LBL-32640
UC-605

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

Information and Computing Sciences Division

Presented at the Workshop on Statistics and Computing in Disease Clustering, Stony Brook, NY, July 23–24, 1992, and to be published in the Proceedings

Density Equalizing Map Projections: Techniques and Applications

D.W. Merrill, S. Selvin, and M.S. Mohr

July 1992

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

DISCLAIMER

LBL-32640

DENSITY EQUALIZING MAP PROJECTIONS: TECHNIQUES AND APPLICATIONS

Deane W. Merrill, Ph.D.[1,2]

Steve Selvin, Ph.D.[1,2]

Michael S. Mohr, M.S.[1,2]

6 August 1992

[1]Information and Computing Sciences Division, Lawrence Berkeley Laboratory, 1 Cyclotron Road, Berkeley CA 94720.

[2]Department of Biomedical and Environmental Health Sciences, University of California, Berkeley CA 94720.

Presented at Workshop on Statistics and Computing in Disease Clustering, Stony Brook NY, 23-24 July 1992. Proceedings to be published in Statistics and Medicine, August 1993.

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

Density Equalizing Map Projections: Techniques and Applications

ABSTRACT

In the statistical investigation of geographic disease clusters, one needs to correct for inevitable variations in population density. Density equalizing map projections (DEMP) have been proposed as an alternative to conventional methods. In a DEMP, a geographic map is transformed so that population density is everywhere equal. Disease cases plotted on the transformed map should have a uniform distribution if risk is everywhere equal. Because one does not need to define arbitrary geographic subareas, the full geographic detail of the original data is preserved. Unlike previous cartogram algorithms, the DEMP algorithm described here is a continuous transformation which provides correct area magnification over the entire map domain. This fact permits analysis of geographic disease distributions by simple statistical methods which rely upon the assumption of uniform population density. Here we describe recent progress, including new objective and constraint functions, diagnostic tests and graphic displays, and estimates of computing requirements. The use and interpretation of the DEMP technique in epidemiologic applications is discussed.

INTRODUCTION

In analyzing the geographic variation of disease, it is almost always necessary to compensate for the confounding effect of non-uniform population density. A conventional method is by comparing rates which are calculated for geographic subareas. However this method is inadequate when the number of cases is small: either the number of cases per subarea is too small to provide stable rates, or geographic detail is lost through combination of subareas. In either case the results can be influenced by the arbitrary choice of subareas. Furthermore, a simple comparison of subarea rates fails to describe smoothly varying trends across adjacent subareas. Other methods have been devised to compensate mathematically for variations in population density. Frequently such models are difficult to comprehend or to represent graphically.

The method to be described here is the technique of density equalizing map projections (DEMP), which have been used to analyze public health data for more than 60 years.[1,2] The geographic map is transformed so as to make population density everywhere equal. On such a map, the distribution of disease cases is expected to be random if risk is everywhere equal. It is relatively simple to recognize a disease cluster on a transformed map, and to evaluate its statistical significance even if the total number of cases is small.

Although geopolitical boundaries are necessarily distorted by the DEMP, physical features and/or coordinate grids can be superimposed on the transformed map to aid in interpretation.

Contours of supposedly equal risk (for example circles around a point source of pollution in the original map) are likewise easily transformed.

The choice of a DEMP transformation is not unique. For aesthetic reasons and to simplify interpretation, it is desirable to choose the unique DEMP which minimizes distortion relative to the original map.

Computer algorithms have been created, at Lawrence Berkeley Laboratory (LBL) and elsewhere, to automatically create density-equalized maps, or *cartograms*.[3,4,5,6,7] Recently, the authors of this paper described a new algorithm which has important advantages over earlier methods.[8,9] The new algorithm (a) avoids illegal overlapping of transformed polygons; (b) finds the unique solution that minimizes map distortion; (c) provides constant magnification over each map polygon; (d) defines a continuous transformation over the entire map domain; (e) defines an inverse transformation; (f) can accept optional constraints such as fixed boundaries; and (g) can use commercially supported minimization software.

In the new DEMP algorithm, the map is stripped of excess detail and represented by polygons, which are further subdivided into triangles. Triangles are chosen because they allow the DEMP to be described by a minimum number of variable parameters. The original map is described by constants (x,y), the original coordinates of the triangle vertices. The variable parameters of the DEMP are the transformed vertex coordinates (u,v); these are adjusted so that the area of each triangle, after transformation, is proportional to its population. A unique solution is obtained by requiring that overall shape distortion, relative to the original map, be minimized. We define an area constraint function $H(u,v)$ and a distortion function $G(u,v)$, both of which are continuous non-linear functions of the transformed coordinates (u,v). The NAG optimization program E04VDF is used to minimize the objective function $G(u,v)$ subject to the condition $H(u,v) = 0$,[10] thereby providing transformed coordinates (u,v) of all the triangle vertices.

The DEMP transformation is a continuous piecewise linear transformation of the entire map; within any triangle i the transformation is

$$u = a_i x + b_i y + e_i$$

$$v = c_i x + d_i y + f_i$$

The constants a_i through f_i for each triangle are determined by the original (x,y) and final (u,v) of its three vertices. The continuous DEMP transformation is applied to the entire area of the original map, in particular the locations of disease cases. Except for triangles having zero population (and therefore zero area in the transformed map) the inverse transformation is also defined.

As an example, in Fig. 1 we present the hypothetical disease distribution of a sample test map, before a DEMP, of an imaginary island and four smaller islands. The empty triangle

near the upper left corner represents a lake. Tract boundaries are represented by thin solid lines. Population density is assumed to be uniform within each tract, and is greatest in the small island near the top of the map. Disease risk is assumed constant over the entire map. Each point represents a disease case, and cases are randomly distributed within each tract.

In Fig. 2 we present the same map after a DEMP. Population density is uniform over the entire transformed map, and cases are randomly distributed over the entire transformed map. A significant cluster of cases, if one were seen in Fig. 2, would provide evidence that disease risk is not constant, and would *not* be attributable to variations in population density.

The purpose of this paper is to describe recent experience and insight obtained from the new DEMP algorithm. Present difficulties and future plans will be discussed.

MAP PREPARATION

The preparation of maps for the DEMP transformation is described in Ref. 11. The process is illustrated in Fig. 3 through Fig. 8 with maps of Alameda county, California.

Fig. 3 is a map of 1980 census tract boundaries for a portion of Alameda county, California, including Berkeley, Oakland, and Alameda cities. The original map was obtained from National Planning Data Corporation.

In Fig. 4, to save computing time. unnecessary detail has been filtered out by removing geographic line features having area less than 1 km^2. In Fig. 5 small triangular islands and lakes have also been removed.

In Fig. 6 a Delaunay triangulation has been applied to every polygon in the map, including water areas wholly or largely surrounded by land areas. The Delaunay triangulation is unique and provides triangles that are as nearly equiangular as possible. The entire map is simply connected and fully triangulated.

In Fig. 7 is the original map for all of Alameda county, California, containing 289 tracts, 305 polygons, 8605 points, and 8897 segments.

In Fig. 8 is the triangulated map for all of Alameda county, containing 289 tracts, 581 points, 1090 triangles, and 1670 segments. As in any simply connected map of triangles, the number of segments is one less than the number of points plus the number of triangles.

In order to illustrate certain features of the DEMP transformation, we consider Fig. 9, which is the triangulated test map (before the DEMP) corresponding to Fig. 1. Fig. 10 is the same triangulated test map (after the DEMP) corresponding to Fig. 2. Land triangles and water triangles are labeled with horizontal and vertical numbers, respectively. Fig. 9 and Fig. 10 each contain 8 tracts, 49 points, 82 triangles, and 130 segments.

AREA CONSTRAINT VIOLATION

Repeated trials led to a new function $H(u,v)$ for area constraint violation, which is superior to the form proposed in Ref. 11. The form finally adopted for $H(u,v)$ is a sum over land triangles of

$$\frac{[\ B_i^{final}(u,v)\ -\ B_i^{target}\ ]^2}{\max(B_i^{target},\ B_i^{min})}$$

plus a sum over water triangles of

$$\frac{[\ \min(\ 0,\ B_i^{final}(u,v)\ )\ ]^2}{B_i^{min}}$$

where $B_i^{final}(u,v)$ is the ratio of the final area of triangle i to total land area; B_i^{target} is the fraction of total population in triangle i, and B_i^{min} is a small constant (which was chosen to be 0.07 for the sample test map). The denominator $\max(B_i^{target},B_i^{min})$ is needed to make small land triangles approach their target areas as rapidly as larger triangles. Target areas of water triangles, which have no population, are not specified. The sum over water triangles is required to prevent the final areas of those triangles from becoming negative, which would correspond to upside-down triangles and self-intersecting polygon boundaries. The area constraint function $H(u,v)$ is multiplied by a constant coefficient h_0, which is taken to be 1. Larger (or smaller) values of h_0 cause the DEMP to terminate after more (or fewer) iterations, causing the constraint $H(u,v) = 0$ to be more (or less) exactly satisfied. Smaller (or larger) values of B_i^{min} cause the area constraint to be more (or less) tightly applied to small-area land triangles and to water triangles.

In Fig. 11 are shown the original (Fig. 9, pre-DEMP) area A_i and the target area B_i^{target} of each land triangle i in the sample test map. (All triangle areas in Fig. 11 through Fig. 15 are expressed as fractions of total land area.) The ratio of target area to original area specifies the area magnification to be applied to each land triangle during the DEMP. The numbers 56,36,13... correspond to triangles in a single large tract, all of which receive the same magnification. Triangle 3 receives the maximum magnification.

In Fig. 12 are shown the final (Fig. 10, post-DEMP) area $B_i^{final}(u,v)$ and the target area B_i^{target} of each land triangle in the sample test map. For every land triangle the final area and target area are equal, showing that land triangle areas in Fig. 10 have been correctly adjusted.

In Fig. 13 are shown the original (Fig. 9, pre-DEMP) area A_i and the final (Fig. 10, post-DEMP) area $B_i^{final}(u,v)$ of each land and water triangle i in the sample test map. As in Fig. 9 and Fig. 10, land and water triangles are represented by horizontal and vertical numbers respectively. Water triangles 4 and 40 have final areas equal to zero, as is evident in Fig.

10. Negative triangle areas, corresponding to upside-down triangles and self-intersecting polygon boundaries, are excluded by the form of the area constraint function $H(u,v)$.

SHAPE DISTORTION

Repeated trials also led to a new shape distortion function, which is superior to the form proposed in Ref. 11. The form finally adopted for the global shape distortion function $G(u,v)$ is a sum over all triangles (land and water) of

$$A_i * [\ dist_i(u,v)\ ]^2$$

where A_i is the ratio of the initial triangle area to total land area and $dist_i(u,v)$, the shape distortion of triangle i, is equal to

$$\arctan \frac{(a_i - d_i)^2 + (b_i + c_i)^2}{a_i d_i - b_i c_i}$$

As described earlier, a_i, b_i, c_i, and d_i are functions of (u,v) that describe the linear transformation of triangle i from initial coordinates (x,y) to final coordinates (u,v). The shape distortion function $dist_i(u,v)$ measures the change in shape of triangle i, neglecting rotation and magnification. The denominator $a_i d_i - b_i c_i$, which is equal to the area magnification $B_i^{final}(u,v)/A_i$, helps to discriminate against triangles that are flattened into zero-area straight lines. The *arctan* function provides continuity of $dist_i(u,v)$ in the neighborhood of regions where $a_i d_i - b_i c_i = 0$. Inclusion of water triangles in the global shape distortion function $G(u,v)$ causes separate land bodies such as islands to maintain approximately constant distance and orientation with respect to each other. Their inclusion in $G(u,v)$ also helps reduce shape distortion of lakes and coastline boundaries. The global shape distortion function $G(u,v)$ is multiplied by a constant coefficient g_0, which was chosen to be 10 in the examples shown. The effect of varying g_0 will now be described.

Note that the distortion functions $dist_i(u,v)$ and $G(u,v)$, unlike the area constraint function $H(u,v)$, depend upon not only the final (Fig. 10, post-DEMP) coordinates (u,v), but also upon the original (Fig. 9, pre-DEMP) coordinates (x,y). Because the "memory" of the initial (Fig. 9, pre-DEMP) coordinates (x,y) is contained in $G(u,v)$, the DEMP converges (if $g_0 > 0$) to the optimum (Fig. 10, post-DEMP) solution (u,v) even if the starting coordinates (u_0,v_0) are not equal to (x,y). This was confirmed by performing repeated DEMPs after random perturbation of the starting coordinates (u_0,v_0).

One cannot guarantee that the Fig. 10 solution is the global minimum of $G(u,v)$ consistent with the constraint $H(u,v) = 0$. However, this solution is consistently reached over a broad range of the parameters h_0, g_0, and B_{min} even after random perturbation of the starting coordinates (u_0,v_0). As an added bonus, the NAG routine E04VDF reaches the same solution even with $g_0 = 0$ (i.e. with *no* distortion function!) provided the starting point (u_0,v_0) is equal

to the original Fig. 9 configuration (x,y), and provided the DEMP is completed in a single run. Attempting to restart from an intermediate iteration does not lead to the same solution, presumably because temporary information calculated during the DEMP has been lost. A different assumed form of the distortion function $G(u,v)$; namely, the one used in Ref. 11, is inferior because it causes the number of required iterations to increase, and the final solution depends upon the value chosen for g_0.

Fig. 14 through Fig. 16 display quantities that are useful for analyzing and comparing the distortion of various DEMP solutions. Fig. 14 shows the shape distortion function $dist_i(u,v)$ as a function of the original (Fig. 9, pre-DEMP) area A_i, for each land and water triangle in the sample test map. Fig. 15 shows the same function $dist_i(u,v)$ as a function of the final (Fig. 10, post-DEMP) area $B_i^{final}(u,v)$. Triangle 80 is unchanged in shape between Fig. 9 and Fig. 10, and has zero distortion, i.e. $dist_i(u,v) = 0$. Triangles 4 and 40 have zero final area (in Fig. 10) and have maximum shape distortion, i.e. $dist_i(u,v) = \pi/2$.

Fig. 16 shows the global shape distortion function $G(u,v)$ along the horizontal axis, and the (base 10) logarithm of the area constraint violation function $H(u,v)$ along the vertical axis, during the DEMP process. Large negative values of $\log_{10}(H(u,v))$ correspond to configurations close to the desired solution $H(u,v) = 0$. Iterations proceed from upper left to lower right; each iteration, marked by a small circle, requires about 1.2 seconds on a VAX 6610 computer. Here we chose $h_0 = 1$, $B_{min} = 0.07$, and $g_0 = 10$. The starting coordinates (u_0,v_0) were equal to (x,y); initial distortion $G(u,v)$ was zero by definition and $\log_{10}(H(u,v))$ was equal to -0.32. After 41 iterations (50 seconds on a VAX 6610) the distortion $G(u,v)$ increased to 0.83 and $\log_{10}(H(u,v))$ was reduced to -5.20. During 37 more iterations, no further significant changes occurred. The program terminated after 78 iterations, where $G(u,v)$ was still 0.83 and $\log_{10}(H(u,v))$ was further reduced to -7.66.

In summary, it appears that stable DEMP solutions, minimally distorted by our definition of $G(u,v)$ and suitable for analytic purposes, are reached under the following conditions:

- NAG routine E04VDF
- single DEMP run with no restart
- iterate until $\log_{10}(H(u,v)) \leq -5$
- area constraint function $H(u,v)$ as specified, with $h_0 = 1$ and $B_{min} = 0.07$
- $g_0 = 0$; i.e. *no* shape distortion function $G(u,v)$
- starting coordinates (u_0,v_0) equal to (x,y)

A priori we did not expect the fortunate discovery that the distortion function $G(u,v)$ can be discarded, since without the objective function the DEMP is underconstrained; i.e. has fewer

constraints than parameters. The obvious next step is to test $H(u,v)$ as an objective function in one or more *un*constrained minimization programs, which remains to be done.

A REAL MAP - SAN FRANCISCO

The revised DEMP algorithm has been tested on a map of San Francisco. For comparison with the following, Fig. 17 is a detailed map derived from the TIGER map files produced by the Bureau of the Census.[11,12] The conversion of TIGER files to other formats is the topic of another paper in this conference.[13]

In Fig. 18 is a hypothetical distribution of 10,000 cases of disease, based on the population of nonwhite children (ages 0 through 20) in 1980. Risk is assumed uniform over the entire map. Within each tract, population density is assumed uniform and cases are randomly distributed. Park and industrial areas, which can be visually identified in Fig. 17, have no population and no cases of disease. The units of analysis are 1980 census tracts. The triangulated map (not shown) has 150 tracts, 270 points, 502 triangles, and 771 segments.

In Fig. 19 is the same map after the DEMP transformation. Over the entire map, risk and population density are assumed uniform and cases are randomly distributed. If any significant clustering of cases were observed in Fig. 19, it would be attributable to unequal risk and not to variations in population density.

(Taking a shortcut for purposes of display, we actually created Fig. 19 by generating random points within its external boundary, then converted back to the "original" map of Fig. 18 via an inverse DEMP transformation.)

COMPUTING REQUIREMENTS

In Fig. 20 we plot computation time per iteration ("major" iteration) as a function of the number of points in the map. The maps *alam, test* and *sf* are those illustrated in this paper; the maps *twotri, vt500* and *vt50* were discussed in Ref. 11. The diagonal lines through the measured points have a slope of 3 on a log-log plot, indicating that computing time per iteration increases as the cube of the number of points in the map. The spacing of the diagonal lines indicates the relative effective speeds in mips (million instructions per second) of the six computers tested. The number of iterations required for a solution increases from about 50 for *vt500* and *test*, to about 100 for *vt50* and about 200 for *sf*.

On the fastest of the six computers tested, the Alameda county map *alam* requires 40 minutes per iteration. Probably 200 to 300 iterations, about 6 to 9 days, are needed to reach a solution. In addition, the memory usage of the NAG program E04VDF increases as the square of the number of points in the map; the Alameda county map requires 30 megabytes, which is close to the maximum available at LBL.

Computational limitations prevent the DEMP algorithm from being used to its full potential; presently, study areas containing more than a few hundred census tracts cannot be analyzed. In the future, several techniques for improving efficiency will be explored, including:

- conversion of the area constraint function $H(u,v)$ to an objective function, permitting *un*constrained minimization techniques to be used;
- successive subdivision of the map into subareas, with boundary points held fixed on some iterations;
- implementation on supercomputers and/or parallel computer architectures.

APPLICATIONS

Despite the present computing limitations of the technique, early prototype analyses [14,15,16] have been completed which have prompted useful discussion and have led to new insights which were not immediately obvious. The following observations suggest research topics which will be investigated and presented in future publications.

Previously described cartogram techniques correctly adjust *total* area of polygons and are useful for display purposes.[3,4,5,6,7] However the algorithm developed at LBL[8,9] is the first true *projection* that provides correct area magnification in a continuous transformation of the entire map domain. This opens the door to statistical techniques that either depend upon, or are greatly simplified by, the assumption of uniform population density. These include, but are not limited to the following (all performed on the density-equalized map):

- distance metrics, for example the average distance of cases from a fixed point;
- nearest neighbor statistics, for example the average distance of each case from its nearest neighbor;
- two-dimensional scan statistics, for example the maximum number of cases observed within a circle of fixed radius;
- approximation of the observed case distribution by a two-dimensional polynomial, whose coefficients can be determined either by likelihood maximization or direct calculation. Expected values and variances of arbitrary moments over an arbitrary polygon can be calculated from exact closed-form expressions.

Given the spatial distortion inherent in density equalized maps, their correct interpretation requires some discussion. The utility of the transformed map is in detecting non-uniformity of risk, either visually or statistically. Critics have correctly noted that distance on the transformed map does not have physical meaning and should not be used to model specific

alternatives to the null hypothesis of uniform risk. Risk, if assumed *a priori* to be non-uniform, should normally be modeled as a function of physical geographic location on the *original* map. Physical features and *assumed* equal-risk contours (for example circles on the original map) may then be projected as necessary onto the transformed map, in order to compare their locations with areas of observed elevated risk.

On the other hand the *transformed* map is ideally suited for analyzing geographic variations in *observed* risk. Moments and their variances can be reliably estimated because the spatial distribution of cases is approximately uniform. The *necessity* of a non-uniform risk model is demonstrated if one or more moments is significantly different from zero. A p-value (the probability that non-uniformity as great as that observed could have arisen by chance) can be easily evaluated on the transformed map. If (and *only* if) a statistically significant departure from uniformity is observed, stable contours of equal observed risk can be calculated on the transformed map (and then projected as necessary back onto the original map).

The DEMP method can also be used to test the *adequacy* of a model in explaining observed geographic variation of disease rates. If non-uniform risk is assumed, the target area B_i of each triangle i is made proportional to the number of *expected cases* E_i (i.e. the product of population and the assumed rate) rather than the population. (As usual, one assumes equal population density within all triangles of a given census tract.) If the observed case distribution on this *revised* DEMP is consistent with uniformity, then the assumed model is sufficient (but not necessarily required) to explain the observed geographic variation of rates.

Statistical power (the probability of detecting an effect if the underlying risk is truly non-uniform) can be estimated on the transformed map. As usual, a power calculation depends on a specific nonuniform risk model and an assumed detection criterion expressed as a p-value. (As mentioned earlier, a specific model of nonuniform risk is most appropriately described with reference to the original geographic map.)

In order to combine population subgroups having different geographic distributions (for example various time-age-sex-race categories) the target area B_i^{target} of each triangle i is again made proportional to the number of *expected cases* E_i rather than the population. The procedure is exactly analogous to an indirect age adjustment or calculation of a standard incidence ratio. One multiplies the average time-age-sex-race-disease-specific rates of the entire study area by the time-age-sex-race-specific populations of each triangle i, and sums E_i over the times, ages, sexes, races and diseases that one wishes to analyze in combination. A separate DEMP must be calculated and analyzed for each such combined group (but not for each component of the group).

In the past, some critics have objected to the non-uniqueness of the DEMP. Even though any DEMP satisfying the area constraint $H(u,v) = 0$ is theoretically valid for testing the null hypothesis of uniform risk, the non-reproducibility of results is aesthetically displeasing and allows possible trial-and-error manipulation by an overzealous analyst. Now, with a reliable

means of obtaining a unique and minimally distorted DEMP, it appears that this objection has been removed.

CONCLUSIONS

In the analysis of geographic disease distributions, the technique of density equalizing map projections (DEMP) is suggested as an alternative to conventional rate calculations. Unlike previous cartogram programs, the DEMP algorithm described in this paper is a continuous map projection that equalizes population density over the entire map domain. The solutions found are unique, reproducible and minimally distorted. The DEMP transformation permits analysis of observed disease distributions by simple statistical techniques that rely upon the assumption of uniform population density; these include but are not limited to: distance metrics, nearest neighbor statistics, two-dimensional scan statistics, and evaluation of two-dimensional polynomial coefficients by likelihood maximization or direct calculation of moments. The ability to use commercially supported minimization software opens up possibilities for future computational enhancements.

ACKNOWLEDGMENTS

The authors are grateful to Mark Rizzardi for Fig. 17; to Michal Weingart for Fig. 18 and Fig. 19; to Mark Durst for helpful suggestions regarding optimization; and to Harvard Holmes and Carl Quong for continuing interest and support. This work was supported by the Director, Office of Epidemiology and Health Surveillance; Office of Health; Office of Environment, Safety and Health; U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

REFERENCES

1. Wallace JW. Population map for health officers. Amer Jour Publ Hlth 16: 1023 (1926).

2. Gillihan AF. Population Maps. Amer Jour Publ Hlth Vol.17: 316-319 (1927).

3. Selvin S, Merrill D, Schulman J, Sacks S, Bedell L and Wong L. Transformations of maps to investigate clusters of disease. Soc Sci Med (26:2) 215-221 (1988).

4. Tobler W. *Cartogram Programs.* Cartographic Laboratory Report, Ann Arbor, Michigan (1974).

5. Dougenik JA, Chrisman NR and Niemeyer DR. An algorithm to construct continuous area cartograms. Professional Geographer 37(1): 75-81 (1985).

6. Rase, W. Bundesforschungsanstalt für Landeskunde und Raumordnung, Bonn-Bad Godesberg, Germany. Private communication (1989).

7. Wesseling, C. Faculty of Geographical Sciences, Utrecht University, Netherlands. Private communication (1991).

8. Merrill DW, Selvin S and Mohr MS. Analyzing geographic clustered response (44 pages); report LBL-30954, June 1991; invited paper presented at 1991 Joint Statistical Meetings of the American Statistical Association, Atlanta GA, August 1991. Summary version (6 pages) in proceedings, Section on Statistics and the Environment, American Statistical Association, pp.96-101, published June 1992.

9. 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.

10. Numerical Algorithms Group Inc., 1101 31st Street, Suite 100, Downers Grove IL 60515.

11. Marx RW, ed. The Census Bureau's TIGER system. Cartography and Geographic Information Systems 17, Vol.1 (1990).

12. Merrill D. Public Census Data on CD-ROM at Lawrence Berkeley Laboratory. Lawrence Berkeley Laboratory Report LBL-32165 (Rev. 1), July 1992.

13. Rizzardi MA, Mohr MS, Merrill DW and Selvin S. Processing 1990 Census TIGER map files for use in epidemiologic applications. Report LBL-32641, July 1992. Presented at Workshop on Statistics and Computing in Disease Clustering, Stony Brook NY, 23-24 July 1992. Proceedings to be published in Statistics and Medicine, August 1993.

14. Schulman J, Selvin S and Merrill DW. Density Equalized Map Projections: a method for analyzing clustering around a fixed point. Statistics in Medicine 7: 491-505 (1988).

15. Schulman J, Selvin S, Shaw G and Merrill D. Detection of excess disease near an exposure point: a case study. Arch Env Hlth 45: 168-174 (1990).

16. Shaw GM, Selvin S, Swan SH, Merrill D and Schulman J. An examination of three disease clustering methodologies. Int Jour Epi (17:4) 913-919 (1988).

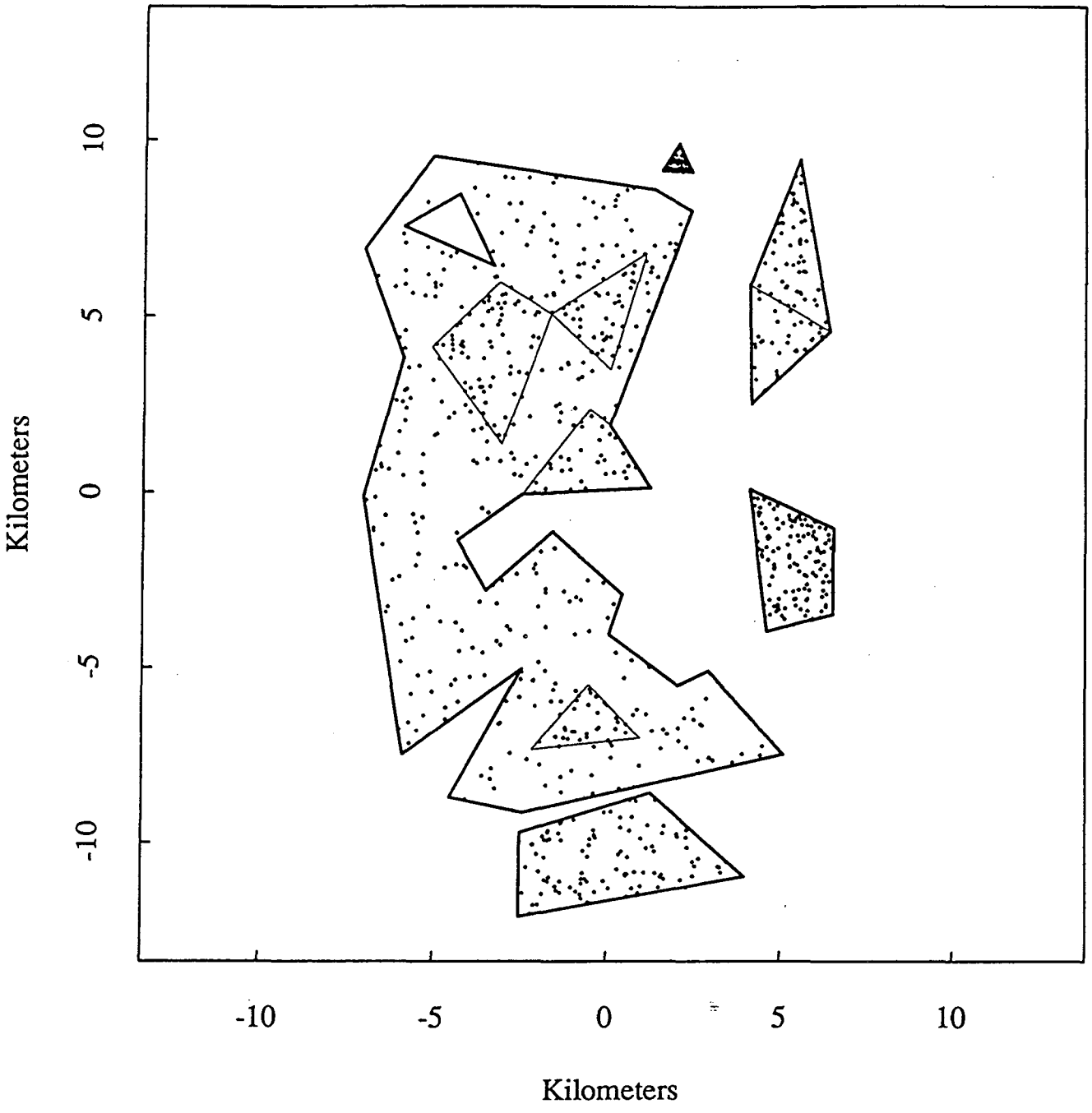

Fig. 1. Hypothetical disease distribution of sample test map, before DEMP. Solid lines are census tract boundaries. Disease risk is assumed constant. Population density is uniform within each tract; however, unequal population densities are assumed for different tracts. Cases are randomly distributed over each census tract. Fig. 1 is obtained by performing an inverse DEMP on Fig. 2.

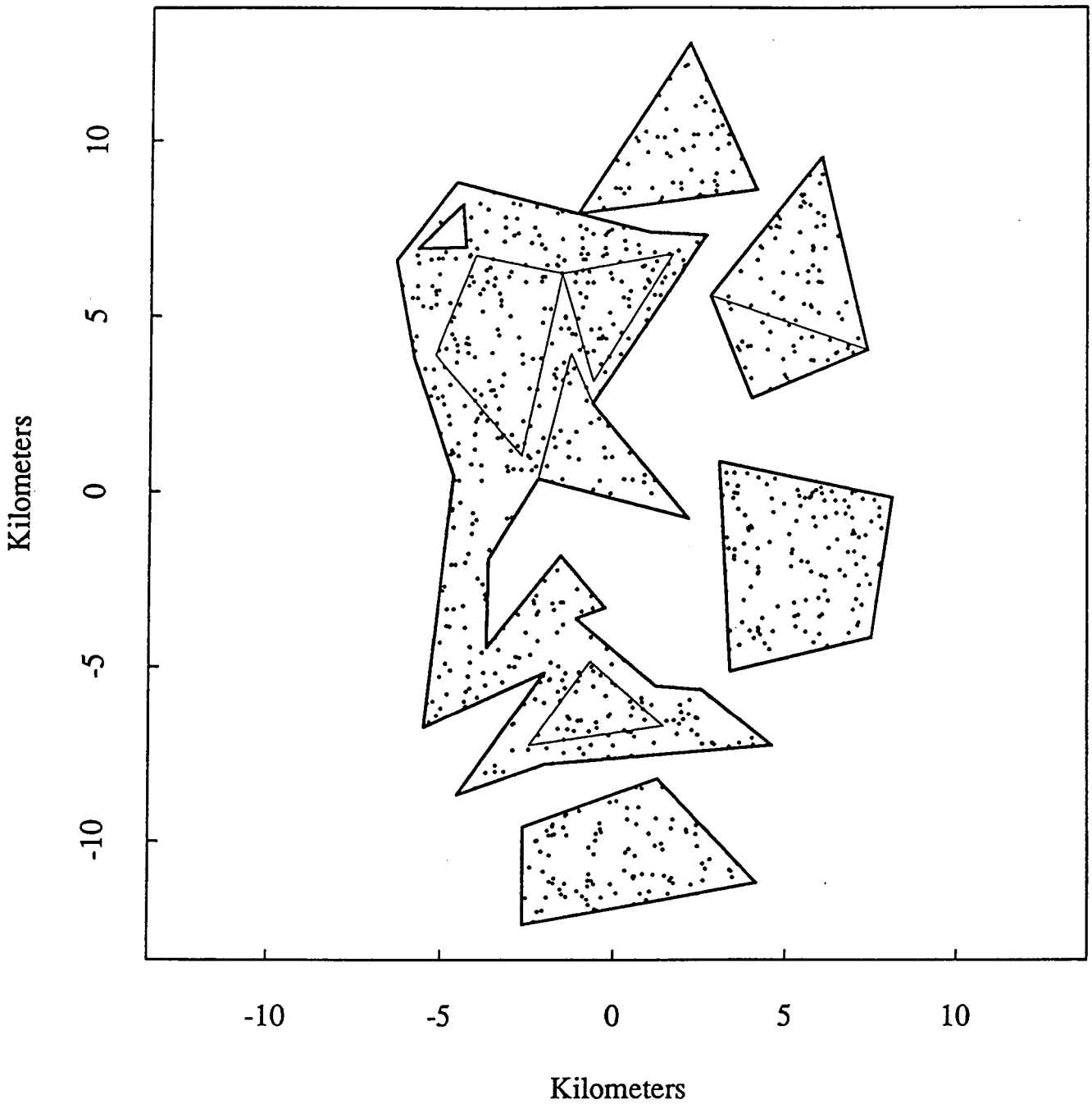

Fig. 2. Hypothetical disease distribution of sample test map, after DEMP. Solid lines are census tract boundaries. Disease risk and population density are uniform over the entire map. Cases are randomly distributed over the entire map. Fig. 2 is obtained by performing a DEMP on Fig. 1.

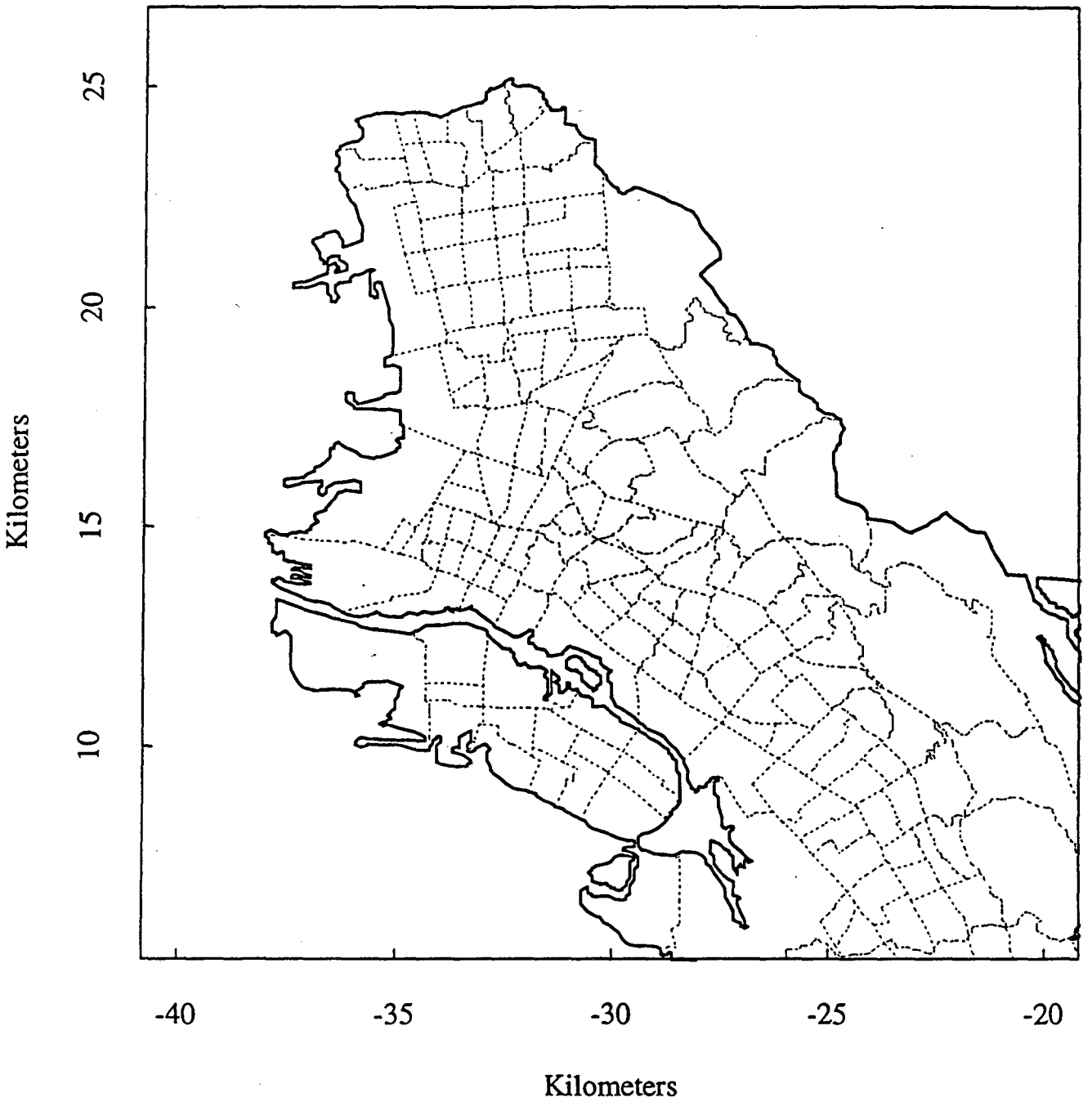

Fig. 3. Part of Alameda county, California, including Berkeley, Oakland and Alameda cities. Original map from National Planning Data Corporation. Dotted lines are 1980 Census tract boundaries.

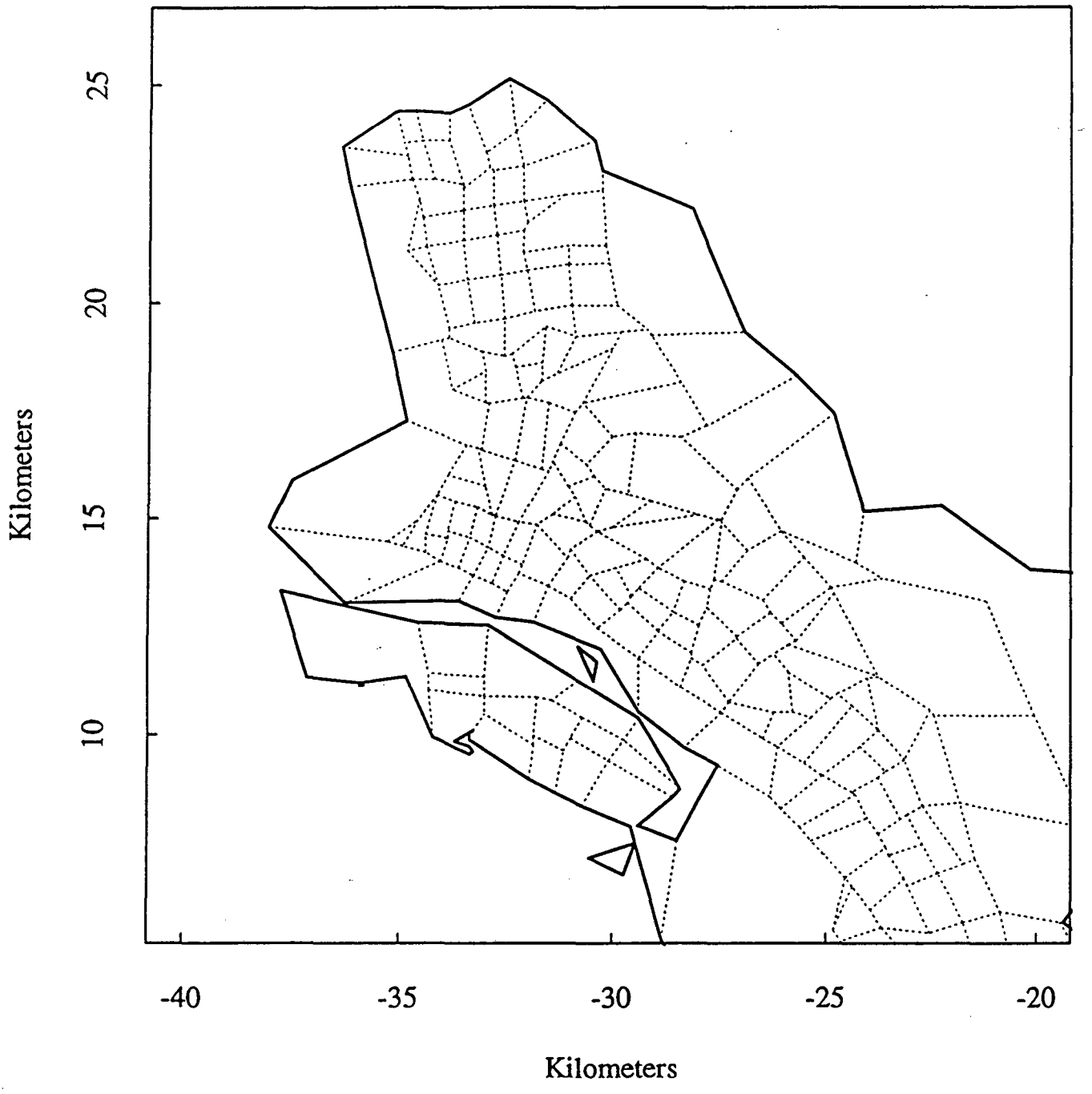

Fig. 4. Part of Alameda county, California, including Berkeley, Oakland and Alameda cities. Geographic line details smaller than one square kilometer are removed, except small triangular lakes and islands. Dotted lines are 1980 Census tract boundaries.

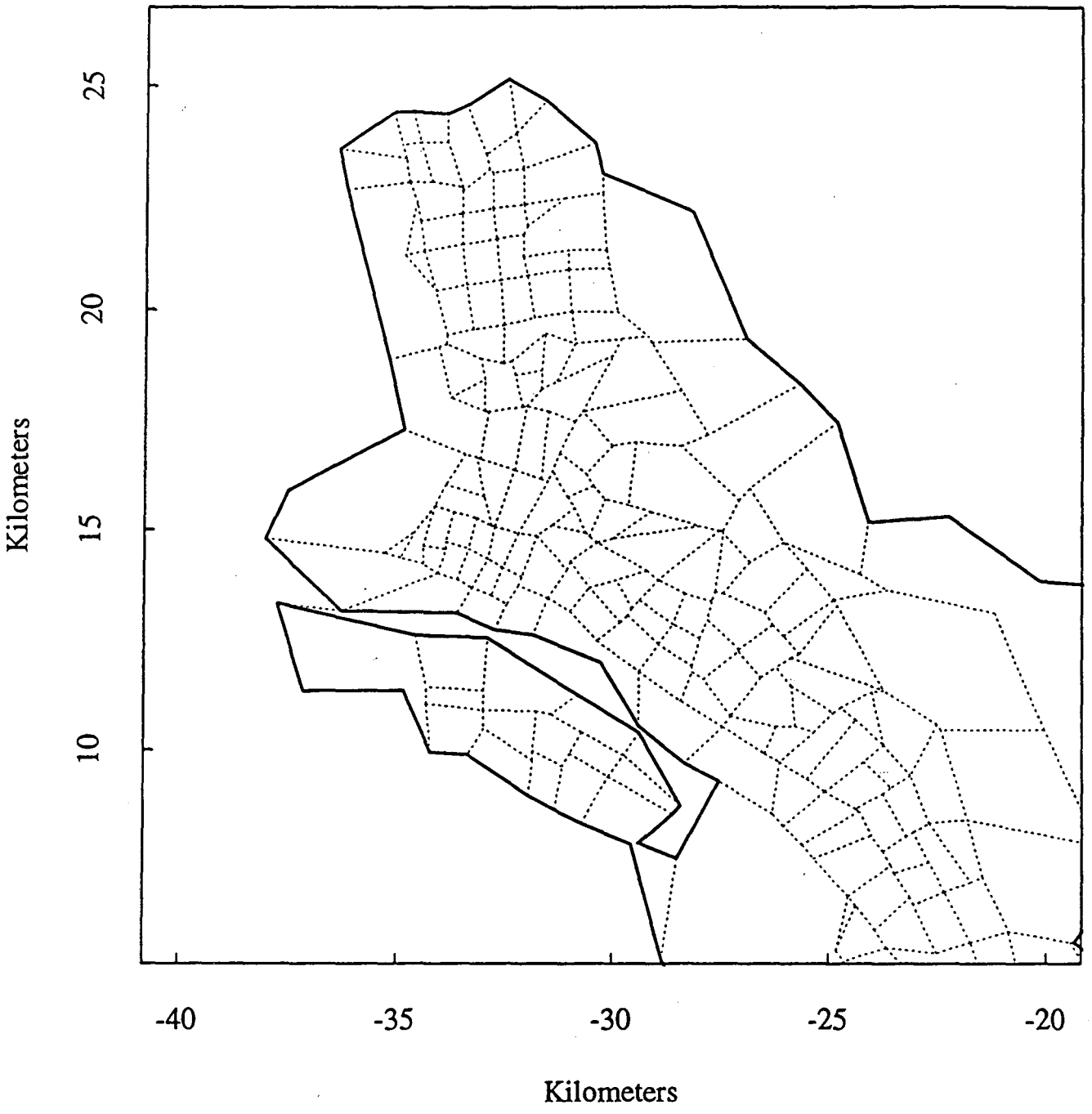

Fig. 5. Part of Alameda county, California, including Berkeley, Oakland and Alameda cities. All geographic details smaller than one square kilometer are removed. Dotted lines are 1980 Census tract boundaries.

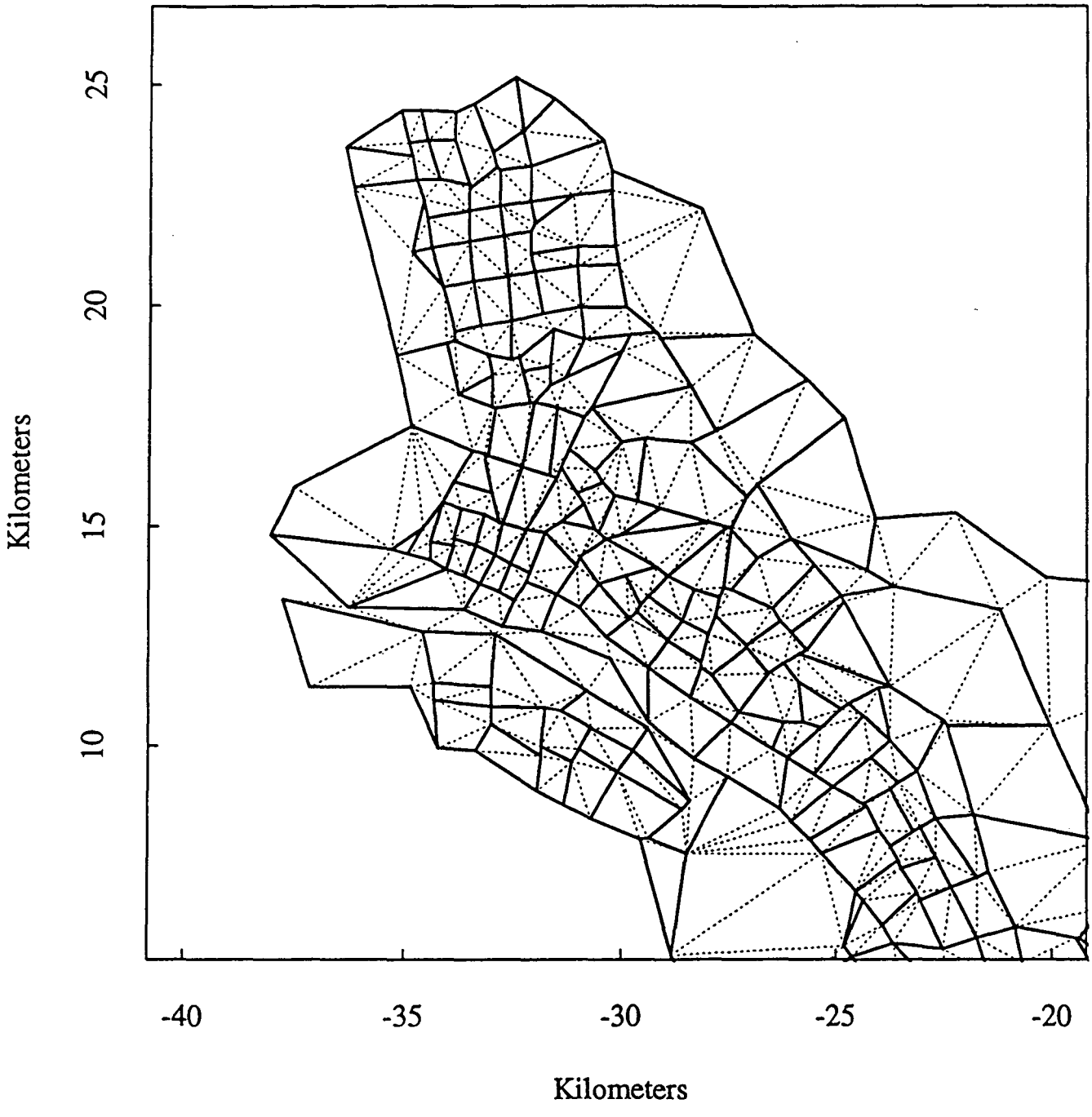

Fig. 6. Part of Alameda county, California, including Berkeley, Oakland and Alameda cities, before DEMP. Land areas and selected water areas are subdivided into triangles, indicated by dotted lines. Solid lines are 1980 Census tract boundaries.

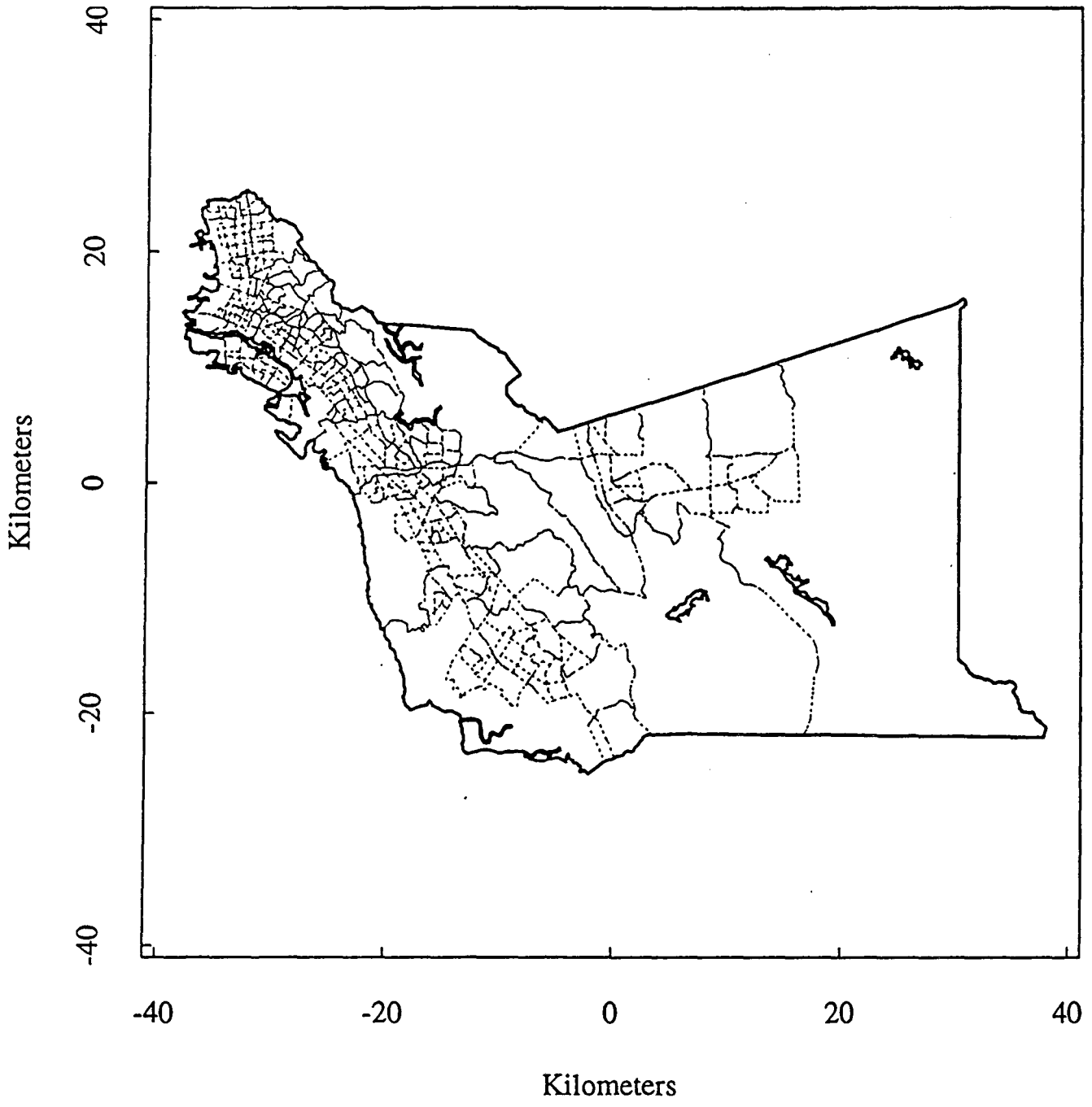

Fig. 7. Alameda county, California. Original map from National Planning Data Corporation. Dotted lines are 1980 Census tract boundaries. The map contains 289 census tracts, 305 polygons, 8605 points, and 8897 segments.

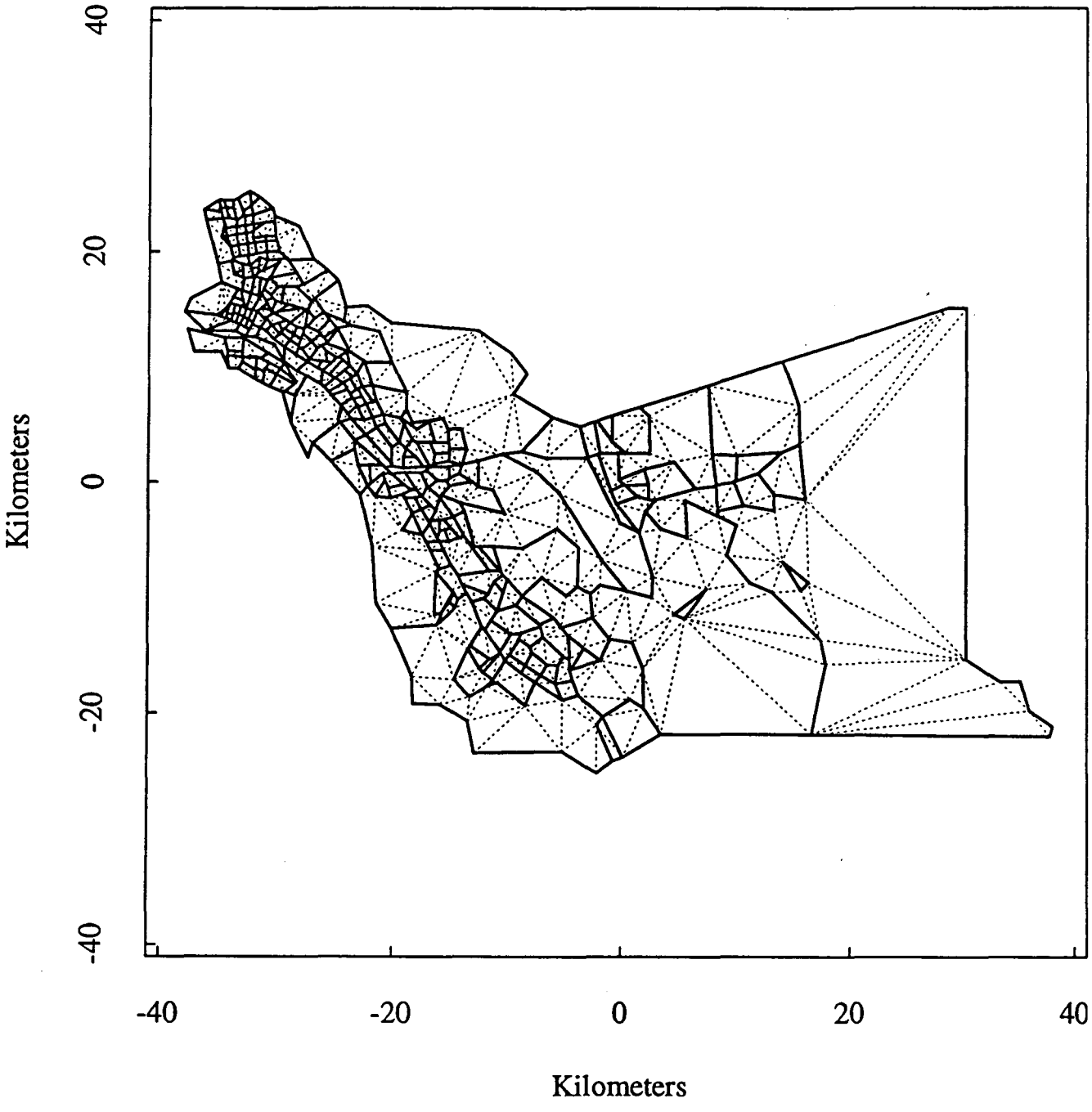

Fig. 8. Alameda county, California, before DEMP. Land areas and selected water areas are subdivided into triangles, indicated by dotted lines. Solid lines are 1980 Census tract boundaries. The map contains 289 census tracts, 581 points, 1090 triangles, and 1670 segments.

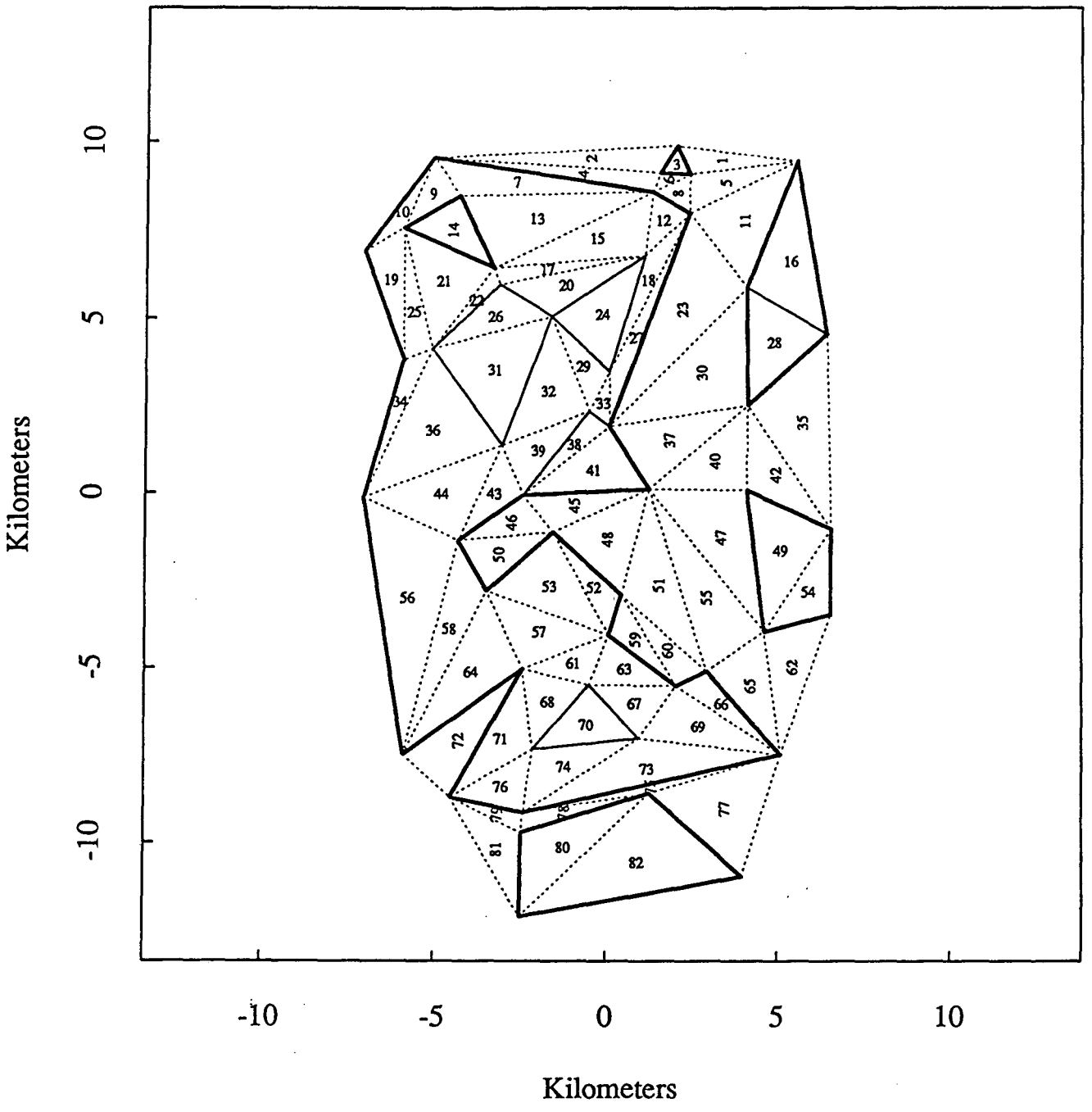

Fig. 9. Sample test map, before DEMP. Land areas and selected water areas are subdivided into triangles, indicated by dotted lines. Land and water triangles are indicated by horizontal and vertical numbers, respectively. Solid lines are census tract boundaries. The map contains 8 tracts, 49 points, 82 triangles, and 130 segments.

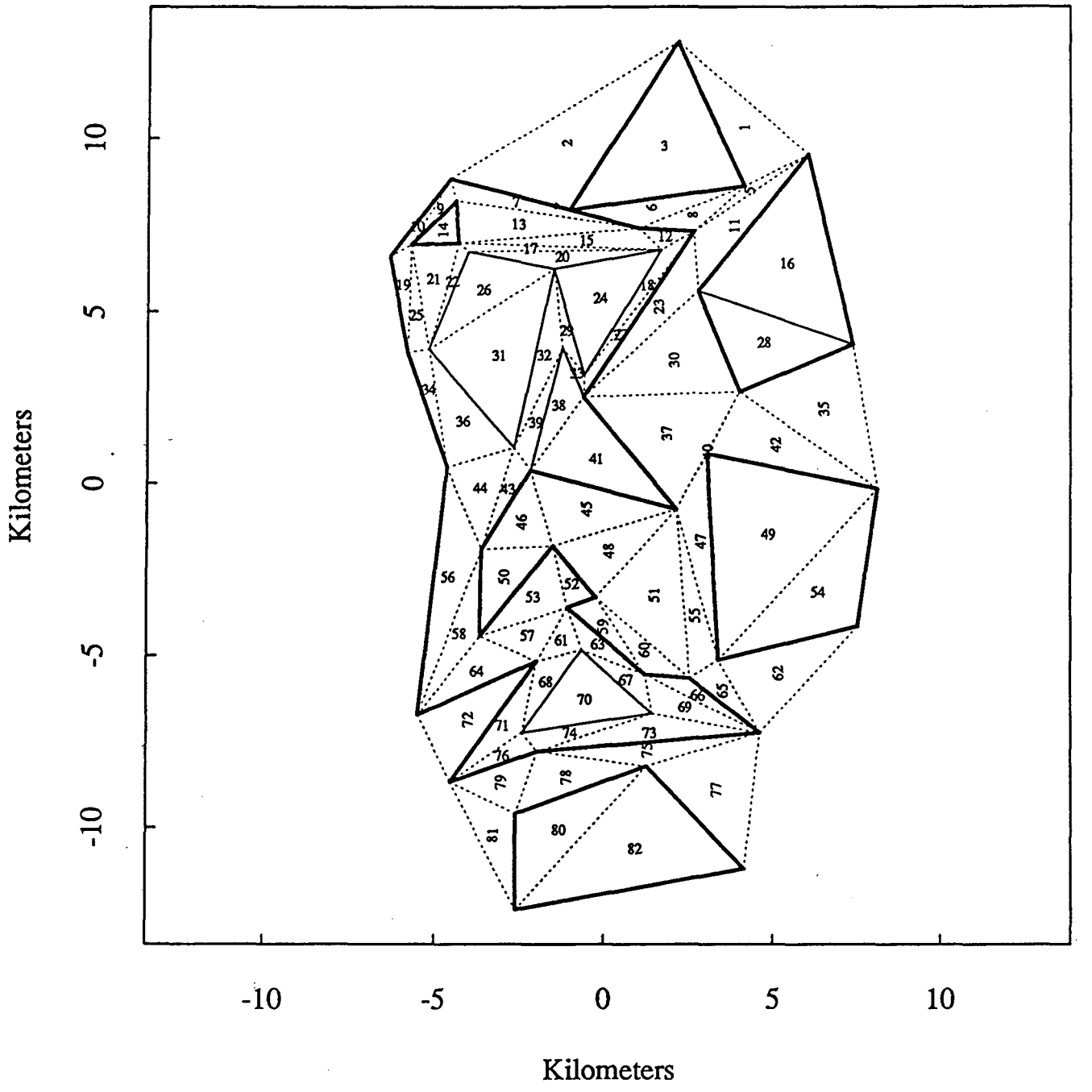

Fig. 10. Sample test map, after DEMP. Land areas and selected water areas are subdivided into triangles, indicated by dotted lines. Land and water triangles are indicated by horizontal and vertical printed numbers, respectively. Solid lines are census tract boundaries. The map contains 8 tracts, 49 points, 82 triangles, and 130 segments.

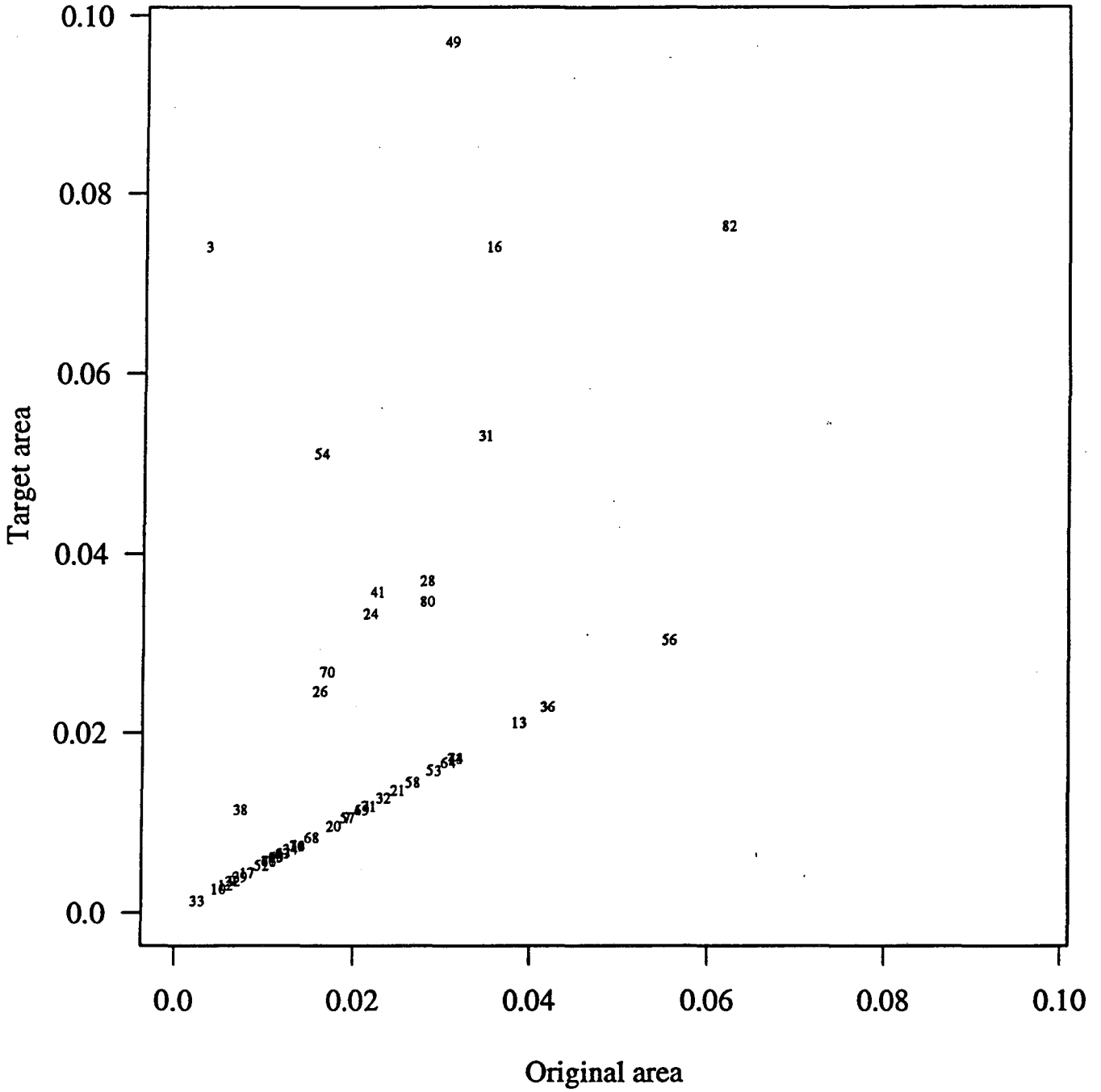

Fig. 11. Original (Fig. 9, pre-DEMP) area A_i and target area B_i^{target} of each land triangle in the sample test map. (Target areas of water triangles are not specified.) The ratio of target area to original area indicates the area magnification to be applied to each land triangle during the DEMP. The numbers 56,36,13... correspond to triangles in a single large tract, all of which receive the same magnification. Triangle 3 receives the maximum magnification.

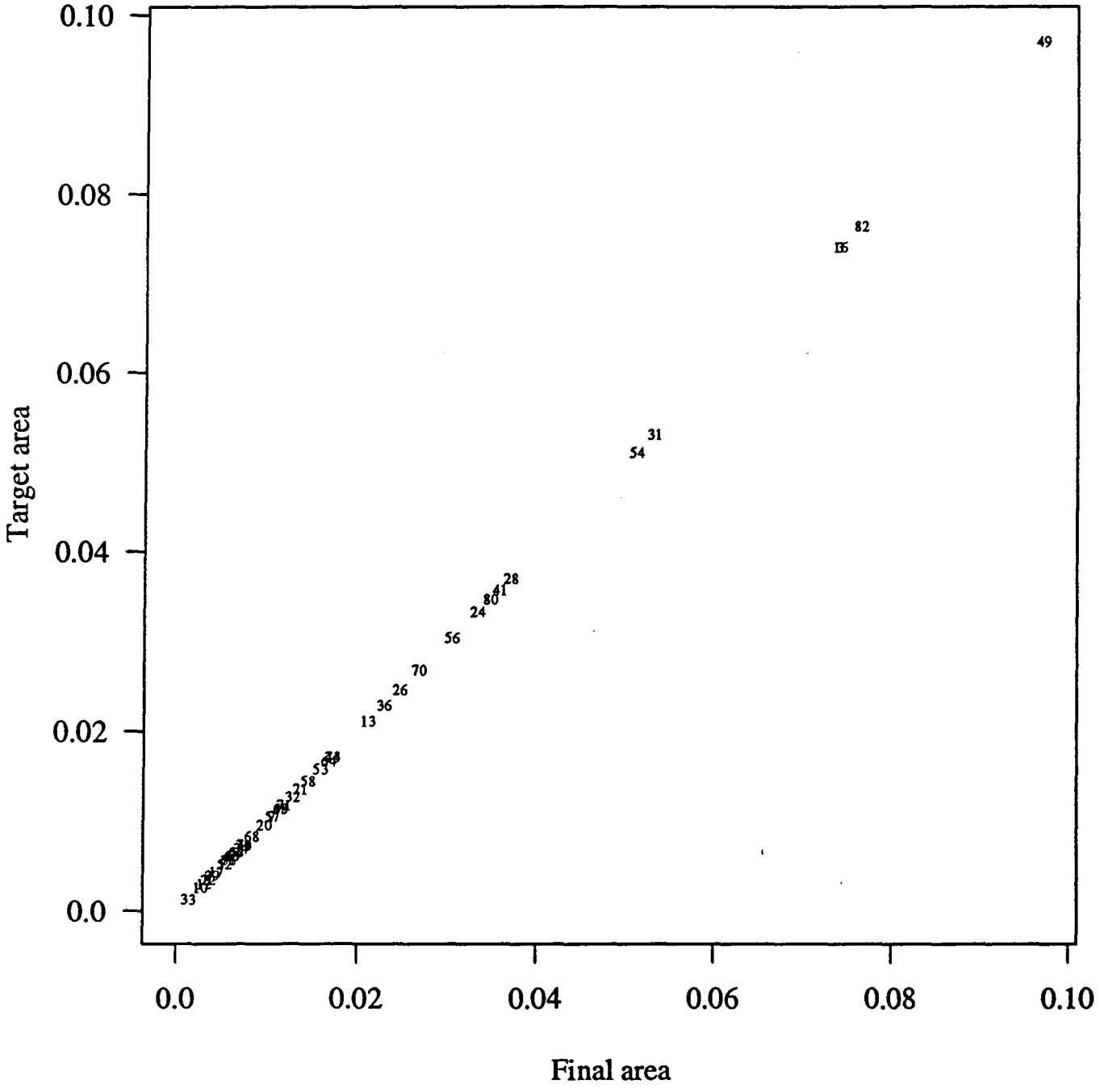

Fig. 12. Final (Fig. 10, post-DEMP) area $B_i^{final}(u,v)$ and target area B_i^{target} of each land triangle in the sample test map. (Target areas of water triangles are not specified.) For every land triangle final area and target area are equal, showing that land triangle areas in Fig. 10 have been correctly adjusted.

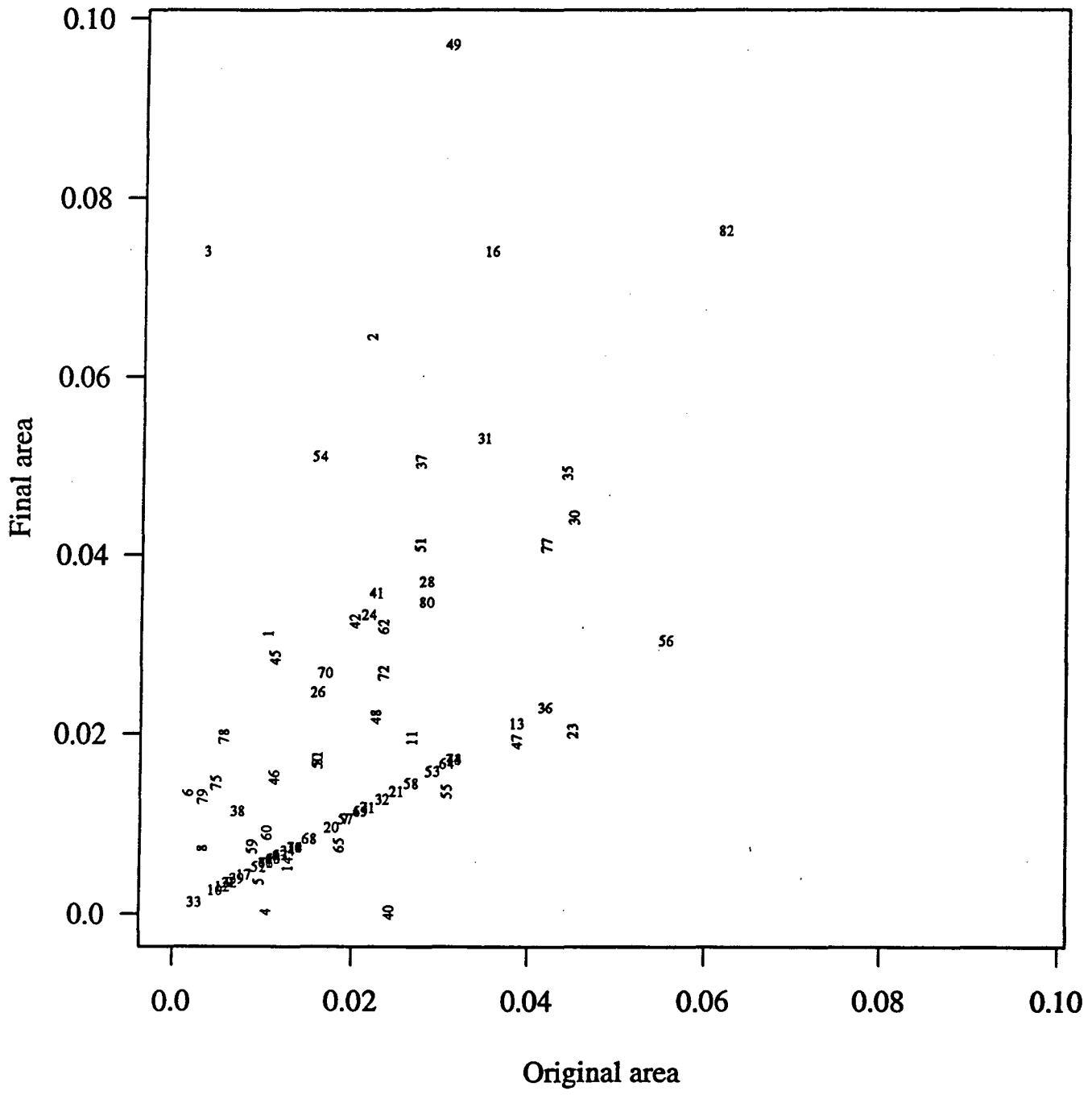

Fig. 13. Original (Fig. 9, pre-DEMP) area A_i and final (Fig. 10, post-DEMP) area $B_i^{\text{final}}(u,v)$ of each land and water triangle in the sample test map. Land and water triangles are indicated by horizontal and vertical numbers, respectively. Water triangles 4 and 40 have final areas equal to zero, as shown in Fig. 10. Negative triangle areas, which would correspond to upside-down triangles and self-crossing polygon boundaries, are not permitted.

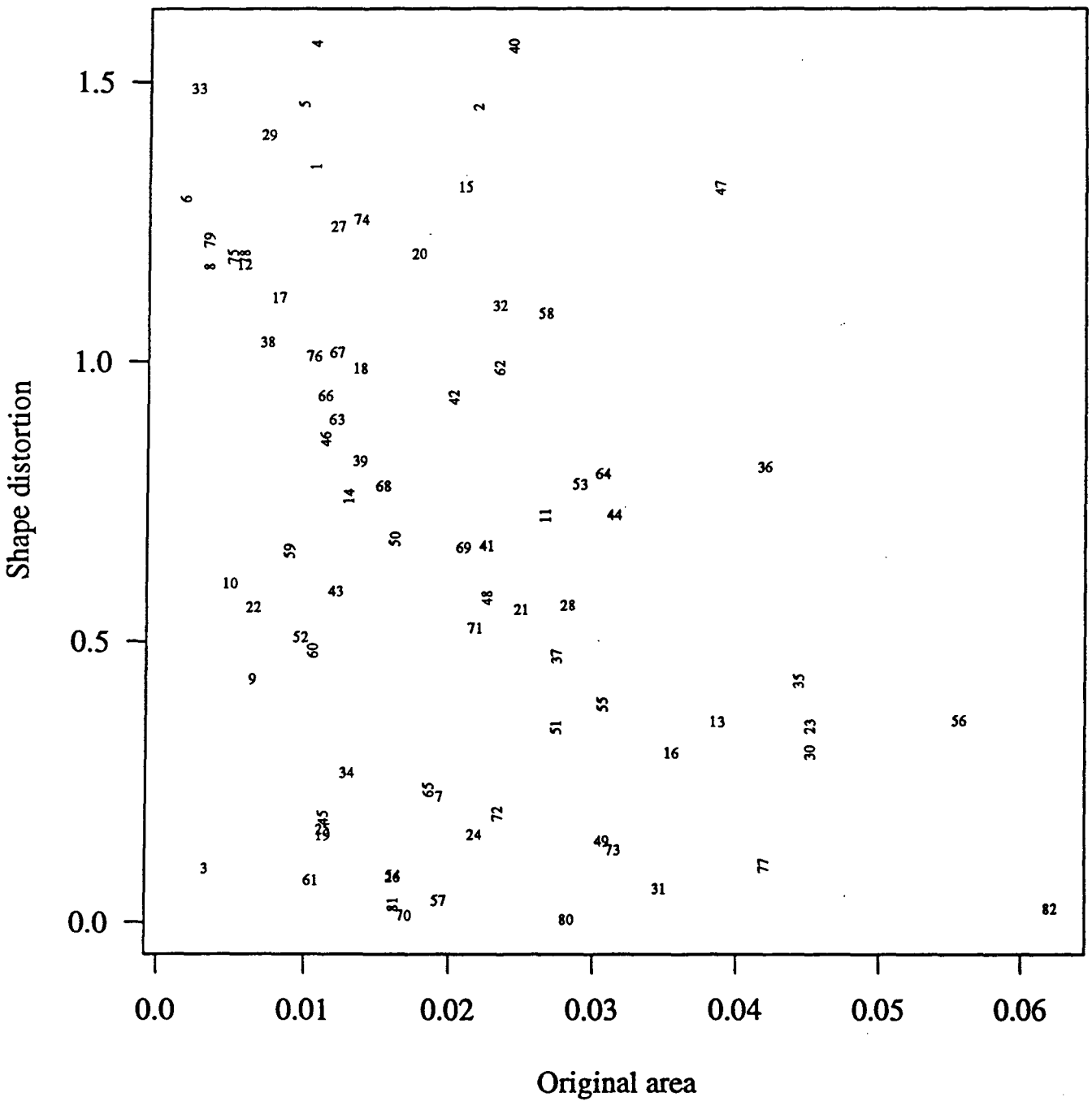

Fig. 14. Original (Fig. 9, pre-DEMP) area A_i and shape distortion $dist_i(u,v)$ of each land and water triangle in the sample test map. Shape distortion measures the change in shape of each triangle, neglecting rotation and magnification. Land and water triangles are indicated by horizontal and vertical numbers, respectively. Triangle 80 is unchanged in shape between Fig. 9 and Fig. 10, and has zero distortion. Triangles 4 and 40 have zero final area (in Fig. 10) and have maximum shape distortion (equal to $\pi/2$).

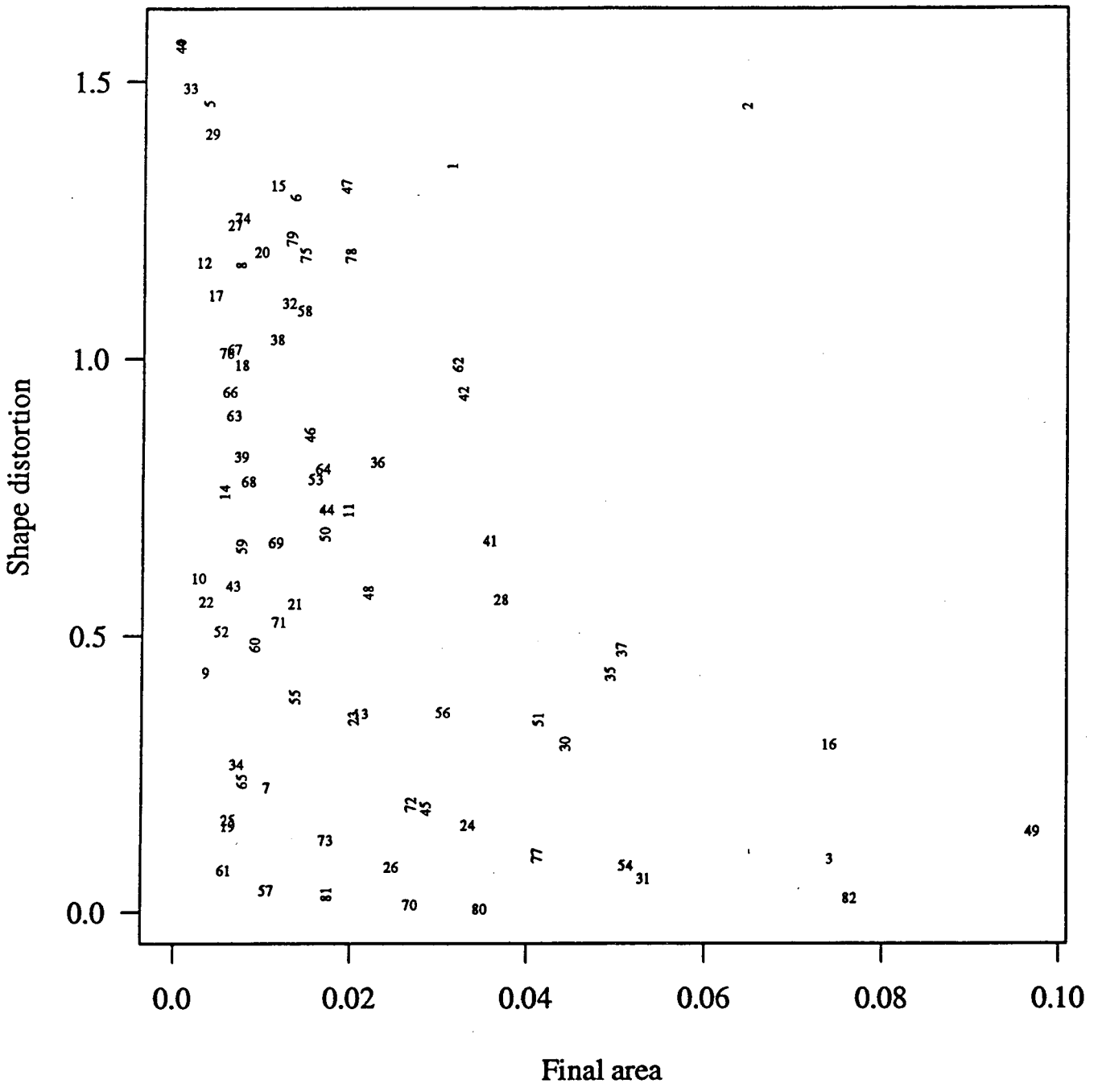

Fig. 15. Final (Fig. 10, post-DEMP) area $B_i^{final}(u,v)$ and shape distortion $dist_i(u,v)$ of each land and water triangle in the sample test map. Shape distortion measures the change in shape of each triangle, neglecting rotation and magnification. Land and water triangles are indicated by horizontal and vertical numbers, respectively. Triangle 80 is unchanged in shape between Fig. 9 and Fig. 10, and has zero shape distortion. Triangles 4 and 40 have zero final area (in Fig. 10) and have maximum shape distortion (equal to $\pi/2$).

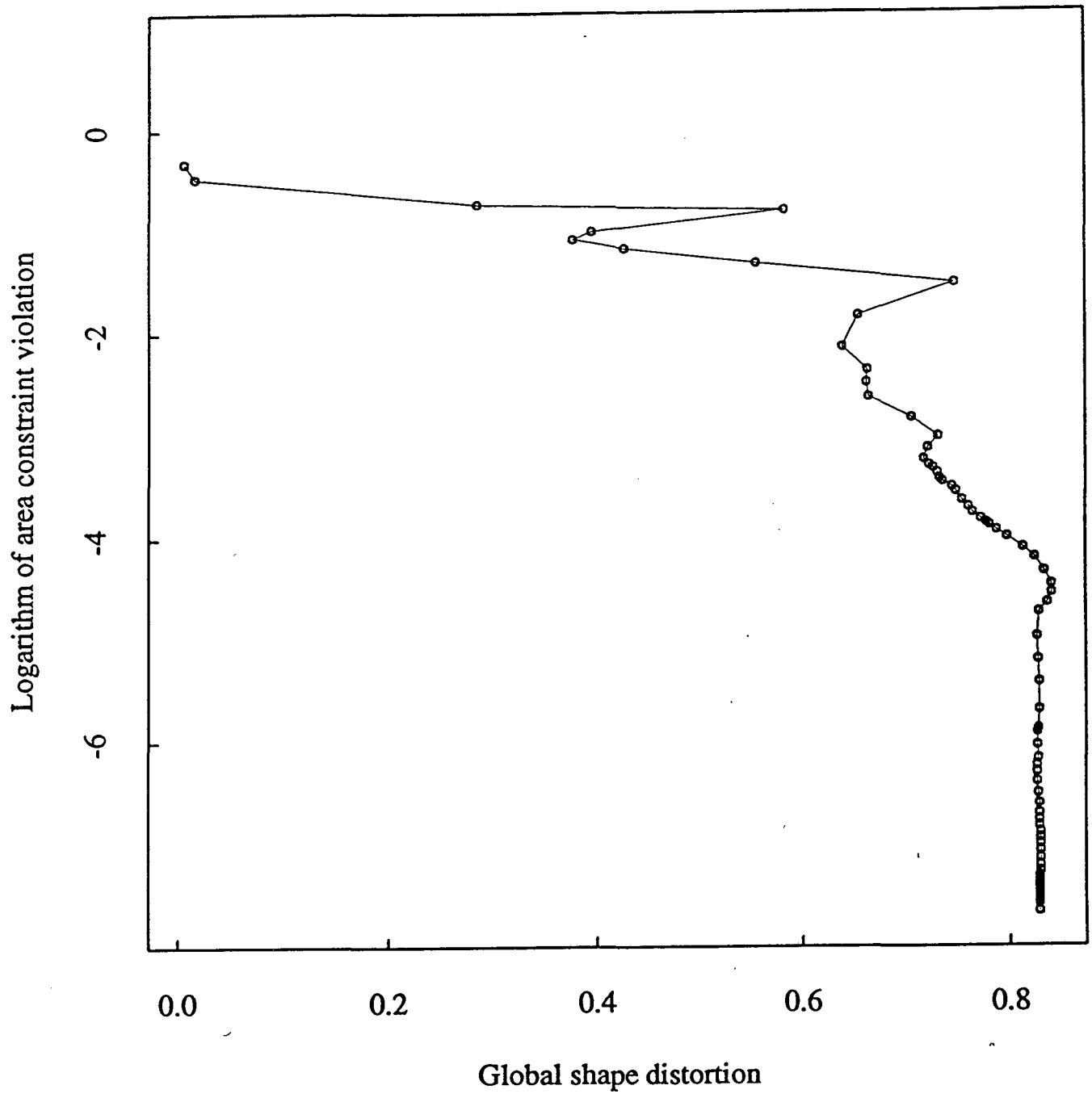

Fig. 16. Global shape distortion function $G(u,v)$ on horizontal axis, and (base 10) logarithm of area violation $H(u,v)$ on vertical axis, during the DEMP of the sample test map. Large negative values of $\log_{10}(H(u,v))$ correspond to configurations increasingly close to the desired solution $H(u,v) = 0$. Each iteration, marked by a small circle, requires about 1.2 seconds on a VAX 6610. Iterations proceed from upper left to lower right. Before the DEMP, $G(u,v)$ is zero by definition and $\log_{10}(H(u,v))$ is -0.32. After 41 iterations, $G(u,v)$ has increased to 0.83 and $\log_{10}(H(u,v))$ has decreased to -5.20. No significant further change is observed up to 78 iterations, where $G(u,v)$ is still 0.83 and $\log_{10}(H(u,v))$ is -7.66.

Fig. 17. Map of San Francisco, derived from 1990 TIGER geographic base map produced by the U.S. Bureau of the Census.

Fig. 18. Hypothetical disease distribution among non-white children in San Francisco, before a DEMP. Disease risk is assumed constant. Population density is assumed to be uniform within each tract; however, population densities are unequal for different tracts. Disease cases are randomly distributed within each tract; the number of cases in each tract is proportional to the 1980 population of non-white children (ages 0 through 21). The map contains (not shown) 150 tracts, 270 points, 502 triangles, and 771 segments. Fig. 18 is obtained by performing an inverse DEMP on Fig. 19.

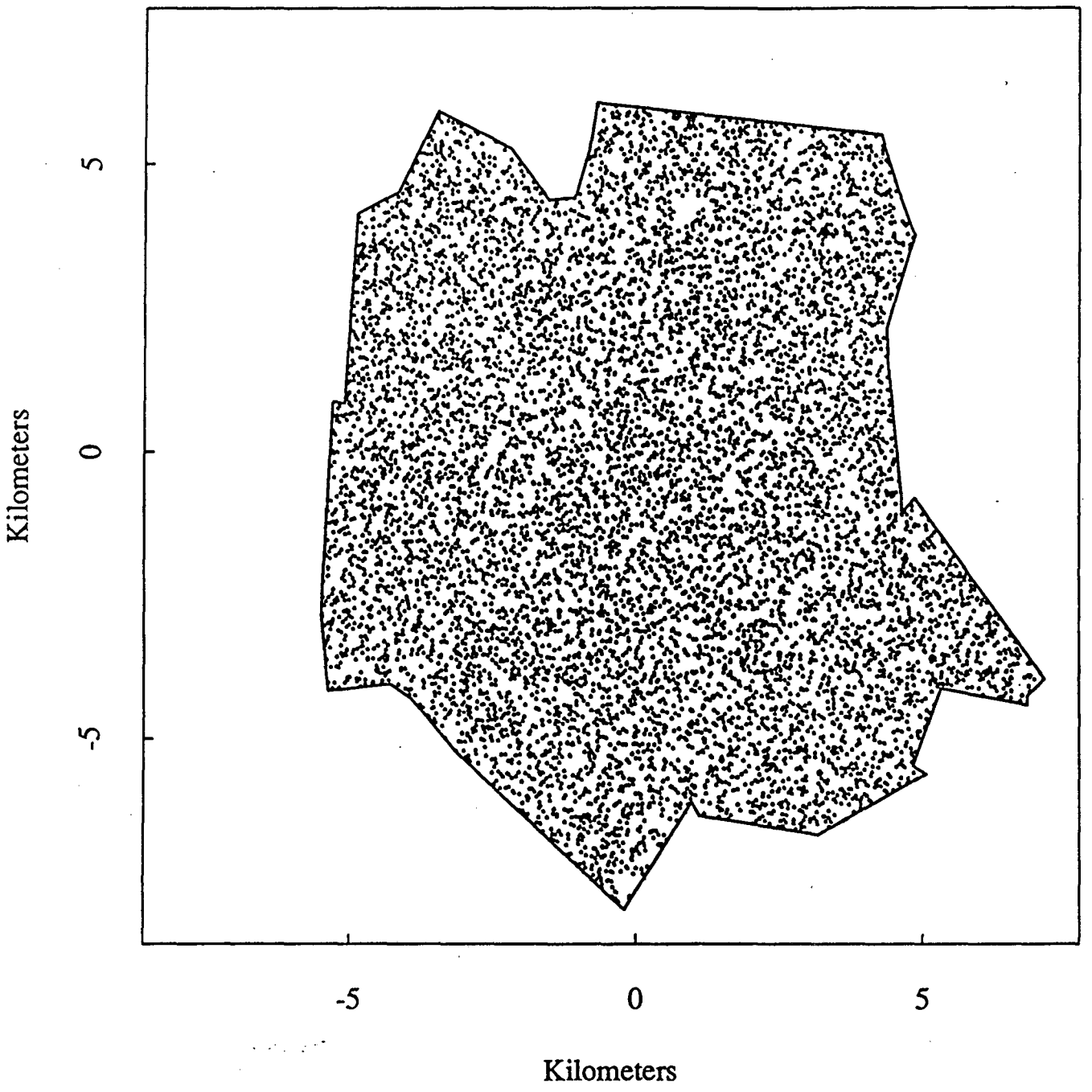

Fig. 19. Hypothetical disease distribution among non-white children in San Francisco, after a DEMP. Disease risk and population density are constant over the entire map. Disease cases are randomly distributed over the entire map. The map contains (not shown) 150 tracts, 270 points, 502 triangles, and 771 segments. Fig. 19 is obtained by performing a DEMP on Fig. 18.

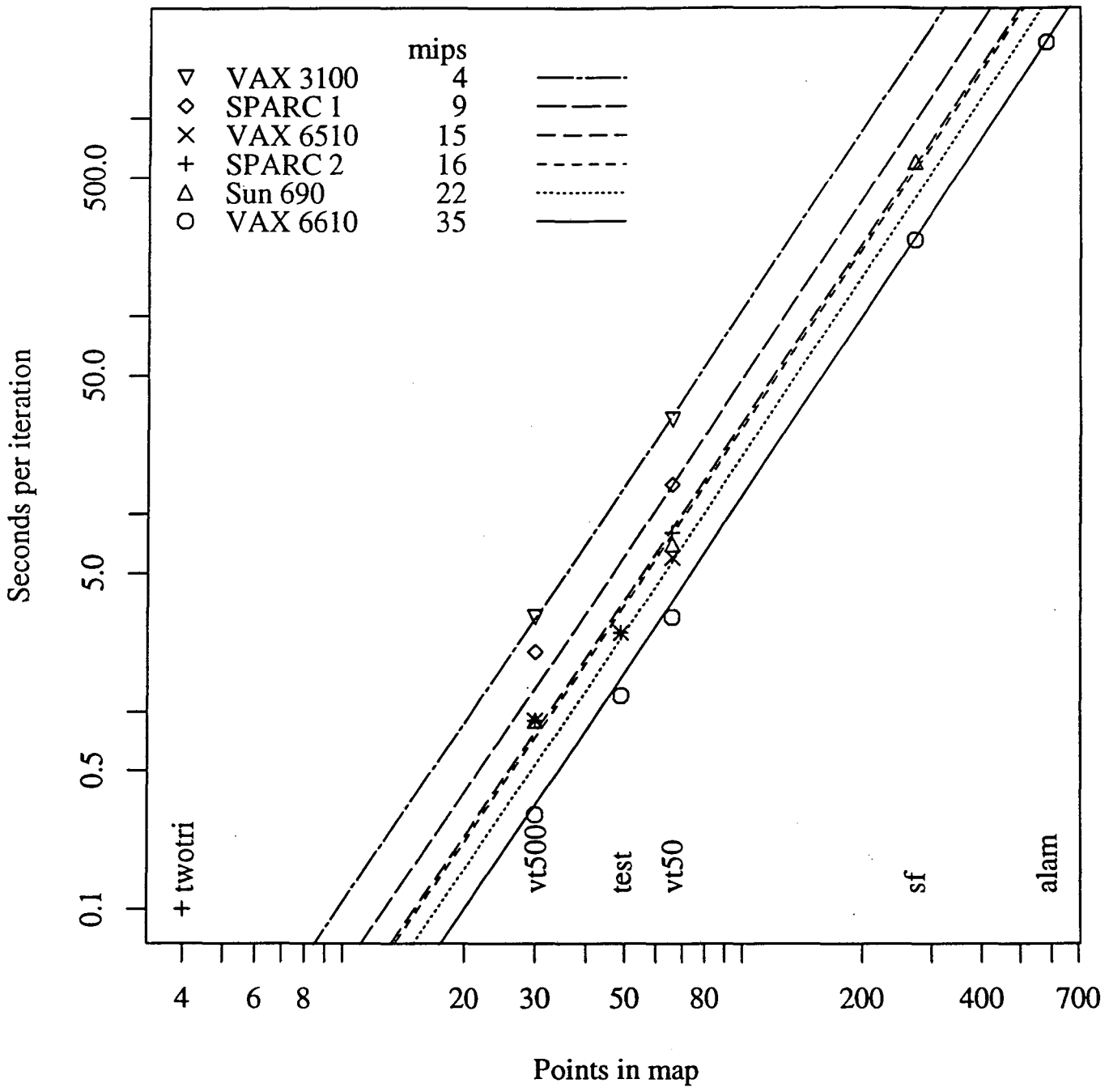

Fig. 20. Computing requirements (seconds per iteration as a function of the number of points in the map), for six different maps and six different computers. For example, the sample map *test* has 49 points and requires 1.2 seconds per iteration on a VAX 6610, a 35 mips (million instructions per second) computer. The diagonal lines have a slope of 3 on a log-log plot, showing that computing time increases as the cube of number of points. The spacing of the diagonal lines indicates the relative observed speeds of the six computers, in performing the DEMP calculation.

SOURCE DATA

Text	csr6	c:\docs\stonybr\stonybr.11 8/6/92
Fig		
1-2	parep2	/home/parep2/data1/merrill/Puff/Version4a/ plot.com /home/parep2/data1/merrill/Plot.routines/ graphpair.com /home/parep2/data1/merrill/Puff/Version4a/ Output/edime.out.0255 Output/edime.out.0256 Output/vars.out.0255 Output/vars.out.0256 .tmp_pts_rnd_demp .tmp_pts_xxx_geo .tmp_area_1 .tmp_area_2
3-8	parep2	/home/mss/icsd/biostat/merrill/demp/alameda/ detail.doc alam.edime alam_1k2.edime alam_1k2_fix8.edime alam_1k2_fix8_tri.edime
9-10	parep2	/home/parep2/data1/merrill/Puff/Version4a/ plot.com /home/parep2/data1/merrill/Plot.routines/ graphpair.com /home/parep2/data1/merrill/Puff/Version4a/ Output/edime.out.0255 Output/edime.out.0256 Output/vars.out.0255 Output/vars.out.0256 .tmp_area_1 .tmp_area_2
11-15	parep2	/home/parep2/data1/merrill/Puff/Version4a/ plot.com Graphics/graph3.com .tmp_graph3_in .tmp_area Graphics/graph3.S Output/vars.out.0256
16	parep2	/home/parep2/data1/merrill/Puff/Version4a/ .tmp_graph2_in Graphics/graph2.com Graphics/graph2.S Output/its.out.0256
17	parep2	/home/parep2/data1/merrill/rizzardi/Smap/fig7/deane.ps
18-19	parep2	/home/parep2/data1/merrill/Puff/Version4a/ sf80_1k2_fix2_tri/nw_ch/detail.doc
20	parep2	/home/parep2/data1/merrill/Puff/Version4a/ .tmp_graph4_in Graphics/graph4.S

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