Lawrence Berkeley National Laboratory
Recent Work

Title

Interpoint Squared Distance as a Measure of Spatial Clustering

Permalink

https://escholarship.org/uc/item/4jh228bv

Authors

Selvin, S.
Schulman, J.
Merrill, D.W.

Publication Date

1991-02-01

Lawrence Berkeley Laboratory

UNIVERSITY OF CALIFORNIA

Information and Computing Sciences Division

Interpoint Squared Distance as a Measure of Spatial Clustering

S. Selvin, J. Schulman, and D.W. Merrill

February 1991

DISCLAIMER

DISCLAIMER

INTERPOINT SQUARED DISTANCE AS A MEASURE OF SPATIAL CLUSTERING

February 1991

Submitted to the Journal of the Royal Statistical Society

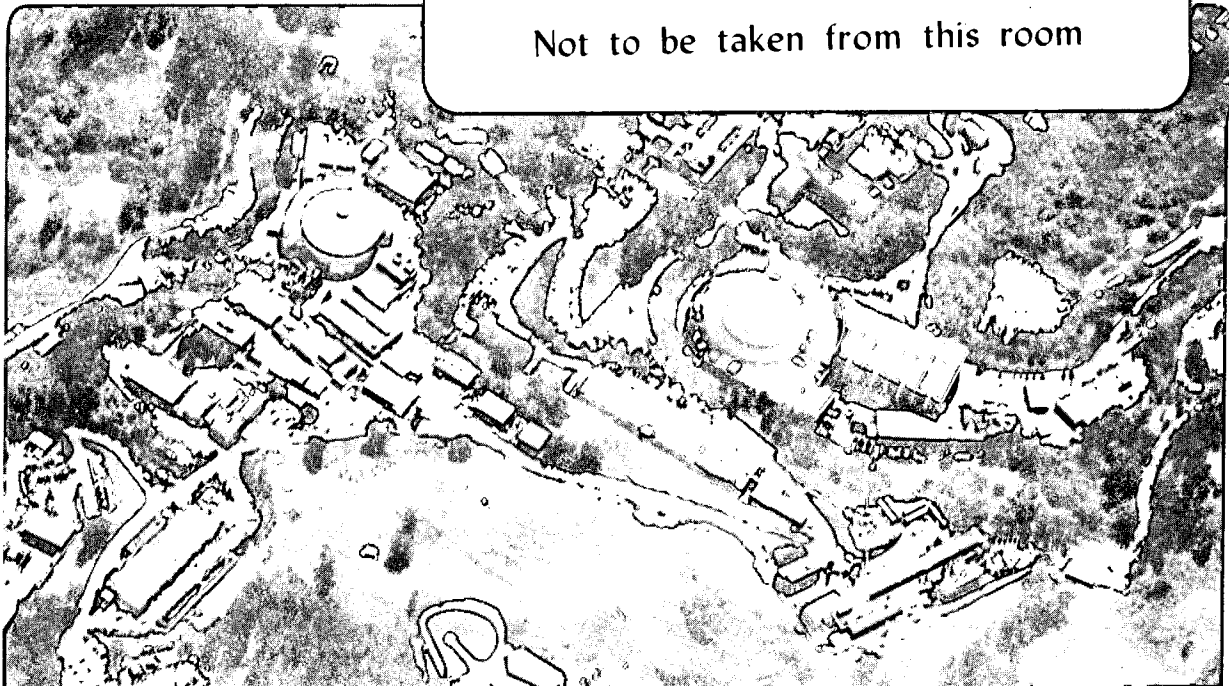# INTERPOINT SQUARED DISTANCE AS A MEASURE OF SPATIAL CLUSTERING

S. Selvin, Ph.D.[1,3]

J. Schulman, Ph.D.[1,2]

D. W. Merrill, Ph.D.[1,3]

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

[2]California Birth Defects Monitoring Program, March of Dimes Birth Defects Foundation, 5900 Hollis Street, Emeryville, CA 94608.

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

Interpoint Squared Distance as a Measure of Spatial Clustering

ABSTRACT

The expectation and variance for the mean interpoint squared distance are presented. In order to evaluate these expressions it is necessary to calculate the moments of a bivariate uniform distribution defined for an arbitrary polygon. Expressions for these moments are also presented, allowing the mean interpoint squared distance to be used as a measure of spatial clustering. The distribution and power of this test statistic is explored on the unit square, and the spatial distribution of 11 cases of non-Hodgkin's lymphoma is investigated to illustrate an application of the technique.

running head: Mean Interpoint Squared Distance

key words: mean interpoint squared distance; moments; power; spatial analysis

1. INTRODUCTION

The study of spatial distributions has produced a variety of statistical measures to examine possible clustering among a series of points. Some methods are based on measuring distances to the nearest neighbor (e.g., Ripley, 1981; Boots, 1988), others on specialized counts (e.g., Cuzick, 1990; Cliff, 1988), and a few directly employ distance (e.g., Schulman, 1988; Whittemore, 1987). In general, these statistics reflect departure from a null hypothesis which states that the spatial distribution of the observed points arises from a homogeneous Poisson process. Measures of spatial clustering are not equally useful. The utility of any measure of clustering depends on the configuration of the points that make up the spatial pattern (null hypothesis false). It is not difficult to postulate patterns where one method works well and another is ineffective.

This paper describes the mean interpoint squared distance among all possible pairs of points as a test statistic for the analysis of spatial data, providing an additional approach to spatial analysis. Also included are simulation results that indicate the accuracy of a normal distribution as a way to assess this test statistic when the observed data are uniformly distributed on a unit square. A simple model is postulated to define the power of the interpoint squared distance and to compare its effectiveness as a measure of spatial clustering to the more usual Poisson and nearest-neighbor approaches. A small set of non-Hodgkin's lymphoma cases is used to illustrate.

2. MEAN INTERPOINT SQUARED DISTANCE

Assume that n independent observations are selected from a predefined region, with locations denoted by (x_i, y_i). A measure of spatial clustering based directly on distance is the mean of the interpoint squared distances among all possible pairs of n points or

$$\overline{d^2} = \frac{2}{n(n-1)} \sum\sum [(x_i - x_j)^2 + (y_i - y_j)^2] \tag{1}$$

for all pairs $i > j$. Under the condition that the points (x_i, y_i) arise independently from a homogeneous Poisson process associated with a defined region, the expectation and variance can be derived (Schulman, 1986). The expressions are

$$expectation = E\overline{D^2} = 2[EX^2 + EY^2] \tag{2}$$

and

$$variance(\overline{D^2}) = \sigma_d^2 = \frac{2}{n(n-1)} \left\{ 2(n-1)\left[EX^4 + EY^4\right] + 4(n-1)\left[EX^2Y^2 - EX^2EY^2\right] \right. \tag{3}$$
$$\left. - 2(n-3)\left[[EX^2]^2 + [EY^2]^2\right] + 8[EXY]^2 \right\}$$

The symbol EX^k and EY^k represents the kth central moment of the distribution of X and Y, respectively.

2.1 MOMENTS ASSOCIATED WITH AN ARBITRARY POLYGON

To evaluate the expectation and variance given by expressions (2) and (3), it is necessary to calculate the moments associated with the variables X and Y over the region of interest. Consider an arbitrary polygon whose boundary is described by the sequence of m points $(x_1,y_1), (x_2,y_2), \cdots ,(x_m,y_m)$. The noncentral moments associated with random variables X and Y (denoted by subscript "o") when the point (X, Y) is uniformly distributed over this polygon are given by the following:

$$area = A = \frac{1}{4}\sum_{i=1}^{m} w_i \tag{4}$$

and

$$EX_o = \frac{1}{12A}\sum_{i=1}^{m} w_i S_{xi}, \tag{5}$$

$$EX_o^2 = \frac{1}{32A}\sum_{i=1}^{m} w_i (S_{xi}^2 + \frac{1}{3}D_{xi}^2),$$

$$EX_o^3 = \frac{1}{80A}\sum_{i=1}^{m} w_i (S_{xi}^3 + S_{xi}D_{xi}^2), \text{ and}$$

$$EX_o^4 = \frac{1}{192A}\sum_{i=1}^{m} w_i (S_{xi}^4 + 2S_{xi}^2 D_{xi}^2 + \frac{1}{5}D_{xi}^4).$$

where $w_i = S_{xi}D_{yi} - S_{yi}D_{xi}$ with $S_{xi} = x_{i+1} + x_i$, $D_{xi} = x_{i+1} - x_i$, $S_{yi} = y_{i+1} + y_i$, and $D_{yi} = y_{i+1} - y_i$. For $i = m$, $x_{m+1} = x_1$, and $y_{m+1} = y_1$. Counterclockwise ordering of the points implies $A > 0$. The non-central moments associated with the variable Y are achieved by replacing x with y in the four expressions for X (5). Furthermore, moments of the joint distribution of X and Y are

$$EX_o Y_o = \frac{1}{32A}\sum_{i=1}^{m} w_i (S_{xi}S_{yi} + \frac{1}{3}D_{xi}D_{yi}), \tag{6}$$

$$EX_o^2 Y_o = \frac{1}{80A}\sum_{i=1}^{m} w_i (S_{xi}^2 S_{yi} + \frac{2}{3}S_{xi}D_{xi}D_{yi} + \frac{1}{3}D_{xi}^2 S_{yi}),$$

$$EX_o Y_o^2 = \frac{1}{80A}\sum_{i=1}^{m} w_i (S_{xi}S_{yi}^2 + \frac{2}{3}S_{yi}D_{xi}D_{yi} + \frac{1}{3}S_{xi}D_{yi}^2), \text{ and}$$

$$EX_o^2 Y_o^2 = \frac{1}{192A}\sum_{i=1}^{m} w_i (S_{xi}^2 S_{yi}^2 + \frac{1}{3}S_{xi}^2 D_{yi}^2 + \frac{4}{3}S_{xi}D_{xi}S_{yi}D_{yi} + \frac{1}{3}D_{xi}^2 S_{yi}^2 + \frac{1}{5}D_{xi}^2 D_{yi}^2).$$

To obtain expressions (4) - (6), observe that the noncentral moment $EX_0^k Y_0^l$ is the integral of a function $x^k y^l$ over the polygon, divided by the polygon area A. By choosing an arbitrary point (x_0, y_0), one can consider an m-sided polygon as the sum of m triangles, where triangle i is formed by (x_i, y_i), (x_{i+1}, y_{i+1}), and (x_0, y_0) for $i = 1, 2, \cdots, m$. The point (x_{m+1}, y_{m+1}) is identical to (x_1, y_1).

The integral of $x^k y^l$ over the polygon is equal to the sum over all m triangles, of the integral of $x^k y^l$ over each triangle; the latter integral is easily evaluated analytically for various values of

k and *l*.

Due to cancellation of terms, the final expression for the integral of $x^k y^l$ over the polygon does not depend upon the arbitrarily chosen point (x_0, y_0). Each of the moments is equal to a sum of *m* terms, divided by the area *A* which is also a sum of *m* terms. Each term in either the numerator or denominator depends only on the two adjacent points (x_i, y_i) and (x_{i+1}, y_{i+1}), which correspond to a single directed line segment.

It is frequently convenient to store an entire map file not as polygons, but as an unsorted collection of directed line segments. Each line segment has an arbitrary direction; namely, a "from" point and a "to" point, and a "left" polygon and a "right" polygon. With such a map file, expressions (4), (5), and (6) are easily evaluated as sums (in any order) over line segments. For a given polygon, one uses only those line segments which lie on the polygon boundary (i.e., which have the desired polygon on one side but not the other). The contribution from each line segment is either added or subtracted depending on whether the polygon lies to the left or right of the directed line segment; this convention implies that $A > 0$. In the calculation, care must be taken so that cancellation of nearly equal terms does not lead to imprecise results.

A simple example is provided by applying these moment expressions to the unit square (i.e., [0,0], [1,0], [1,1], [0,1]). Table 1 shows the specific values of S, D, and w.

Table 1
Example moments for a unit square

i	x_i	y_i	S_{xi}	S_{yi}	D_{xi}	D_{yi}	w_i
1	0	0	1	0	1	0	0
2	1	0	2	1	0	1	2
3	1	1	1	2	-1	0	2
4	0	1	0	1	0	-1	0

Using the values in Table 1 and expressions (4) - (6), then

$$area = A = \frac{0 + 2 + 2 + 0}{4} = 1$$

$$EX_o = \frac{0(1) + 2(2) + 2(1) + 0(0)}{12} = \frac{1}{2}$$

$$EX_o^2 = \frac{0(1 + 1/3) + 2(4 + 0/3) + 2(1 + 1/3) + 0(0 + 0/3)}{32} = \frac{1}{3}$$

$$EX_o Y_o = \frac{0(0 + 0/3) + 2(2 + 0/3) + 2(2 + 0/3) + 0(0 + 0/3)}{32} = \frac{1}{4}$$

etc.

The moments of the bivariate uniform variable (X, Y) for any polygon (degree < 5), regular or arbitrary, can be calculated from expressions (5) and (6). To further illustrate, Table 2 gives the central moments, the expected interpoint squared distance, and its variance for five polygons each with centroid at (0, 0) and with area = 1 (a square, a right triangle, an equilateral triangle, a circle and an irregular polygon).

Table 2
Central moments for uniformly distributed variable (X, Y)

	square	right	equilateral	circle[*]	S.F.[**]
EX	0.0	0.0	0.0	0.0	0.0
EY	0.0	0.0	0.0	0.0	0.0
EX^2	0.0833	0.0555	0.0962	0.0796	0.0756
EY^2	0.0833	0.2222	0.0962	0.0796	0.0941
EXY	0.0	0.5555	0.0	0.0	-0.0075
EX^2Y	0.0	-0.0074	-0.0169	0.0	0.0031
EXY^2	0.0	0.0148	0.0	0.0	0.0004
EX^3	0.0	-0.0074	0.0	0.0	-0.0009
EY^3	0.0	0.0593	0.0169	0.0	-0.0051
EX^2Y^2	0.0069	0.0148	0.0074	0.0422	0.0057
EX^4	0.0125	0.0074	0.0222	0.1266	0.0116
EY^4	0.0125	0.1185	0.0222	0.1266	0.0172
n	100	100	100	100	100
ED^2	0.3333	0.5555	0.3849	0.3183	0.3396
σ_d^2	0.0004	0.0022	0.0009	0.0003	0.0004
σ_d	0.0213	0.0470	0.0301	0.0187	0.0211

* approximated by a 360 sided regular polygon
** = the outline of San Francisco city/county normalized to have area = 1 (Figure 1)

Central moments can be calculated from noncentral moments (Kendall, 1963); specifically

$$EX^k = \sum_{j=0}^{k} \binom{k}{j} EX_o^{k-j} (-EX_o)^j .$$ (7)

More simply, the central moments can be calculated directly from expressions (5) and (6) by shifting the coordinate system so that $EX = 0$ and $EY = 0$. That is, the noncentral moment expressions applied to $X - EX_o$ and $Y - EY_o$ yield values for the central moments of the distribution of (X, Y) over the bounded region (e.g., Figure 1).

2.2 RESULTS FROM SIMULATION ON A UNIT SQUARE

The mean interpoint squared distance $\overline{d^2}$ can be readily evaluated since a version of the central limit theorem applies (Silverman, 1976). The test statistic

$$z = \frac{\overline{d^2} - \overline{ED^2}}{\sigma_d} \tag{8}$$

has an approximate standard normal distribution under the null hypothesis when $\overline{ED^2}$ and σ_d^2 are calculated from expressions (2) - (6). The test statistic represented by z has an approximate normal distribution when the sample size n is large.

To get some idea of the accuracy of this approach for small sample sizes, samples of $n = 5$, 10, 15, 20, and 50 were selected from a uniform distribution over the unit square and the distribution of z simulated. From expressions (2) and (3), the expectation and variance when no spatial pattern exists are $\overline{ED^2} = 1/3$ and $\sigma_d^2 = (2n - 3)/[45n(n-1)]$ where $EX_0^k = EY_0^k = 1/(k+1)$. Simulation results based on these two values and 10,000 iterations for each value of n are given in Table 3.

Table 3
Percentiles from simulation results on a unit square

	0.01	0.025	0.05	0.10	0.90	0.95	0.975	0.99
$n=5$	0.003	0.008	0.017	0.044	0.964	0.991	0.998	1.000
$n=10$	0.006	0.015	0.031	0.067	0.934	0.975	0.991	0.998
$n=15$	0.007	0.018	0.037	0.081	0.924	0.966	0.985	0.995
$n=20$	0.007	0.019	0.040	0.085	0.917	0.960	0.981	0.993
$n=50$	0.009	0.024	0.047	0.094	0.905	0.955	0.980	0.993

The normal distribution is not an accurate approximation for the distribution of the mean interpoint squared distance for sample sizes less than 15, but steadily improves as n increases. For $n = 50$, the percentiles obtained from the simulated data are close to the expected values from a standard normal distribution.

3. APPLIED EXAMPLE

Data for non-Hodgkin's lymphoma (*International Classification of Disease for Oncology*, code 169.1) were abstracted from the Surveillance, Epidemiology and End Results cancer registry for the city/county of San Francisco, California. Eleven cases were observed over a 16-year period (1973-88) among individuals under 21 years of age. Based on the 1980 U.S. Census counts, an estimate of 50,086 white individuals were at risk for this period, corresponding to an incidence rate of 1.36 cases per 100,000 person-years. These incident cases are plotted on a transformed map, shown in Figure 1. The map is transformed so that the population at risk is uniformly distributed over San Francisco city/county (Selvin, 1988); the transformed map area is normalized to the area of the original map. When a map is transformed so the the population at risk is uniformly distributed, then the spatial distribution of cases of a specific disease will also be uniformly distributed when no spatial pattern exists. Since the exact geographic location of each was not available, each case was plotted at the centroid of the census tract of residence. Table 4 summarizes the observed data.

Table 4
Non-Hodgkin's lymphoma

locations (*km*)

	1	2	3	4	5	6	7	8	9	10	11
x	-0.998	-0.950	-2.075	-2.008	-2.368	-2.506	-4.001	-3.574	-5.102	2.769	3.008
y	4.031	1.999	-5.288	-2.536	3.779	-0.530	-2.730	3.883	-2.248	-1.877	0.184

Summary values

n	11
pairs	55
d^2	33.002 km^2
ED^2	36.624 km^2
σ_d	7.639 km^2
z	-0.474
p-value	0.318

Using the mean intercase squared distance (33.002 km^2) as a measure of clustering produces no strong evidence of a nonrandom spatial pattern. The probability of observing a smaller inter-

point squared distance when no spatial pattern exists is approximately 0.318. The moments used to calculate this approximate p-value come from expressions (4) - (6). Note that the moments associated with the polygon describing the outline of the San Francisco transformed map are close to those of a square (Table 2).

4. POWER AGAINST A SPECIFIC ALTERNATIVE
ON THE UNIT SQUARE

The following simple model is postulated to explore the statistical power of using the mean interpoint squared distance to evaluate possible clustering within the unit square. A series of n independent points (x_i, y_i) is generated, where a fraction q are chosen from a bivariate circular normal distribution with center located at $(x_0 = 0.4, y_0 = 0.4)$ and the remainder uniformly distributed over the unit square. The variance of the bivariate normal distribution is set at three illustrative values, $\sigma_x^2 = \sigma_y^2 \equiv \sigma^2 = 0.01$, 0.001, and 0.0001. Each point in the power calculation is based on a sample size of $n = 50$ points with 1000 iterations (one tail, $\alpha = 0.05$; displayed in Figure 2a). Little difference in power exists when the variance is less than 0.001 (both dotted lines). A typical power curve emerges for $\sigma^2 = 0.01$ (solid line) showing, for example, that q must be greater than 0.33 to achieve a power greater than 0.90. Also, to illustrate the power characteristics of $\overline{d^2}$, Figure 2b shows the difference in power for four sample sizes ($n = 10, 20, 50$, and 100) generated under the postulated unit-square model when $\sigma^2 = 0.01$. Not surprisingly, the sample size greatly affects the power to detect a nonuniform spatial pattern. For example, when $n = 20$ the power is 0.23 and when $n = 100$ the power increases to 0.80, for $q = 0.2$.

A typical approach to analyzing spatial data is to divide the region under study into a number of sub-areas of equal size and count the number of points falling into each of these regions. When no spatial pattern exists among a series of independent points, these counts have a Poisson distribution. A series of data sets consisting of $n = 50$ random points were generated under the conditions of the unit-square model, and the power was calculated for a range of q-values. Specifically, the unit square was divided into 25 equal sub-squares and the expected

counts (expected points per sub-unit = 2) were compared to the observed counts with the use of a chi-square statistic for varying degrees of clustering. The results of 5000 samples of n = 50 points for each value of q are shown in Figure 3 (σ^2 = 0.01). The power associated with the mean interpoint squared distance is superior to that of the chi-square statistic, for all values of q.

The power to detect a nonrandom spatial distribution using a nearest-neighbor approach was also computed for the unit-square model and contrasted to the power of $\overline{d^2}$ (n = 50 and σ^2 = 0.01) in Figure 3. The expectation and variance for a mean estimated from a set of nearest-neighbor data are quoted by various authors (e.g., Ripley, 1981; Boots 1988) and can be corrected for edge effects (Donnelly, 1978). A corrected test statistic is

$$z = \frac{\overline{r} - [.5/\sqrt{n} + c_1]}{\sqrt{0.0683/n + c_2}} \quad with \quad c_1 = \frac{0.206 + 0.164/\sqrt{n}}{n} \quad and \quad c_2 = \frac{0.148}{\sqrt{n^5}}$$

where $\overline{r}$ is the mean nearest-neighbor distance among n points. Again, z is assumed to have at least approximately a standard normal distribution when no spatial pattern exists. Donnelly (1978) demonstrated that the normal distribution adequately serves as an approximation for the distribution of $\overline{r}$ for samples sizes greater than six (Donnelly, 1978). For the specific comparison shown in Figure 3, the mean interpoint squared distance again has uniformly greater power.

The fact that the interpoint squared distance has uniformly more power than either the Poisson or nearest neighbor methods is not surprising since $\overline{d^2}$ more appropriately reflects a continuous measure of distance. The Poisson approach looses power by categorizing a continuous variable and the power of the nearest neighbor method is strongly influenced by noninformative points. As mentioned, the power of a spatial statistic depends critically on the spatial pattern underlying the data; other models can be envisioned that would produce different results. For example, a mean distance measure is a poor choice to evaluate a spatial distribution containing several discrete clusters.

REFERENCES

Boots, N. B., and Getis, A., (1988) *Point Pattern Analysis*, chapter 4. Newbury Park: Sage.

Cliff, A. D., and Haggett, P., (1988) *Atlas of Disease Distributions*, Oxford: Blackwell.

Cuzick, J., and Edwards, R., (1990) Spatial clustering for inhomogeneous populations. *J.R. Statist. Soc B.*, 52, 73-104.

Donnelly, K. P., (1978) Simulations to determine the variance and edge effect of total nearest neighbor distance. In *Simulation methods in archaeology*, ed. I. Hodder p.91-95. Cambridge: Cambridge University Press.

Kendall, M. G., and Stuart, A. S., (1963) *The Advanced Theory of Statistics*, p. 56. London: Griffin.

Ripley, B. D., (1981) *Spatial Statistics*, chapter 7. New York: Wiley.

Schulman, J., (1986) *The Statistical Analysis of Density Equalized Maps*. Ph.D thesis. University of California, Berkeley.

Schulman, J., Selvin, S., and Merrill, D. W., (1988) Density equalized map projections: a method for analyzing clustering around a fixed point. *Stat. Med.*, 7, 491-505.

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

Silverman, B. W., (1976) Limit theorem for dissociated random variables. *Advances in Applied Probability*, 8, 806-819.

Whittemore, A. S., Friend, D., Brown, B., and Holly, E. A., (1987) A test to detect clusters of disease. *Biometrika*, 74, 631-635.
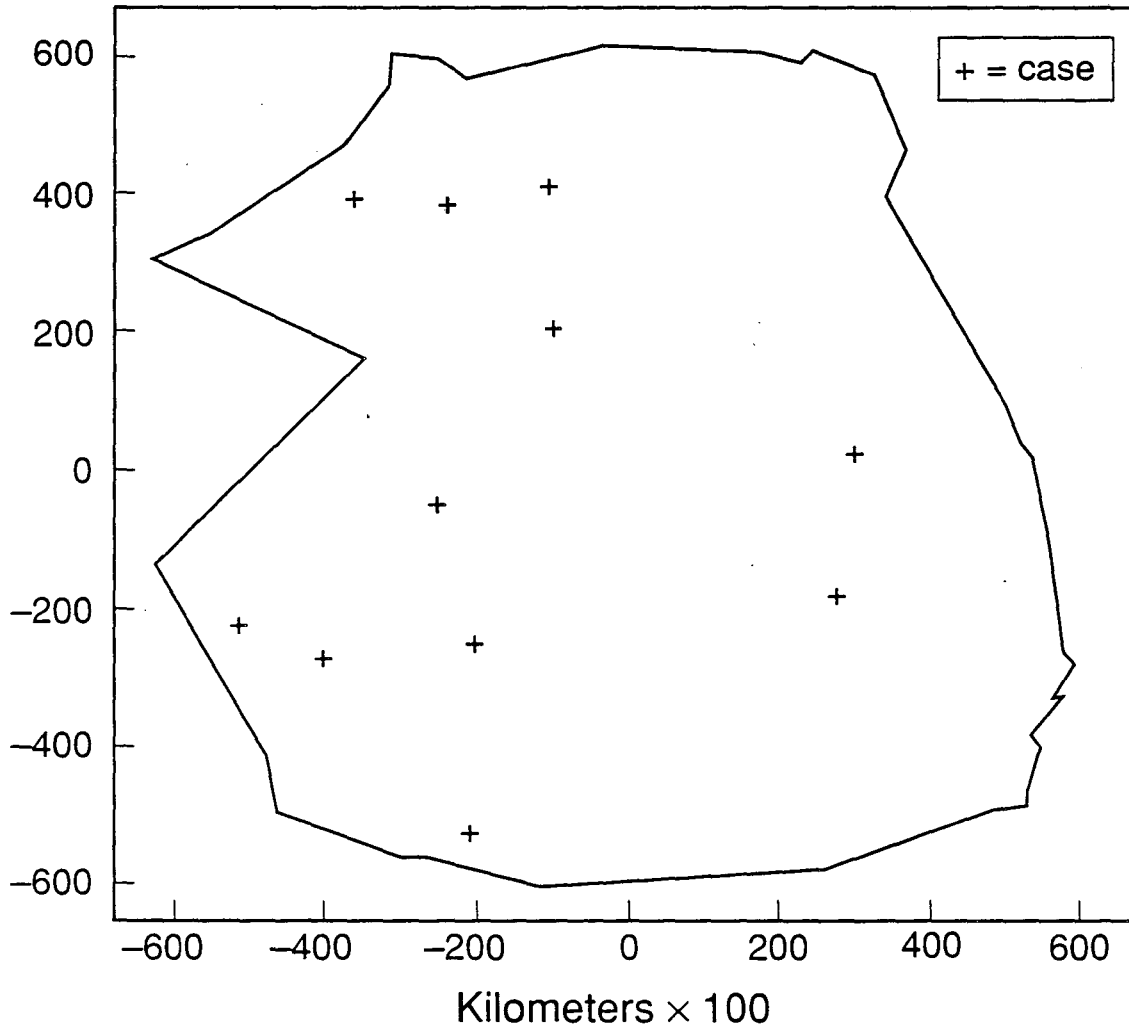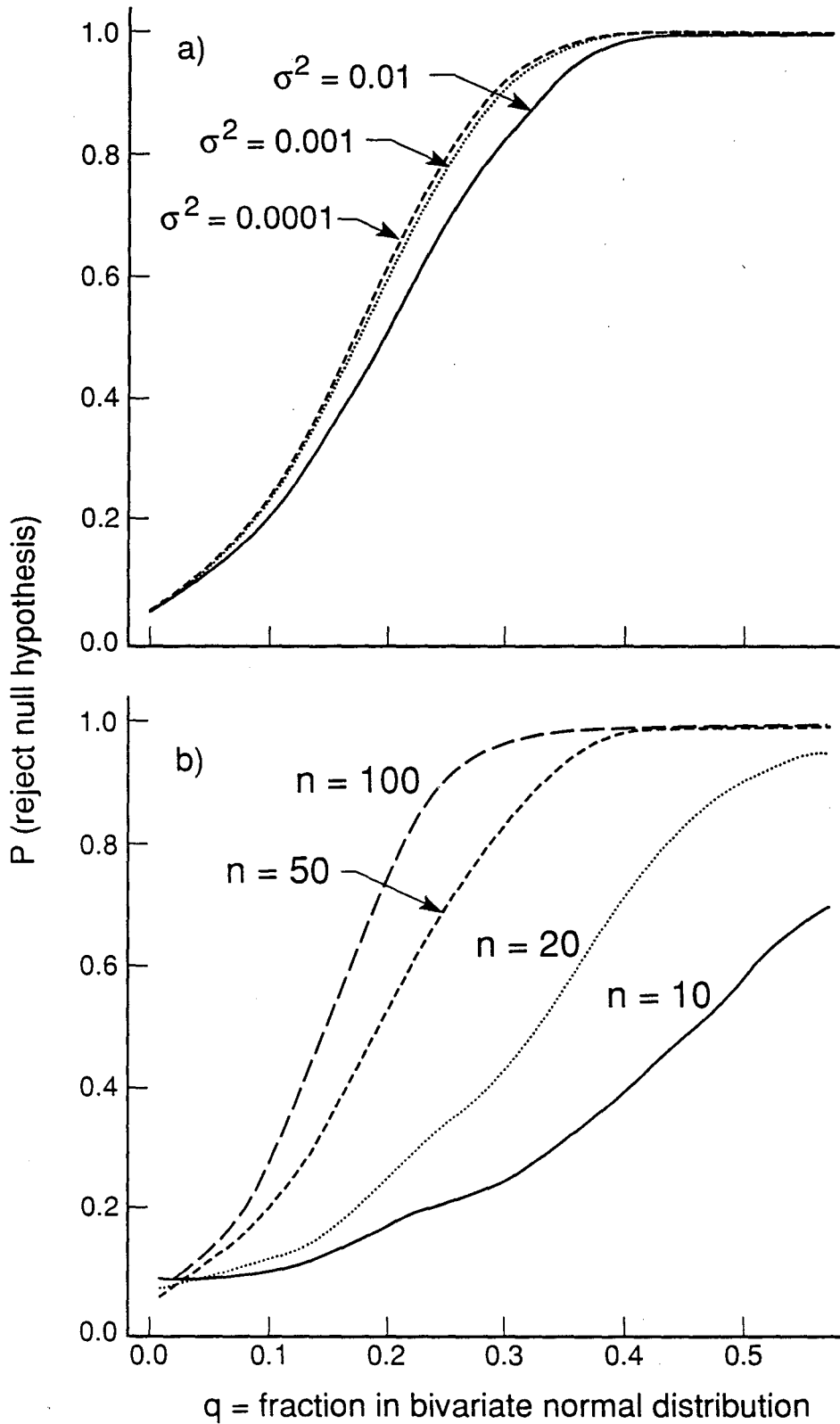
Figure 1.   Eleven cases of non-Hodgkin's lymphoma plotted on a transformed map of the city/county of San Francisco, Califonia.

Figures 2a and 2b. Power curves (smoothed) for $\overline{d^2}$ from simulated data sampled from the unit-square model for three variances (figure 2a; $\sigma^2 = 0.01$, 0.001, and 0.0001 with $n = 50$) and four sample sizes (figure 2b; $n = 10$, 20, 50, and 100 with $\sigma^2 = 0.01$).
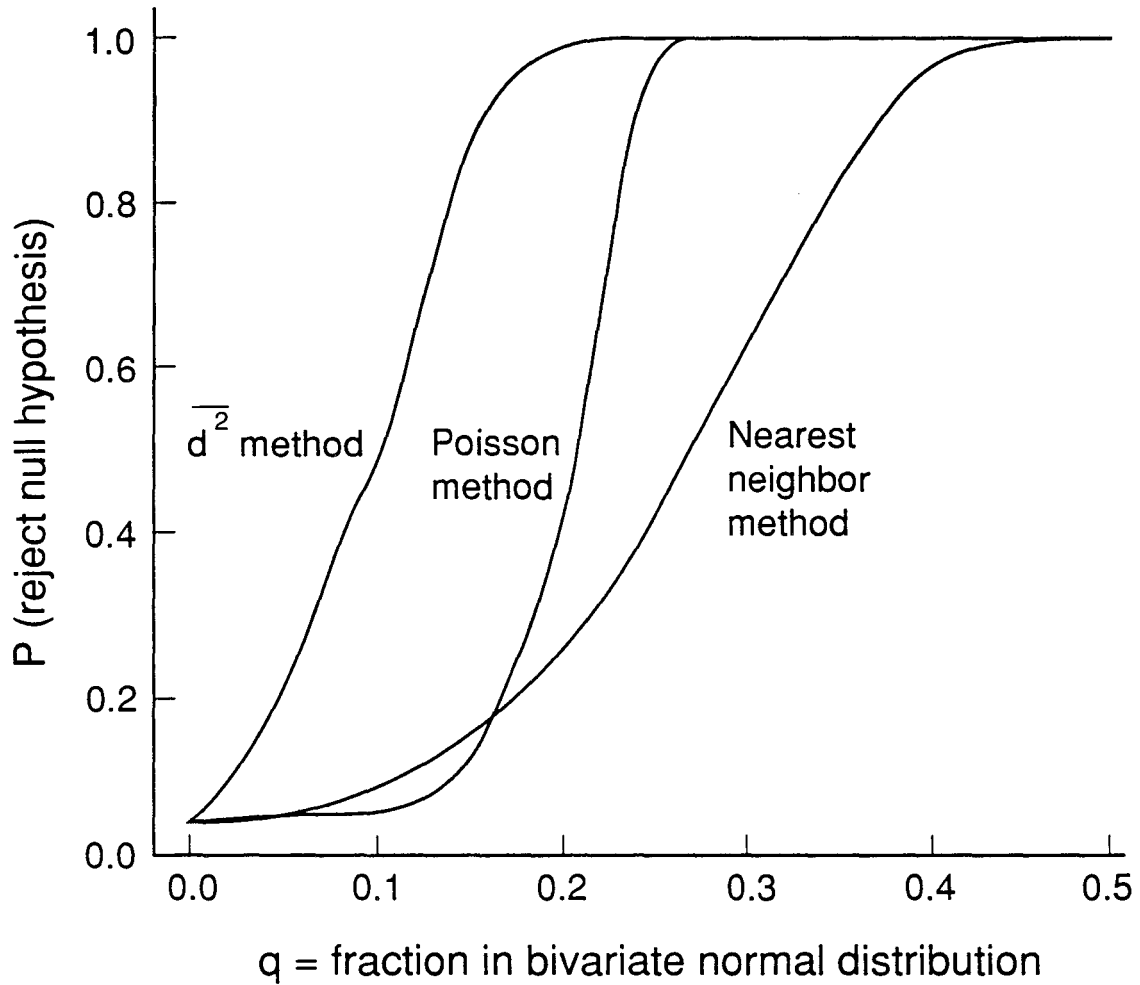
Figure 3.   Power curves (smoothed) associated wth three methods of spatial analysis ($\overline{d^2}$, Poisson counts, and nearest-neighbor) using data sampled from the unit square with $\sigma^2 = 0.01$ and $n = 50$ for varying degrees of clustering (q).

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