronic File Locations: Figures.
describes: = ID(s) described by this document.
html = HyperText Markup Language.
ID = identifier of this document. All the documents ad1-ad12 are included in MERR94B.

Electronic Documentation:

Map files: 1990 Census tracts: pre-DEMP maps:

ID	describes:	title	location, or cgi parameters	date	format
ad1	ad2, ad13	Preparation of Geographic Map Files for DEMP Transformation: Figures. Included in MERR94B.	$tr94/97figs.html	11/5/95	html
ad13	ad14	Four California Counties: 1990 Census Tracts	$tr94/97figs4county90.html	4/25/95	html
ad14	ad15-ad22	1990 Census Tracts: Pre-DEMP Maps	$tr94/97figs4county90_predemp.html	4/28/95	html
ad15	af15	1990 Census Tracts: Pre-DEMP Maps: Original map, from GDT	$tr94/97106fig106.html	4/22/95	html
ad16	af16	1990 Census Tracts: Pre-DEMP Maps: Reduced (20 percent, 40 km)	$tr94/97107fig107.html	4/25/95	html
ad17	af17	1990 Census Tracts: Pre-DEMP Maps: Land-cleaned (0.5 sq km)	$tr94/97108fig108.html	4/25/95	html
ad18	af18	1990 Census Tracts: Pre-DEMP Maps: Lakes removed (40 sq km)	$tr94/97109fig109.html	4/27/95	html
ad19	af19	1990 Census Tracts: Pre-DEMP Maps: Tracts reassembled (20 percent, 40 km)	$tr94/97110fig110.html	4/28/95	html
ad20	af20	1990 Census Tracts: Pre-DEMP Maps: Simply connected	$tr94/97111fig111.html	4/28/95	html
ad21	af21	1990 Census Tracts: Pre-DEMP Maps: Triangulated	$tr94/97112fig112.html	4/28/95	html
ad22	af22old	1990 Census Tracts: Pre-DEMP Maps: Hexagons	$tr94/97113fig113.html	5/16/95	html
cgi22	af22	1990 Census Tracts: Pre-DEMP Maps: Hexagons	$cgi: Fresno etc; 1990 tracts; original map; no cases; geopolitical (ps, gif)	10/8/96	html

$tr94 = http://parep2.lbl.gov/pdocs/tr940115 = parep2.lbl.gov /data9/old/parep2/merrill/docs/parep/tr940115.
ad1 and ad13-ad22 = HTML documents included in MERR94B: Preparation of Geographic Map Files for DEMP Transformation: Figures. $tr94/97figs. html.
af15-af22 = figures in electronic format, listed in Appendix A.1: Electronic File Locations: Figures.

$cgi = DEMP Common Gateway Interface http://parep2.lbl.gov/cgi-bin/merrilltest4. The cgi parameters indicate the options to be chosen by the user, to display the item shown in the "describes:" field. All the CGI options display reduced hexagon maps.
cgi22 = the particular set of cgi parameters to obtain the figure of af22.
"describes:" = ID(s) described by this document.
GDT = Geographic Data Technology, Lebanon NH.
html = HyperText Markup Language.
ID = identifier of this document. All the documents ad1, ad13-ad22 are included in MERR94B.

Electronic Documentation:

Map files: modified 1980 Census tracts: pre-DEMP maps:

ID	describes:	title	cgi parameters	date	format
cgi23	af23	mod 1980 Census Tracts: Pre-DEMP Maps: Hexagons	$cgi: Fresno etc; hexagons in mod 1980 tracts; original map; no cases; geopolitical (ps, gif)	10/10/96	html

af23 = figure in electronic format, listed in Appendix A: Electronic Documentation: Figures.
cgi23 = the particular set of cgi parameters to obtain the figure of af23.
$cgi = DEMP Common Gateway Interface program, at http://parep2.lbl.gov/cgi-bin/merrilltest4. The cgi parameters indicate the options to be chosen by the user, to display the item shown in the "describes:" field. All the CGI options display reduced hexagon maps..
describes: = ID(s) described by this document.
html = HyperText Markup Language.
ID = identifier of this document.

Electronic Documentation:

Map files: modified 1980 Census tracts: density equalized maps: no cases:

ID	describes:	title	cgi parameters	date	format
cgi24-cgi33	af24-af33	mod 1980 Census Tracts: density equalized maps without cases: (10 separate maps with different population denominators, described under af24-a33)	$cgi; then select Fresno etc; hexagons in mod 1980 tracts; then select one of the 10 options listed under af24-af33; then select no cases; then select geopolitical (ps, gif).	10/96 through 1/97	html

af24 etc = figure in electronic format, listed in Appendix A.1: Electronic File Locations: Figures.
cgi24 etc. = the particular set of cgi parameters to obtain the figure af24.
$cgi = DEMP Common Gateway Interface program, at http://parep2.lbl.gov/cgi-bin/merrilltest4. The cgi parameters indicate the options to be chosen by the user, to display the item shown in the "describes:" field. All the CGI options display reduced hexagon maps.
describes: = ID(s) described by this document.
html = HyperText Markup Language.
ID = identifier of this document.

Electronic Documentation:

Population data files: LBNL tape library:

ID	operation	description	location	date	format
dp011a	hand edit	contents and locations of active and high use tapes from LBL tape library	$tapes/tapes.zip, subset newdean.xls	2/21/98	Excel 95 Worksheet in ZIP archive
dp011b	and edit	contents and locations of active and high use tapes from LBL tape library	$tapes/tapes.zip, subset newdean.txt	2/21/98	tab delimited text in ZIP archive

$tapes = http://parep2.lbl.gov/pdocs/tapes
dp011a etc = electronic documentation files listed in Appendix A.2: Electronic File Locations: Electronic Documentation.
Excel 95 = 1995 version of Microsoft Excel.
ZIP = compressed data archive format used by WinZip or PKZIP.

Electronic Documentation:

Population data files: SEEDIS:

ID	operation	description	location	date	format
dp01	hand edit	SEEDIS databases	$seedis/databases.html	1/14/92	hel
dp02	hand edit	SEEDIS geographic levels	$seedis/level.html	1/14/92	hel
dp03	cotools	SEEDIS data locations, single-tape databases	$level:series.cod	11/20/96	cod
dp041	cotools	SEEDIS data locations, multi-tape databases at level TRACT80PT	$level:[.series.st]tract80pt.cod	11/20/96	cod
dp042	cotools	SEEDIS data locations, multi-tape databases at level TRACT80PT	$level:[.series.st]tract80pt.cod	11/20/97	cod
dp043	cotools	SEEDIS data locations, former slot locations of high-use BCK tapes	$bck:backup.cod	5/16/91	cod

$bck = seedis.census.gov::sy$seedis:[seedis.csa3.lsttape.bck]
$level = seedis.census.gov::sy$cache:[cache.perm.dbname.ftype.level]
$seedis = http://parep2.lbl.gov/mdocs/seedis
cod: self-describing ASCII CODATA format for rectangular data files, developed at LBNL. See http://parep2.lbl.gov/mdocs/seedis.codata.html.
cotools: VMS programs for manipulating codata files. For documentation in seedis.census.gov, type "cotools".
dp01 etc = electronic documentation files listed in Appendix A.2: Electronic File Locations: Electronic Documentation.

Electronic Documentation:

Population data files: SEEDIS: 1980 Census data:

ID	operation	description	location	date	format
dp05	hand edit	SEEDIS database STF1. This analysis used TAB1 (total pop) and TAB7 (pop by race) and TAB8 (pop by ethnicity), for levels TRACT80PT and COUNTY80.	$seedict:stf1.hel	5/7/88	hel
dp06	hand edit	SEEDIS database STF2A. This analysis used TABA1 (total pop) and TABA11 (pop by race) and TABA12 (pop by ethnicity), for level PLTRACT80.	$seedict:stf2a.hel	11/11/88	hel
dp07	hand edit	SEEDIS database STF2B7R. This analysis used B1 (pop by race/ethnicity) and B8 (pop by sex by age by race/ethnicity), for level PLTRACT80.	$seedict:stf2b7r.hel	11/11/88	hel
dp08	hand edit	SEEDIS database STF2C. This analysis used TABA1 (total pop) and TAB TABA11 (pop by race) and TABA12 (pop by ethnicity), for level COUNTY80.	$seedict:stf2c.hel	2/9/91	hel
dp09	hand edit	SEEDIS database STF2B28R. This analysis used B1 (pop by race/ethnicity) and B8 (pop by sex by age by race/ethnicity), for level COUNTY80.	$seedict:stf2b28r.hel	10/31/90	hel

$seedict = seedis.census.gov::sy$seedis:[seedis.seedict].
COUNTY80 = 1980 Census counties. In California 1980 and 1990 Census counties are identical. Counties nest within states.
CYPL80 = COUNTY80/PLACE80I parts. CYPL80s nest within COUNTY80s and within PLACE80Is.
dp05 etc = electronic documentation files listed in Appendix A.2: Electronic File Locations: Electronic Documentation.
hel = format used by SEEDIS "page" program. For example, to view stf1.hel, log into seedis.census.gov and type "page stf1".
MCD80 = 1980 Census Minor Civil Divisions. MCD80's nest within COUNTY80s but may overlap TRACT80s, CYPL80s, PLACE80Is, and PLACE80s.
PLACE80 = 1980 Census places. PLACE80s nest within states and PLACE80Is, but may overlap COUNTY80s, TRACT80s, and MCD80s.
PLACE80I = 1980 Census places with population 10,000 or greater. PLACE80Is nest within states but may overlap COUNTY80s, PLACE80s, TRACT80s, and MCD80s. Each state has a PLACE80I for remainder of state, including all the places under 10,000 population.
PLTRACT80 = 1980 Census PLACE80I/TRACT80 parts. PLTRACT80s nest within PLACE80Is and TRACT80s, but may overlap MCD80s and PLACE80s.
SEEDIS = LBNL Socio-Economic Environmental Demographic Information System. See http://parep2.lbl.gov/mdocs/seedis/seedis.html.
STF = Summary Tape File. 1980 Census STF1 and STF2 are complete count data for summary tables and detailed tables respectively.
STF1 = STF1, file A (COUNTY80 level and below). No race detail. 342 variables. This document also describes STF1, file C (COUNTY80 level and above), which was not used.
STF2A = STF2, file A (CYPL80 level and below), record A (no race detail). 1098 variables.

STF2B28R = STF2, file C3 (COUNTY80 level and above), record B (28 race/ethnic groups). 28*968 = 27104 variables.
STF2B7R = STF2, file A (CYPL80 level and below), record B (seven race/ethnic groups). 7*968 = 6776 variables.
STF2C = STF2, file C3 (COUNTY80 level and above), record A (no race detail). 1364 variables.
TRACT80 = 1980 Census tracts. TRACT80s nest within COUNTY80s but may overlap MCD80s, PLACE80Is and PLACE80s.
TRACT80PT = 1980 Census MCD80/PLACE80/TRACT80 parts. TRACT80PTs nest within MCD80s, PLACE80s, PLACE80Is, and TRACT80s.

Electronic Documentation:

Population data files: 1980-88 person-years by age, sex, race, and tract:

ID	operation	description	location	date	format
dp10	hand edit	age-sex-race-specific census tract populations in 1980	$feas/pop1980.html	6/11/96	html
dp11	hand edit	Files used to estimate population in missing tracts of 4county study	$pop/pop_work.html	7/17/94	html
dp12	hand edit	age-specific (0-4,5-17,0-17) population estimates	$pop/pop.html	8/6/94	html
dp21	hand edit	Estimation of 1980 population by race and ethnicity	$pop/tract80pt.html	4/14/95	html
dp22	hand edit	Estimation of 1980 population for ages 0-14 and 5-14	$pop/age0014.html	4/14/95	html
dp23	hand edit	1980-88 population estimates for four-county study	$pop/pop8088.html	4/17/95	html
dp24	hand edit	1980-88 population estimates: detailed calculations	$pop/details.html	4/28/95	html
dp25	hand edit	1980-88 population estimates for modified 1980 census tracts	$8090/readme.dwm	10/30/96	txt

$feas = http://parep2.lbl.gov/pdocs/feas.
$pop = http://parep2.lbl.gov/~merrill/docs/parep/4county/pop.
$8090 = http://parep2.lbl.gov/~merrill/maps/tr940115/4county8090.
dp11 etc = electronic documentation about population calculation, listed in Appendix A.2: Electronic File Locations: Electronic Documentation.
ethnicity = Hispanic or non-Hispanic. In the Census, race and ethnicity are independent; Hispanics may be of any race.
html = HyperText Markup Language.
race = white, black, Native American, Asian and Pacific Islander, and other (5 groups). In the Census, race and ethnicity are independent; Hispanics may be of any race.
txt = ASCII text file.

APPENDIX A. ELECTRONIC FILE LOCATIONS (CONTINUED)

A.3. DATA FILES

Data files:

Map files: 1980 map files: tapes from NPDC: *(see copyright notice in Appendix B.1)*

from	ID	operation	description	location	date	format
NPDC	m2a	purchase	formatted polygon files for 1980 census tracts in SMSAs. NPDC tape no. LBL001. 318 files, SMSA codes 0040 through 9340.	Nor-Cal box 558129, LBL tape no. 40053. 9-track, 6250bpi, ASCII unlabeled, record length=80, block size=32000.	6/16/86	txt
NPDC	m2b	purchase	formatted polygon files for 1980 census tracts in non-SMSA tracted counties. NPDC tape no. LBL003. 219 files, FIPS counties 01009 through 56021.	Nor-Cal box 558129, LBL tape no. 40055. 9-track, 6250bpi, ASCII unlabeled, record length=80, block size=32000.	6/16/86	txt
m2a	m3a	VMS tapecopy	formatted polygon files for 1980 census tracts in SMSAs. 318 files, SMSA codes 0040 through 9340.	Nor-Cal box 558129, LBL tape no. 40060. 9-track, 6250bpi, VMS ascii tape. Files p1.index and p1.dat.*	1986	txt
m2b	m3b	VMS tapecopy	formatted polygon files for 1980 census tracts in non-SMSA tracted counties. 219 files, FIPS counties 01009 through 56021.	Nor-Cal box 558130, LBL tape no. 40061. 9-track, 6250bpi, VMS ascii tape. Files p2.index and p2.dat.*	1986	txt
m3a	m6a	DECNET-DOS; VMS backup	formatted polygon files for 1980 census tracts in SMSAs. 318 files, SMSA codes 0040 through 9340. VMS ascii format.	Nor-Cal box 558134, VMS backup tape. LBL tape no. 40134, files [cache.junk.npdc.polygons.smsatr] sm<nnnn>.dat	2/22/89	txt
m3b	m6b	DECNET-DOS; VMS backup	formatted polygon files for 1980 census tracts in non-SMSA tracted counties. 219 files, FIPS counties 01009 through 56021. VMS ascii format.	Nor-Cal box 558134, VMS backup tape. LBL tape no. 40134, files [cache.junk.npdc.polygons.othertr] s<nn>c<nnn>.dat	2/22/89	txt

<nn> = FIPS (Federal Information Processing System) state code.
<nnn> = 1980 Census county code.
<nnnn> = 1981 SMSA (Standard Metropolitan Statistical Area) code.
m2a, etc = electronic map files listed in Appendix A.3: Electronic File Locations: Data Files: Map Files.
Nor-Cal = 117*17=1989 active tapes from LBL tape library. Archived January 1997, at Nor-Cal Records Management, 10901 Bigge Street, San Leandro CA 94577. Nor-Cal contact: Julie Miller, 510-635-1944, X232. Owner contact: Deane Merrill, merrill@crocker.com, or Val Gregg, U.S. Bureau of the Census, 301-457-4102, vgregg@census.gov; or Don James, U.S. Bureau of the Census, 301-457-1758, Donald.R.James@ccmail.census.gov, project no. 20-00-70-0301-00-2590.
NPDC = National Planning Data Corporation. LBL purchase order 3800402, 6/19/86. LBL contact: Deane Merrill. NPDC contact in 1986: Bruce Harris, Los Angeles Office, (213) 657-0158. In 1992 NPDC merged with Claritas Corporation. See copyright notice in Appendix B.1.
txt = ascii text

Data files:

Map files: 1980 map files: installation in SEEDIS: *(see copyright notice in Appendix B.1)*

from	ID	operation	description	location	date	format
m6a	m7a	VMS copy	formatted polygon files for 1980 census tracts in SMSAs. 318 files, SMSAcodes 0040 through 9340. VMS ascii format.	$npdc1:[.smsatr]sm<nnnn>.dat	12/20/91	txt
m6b	m7b	VMS copy	formatted polygon files for 1980 census tracts in non-SMSA tracted counties. 219 files, FIPS counties 01009 through 56021. VMS ascii format.	$npdc1:[.othertr]s<nn>c<nnn>.dat	12/20/91	txt

$npdc1 = seedis.lbl.gov::disk$seedis004:[seedis.npdc]. Moved in 1997 to $npdc2. See copyright notice in Appendiex B.1.
$npdc2 = seedis.census.gov::disk$seedis004:[seedis.npdc]
<nn> = FIPS (Federal Information Processing System) state code.
<nnn> = 1980 Census county code.
<nnnn> = 1981 SMSA (Standard Metropolitan Statistical Area) code.
m6a, etc. = electronic map files listed in Appendix A.3: Electronic File Locations: Data Files: Map Files.
txt = ascii text

Data files:

Map files: 1980 map files: map files for future use: *(see copyright notice in Appendix B.1)*

from	ID	operation	description	location	date	format
m7a, m7b	n10	VMS zip; FTP	contents of csa.lbl.gov:: disk$seedis004	seedis.census.gov:: dka400:[seedis.tarfiles] seedis004_seedis.zip	11/22/96	zip
n10	n11a	VMS unzip	formatted polygon files for 1980 census tracts in SMSAs. 318 files, SMSA codes 0040 through 9340. VMS ascii format.	$npdc2:[.smsatr]sm<nnnn>.dat	11/22/96	txt
n10	n11b	VMS unzip	formatted polygon files for 1980 census tracts in non-SMSA tracted counties. 219 files, FIPS counties 01009 through 56021. VMS ascii format	$npdc2:[.othertr]s<nn>c<nnn>.dat	11/22/96	txt

$npdc1 = seedis.lbl.gov::disk$seedis004:[seedis.npdc]
$npdc2 = seedis.census.gov::disk$seedis004:[seedis.npdc]
<nn> = FIPS (Federal Information Processing System) state code.
<nnn> = 1980 Census county code.
<nnnn> = 1981 SMSA (Standard Metropolitan Statistical Area) code.
FTP = File Transfer Protocol.
SEEDIS = LBL Socio-Economic Environmental Demographic Information System. See http://parep2/lbl.gov/mdocs/seedis/seedis.html.
txt = ascii text
VMS = operating system of Digital VAX computers.
zip = ZIP archives (MS-DOS or VMS or UNIX)

Data files:

Map files: 1980 map files: pre-DEMP map files: *(see copyright notice in Appendix B.1)*

from	ID	operation	description	location	date	format
m7a, m7b	m8	SEEDIS	1980 census tracts from SEEDIS. level=nmcdtr80; state=California; counties=Fresno, Kern, Kings, Tulare.	(temporary) seedis.map, seedis.dime	4/30/93	map, dime
m8	m9	dime_to_edime, edime_area_calc	1980 census tracts from SEEDIS. 308 polygons, 8803 points, 9086 segments.	$Maps_Cod/ 4county_orig.fdime	4/17/93	fdime
	m11		1980 census tracts from SEEDIS. 308 polygons, 8803 points, 9086 segments.	$Maps_Cod/ 4county_308poly.fdime	4/30/93	fdime
			1980 census tracts 309 polygons, 8803points, 9086 segments	$Maps_Cod/ 4county_309poly.fdime	10/18/94	fdime
m11	m12	hand edit to fix errors	1980 census tracts. 277 polygons, 8803 points, 9054 segments	$Maps_Cod/ 4county_277poly.fdime	1/18/94	fdime
m12	m13	edime_point_reduction	1980 census tracts, reduced (20%). 277 polygons, 610 points, 861 segments	$Maps_Cod/ 4county_277poly_20pct.fdime	1/21/94	fdime
m13	m14	triangulate	1980 census tracts, triangles. 1108 triangles, 610 points, 1692 segments	$Maps_Cod/ 4county_610tri_20pct.fdime	1/21/94	fdime
	m16	cut doughnut tracts with causeways	1980 census tracts, simply connected 265 polygons, 610 points, 873 segments	$Maps_Cod/ 4county_265poly.fdime	7/11/94	fdime
m17	m18	triangulate	1980 census tracts, triangles 1121 triangles, 610 points, 1729 segments	$Maps_Cod/ 4county_1121tri.fdime	7/11/94	fdime
m18	m19	hand edit to fix errors	1980 census tracts, 1121 triangles, 610 points, 1729 segments	$Maps_Cod/ 4county_1121tri_fix1.fdime	7/17/94	fdime
m19	m20	nickel_divide	1980 census tracts, hexagons 1121 hexagons, 2336 points, 3458 segments	$Maps_Cod/ 4county_1121hex.fdime	8/7/94	fdime
			1980 census tracts, triangles 1126 hexagons, 2336 points, 3473 segments	$Maps_Cod/ 4county_from_cl3.fdime	7/17/94	fdime

$Maps_Cod = http://parep2.lbl.gov/mpub/Puff/Maps_Cod = parep2.lbl.gov/data9/old/parep2/merrill/Puff/Maps_Cod

dime = Dual Independent Map Encoding (DIME) map format. Ascii, one record per segment, specifying right geocodes, left geocodes, from-point and to-point of each segment.
edime = extended dime format. Ascii, in three sections. Section 1 defines geocodes of each polygon; section 2 defines coordinates of each point; section 3, for each segment, provides sequence numbers of right polygon, left polygon, from-point, and to-point.
edime_area_calc = VMS program to convert from edime to fdime map format.
edime_point_reduction = VMS program to remove unnecessary geographic detail from a dime map.
fdime = enhanced edime format. Ascii, in three sections. Same as edime format, with the following additional information: total map area, number of points and area of each polygon; and the number of segments associated with each point.
dime_to_edime = VMS format conversion program by Michael Mohr. In seedis.census.gov::sy$seedis:[seedis.map_routines]dime_to_edime.com.
map = binary map format produced by SEEDIS.
m8, etc = map files described in Appendix A.3: Electronic locations: Data files.
nickel_divide = VMS program to convert triangles to hexagons.
RLInt = Russian Line Integral program.
SEEDIS = LBL Socio-Economic Environmental Demographic Information System. See http://parep2.lbl.gov/mpub/seedis/seedis.html.
triangulate = routine to perform Delaunay triangulation of a polygon file.
txt = ascii text

Data files:

Map files: 1980 map files: pre-DEMP map files, continued: *(see copyright notice in Appendix B.1)*

from	ID	operation	description	location	date	format
m13	m23	edime_to_nickel, nickel_to_cl3	1980 census tracts, hexagons	$age0017/old/ 4county_277poly_20pct.cl3	1/31/94	cl3
m16	m24	edime_to_nickel, nickel_to_cl3	1980 census tracts, hexagons	$age0017/old/ 4county_265poly_20pct.cl3	2/1/94	cl3

$age0017 = $p2/4county/age0017
$p2 = $demp = $dempcedr2 = parep2.lbl.gov/work/merrilldg/Puff/Version5 (4county/(age0017, age0014, age0004, age0514))
cl3 = input format for RLInt (1/94 version)
m23, etc = map files described in Appendix A.3: Electronic locations: Data Files: Map Files.

Data files:

Map files: 1980 map files: density equalized map files:

from	ID	operation	description	location	date	format
m24	m25a	copy	1980 tracts, map, step 0	$itstp10/RLInt.out.0000	11/15/94	cl3
m25	m26a	RLInt	1980 tracts, map, step5	$itstp10/RLInt.out.0005	11/15/94	cl3
m25	m27a	RLInt	1980 tracts, map, step10	$itstp10/RLInt.out.0010	11/15/94	cl3
m25	m25b	RLInt	1980 tracts, areas, step 0	$itstp10/RLInt.sum.0000	11/15/94	cl3
m25	m26b	RLInt	1980 tracts, areas, step5	$itstp10/RLInt.sum.0005	11/15/94	cl3
m25	m27b	RLInt	1980 tracts, areas, step10	$itstp10/RLInt.sum.0010	11/15/94	cl3

$age0017 = $p2/4county/age0017
$itstp10 = $p2/4county/age0017/itstp10 (11/94)
$p2 = $demp = $dempcedr2 = parep2.lbl.gov/work/merrilldg/Puff/Version5 (4county/(age0017, age0014, age0004, age0514))
cl3 = input format for RLInt (1/94 version)
m25a, etc = map files described in Appendix A.3: Electronic locations: Data files.
RLInt = Russian Line Integral program.

Data files:

Map files: 1990 map files: *(see copyright notice in Appendix B.2)*

from	ID	operation	description	location	date	for mat
GDT	m30	purchase	MS-DOS installation file	$gdt\4county\disk1of1\install.bat	4/17/95	exe
GDT	m31	purchase	1990 Census tract boundaries for four California counties: Bakersfield, Fresno, Kern, Kings	$gdt\4county\disk1of1\Gdt.exe	4/17/95	exe
m30, m31	m32	m30	1990 Census tract boundaries for four California counties: Bakersfield, Fresno, Kern, Kings: DIME Boundary file	$gdt\Trt210\Dime\Td3<nnnnn>.t21	7/23/93	t21
m30, m31	m33	m30	1990 Census tract boundaries for four California counties: Bakersfield, Fresno, Kern, Kings: Tract Inventory file	$gdt\Trt210\Inv\Ti0<nnnnn>.t21	7/19/93	t21
	m34	m30	1990 State of California Census Tract Inventory File	$gdt\Trt210\Inv\Ti006xxx.t21	5/24/93	t21
m33	m35	hand edit	area and centroid lat/log; of 5858 segments in GDT DIME file	$G:gdtinv.cod	4/18/95	cod
m35	m36	coroagg	area and centroid lat/long; of 310 TRACT90_4s	$G:gdt4county.cod	4/18/95	cod

$gdt = csr6.lbl.gov::c:\gdt
$G = seedis.census.gov::dka300:[users.merrill.fromlbl.merrill.4county.tract90]
cod = LBNL Codata format. See http://parep2.lbl.gov/mdocs/seedis/codata.html.
coroagg = Codata tool for aggregating Codata files according to geocode values.
exe = self-extracting ZIP archive
FIPS = Federal Information Processing System
GDT = Geographic Data Technology, 11 Lafayette Street, Lebanon NH 03766-1445. Phone 1-800-331-7881. Fax 1-603-643-6868. Lyme NH. Files purchased 4/6/95 by Deane Merrill. See Appendix B.2 for License Agreement.
<nnnnn> = state/county FIPS code: 06019, 06029, 06031, 06107
t21 = GDT DIME format. Each record contains (state, county, tract code) of left polygon; (state, county, tract code) of right polygon; lat/long of start point; lat/long of end point Documentation in binder "SEEDIS - GEOG - MAP FILES - GDT."

Data files:

Population data files: 1980 Census: SEEDIS ddx and ddf files:

ID	oper ation	description	location	date	format
p021	cpr	STF1: 342 variables. This analysis used TAB1 (total pop) and TAB7 (pop by race) and TAB9 (Hispanic pop by race) and TAB12 (pop by age and race) and TAB13 (Hisp pop by age and race) at levels TRACT80PT and COUNTY80.	$seedata:[.census80.stf1]ddf.ddx	5/23/90	ddx, ddf
p022	cpr	STF2C: rec A, 1364 variables. This analysis used TABA1 (total pop) and TAB TABA11 (pop by race) and TABA12 (pop by ethnicity), at level COUNTY80.	$seedata:[.stf2]stf2c.edx	11/8/84	ddx, ddf
p023	cpr	STF2B28R: rec B, 28 races, 28*968 = 27104 variables. This analysis used B1 (pop by race/ethnicity) and B10 (pop by sex by age by race/ethnicity), at level COUNTY80.	$seedata:[.stf2]stf2b28r.edx	11/7/84	ddx, ddf
p024	cpr	STF2A: rec A, 1098 variables. This analysis used TABA1 (total pop) and TABA11 (pop by race) and TABA12 (pop by ethnicity), at level PLTRACT80.	$seedata:[.stf2]stf2a.edx	2/10/91	ddx, ddf
p025	cpr	STF2B7R: rec B, 7 races, 7*968 = 6776 variables. This analysis used B1 (pop by race/ethnicity) and B10 (pop by sex by age by race/ethnicity), at level PLTRACT80.	$seedata:[.stf2]stf2b7r.edx	11/7/84	ddx, ddf

$seedata = seedis.census.gov::disk$seedis004:[seedis.seedata]
cod: self-describing ASCII CODATA format for rectangular data files, developed at LBNL. See http://parep2.lbl.gov/mdocs/seedis.codata.html.
COUNTY80 = 1980 Census counties. In California 1980 and 1990 Census counties are identical. Counties nest within states.
CYPL80 = COUNTY80/PLACE80I parts. CYPL80s nest within COUNTY80s and within PLACE80Is.
cpr = VAX program for creating data files in SEEDIS compressed format. Used in 1980-85 to compress the original Census Summary Tape Files.
ddf = SEEDIS data definition file. An extended version of the cod format.
ddx = index to the SEEDIS data definition file.
PLACE80 = 1980 Census places. PLACE80s nest within states and PLACE80Is, but may overlap COUNTY80s.

PLACE80I = 1980 Census places with population 10,000 or greater. PLACE80Is nest within states but may overlap COUNTY80s. Each state has a PLACE80I for remainder of state, including all the places under 10,000 population.

rec: record A has no racial/ethnic breakdown. Record B has is available for 7 race/ethnic groups in file A and 28 race/ethnic groups in file B.

STF = Summary Tape File. 1980 Census STF1 and STF2 are complete count data for summary tables and detailed tables respectively.

STF1 = STF1, file A (COUNTY80 level and below). No race detail. 342 variables. This document also describes STF1, file C (COUNTY80 level and above), which was not used in this analysis.

STF2A = STF2, file A (CYPL80 level and below), record A (no race detail). 1098 variables.

STF2B28R = STF2, file C3 (COUNTY80 level and above), record B (28 race/ethnic groups). 28*968 = 27104 variables.

STF2B7R = STF2, file A (CYPL80 level and below), record B (seven race/ethnic groups). 7*968 = 6776 variables.

STF2C = STF2, file C3 (COUNTY80 level and above), record A (no race detail). 1364 variables.

Data files:

Population data files: 1980 Census: SEEDIS ndx and dat files: COUNTY80 level:

ID	operation	description	location	date	format
p020	cpr	STF1 (file A) at COUNTY80 level, series 6	$stf1:[.county80]s06.ndx, s06.dat	2/19/82	ndx, dat
p021a	cpr	STF2C (file C3, rec A) at COUNTY80 level, series 1	Nor-Cal box 558053, GSS tape 05780 (ASTF2CCY1) (slot I15)	6/21/85	ndx, dat
p021b	cpr	STF2C (file C3, rec A) at COUNTY80 level, series 2	Nor-Cal box 558063, GSS tape 11246	6/21/85	ndx, dat
p021c	cpr	STF2C (file C3, rec A) at COUNTY80 level for California, series 4	UCDATA box 558170, VMS backup tape 50080 (slot J22)	5/5/84	ndx, dat
p021d	cpr	STF2C (file C3, rec A) at COUNTY80 level for California, series 5	Nor-Cal box 558134, VMS backup tape 40138	5/5/84	ndx, dat
p023a	cpr	STF2B28R (file C3, rec B) at COUNTY80 level, series 1	Nor-Cal box 558053, GSS tape 05780 (ASTF2CCY1) (slot I15)	6/21/85	ndx, dat
p023b	cpr	STF2B28R (file C3, rec B) at COUNTY80 level, series 2	Nor-Cal box 558063, GSS tape 11246	6/21/85	ndx, dat
p023c	cpr	STF2B28R (file C3, rec B) at COUNTY80 level for California, series 4	UCDATA box 558170, VMS backup tape 50080 (slot J22)	5/5/84	ndx, dat
p023d	cpr	STF2B28R (file C3, rec B) at COUNTY80 level for California, series 5	Nor-Cal box 558134, VMS backup tape 40138	5/5/84	ndx, dat

$stf1 = seedis.census.gov::disk$seedis004:[seedis.seedata.census80.stf1]
COUNTY80 = 1980 Census counties.
cpr = VAX program for creating data files in SEEDIS compressed format. Used in 1980-85 to compress original Census Summary Tape Files.
dat = SEEDIS file in compressed binary format. See http://parep2.lbl.gov/mdocs/seedis/compressed/compressed.dat. A stand-alone program SEED2TXT, for converting the contents of a dat file to ASCII text, is at http://venus.census.gov/seedis/software/seed2txt/readme.htm.
file A = detailed Census geography at the county level and lower.
file C3 = summary Census geography at the county level and higher.
GSS = former LBNL gettape/stotape system.

ndx = ascii index file in CODATA format, pointing to location of SEEDIS dat file, and providing block location of each geographic record. See http://parep2.lbl.gov/mdocs/seedis/compressed/compressed.dat.

Nor-Cal = 117 boxes of tapes (17 tapes per box) archived January 1997, and 159 boxes of tapes (8 tapes per box) archived May 1997, at Nor-Cal Records Management, 10901 Bigge Street, San Leandro CA 94577. Nor-Cal contact: Julie Miller, 510-635-1944, X232. Owner contact: Deane Merrill, merrill@crocker.com, or Val Gregg, U.S. Bureau of the Census, 301-457-4102, vgregg@census.gov; or Don James, U.S. Bureau of the Census, 301-457-1758, Donald.R.James@ccmail.census.gov, project no. 20-00-70-0301-00-2590.

series: series 1 and 2 are duplicate copies of tapes in GSS format. Series 4 and 5 are duplicate copies of tapes in VMS backup format. Series 6 is on VMS disk.

rec: record A has no racial/ethnic breakdown; record B has is broken down for 7 race/ethnic groups in file A and 28 race/ethnic groups in file B.

STF: Summary Tape File. STF1 and STF2 are complete count data for summary tables and detailed tables respectively.

UCDATA = 12 boxes of tapes (17 tapes per box) archived January 1997 at UCDATA, University of California. Owner contact: Fred Gey, gey@ucdata.berkeley.edu.

VMS = operating system of Digital VAX computers.

Data files:

Population data files: 1980 Census: SEEDIS ndx and dat files: TRACT80PT and PLTRACT80 level:

ID	operation	description	location	date	format
p012a	cpr	STF1 (file A) at TRACT80PT level for California, series 1	Nor-Cal box 558124, GSS tape 38831 (ASTF1MCDPL) (slot I64)	10/5/84	ndx, dat
p012b	cpr	STF1 (file A) at TRACT80PT level for California, series 2	not available	NA	ndx, dat
p012c	cpr	STF1 (file A) at TRACT80PT level, series 4	UCDATA box 558141, VMS backup tape 50040 (slot G45)	3/18/89	ndx, dat
p012d	cpr	STF1 (file A) at TRACT80PT level, series 5	Nor-Cal box 558133, VMS backup tape 40119	3/18/89	ndx, dat
p025a	cpr	STF2A (file A, rec A) at PLTRACT80 level for California, series 1	Nor-Cal box 558056, GSS tape 09367 (ASTF2ATR1) (slot I19)	7/5/84	ndx, dat
p025b	cpr	STF2A (file A, rec A) at PLTRACT80 level for California, series 2	Nor-Cal box 558061, GSS tape 10673 (BSTF2ATR1)	7/5/84	ndx, dat
p027a	cpr	STF2B7R (file A, rec B) at PLTRACT80 level for California, series 1	Nor-Cal box 558056, GSS tape 09367	7/5/84	ndx, dat
p027b	cpr	STF2B7R (file A, rec B) at PLTRACT80 level for California, series 2	Nor-Cal box 558061, GSS tape 10673	7/5/84	ndx, dat

COUNTY80 = 1980 Census counties. In California 1980 and 1990 Census counties are identical. Counties nest within states.
cpr = VAX program for creating data files in SEEDIS compressed format. Used in 1980-85 to compress original Census Summary Tape Files.
dat = SEEDIS file in compressed binary format. See http://parep2.lbl.gov/mdocs/seedis/compressed/compressed.dat.
file A = detailed Census geography including tracts and tract parts.
GSS = former LBNL gettape/stotape system.
MCD80 = 1980 Census Minor Civil Divisions. MCD80's nest within counties but may overlap TRACT80s, PLACE80Is, and PLACE80s.
ndx = ascii index file in CODATA format, pointing to location of SEEDIS dat file, and providing block location of each geographic record. See http://parep2.lbl.gov/mdocs/seedis/compressed/compressed.dat.
Nor-Cal = 117 boxes of tapes (17 tapes per box) archived January 1997, and 159 boxes of tapes (8 tapes per box) archived May 1997, at Nor-Cal Records Management, 10901 Bigge Street, San Leandro CA 94577. Nor-Cal contact: Julie Miller, 510-635-1944, X232. Owner contact: Deane Merrill,

merrill@crocker.com, or Val Gregg, U.S. Bureau of the Census, 301-457-4102, vgregg@census.gov; or Don James, U.S. Bureau of the Census, 301-457-1758, Donald.R.James@ccmail.census.gov, project no. 20-00-70-0301-00-2590.

PLACE80 = 1980 Census places. PLACE80s nest within states and PLACE80Is, but may overlap COUNTY80s, TRACT80s, and MCD80s.

PLACE80I = 1980 Census places with population 10,000 or greater. PLACE80Is nest within states but may overlap COUNTY80s, PLACE80s, TRACT80s, and MCD80s. Each state has a PLACE80I for remainder of state, including all the places under 10,000 population.

PLTRACT80 = 1980 Census PLACE80I/TRACT80 parts. PLTRACT80s nest within PLACE80Is and TRACT80s, but may overlap MCD80s and PLACE80s.

rec: record A has no racial/ethnic breakdown; record B has is broken down for 7 race/ethnic groups in file A and 28 race/ethnic groups in file B.

series: series 1 and 2 are duplicate copies of tapes in GSS format. Series 4 and 5 are duplicate copies of tapes in VMS backup format.

STF: Summary Tape File. STF1 and STF2 are complete count data for summary tables and detailed tables respectively.

TRACT80 = 1980 Census tracts. TRACT80s nest within COUNTY80s but may overlap MCD80s, PLACE80Is and PLACE80s.

TRACT80PT = 1980 Census MCD80/PLACE80/TRACT80 parts. TRACT80PTs nest within MCD80s, PLACE80s, PLACE80Is, and TRACT80s.

UCDATA = 12 boxes of tapes (17 tapes per box) archived January 1997 at UCDATA, University of California. Owner contact: Fred Gey, gey@ucdata.berkeley.edu.

Data files:

Population data files: 1980 Census: derived population estimates:

from	ID	operation	description	location	date	format
p012, p025	p041	dp12	1980 pop, age 0-4 and 5-17, for 486 PLTRACT80As	$pop/pltract80a.cod.html	8/6/94	cod
p041	p042	dp12	1980 pop, age 0-4 and 5-17, for 393 PLTRACT80Is	$pop/pltract80i.cod.html	8/6/94	cod
p042	p043	dp12	1980 pop, ages 0-4 and 5-17, for 262 TRACT80s	$pop/tract80.cod.html	8/6/94	cod
p043	p044	dp12	1980 pop, ages 0-4 and 5-17, for 262 NMCDTR80s	$pop/nmcdtr80.cod.html	8/6/94	cod
p012	p051	dp21	1980 pop by race and ethnicity, for 486 TRACT80PTs	$pop/tract80pt.cod	4/14/95	cod
p012, p025, p027, p021, p023	p052	dp22	1980 pop by race and ethnicity, males and females, ages 0-4 and 5-14, for 393 PLTRACT80s			cod
p012	p053	dp24	1980 pop by race (total,white) by ethnicity (Hisp, non-Hisp) by age (0-4,5-17) for 486 TRACT80PTs	tract80pt.A = $A:[.tract80pt]codata.dat	4/9/95	cod
p020	p054	dp24	1980 pop by race (total,white) by ethnicity (Hisp, non-Hisp) by age (0-4,5-17) for 4COUNTY80s	county80.A = $A:[.county80]codata.dat	10/14/93	cod
p027	p055	dp24	1980 pop by ethnicity (Hisp, non-Hisp) by age (0-4,5-17,5-14) for 393 PLTRACT80s	pltract80.A = $A:[.pltract80]codata.dat	4/9/95	cod
p027	p056	dp24	1980 pop by race/ethnicty (total, white, Hisp) by age (0,1,...17) for 262 TRACT80s	tract80.A = $A:[.tract80]codata.dat	4/19/93	cod
	p057		land area and centroid lat/long, for 262 NMCDTR80s	nmcdtr80.A = $A:[.nmcdtr80]codata.dat	4/17/93	cod

$A = seedis.census.gov::dka300:[users.seedtest.fromlbl.seedtest.merrill.4county]
$pop = http://parep2.lbl.gov/~merrill/docs/parep/4county/pop/pltract80a.cod.html.
cod = LBNL Codata format. See http://parep2.lbl.gov/mdocs/seedis/codata.html.

dp12 etc = electronic documentation files listed in Appendix A.2.
p012 etc = population data files listed in Appendix A.3.
MCD80 = 1980 Census Minor Civil Divisions. MCD80's nest within counties but may overlap TRACT80s, PLACE80Is, and PLACE80s.
NPDC = National Planning Data Corporation.
NMCDTR80 = Geographic units defined by NPDC. Either MCD80s in rural counties; or TRACT80s in urban counties. The "tracts" in the 1980 NPDC map files are NMCDTR80s. In the four-county area, all the NMCDTR80s are synonymous with TRACT80s.
PLACE80 = 1980 Census places. PLACE80s nest within states and PLACE80Is, but may overlap COUNTY80s, TRACT80s, and MCD80s.
PLACE80I = 1980 Census places with population 10,000 or greater. PLACE80Is nest within states but may overlap COUNTY80s, PLACE80s, TRACT80s, and MCD80s. Each state has a PLACE80I for remainder of state, including all the places under 10,000 population.
PLTRACT80A = 1980 Census PLACE80/TRACT80 parts. PLTRACT80As nest within TRACT80s and PLACE80s. In the four-county study area, PLTRACT80As are synonymous with TRACT80PTs; in other words, PLTRACT80s nest within MCD80s. This is not true for the U.S. in general.
PLTRACT80 = 1980 Census PLACE80I/TRACT80 parts. PLTRACT80s nest within TRACT80s and PLACE80Is, but may overlap PLACE80s.
TRACT80 = 1980 Census tracts. TRACT80s nest within COUNTY80s, but may overlap MCD80s and PLACE80s.
TRACT80_4 = four-digit 1980 Census tracts. Aggregates of TRACT80s, ignoring the two-digit tract suffix. In the four-county study area, 4-digit 1980 and 1990 census tracts are identical.
TRACT80PT = 1980 Census MCD80/PLACE80/TRACT80 parts. TRACT80PTs nest within MCD80s, PLACE80s, PLACE80Is, and TRACT80s.

Data files:

Population data files: 1980 Census: derived population estimates (continued):

from	ID	operation	description	location	date	format
p012	p061	dp24	1980 pop by race (total,white) by ethnicity (Hisp, non-Hisp) by age (0-4, 5-17) for 486 TRACT80PTs	tract80pt.B = $B:tract80pt.cod	4/13/95	cod
p012, p020	p062	dp24	1980 pop by age (0-4, 5-17) for 486 PLTRACT80As, with no missing data	pltract80a.B = $B:pltract80a.cod	7/16/94	cod
p062	p063	dp24	1980 pop by age (0-4, 5-17) for 393 PLTRACT80s, with no missing data	pltract80i.B = $B:pltract80i.cod	7/16/94	cod
p057, p063	p064	dp24	1980 pop by age (0-4, 5-17) with no missing data; and land area, centroid lat/long; for 262 TRACT80s	tract80.B = $B:tract80.cod	4/18/95	cod
p057, p064	p065	dp24	1980 pop by age (0-4, 5-17) with no missing data; and land area, centroid lat/long; for 262 NMCDTR80s	nmcdtr80.B = $B:nmcdtr80.cod	4/18/95	cod
p023	p071	dp24	1980 Hisp pop by age (0-4, 5-17, 15-17); and Hisp ratio (15-17)/(5-17); for 4 COUNTY80s	county80.C = $C:county80.cod	4/9/95	cod
p071	p072	dp24	1980 Hisp pop by age (0-4,5-17,5-14) for 4 COUNTY80s	county80.D = $D:small.cod	4/9/95	cod
	p073	dp24	1980 pop by age (0-4,5-14), land area, observed cases, and expected cases; for 262 TRACT80s	tract80.E1 = $E1:tract80.cod	4/18/95	cod
p073	p074	dp24	1980 pop by age (0-4,5-14), land area, observed cases, and expected cases; for 210 TRACT80_4s	tract80_4.E1 = $E1:tract80_4.cod	4/18/95	cod
	p075	dp24	1980 pop age 0-17, observed and expected cases, and number of random cases from nullr1.dat; for 210 TRACT80_4s	tract80_4.E2 = $E2:tract80_4.cod	4/8/95	cod

$B = seedis.census.gov::dka300:[users.seedtest.fromlbl.seedtest.merrill.4county]
$C = $D = seedis.census.gov::dka300:[users.seedtest.fromlbl.seedtest.merrill.4county.county80.age0017]
$E1 = private location. Files are locked to protect confidentiality of case data.
$E2 = private location. Files are locked to protect confidentiality of case data.
cod = LBNL Codata format. See http://parep2.lbl.gov/mdocs/seedis/codata.html.
dp24 etc = electronic documentation files listed in Appendix A.2.

p061 etc = population data files listed in Appendix A.3.

MCD80 = 1980 Census Minor Civil Divisions. MCD80's nest within counties but may overlap TRACT80s, PLACE80Is, and PLACE80s.

NPDC = National Planning Data Corporation.

NMCDTR80 = Geographic units defined by NPDC. Either MCD80s in rural counties; or TRACT80s in urban counties. The "tracts" in the 1980 NPDC map files are NMCDTR80s. In the four-county area, all the NMCDTR80s are synonymous with TRACT80s.

PLACE80 = 1980 Census places. PLACE80s nest within states and PLACE80Is, but may overlap COUNTY80s, TRACT80s, and MCD80s.

PLACE80I = 1980 Census places with population 10,000 or greater. PLACE80Is nest within states but may overlap COUNTY80s, PLACE80s, TRACT80s, and MCD80s. Each state has a PLACE80I for remainder of state, including all the places under 10,000 population.

PLTRACT80A = 1980 Census PLACE80/TRACT80 parts. PLTRACT80As nest within TRACT80s and PLACE80s. In the four-county study area, PLTRACT80As are synonymous with TRACT80PTs; in other words, PLTRACT80s nest within MCD80s. This is not true for the U.S. in general.

PLTRACT80 = 1980 Census PLACE80I/TRACT80 parts. PLTRACT80s nest within TRACT80s and PLACE80Is, but may overlap PLACE80s.

TRACT80 = 1980 Census tracts. TRACT80s nest within COUNTY80s, but may overlap MCD80s and PLACE80s.

TRACT80_4 = four-digit 1980 Census tracts. Aggregates of TRACT80s, ignoring the two-digit tract suffix. In the four-county study area, 4-digit 1980 and 1990 census tracts are identical.

TRACT80PT = 1980 Census MCD80/PLACE80/TRACT80 parts. TRACT80PTs nest within MCD80s, PLACE80s, PLACE80Is, and TRACT80s.

Data files:

Population data files: 1990 Census: Summary Tape File 1A:

from	ID	operation	description	location	date	format
census	p051	purchase	STF1 (file A) at TRACT90 level for California	CD-ROM diskette CD90-1A-9-2	August 1991	dbf

census = U.S. Bureau of the Census.
COUNTY90 = 1990 Census counties. In California 1980 and 1990 Census counties are identical. Counties nest within states.
dbf = dBase III (copyright Borland, previously Ashton-Tate).
MCD90 = 1990 Census Minor Civil Divisions. MCD90s nest within COUNTY90s but not within PLACE90s or TRACT90s.
p051 etc = population data files listed in Appendix A.3.
PLACE90 = 1990 Census places. PLACE90s nest within states but not within MCD90s or TRACT90s.
TRACT90 = 1990 Census tracts. TRACT90s nest within COUNTY90s.
TRACT90PT = 1990 Census MCD90/PLACE90/TRACT90 parts. TRACT90PTs nest within MCD90s, PLACE890s, and TRACT90s.

Data files:

Population data files: 1990 Census: derived population estimates:

from	ID	operation	description	location	date	format
p051	p081	dbutil	1990 population by age, and land area and water area; for 310 TRACT90s	tract90.F = $F/age0017_4county.out	4/8/95	cod
p081	p082	dp24	1990 pop by age (0-4,5-17); land area, water area and total area; for 310 TRACT90s	tract90.G = $G:tract90.cod	4/28/95	cod
p082	p083	dp24	1990 pop by age (0-4,5-14,5-17); land area, water area and total area; for 213 TRACT90_4s	tract90_4.G = $G:tract90_4.cod	4/18/95	cod
p082	p084	dp24	1990 pop by age (0-4, 5-14, 5-17); land area, water area, total area, and centroid lat/long; for 306 MCDTRACT90Xs	mcdtract90x.G = $G:mcdtract90x.cod	4/28/95	cod

$F = http://parep2.lbl.gov/mpub/census90.
$G = seedis.census.gov::dka300:[users.merrill.fromlbl.merrill.4county.tract90]
cod = LBNL Codata format. See http://parep2.lbl.gov/mdocs/seedis/codata.html.
dBase = copyright Borland, previously Ashton-Tate.
dbutil = software routines for extracting data from 1990 STFs in dBase format, by Nathan Parker. See http://parep2.lbl.gov/mpub/cdrom/doc/dbutil.html.
GDT = Geographic Data Technology, Inc. See copyright notice in Appendix B.2.
MCDTRACT90X = the 1990 analog of NMCDTR80. TRACT90s with 2-digit suffix equal to 99 are included with their parent TRACT90, and a dummy MCD90 code has been added.
TRACT90 = 1990 Census tracts. TRACT90s nest within COUNTY90s.
TRACT90_4 = four-digit 1990 Census tracts. Aggregates of TRACT90s, ignoring the two-digit tract suffix. In the four-county study area, 4-digit 1980 and 1990 census tracts are identical.

Data files:

Population data files: 1980-88 person-year estimates:

from	ID	operation	description	location	date	format
p102	p101	splus	person-years for (259) modified 1980 Census tracts. Col (2,5,6,8) = person-years, state, county, tract.	$8090/demp_hex_total.geopop	10/21/96	txt
	p102	dp25	person-years for modified 1980 Census tracts	$8090/py.cod	10/21/96	cod

$8090 = http://parep2.lbl.gov/~merrill/maps/tr940115/4county8090.
cod = LBNL Codata format. See http://parep2.lbl.gov/mdocs/seedis/codata.html.
dp25, etc. = documentation files in Appendix A.2.
p101, etc = population data files in Appendix A.3.

Data files:

Case data files: source data from California DHS: *(see non-disclosure notice in Appendix B.4)*

ID	operation	description	location	date	format
c0	hand edit	documentation for c1-c3	$cases/ readme.dwm	10/30/96	txt
c1		data for 401 cases, excluding cancer site codes	$cases/ fourclbl.dat	4/7/93	txt
c2		cancer site codes of 401 cases	$cases/vonbehren.dat	1/31/95	txt

$cases = private location. These files are locked to protect confidentiality of the study stubjects.
txt = DOS ASCII text.

Data files:

Case data files: derived data: *(see non-disclosure notice in Appendix B.4)*

from	ID	operation	description	location	date	format
c1	c1.3	hand edit	coordinates of 401 cases, in lat-long coordinates	$cases2/ 4county_caseloc.ll	7/14/94	
c1.3	c1.5	Splus	coordinates of 401 cases, in km	$cases2/ 4county_caseloc. km	7/14/94	txt
c1,c2	c3	hand edit	combined data for 401 cases	$cases/ cases.dat	10/12/96	txt

$cases2 = private location. These files are locked to protect confidentiality of the study stubjects.
txt = DOS ASCII text

Data files:

Case data files: 1980 Census tracts: *(see non-disclosure notice in Appendix B.4)*

from	ID	operation	description	location	date	format
selvin	c5	hand edit with splus	actual, random, and expected cases for 1980 Census tract (based on 1980 pop, age 0-17)	$S1/tract80.df	1/18/95	txt
c7, p101	c6	hand edit with splus	actual, random and expected cases for (259) modified Census tracts (based on 1980-88 PY, age 0-14)	$S1/tract8090_total.df	3/27/98	txt
	c7	hand edit with cotools	actual, random and expected cases by modified 1980 Census tract	$S2/four_total_casegeo	10/30/96	txt

$S1 = private location. Files locked to protect confidentiality of study subjects.
$S2 = private location. Files locked to protect confidentiality of study subjects.
c5, etc = case data files listed in Appendex A.3.
p101, etc = population data files listed in Appendix A.3.
cotools = Codata tools used in SEEDIS. See http://parep2.lbl.gov/mdocs/seedis/codata.html.
txt = DOS ASCII text.

A.4. SCRIPTS AND PROGRAM FILES

Scripts and program files:

scripts for plotting maps:

inputs	ID	outputs	description	location	date	format
s101, m11	s02	af4old	plot 4-county SEEDIS map	$figs/fig2.csh	7/6/94	csh
s101, m12	s03	af6old	plot: 4-county map with errors removed	$figs/fig3.csh	7/6/94	csh
	s04	af7	plot: reduction of map complexity	$figs/fig4.csh	7/6/94	csh
s101, m13	s05	af8	plot: reduced 4-county map (20 percent)	$figs/fig5.csh	7/6/94	csh
s105, m25a	s31	a31old	plot: 4-county map, age 0-17, iteration 0	$figs/fig31.csh	12/22/94	csh
s105, --	s41	af41	plot: 4-county map, age 0-17, iteration 1 of 1	$figs/fig41.csh	7/18/94	csh
s105, m27a	s42	af42old	plot: 4-county map, age 0-17, iteration 10 of 10	$figs/fig42.csh	7/19/94	csh
s105, --	s43	af43	plot: 4-county map, age 0-17, iteration 1 of 10	$figs/fig43.csh	7/18/94	csh
s105, --	s44	af44	plot: 4-county map, age 0-17, iteration 9 of 10	$figs/fig44.csh	7/18/94	csh
s105, m26a	s45	af45	plot: 4-county map, age 0-17, iteration 5 of 10	$figs/fig45.csh	7/18/94	csh
s109, m27b	s50	af50old	plot: present vs target areas, iteration 10 of 10	$figs/fig50.csh	7/19/94	csh

$figs = http://parep2.lbl.gov/pdocs/tr940115/figs = parep2.lbl.gov/data9/old/parep2/merrill/docs/parep/tr940115/figs
af4old, etc = figures in Appendix A.1.
csh = UNIX C shell script.
m11, etc = map files in Appendix A.3.
s101 etc = scripts and programs in Appendix A.4.

Scripts and program files:

subroutines for plotting maps:

inputs	ID	description	location	date	format
s102old	s101old	plot a map in EDIME format (old)	$plot/graph.com	1/17/94	csh
s102	s101	plot a map in EDIME format	$figs/graph.com	8/23/96	csh
s106	s105	plot a DEMP map	$figs/figxx/csh	7/19/94	csh
s107,s108	s106	plot a DEMP map	$figs/dempmap.csh	7/20/94	csh
	s107	post-process a DEMP map	$figs/postdemp.com	11/15/94	com
s110	s109	plot present vs target areas	$figs/areaxx.csh	7/19/94	csh
s111,s108	s110	plot present vs target areas	$figs/areas.csh	7/19/94	csh
s103	s102old	plot a map in EDIME format (old)	$plot/graph.S	7/15/93	splus
	s102	plot a map in EDIME format	$figs/graph.S	12/28/94	splus
	s111	plot present vs target areas	$figs/areas.s	12/22/94	splus
		plot reduction of map complexity	$figs/reduce.s	7/6/94	splus
	s108	plot a postscript file	$figs/plotps.csh	7/16/94	csh
	s103	Splus version 1.69 (old)	$splus1/Splus	3/17/95	sh
		S-Plus version 1.75 v.4.3.1 u1	$splus2/Splus	3/4/98	sh

$figs = http://parep2.lbl.gov/pdocs/tr940115/figs = parep2.lbl.gov/data9/old/parep2/merrill/docs/parep/tr940115/figs
$plot = http://parep2.lbl.gov/mpub/Plot.routines = parep2.lbl.gov/data9/old/parep2/merrill/Plot.routines
$splus1 = birks.lbl.gov:/cedrvol/splus
$splus2 = birks.lbl.gov:/usr/splus/bin
af4old, etc = figures in Appendix A.1.
csh = UNIX C shell script.
com = DEC VAX command file
s101 etc = scripts and programs in Appendix A.4.
sh = UNIX shell script.
splus = Splus script.
Splus = S-Plus, copyright MathSoft, Seattle WA.

Scripts and program files:

other programs:

ID	outputs	description	location	date	format
s201	af3a	Poisson based test, 1980 Census tracts, real cases	$spatial/sum.s	5/4/98	splus
s201	af3b	Poisson based test, 1980 Census tracts, random cases	$spatial/sum.s	5/4/98	splus
s202	af22a	Poisson based test, modified 1980 Census tracts, real cases	$spatial/sum8090.s	5/20/98	splus
s202	af22b	Poisson based test, modified 1980 Census tracts, random cases	$spatial/sum8090.s	5/20/98	splus

$spatial = parep2.lbl.gov/~merrill/selvin/parep/spatial
af3a, etc = figures in Appendix A.1
s201, etc = programs in Appendex A.4
splus = Splus script for SunOS
Splus = copyright, MathSoft, Seattle WA

Scripts and program files:

other programs, continued:

ID	outputs	description	location	date	format
s203	af58	T = fraction of log RR in tail, two real and 20 uniform samples	$splus4\uniform.ssc	5/19/98	splus4
s204	af59	T = fraction of log RR in tail, two real and 20 random samples	$splus4\random.ssc	5/19/98	splus4

$splus4 = csr6.lbl.gov: c:\Program Files\splus4\users\merrill
af58a, etc = figures in Appendix A.1
s203, etc = programs in Appendex A.4
splus4 = Splus script for Windows95
Splus = copyright, MathSoft, Seattle WA

APPENDIX B. COPYRIGHT AND NON-DISCLOSURE NOTICES

B.1. National Planning Data Corporation: 1980 Census tract map files

1980 Census tract map files, including those used in this report, were licensed in 1986 from National Planning Data Corporation. Their use is governed by the following copyright notice, signed by Deane Merrill at the time of purchase. The relevant project at LBNL having terminated, the authorized LBNL copy of these files has been moved to a computer at the U.S. Bureau of the Census, for use exclusively by Deane Merrill.

> This file is a proprietary asset of National Planning Data Corporation, with corporate headquarters at P.O. Box 610, Ithaca NY 14851-0610. 607-273-8208. This file is available only for use by agencies or individuals who have signed a license agreement with National Planning Data Corporation. This material is an unpublished work under the copyright act of the United States. Certain ideas and concepts also contained in this material are trade secrets of National Planning Data Corporation. Unauthorized copying or other disclosure of this material will make you liable for substantial penalties.

In 1992, National Planning Data Corporation merged with its sister company Claritas to form Claritas, Inc.

B.2. Geographic Data Technology: 1990 Census tract map files

The 1990 Census tract map files used in this report were licensed on 4/6/95, from: Geographic Data Technology, 11 Lafayette Street, Lebanon, NH 03766-1445, phone 1-800-331-7881, fax 1-603-643-6808; to Deane Merrill, Lawrence Berkeley Laboratory. The following license agreement was signed by Deane Merrill. The relevant project at LBNL having terminated, the authorized LBNL copy of these files has been moved to a computer at the U.S. Bureau of the Census, for use exclusively by Deane Merrill.

Grant of License:

GDT hereby grants to the above named customer ("Licensee") and Licensee hereby accepts, a non-exclusive license to use, for its own internal purposes the Product(s) identified above. Licensee is granted a license to use the Product(s) provided, or any derivation thereof, solely for the internal purposes of the Licensee at the site identified above ("Lawrence Berkeley Laboratory") by a number of users equal to the total number of units of the Product(s) ordered, as indicated in the quantity column above ("1"). No part of the Product(s), or any derivation thereof, shall be disclosed to third parties or used for the benefit of third parties, other than authorized agents of the Licensee. In no event shall Product(s) be used for the benefit of third parties, without GDT's express written permission. Any Product(s) generated from the Product(s) shall not be disclosed, licensed or sold, in whole or in part, to any third party, without GDT's express written permission. Other than copies for backup and archival purposes, Licensee shall make no copies of the Product(s), or any part thereof, without the express written consent of GDT. All copies made shall remain the property of GDT under the terms of this Agreement. Licensee shall pay GDT the initial license fee(s) indicated above for the perpetual use of the Product(s). As an option, licensee may receive updates for the fee(s) as specified above.

B.3. Regents of the University of California: DEMP program

The DEMP (Density Equalizing Map Projections) program used in this report is copyrighted by the Regents of the University of California (registration effective 1/11/96). Unauthorized copying or use of the program is prohibited.

B.4. State of California Department of Health Services: case data

In an "Informal Agreement between Lawrence Berkeley Laboratory (LBL) and State of California Department of Health Services (DHS)," signed by both parties on 4/14/93, each party promised to the other:

> not to release to any third party any data file containing proprietary or confidential information. This restriction applies to unmodified proprietary files purchased from private vendors (1980 map files from NPDC, Appendix A.3 and B.1 of this report; and 1990 map files from GDT, Appendix A.3 and B.2 of this report), and confidential cancer case data (Appendix A.3 of this report).

APPENDIX C. CHECKING DENSITY EQUALIZATION

A number of checks were performed on the density equalized maps in Figures 9 through 18, to determine whether the population density had been properly equalized. In Figure C-1 is a scatter plot of adjusted tract area versus actual tract area, for the 259 tracts in the density equalized map of Figure 9. If density equalization were perfect, all points would lie along the 45 degree line. For this map, hsum = 0.0148, where hsum is the average over tracts of the squared relative error. Similar plots (not shown here) were examined for all ten of the density equalized maps in Figures 9 through 18.

In Figure C-2 are shown the locations of 8020 artificial random cases, plotted under the assumption of the null hypothesis, that risk is everywhere equal. In Figure C-3 are the locations of the same 8020 artificial random cases, plotted on the density equalized map. If density equalization were perfect, the distribution of the artificial cases in Figure C-3 would be completely random and uniform. The few small areas of low or high density are areas where the density equalization was unsuccessful. For our data set with only 401 cases, the effect of imperfect density equalization is negligible.

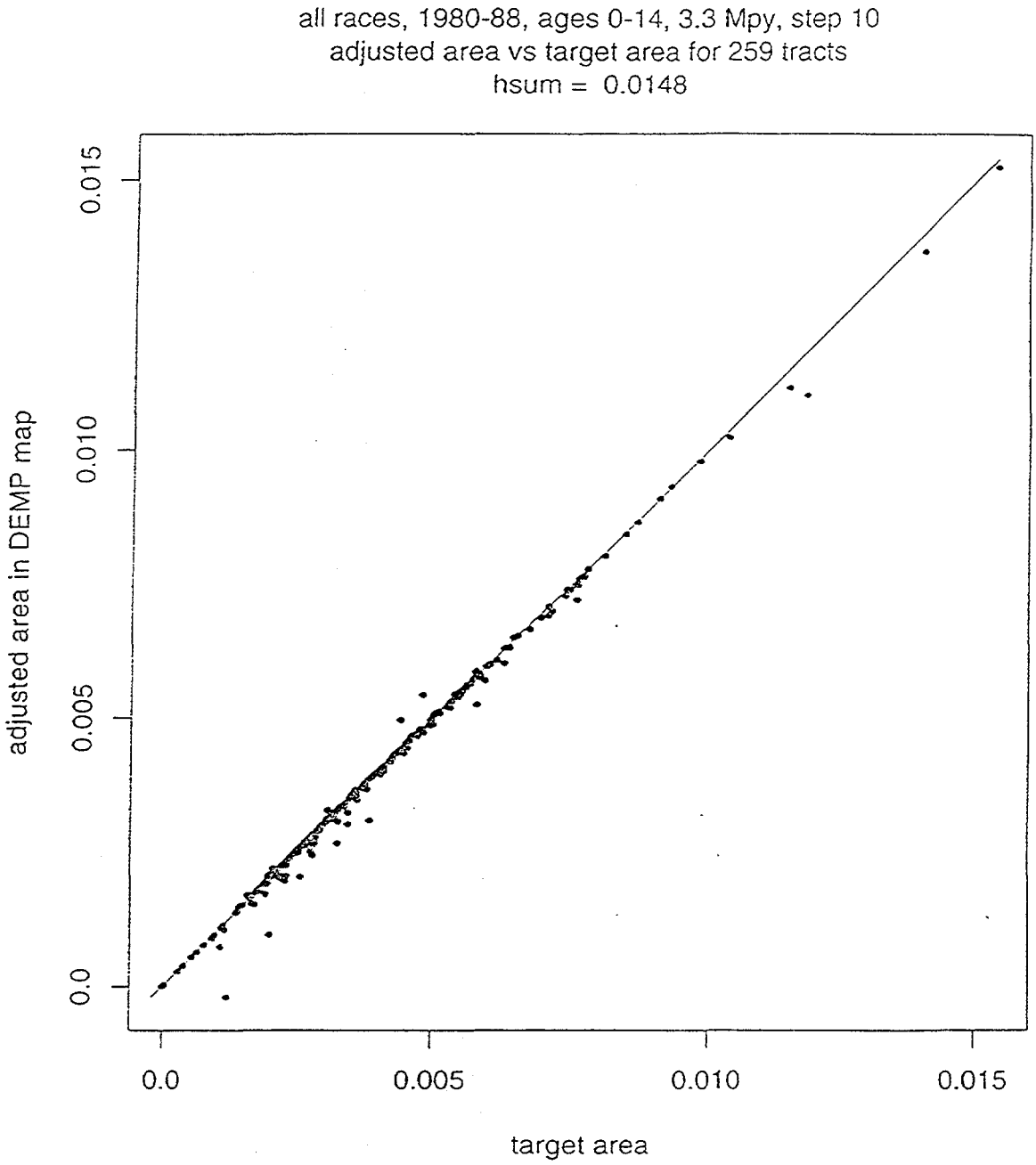


Figure C-1. Adjusted tract areas versus target areas, for the 259 tracts in the density equalized map of Figure 9. If the density equalization were perfect, all points would lie on the 45 degree line.

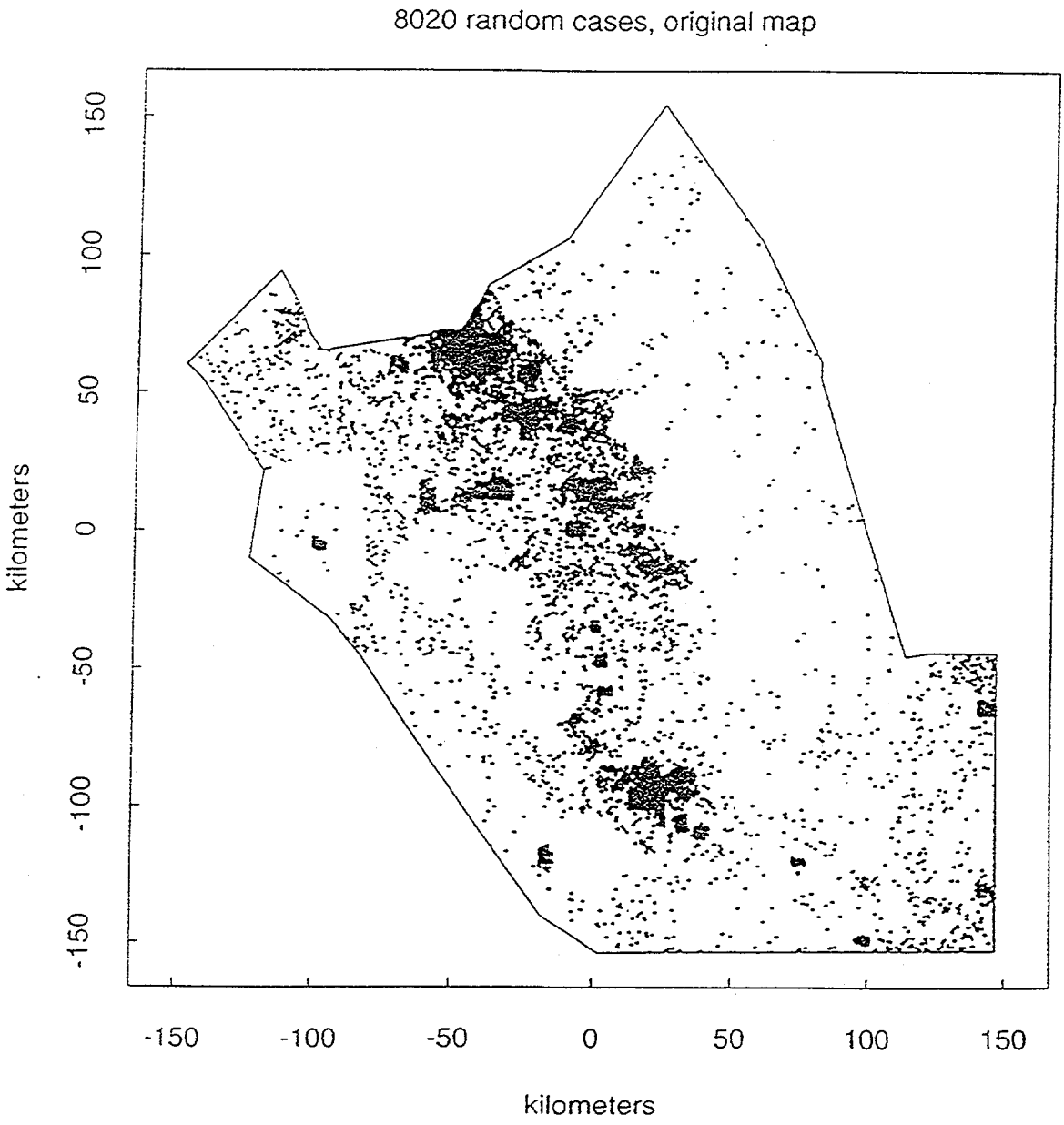


Figure C-2. Locations of 8020 artificial cases in the original geopolitical map of Figure 8. The locations were generated under the assumption that that risk is everywhere equal.

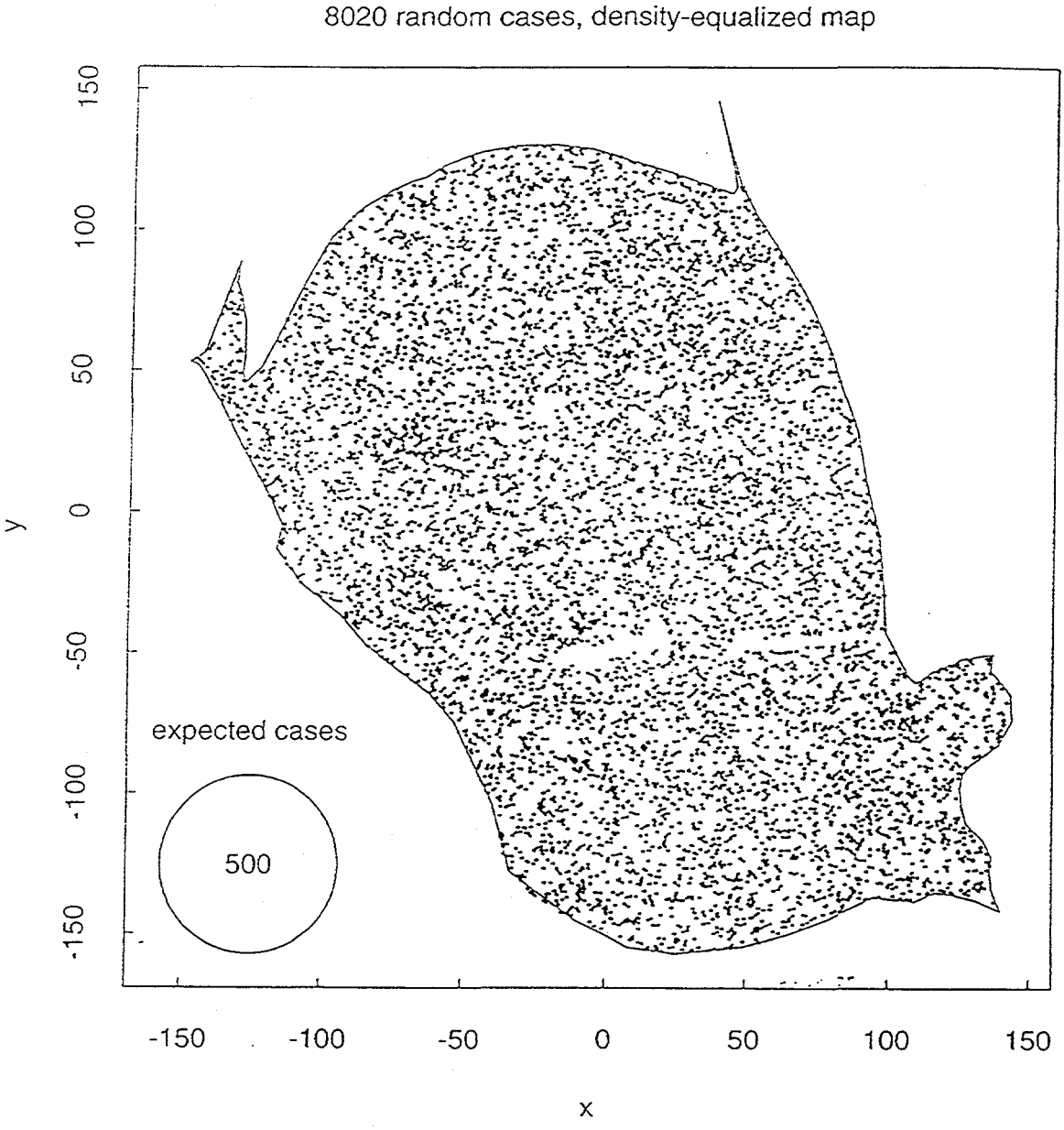


Figure C-3. Locations of the 8020 artificial cases in Figure 8, transformed onto the density equalized map of Figure 9. If the density equalization were perfect, the distribution of points would be completely uniform.

APPENDIX D. COMPARISON WITH EARLIER RESULTS

Geographic variation of rates in the Four County Data Set was previously analyzed by the California Department of Health Services (DHS) [REYN91, REYN96], and in preliminary investigations at Lawrence Berkeley National Laboratory [MERR95A, MERR95B, MERR96A, MERR96B]. Here the conclusions of the successive analyses are summarized, and differences in the data and statistical metrics are described. A subtle but important statistical blunder is discussed, which led to incorrect conclusions in the earlier LBNL analyses.

Summary of conclusions, and differences in the data used

In [REYN91] it was stated, "The distribution of childhood cancers across communities of the entire Four County area during 1980-88 is not substantially different than that which would be expected (based on a Poisson distribution)." In [REYN96] the conclusions were stated as follows: "...there were no previously undiscovered communities with excess rates, although the index community which prompted the initial investigation does stand out as unusual." The analysis in [REYN91] and [REYN96] was based on 101 communities; 1980-88 population estimates were derived from the 1980 Census and post-1980 race-specific county-level projections by the California Department of Finance.

In [MERR95A] it was stated, "...the negative findings of the earlier DHS report are basically confirmed. However, epidemiologic findings 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 analysis in [MERR95A] was based on the 262 tracts of the 1980 Census. Population data were taken directly from the 1980 Census, for children 0-17 years of age.

In [MERR95B] it was stated, "...very significant (4 s.d.) non-uniformity among different tracts, including an excess of cases in sparsely populated areas. The non-uniformity ... may be due to (a) 1980-88 population changes (b) Census undercount of certain populations (c) random uncertainty in population estimates (d) non-uniformity of demographic characteristics (e) other." The data used were the same as in [MERR95A]. In addition, a kth nearest neighbor analysis was performed, which produced the reported 4 s.d. effect.

In [MERR96A] it was stated, "A kth nearest neighbor analysis provides strong evidence for geographic non-uniformity in tract rates ($p < 10^{-4}$)... Work is in progress to repeat the analysis with improved population estimates derived from both 1980 and 1990 Census data. Final epidemiologic conclusions will be reported when that analysis is complete." The data used in [MERR96A], and the kth nearest neighbor analysis, were the same as in [MERR95B].

In [MERR96B] it was stated, "Childhood cancer rates in the four-county area display measurable geographic variation that is portrayed in the contour maps. Some consistency is observed between independent subsamples in the five stratified analysis (not shown here). Overall, the geographic variability of rates is somewhat greater than expected from chance alone; however, no single region has rates sufficiently high or low

to be identified as statistically significant." The units of analysis in [MERR96B] were 259 modified 1980 Census tracts, which can be used with both 1980 and 1990 Census data. Estimates of 1980-88 population at risk, for children 0-14 years of age, were obtained by linear interpolation of age-sex-race-tract specific data from the 1980 and 1990 Census. The stated conclusions were reached after visual examination of RR contour plots.

In [MERR98] (the present work) it is stated, "In agreement with [REYN91] and [REYN96], the findings are consistent with the null hypothesis, that rates are geographically uniform over the four-county area... It is still an open question, as to whether there is any geographic variation of rates in the four-county data set. One can only state that either (a) no such variations exist, or (b) the analysis was not sensitive enough to detect them."

Differences in the statistical metrics used

In [MERR95B] and [MERR96A], the metric that was used to measure non-uniformity of cases in the density equalized map was the mean kth nearest neighbor distance, among cases plotted on the density equalized map. As in the present work, that metric was obtained for real cases and for random artificial cases, and the two results were compared. A measurement of the mean kth nearest neighbor distance is roughly equivalent to a measurement of the mean of log RR(k), i.e. the mean of the Gaussian distribution in Figure 33 (or 34) of the present work.

In the present work the individual measurements of log RR are made on a fixed grid. The presence or absence of clustering does not systematically affect the expected

mean of the distribution. But in the earlier reports the individual measurements of the kth nearest neighbor distance were taken *at the locations of the cases themselves,* which is the usual definition of kth nearest neighbor. In the presence of clustering, the peak is shifted downward because *a greater proportion* of the measurements occur in regions where the inter-case distances are reduced. The shift is a very small second order effect, whose magnitude has been investigated theoretically [CRES91].

The mean kth nearest neighbor distance is not a very sensitive measure of clustering. A more sensitive measure is the *variance* of the individual measurements of log RR. The metric used in this report; namely T, the fraction of log RR measurements in the tail (beyond ±1 s.d.) of the log RR distribution, is closely correlated with the variance. Clustering causes an upward shift in T, because there is a relative increase of both densely and sparsely populated areas. Unlike the mean, T is systematically shifted even if the individual measurements of log RR are made on a fixed grid, which has been done in the present work.

The statistical blunder in the earlier work

The statistical blunder that was made in the earlier work will be discussed here from several different points of view. First, refer to Figure 20 of the present work. For all practical purposes, the distributions of case locations presented there are identical to those presented in earlier reports. In the two upper insets, each real case is plotted at two different random locations within its own census tract. In the two lower insets, the 401 cases are randomly assigned to individual tracts under the assumption of uniform risk, and plotted at two different locations, as in the upper insets. Significantly, for

consistency with the real cases, each artificial case is in the *same* tract in the lower right inset as in the lower left inset.

In [MERR95A] and [MERR95B] it was shown that the actual case locations within tracts provide no useful information, unless one has population data with detail below the tract level. For statistical analysis one must remove the within-tract clustering that cannot be equalized by the DEMP process. One simple method is to plot each case at a random location within its tract; unfortunately this introduction of random noise reduces the sensitivity of the metric. In [MERR95B] and [MERR96A] it was decided to reduce the random noise by plotting each case at not just two, but 20, random locations within its tract, and averaging the results from 20 independent measurements. Imagine Figure 20 with the number of insets in the top row increased from two to 20. For each of the 20 plots in the upper row (the real cases) the kth mean nearest neighbor distance was measured, and then the average of the 20 values was calculated.

In the insets of the bottom row, artificial cases were generated, for comparison with the real cases in the top row. Since on a density equalized map population density is uniform, the cases in the bottom row were simply plotted at random within the external boundary of the density equalized map. Twenty such samples, of 401 cases each, were generated in the same way and then analyzed exactly as the real cases in the top row.

That was the blunder. The analysis was correct but the construction of the random samples was not. For the lower row (the artificial cases) each case *should have been* randomly assigned to a tract, under the null hypothesis of equal risk. Then 20 random locations should have been chosen *in that tract* for each case, as was done for

the real cases. The whole process can then be repeated for a *different* random assignment of cases into tracts, with 20 locations chosen in *those* tracts, and so on.

Graphical presentation

The results of the present analysis were summarized earlier, in Table III. In Figures D-1 and D-2 those results are illustrated, with the artificial samples generated in two ways: incorrectly in D-1 and correctly in D-2. In both figures the metric T, which is a measure of observed clustering, is the fraction of log RR measurements falling in the tail beyond ±1 s.d., of distributions similar to that in Figure 35. The metric used to compare real cases and artificial cases is T_{AV}, the average from two independent plot locations. As shown in Table III and Figures D-1 and D-2, the real cases yielded T=0.36 and T=0.32, with $T_{AV} = 0.34$. (In Figure 35 and in Table II, a different computer run yielded slightly different values). In both Figures D-1 and D-2, the black dot plotted at (.36,.32) represents the measurement from the real cases. (The two black dots do not correspond individually to the two separate measurements of T; rather, the x-y coordinates of *either* black dot represent the two values of T. The dots are simply reflected about the 45 degree line since there is no essential difference between the two measurements.)

In Figures D-1 and D-2, the empty circles represent the results from ten artificial samples, analyzed in the same way as the real cases. In Figure D-1, denoted "uniform," the artificial case locations were *improperly* created, as was done in the analyses of [MERR95B] and [MERR96A]. 20 samples of 401 cases each were randomly generated on the density-equalized map; pairs of these samples produced ten values of $T_{AV,}$ for

comparison with the observed value of 0.34 from the real cases. The ten values of T_{AV} have a sample mean of 0.304 and a sample variance of (0.024)**2. As expected (because the 20 samples are independent) the paired values of T are not correlated. With this improper analysis, the observed value of T_{AV} is about 1.5 standard deviations above the expected value, with a p-value about 0.06.

In Figure D-2, denoted "random," the artificial case locations were *properly* created, as was done in the present analysis. (The real case locations are identical to those in Figure D-1.) In ten separate runs, 401 artificial cases were randomly assigned to tracts; in each of those runs, each case was plotted at two different locations within its tract. In Figure D-2, the ten values of T_{AV} have a sample mean of 0.307 and a sample variance equal to $(0.037)^2$. With this proper analysis, the observed value of T_{AV} is only 0.9 standard deviations above the expected value, with a p-value about 0.18. (The same results appear in Table III).

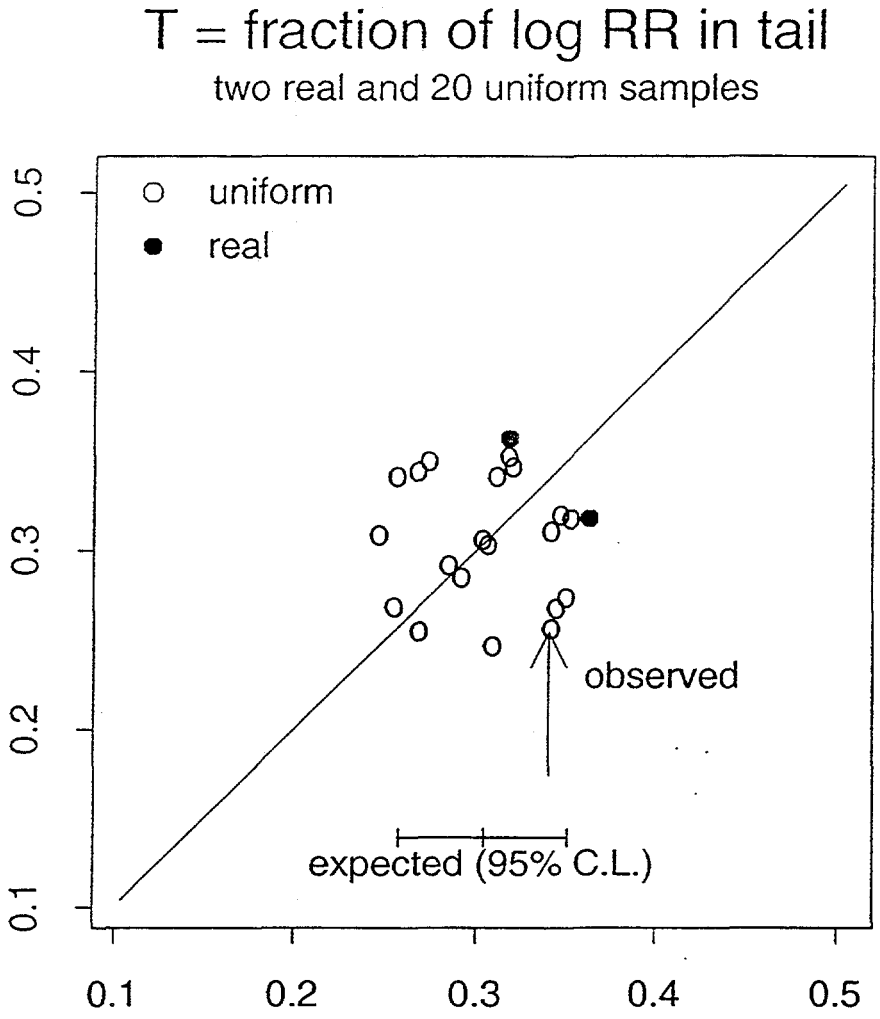


Figure D-1. Values of T, for two real samples and 20 "uniform" samples of artificial cases. The artificial samples were deliberately constructed *improperly*, as was done in previous analyses. Figure D-1 should be compared with Figure D-2, where the samples of artificial cases were constructed *properly*.

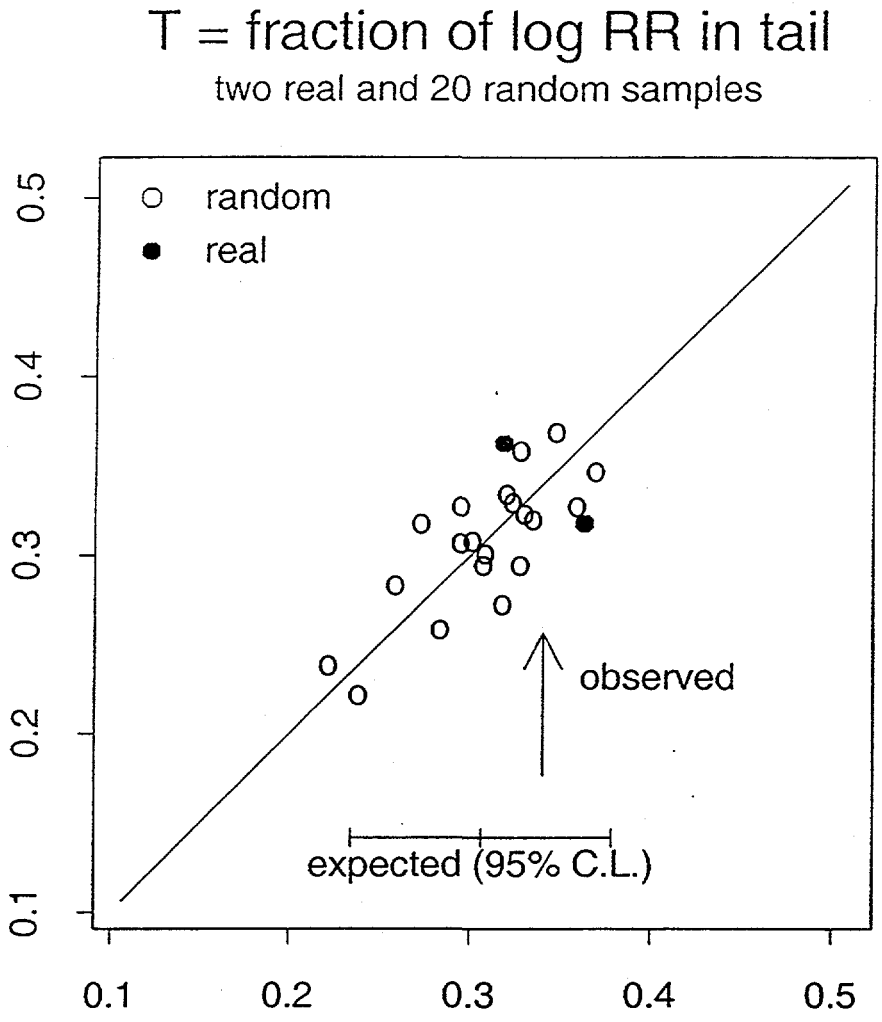


Figure D-2. Values of T, for two real samples and 20 "random" samples of artificial cases. The artificial samples were constructed *properly*, as was done in the present analysis. Figure D-2 should be compared with Figure D-1, where the samples of artificial cases were constructed *improperly*.

Reduction of random noise

From the data presented in Figure D-2 and Table III one can separate the components of statistical error due to (a) allocation of cases into tracts and (b) plotting cases at random locations within their tracts. For each of the ten values of $T_{AV,}$ the variance P^2 due to the plotting error *alone* can be estimated from the two separate values of T entering into T_{AV}. The average of the ten separate P^2 values is $(0.017)^2$. In the preceding section, the total variance due to both causes was estimated to be $V^2=(0.037)^2$. If the two effects produce errors that are independent and normally distributed, then the variance due alone to the random allocation of cases into tracts is $R^2 = V^2-P^2 = (0.037)^2-(0.017)^2=(0.032)^2$. (In Table III, allocation error (alloc err), plot error (plot err), and total error (tot err), correspond to R, P, and V respectively.)

The variation P^2 due to random plotting could be reduced to zero by plotting each case not twice but infinitely many times in its tract or, equivalently, evaluating an area integral. That would increase the statistical power of our method, effectively reducing the standard deviation error of T_{AV} from 0.037 to 0.032. The remaining statistical error (0.032) is the unavoidable result of the size of the data set. The author has derived a general expression for the integral of a polynomial over the area of a polygon [MERR91]; the result is presented in Appendix G. The log RR analysis of this report can be repeated, replacing the randomly plotted cases with a closed-form area integral that will not be too computationally demanding; however, the derivation (which uses the result in Appendix G) is algebraically complex and was too ambitious for the scope of the current work.

Theoretical discussion

Professor Kenneth Wachter has kindly provided a theoretical discussion of the points described above, which provides a more rigorous framework for further investigation. He specifically describes the k-th nearest neighbor algorithm, which was used in previous analyses but not in the present work; however, the conclusions are applicable also to the metric T used in the present work. He correctly predicted, before the calculations were completed, that the random Poisson assignment of cases to tracts is more significant than the randomization of case locations within tracts. The notation below is different than that used elsewhere in this report.

X_0 is the original stochastic point process in the plane.

X is the point process obtained by once randomizing the positions of points of X_0 within tracts of the density-equalized map, maintaining the assignment of points to tracts.

$R_j(X)$ are re-randomizations of X within the tracts in the density-equalized map, independent for different j.

G is a fixed grid of points G_1 , G_2 ... from which distances to points of X will be computed.

d_k (G_g , R_j (X)) is the k-th smallest of the distances from Gg to the points of R_j (X). (Actually, the k-th and (k+1)-st squared distances are averaged, but that feature will be ignored.)

H_0 , the Null Hypothesis, is that X is a uniform process in the plane.

The test statistic is

$$M = \sum_{j} \sum_{g} d_j^2 \, (G_g^2 , R_j(X))$$

Each term in this sum has a Gamma distribution. However, the terms for any two grid points are correlated with each other, because distances are being calculated to points from the same realization of the assumed Poisson Process. Furthermore, the terms for any two randomizations are correlated with each other, since the assignment of points of X to tracts is invariant over randomizations.

The mean, variance, and other moments of the test statistic M involve taking expectations over both the Poisson process X and the collection of independent randomizations R_j. The correlations across values of g and values of j do not affect the mean E M. They do affect the variance and higher moments, and thus the null distribution of M.

As indicated in this Appendix, it is certainly erroneous to do any calculation in which the different terms inside the sum over j are uncorrelated (as would be produced by taking independent realizations of X in the different terms inside the sum instead of the same X and independent realizations of R_j. The variability of a sum of independent terms will be much less than the variability of a sum of positively correlated terms. [This is the point that is made in the comparison of Figures D-1 and D-2.] Each realization of M should be generated with one realization of X and one realization of twenty randomizations R_1 , ... R_{20} within the sum which defines M.

To generate the null distribution for M, then, we should repeat this process by selecting X and R_1 , ... R_{20} over and over independently at random, each time generating one realization of M. The new choices of the set of randomizations do not probably

matter very much, since we are averaging over 20 of them inside the sum, so the principal source of variability that generates a null distribution for M is the variability from successive choices of the process X (that is, specifically, the random Poisson assignment of cases to tracts.) In other words, in the "row-column" language of Figure 20, for each single realization of M we need cases from the different columns to go into the different terms in one calculation of the sum for M, but for the null distribution of M we need to repeat this process over and over with different rows.

APPENDIX E. ESTIMATION OF 1980-88 POPULATION AT RISK

Estimates of 1980-88 population at risk for each census tract were obtained by straight-line interpolation, between age-sex-race-specific populations from the 1980 and 1990 Census. Estimation of the age-sex-race-specific Census populations is documented in detail in the files that are listed in Appendix A. Several difficulties had to be circumvented:

(a) The list of tracts in the three data inputs (cases, map files, and Census populations) did not match exactly. The differences were resolved by consulting maps, and by associating tract codes ending in 99 (persons resident in ships) with the corresponding land-based codes having other 2-digit suffixes.

(b) The list of tracts in 1980 and 1990 do not match. Many 1980 tracts were subdivided in 1990. By combining a small number of 1980 tracts, a list of 259 "modified 1980 tracts" was obtained which are aggregates of individual tracts in either 1980 or 1990.

(c) The geographic levels in Census data are not entire tracts, but rather place/tract pieces or MCD/place/tract pieces, as summarized in Table E-1. These had to be aggregated into entire tracts.

(d) The age ranges provided in the Census Summary Tape Files did not match exactly the age groups 0-4 and 5-14 in the case data. The tract level data were apportioned according to age ratios found in county level data.

(e) In 1980, data were suppressed when the number of persons of a given race in a tract was so small that confidentiality of individuals would have been compromised. (In 1990, a random rounding procedure was used.) The missing data in 1980 were estimated by determining the total population and the sex-age-specific populations for the tracts that were *not* suppressed. Subtracting from the county totals provided the same information for the suppressed tracts as a group. Then the sex-age distributions were assumed to be identical in all of the suppressed tracts.

Table E-1. Geographic levels of detail in Census data

SEEDIS level	where used	description	nests within	number in four-county area
COUNTY80	STF1 STF2C STF2B28R	counties	STATE	4
PLACE80I		PLACE80 with population 10,000 or greater (PLACE80I = 9999 is the remainder of state)	STATE	14
PLACE80A	STF1	places (cities)	STATE PLACE80I	86
MCD80	STF1	minor civil divisions	COUNTY80	
TRACT80	case data file	Census tracts	COUNTY80	262
NMCDTR80	NPDC map file	units of NPDC map file (equivalent to TRACT80 in the four-county area)	COUNTY80	262
PLTRACT80I	STF2A STF2B7R	PLACE80I/TRACT80 pieces	PLACE80I TRACT80	393
TRACT80PT	STF1	MCD80/PLACE80/TRACT 80 pieces	MCD80 PLACE80 TRACT80	486
PLTRACT80A		PLACE80A/TRACT80 pieces (equivalent to TRACT80PT in the four-county area)	PLACE80A TRACT80	486

Race and ethnicity:

Race and ethnicity are two separate questions on the Census form, so that a "Hispanic" person can be of any race, and a "non-Hispanic" persons can be of any race. At subcounty levels of geography, population counts are provided for five racial classifications and two ethnic classifications, *but not for the two-way cross-tabulation.* In Table E-2, only the filled-in cells are provided for individual census tracts.

Table E-2. Race and ethnic classifications in Census data.

	Hispanic	non-Hispanic	total
white			W
black			B
native American			N
Asian and Pacific Islander			A
other			O
total	H	NH	T = (W+B+N+A+O) = (H+NH)

Studies of Census data show that Hispanics are increasingly proud of their ethnic identity. Statistically, many people who classified themselves as (Hispanic, white) in the 1970 Census reclassified themselves as (Hispanic, other) in 1980 [CENS84]. The trend continued into 1990, though less markedly. The observed changes is an artifact of self-identification, not a real demographic change of the population.

In the case data from DHS, the classifications are "white," "Hispanic" and "other" (and total = white + Hispanic + other) which are quite different from the

Census classifications of the same name. If one needed race-specific rate estimates one would need to study very carefully the bias due to different race classification in the case data (numerators) and the population data (denominators). In the present analysis, however, race and ethnicity are used only as a stratifying variable. Here it is assumed that all Hispanics are either white or "other." In other words, (a) "Hispanics" in the case data correspond to (Hispanic,total) in the Census data; (b) all Hispanics are either (Hispanic,white) or (Hispanic,other) in the Census data; i.e. (Hispanic,total) = (Hispanic,white) + (Hispanic,other). With this simplification, the Census classifications become:

Table E-3. Modified race and ethnic classifications.			
	Hispanic	non-Hispanic	total
white+other	"Hispanic" = H	"white" = T-H-B-N-A	W+O
black	assumed zero	B	B
native American	assumed zero	N	N
Asian and Pacific Islander	assumed zero	A	A
total	"Hispanic" = H	NH=T-H	"total" = T = (W+B+N+A+O) = (H+NH)

The DHS race classifications are those in quotes. Now population can be estimated for the DHS classifications, in terms of the Census data: "total" = T; "Hispanic" = H; "white" = T-H-B-N-A; "other" = B+N+A.

APPENDIX F. PREPARATION OF GEOGRAPHIC MAP FILES

1980 Census Tracts:

The 1980 Census tract maps used in this project were purchased from National Planning Data Corporation in 1986. They were installed in SEEDIS (which was earlier at LBNL and has now been moved to the Census Bureau). *The NPDC map files are proprietary and may not be copied. See the license agreement in Appendix B-1.*

Editing was required to repair topological errors, to remove small lakes, to "sew" together the four separate county maps. The editing was partially automated with the use of routines written at LBNL.

1990 Census Tracts:

Maps of 1990 Census tract boundaries, for the four counties, were purchased from Geographic Data Technology, Inc. Some editing was required.

Modifed 1980 Census Tracts:

Unnecessary geographic detail was removed from the 1980 and 1990 geographic map files. The simplified maps are shown in Figures 6 and 7. Visual inspection, aided by routines written at LBNL, provided the proper correspondence between the five input data files: 1980 and 1990 map files, 1980 and 1990 Census data, and the 1980-88 cancer case data. The geographic units chosen for the analysis

are 259 modified 1980 Census tracts, which are shown in Figure 8. These are identical to the 262 original 1980 Census tracts shown in Figure 6; except that five large 1980 tracts had to be aggregated into two larger "modifed tracts," in order to achieve correspondence with the 1990 Census tract definitions. Further map modifications were made: every polygon was subdivided by a Delaunay triangulation [BOOT87]; then every segment in the map was subdivided, thereby converting every triangle into a hexagon. This provided enough degrees of freedom that the DEMP calculation converged successfully.

APPENDIX G. INTEGRAL OF A POLYNOMIAL OVER A POLYGON

In 1991 the author derived a general expression for the integral of a polynomial, over the area of an arbitrary polygon [MERR91B]. As a check on the result, the formula is evaluated for a few special cases. The general result has not been found elsewhere in the literature. Appendix D describes how the sensitivity of the log RR measurement can be improved, if the summation over random case locations is replaced by an exact area integral. The equations given in this Appendix will simplify the required derivation.

In the expressions below, u(x,y) is an arbitrary polynomial of order n, described by coefficients $b_{p,n-p}$. The points (x_k , y_k) are the points describing the boundary of the polygon, with k increasing in the *counterclockwise* direction.

VI.K. Expectation of a real function

If $f(z)$ is a real function $u(x,y)$, the numerator of (VI.J.1) is given by (VI.I.5b):

$$\int_{ABC..Z} u(x,y)\, dA = \frac{1}{2}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk}-\Sigma_{yk}\Delta_{xk})\sum_{n=0}^{\infty}\frac{1}{n+2}\left[\frac{1}{2}\right]^n\sum_{p=0}^{n} b_{p,n-p}\sum_{i=0}^{p}\binom{p}{i}\Sigma_{xk}{}^{i}$$

$$\Delta_{xk}{}^{p-i}\sum_{\substack{j=0\\ n-i-j\ \text{even}}}^{n-p}\binom{n-p}{j}\Sigma_{yk}{}^{j}\Delta_{yk}{}^{n-p-j}\frac{1}{n-i-j+1}$$

where $\Delta_{xk}\equiv x_{k+1}-x_k$, $\Sigma_{xk}\equiv x_{k+1}+x_k$, $\Delta_{yk}\equiv y_{k+1}-y_k$, $\Sigma_{yk}\equiv y_{k+1}+y_k$, and

$$u(x,y) = \sum_{n=0}^{\infty}\sum_{p=0}^{n} b_{p,n-p}x^p y^{n-p}$$

The $b_{p,n-p}$ are real constants equal to

$$b_{p,n-p} = \frac{1}{p!\,(n-p)!}\left.\frac{\partial^n u}{\partial x^p \partial y^{p-n}}\right|_{x=0,y=0}$$

The denominator of (VI.J.1) is given by (VI.I.3d):

$$A(\overline{ABC..Z}) = \int_{ABC..Z} dA = \frac{1}{4}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk}-\Sigma_{yk}\Delta_{xk})$$

VI.K.1. Special case: $n = 1$:

$$\int_{ABC..Z} u(x,y)\, dA = \frac{1}{12}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk}-\Sigma_{yk}\Delta_{xk})$$

$$\left[\sum_{p=0}^{1} b_{p,1-p}\sum_{\substack{i=0\\ i+j\ \text{odd}}}^{p}\sum_{j=0}^{1-p}\binom{p}{i}\binom{1-p}{j}\frac{1}{2-i-j}\Sigma_{xk}{}^{i}\Delta_{xk}{}^{p-i}\Sigma_{yk}{}^{j}\Delta_{yk}{}^{1-p-j}\right]$$

VI.K.1.1. Special case: $p = 1$:

$u(x,y) = x$; ($b_{1,0} = 1$; other $b_{p,n-p} = 0$):

$$\Xi(x) = \frac{\int_{ABC..Z} x \, dA}{\int_{ABC..Z} dA} = \frac{\frac{1}{12}\sum_{k=1}^{m}(\Sigma_{zk}\Delta_{yk} - \Sigma_{yk}\Delta_{zk})\ \Sigma_{zk}}{\frac{1}{4}\sum_{k=1}^{m}(\Sigma_{zk}\Delta_{yk} - \Sigma_{yk}\Delta_{zk})}$$

VI.K.1.2. Special case: $p = 0$:

$u(x,y) = y$; ($b_{0,1} = 1$; other $b_{p,n-p} = 0$):

$$\Xi(y) = \frac{\int_{ABC..Z} y \, dA}{\int_{ABC..Z} dA} = \frac{\frac{1}{12}\sum_{k=1}^{m}(\Sigma_{zk}\Delta_{yk} - \Sigma_{yk}\Delta_{zk})\ \Sigma_{yk}}{\frac{1}{4}\sum_{k=1}^{m}(\Sigma_{zk}\Delta_{yk} - \Sigma_{yk}\Delta_{zk})}$$

VI.K.2. Special case: $n = 2$:

$$\int_{ABC..Z} u(x,y)\, dA = \frac{1}{32}\sum_{k=1}^{m}(\Sigma_{zk}\Delta_{yk} - \Sigma_{yk}\Delta_{zk})$$

$$\left[\sum_{p=0}^{2} b_{p,2-p} \sum_{\substack{i=0 \\ i+j \text{ even}}}^{p} \sum_{j=0}^{2-p} \binom{p}{i}\binom{2-p}{j} \frac{1}{3-i-j} \Sigma_{zk}{}^{i}\Delta_{zk}{}^{p-i} \Sigma_{yk}{}^{j}\Delta_{yk}{}^{2-p-j}\right]$$

VI.K.2.1. Special case: $p = 2$:

$u(x,y) = x^2$; $(b_{2,0} = 1$; other $b_{p,n-p} = 0)$:

$$E(x^2) = \frac{\int_{ABC..Z} x^2 \, dA}{\int_{ABC..Z} dA} = \frac{\frac{1}{32}\sum_{k=1}^{m}(\Sigma_{zk}\Delta_{yk} - \Sigma_{yk}\Delta_{zk})\left[\Sigma_{zk}^{\,2} + \frac{1}{3}\Delta_{zk}^{\,2}\right]}{\frac{1}{4}\sum_{k=1}^{m}(\Sigma_{zk}\Delta_{yk} - \Sigma_{yk}\Delta_{zk})}$$

VI.K.2.2. Special case: $p = 1$:

$u(x,y) = xy$; $(b_{1,1} = 1$; other $b_{p,n-p} = 0)$:

$$E(xy) = \frac{\int_{ABC..Z} xy \, dA}{\int_{ABC..Z} dA} = \frac{\frac{1}{32}\sum_{k=1}^{m}(\Sigma_{zk}\Delta_{yk} - \Sigma_{yk}\Delta_{zk})\left[\Sigma_{zk}\Sigma_{yk} + \frac{1}{3}\Delta_{zk}\Delta_{yk}\right]}{\frac{1}{4}\sum_{k=1}^{m}(\Sigma_{zk}\Delta_{yk} - \Sigma_{yk}\Delta_{zk})}$$

VI.K.2.3. Special case: $p = 0$:

$u(x,y) = y^2$; $(b_{0,2} = 1$; other $b_{p,n-p} = 0)$:

$$E(y^2) = \frac{\int_{ABC..Z} y^2 \, dA}{\int_{ABC..Z} dA} = \frac{\frac{1}{32}\sum_{k=1}^{m}(\Sigma_{zk}\Delta_{yk} - \Sigma_{yk}\Delta_{zk})\left[\Sigma_{yk}^{\,2} + \frac{1}{3}\Delta_{yk}^{\,2}\right]}{\frac{1}{4}\sum_{k=1}^{m}(\Sigma_{zk}\Delta_{yk} - \Sigma_{yk}\Delta_{zk})}$$

VI.K.3. Special case: $n = 3$:

$$\int_{ABC..Z} u(x,y)\ dA = \frac{1}{80}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})$$

$$\left[\sum_{p=0}^{3} b_{p,3-p} \sum_{\substack{i=0 \\ i+j\ \text{odd}}}^{p} \sum_{j=0}^{3-p} \binom{p}{i}\binom{3-p}{j} \frac{1}{4-i-j}\, \Sigma_{xk}{}^{i}\Delta_{xk}{}^{p-i}\Sigma_{yk}{}^{j}\Delta_{yk}{}^{3-p-j}\right]$$

VI.K.3.1. Special case: $p = 3$:

$u(x,y) = x^3$; ($b_{3,0} = 1$; other $b_{p,n-p} = 0$):

$$E(x^3) = \frac{\int_{ABC..Z} x^3\ dA}{\int_{ABC..Z} dA} = \frac{\frac{1}{80}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})\left[\Sigma_{xk}{}^{3} + \Sigma_{xk}\Delta_{xk}{}^{2}\right]}{\frac{1}{4}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})}$$

VI.K.3.2. Special case: $p = 2$:

$u(x,y) = x^2y$; ($b_{2,1} = 1$; other $b_{p,n-p} = 0$):

$$E(x^2y) = \frac{\int_{ABC..Z} x^2y\ dA}{\int_{ABC..Z} dA}$$

$$= \frac{\frac{1}{80}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})\left[\Sigma_{xk}{}^{2}\Sigma_{yk} + \frac{2}{3}\Sigma_{xk}\Delta_{xk}\Delta_{yk} + \frac{1}{3}\Sigma_{yk}\Delta_{xk}{}^{2}\right]}{\frac{1}{4}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})}$$

VI.K.3.3. Special case: $p = 1$:

$u(x,y) = xy^2$; ($b_{1,2} = 1$; other $b_{p,n-p} = 0$):

$$E(xy^2) = \frac{\int_{ABC..Z} xy^2 \, dA}{\int_{ABC..Z} dA}$$

$$= \frac{\frac{1}{80}\sum_{k=1}^{m}(\Sigma_{zk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})\left[\Sigma_{yk}^{2}\,\Sigma_{zk} + \frac{2}{3}\Sigma_{yk}\Delta_{yk}\Delta_{zk} + \frac{1}{3}\Sigma_{zk}\Delta_{yk}^{2}\right]}{\frac{1}{4}\sum_{k=1}^{m}(\Sigma_{zk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})}$$

VI.K.3.4. Special case: $p = 0$:

$u(x,y) = y^3$; ($b_{0,3} = 1$; other $b_{p,n-p} = 0$):

$$E(y^3) = \frac{\int_{ABC..Z} y^3 \, dA}{\int_{ABC..Z} dA} = \frac{\frac{1}{80}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})\left[\Sigma_{yk}^{\,3} + \Sigma_{yk}\Delta_{yk}^{\,2}\right]}{\frac{1}{4}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})}$$

VI.K.4. Special case: $n = 4$:

$$\int_{ABC..Z} u(x,y)\, dA = \frac{1}{192}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})$$

$$\left[\sum_{p=0}^{4} b_{p,4-p} \sum_{\substack{i=0 \\ i+j \text{ even}}}^{p} \sum_{j=0}^{4-p} \binom{p}{i}\binom{4-p}{j} \frac{1}{5-i-j} \Sigma_{xk}^{\,i} \Delta_{xk}^{\,p-i} \Sigma_{yk}^{\,j} \Delta_{yk}^{\,4-p-j}\right]$$

VI.K.4.1. Special case: $p = 4$:

$u(x,y) = x^4$; ($b_{4,0} = 1$; other $b_{p,n-p} = 0$):

$$\int_{ABC..Z} x^4 \, dA = \frac{1}{192}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})\left[\Sigma_{xk}^{\,4} + 2\Sigma_{xk}^{\,2}\Delta_{xk}^{\,2} + \frac{1}{5}\Delta_{xk}^{\,4}\right]$$

$$E(x^4) = \frac{\int_{ABC..Z} x^4 \, dA}{\frac{1}{4}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})}$$

VI.K.4.2. Special case: $p = 3$:

$u(x,y) = x^3y$; $(b_{3,1} = 1$; other $b_{p,n-p} = 0)$:

$$\int_{ABC..Z} x^3y\,dA = \frac{1}{192}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})\left[\Sigma_{xk}^{3}\Sigma_{yk} + \Sigma_{xk}^{2}\Delta_{xk}\Delta_{yk} + \Sigma_{xk}\Delta_{xk}^{2}\Sigma_{yk} + \frac{1}{5}\Delta_{xk}^{3}\Delta_{yk}\right]$$

$$E(x^3y) = \frac{\int_{ABC..Z} x^3y\,dA}{\frac{1}{4}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})}$$

VI.K.4.3. Special case: $p = 2$:

$u(x,y) = x^2y^2$; $(b_{2,2} = 1$; other $b_{p,n-p} = 0)$:

$$\int_{ABC..Z} x^2y^2\,dA = \frac{1}{192}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})\left[\Sigma_{xk}^{2}\Sigma_{yk}^{2} + \frac{1}{3}\Sigma_{xk}^{2}\Delta_{yk}^{2} + \frac{4}{3}\Sigma_{xk}\Delta_{xk}\Sigma_{yk}\Delta_{yk} + \frac{1}{3}\Delta_{xk}^{2}\Sigma_{yk}^{2} + \frac{1}{5}\Delta_{xk}^{2}\Delta_{y}^{2}\right]$$

$$E(x^2y^2) = \frac{\int_{ABC..Z} x^2y^2\,dA}{\frac{1}{4}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})}$$

VI.K.4.4. Special case: $p = 1$:

$u(x,y) = xy^3$; $(b_{1,3} = 1$; other $b_{p,n-p} = 0)$:

$$\int_{ABC..Z} xy^3\,dA = \frac{1}{192}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})\left[\Sigma_{xk}\Sigma_{yk}^{3} + \Delta_{xk}\Sigma_{yk}^{2}\Delta_{yk} + \Sigma_{xk}\Sigma_{yk}\Delta_{yk}^{2} + \frac{1}{5}\Delta_{xk}\Delta_{yk}^{3}\right]$$

$$E(xy^3) = \frac{\int_{ABC..Z} xy^3\,dA}{\frac{1}{4}\sum_{k=1}^{m}(\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})}$$

VI.K.4.5. Special case: $p = 0$:

$u(x,y) = y^4$; ($b_{0,4} = 1$; other $b_{p,n-p} = 0$):

$$\int_{ABC..Z} y^4 \, dA = \frac{1}{192} \sum_{k=1}^{m} (\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk}) \left\{ \Sigma_{yk}^4 + 2\Sigma_{yk}^2\Delta_{yk}^2 + \frac{1}{5}\Delta_{yk}^4 \right\}$$

$$E(y^4) = \frac{\int_{ABC..Z} y^4 \, dA}{\frac{1}{4} \sum_{k=1}^{m} (\Sigma_{xk}\Delta_{yk} - \Sigma_{yk}\Delta_{xk})}$$