J. K. Ord
Arthur Getis

Local Spatial Autocorrelation Statistics: Distributional Issues and an Application

The statistics $G_i(d)$ *and* $G_i^*(d)$, *introduced in Getis and Ord (1992) for the study of local pattern in spatial data, are extended and their properties further explored. In particular, nonbinary weights are allowed and the statistics are related to Moran's autocorrelation statistic,* I. *The correlations between nearby values of the statistics are derived and verified by simulation. A Bonferroni criterion is used to approximate significance levels when testing extreme values from the set of statistics. An example of the use of the statistics is given using spatial-temporal data on the AIDS epidemic centering on San Francisco. Results indicate that in recent years the disease is intensifying in the counties surrounding the city.*

1. INTRODUCTION

In spatial data analysis, it is often necessary to determine whether or not identifiable spatial patterns exist. For example, we may test for spatial pattern by focusing on the locations of the sample points, by studying the values associated with these locations given the sampling pattern, or by combining these analyses. There are many ways to test for the existence of such patterns; perhaps the most popular is Moran's I statistic, which is used to test the null hypothesis that the spatial autocorrelation of a variable is zero. If the null hypothesis is rejected, the variable is said to be spatially autocorrelated. Traditional analyses such as nearest neighbor, k-function analysis, and the semivariogram are all widely used to study spatial patterns. All of these statistics are global in the sense that they require measurements from all or many geo-referenced points in the sample.

This research was supported by the National Science Foundation, grant no. SES-9123832. The authors thank Cleridy Lennert of Scripps Institution of Oceanography and Serge Rey of San Diego State University for making programming suggestions and Xioake Zhang and Long Gen Ying of San Diego State University for gathering data and carrying out the simulations. They are grateful to Luc Anselin and the referees for their insightful comments which led to considerable improvements in the paper.

J. K. Ord is the David H. McKinley Professor of Business Administration in the department of management science and information systems at The Pennsylvania State University. Arthur Getis is Stephen and Mary Birch Professor of Geographical Studies in the department of geography at San Diego State University.

Geographical Analysis, Vol. 27, No. 4 (October 1995)

In recent years, there has been growing interest in local measures of spatial dependence. Much of this work has been inspired by the search for valid tests for the clustering of cases of rare diseases; see, for example Stone (1988), Cuzik and Edwards (1990), and Cressie (1992). Besag and Newell (1991) draw a useful distinction between *general* and *focused* tests. General tests are concerned with overall patterns in a large region, whereas focused tests "concentrate upon one or more smaller regions selected ostensibly because of some factor (for example, the location of a nuclear installation) that has been previously hypothesized to be associated with the disease." Besag and Newell (1991) go on to discuss tests for the detection of clusters, whose purpose is to identify "hot spots" without any preconceptions about their locations.

It is apparent that a test for hot spots could be used to serve the same role as a focused test, in that the hot spot should emerge from the pack if its local structure is sufficiently unusual. Furthermore, such an approach affords some protection against the biases that may arise when only selected areas are tested. Indeed, focused tests must rely upon either the availability of reference data for similar areas well removed from the putative source (Cuzik and Edwards 1990) or an adjustment to the distribution of the test statistic to compensate for the search for hot spots (cf. Stone 1988). In this regard, Besag and Newell (1991) point out the difficulties inherent in the original version of the Geographical Analysis Machine (GAM) introduced by Openshaw et al. (1987); they provide a modified analysis to overcome these difficulties. Besag and Newell also point out that when region i has n_i cases of a disease in a population of t_i, the random permutations distribution for the Moran statistic may not be an appropriate frame of reference. This difficulty may arise when urban (high t_i) areas tend to be clustered, and likewise rural (low t_i) areas. Provided that the $\{n_i\}$ and $\{t_i\}$ are not too small, the difficulty may be resolved by using the pooled incidence estimator $p = \sum n_i / \sum t_i$ and then computing the standard scores for each region as

$$z_i = (n_i - pt_i)/[t_i p(1-p)]^{\frac{1}{2}}. \tag{1}$$

The focus of this paper is a pair of tests for the detection of clusters, introduced by Getis and Ord (1992). These statistics are especially useful in cases where global statistics may fail to alert the researcher to significant pockets of clustering. For example, Getis and Ord showed that the distribution of Sudden Infant Death Syndrome in North Carolina for the period 1979–84 did not display any global spatial pattern, but that a few counties in the southern part of the state displayed a clustering of cases.

2. STATEMENT OF THE PROBLEM

Consider an area subdivided into n regions, $i = 1, 2, \ldots, n$, where each region is identified with a point whose Cartesian coordinates are known. Each i has associated with it a value x_i that represents an observation upon the random variable X_i. Typically, it will be assumed that the X_i have identical marginal distributions; further, if the X_i are independent, we say that there is no spatial structure. Independence implies the absence of spatial autocorrelation, but the converse is not necessarily true. Nevertheless, tests for spatial autocorrelation are typically viewed as adequate assessments of dependence.

Usually, if spatial autocorrelation exists, it will be exhibited by similarities between contiguous regions, although negative patterns of dependence are also possible. The revised statistics considered in this paper may be used to search

for either positive or negative dependence. Further, we focus upon physical distances, but "distance" may be interpreted as travel time, conceptual distance, or any other measure that enables the n points to be located in a space of one or more dimensions.

Getis and Ord (1992) introduced a family of statistics, G, that can be used as measures of spatial association in a number of circumstances. The local statistics, G_i and G_i^*, enable us to detect pockets of spatial association that may not be evident when using global statistics. In this paper, the statistics G_i and G_i^* are extended to include variables that do not have a natural origin. The cost of this move is that the statistics lose some intuitive appeal, but the benefit is that the earlier restriction no longer applies. In addition, the statistics may incorporate nonbinary weight matrices. The new form increases the statistics' flexibility, and, therefore, their usefulness.

In section 3, we provide the results of a series of simulations designed to show the distributional and small sample properties of the statistics in different circumstances. We show how the statistics are related to Moran's autocorrelation statistic, I, in section 4. In section 5, we address questions of edge effects and the correlation of G_i^* values with one another. Section 6 contains an approximate procedure that allows us to test the most extreme of the observed G_i values, as a test for hot spots; section 7 examines the effect of global autocorrelation upon these local tests. Finally, in section 8, we give an example of the use of the G_i statistics with regard to the spatial analysis of the location of the those suffering from the AIDS disease in San Francisco and neighboring areas for the period 1989–1993.

3. THE REWRITTEN STATISTICS

In Getis and Ord (1992), the statistic $G_i(d)$ is defined as

$$G_i(d) = \frac{\sum_j w_{ij}(d)x_j}{\sum_j x_j}, \quad j \text{ not equal to } i, \tag{2}$$

where $\{w_{ij}(d)\}$ is a symmetric one/zero spatial weight matrix with ones for all links defined as being within distance d of a given i; all other links are zero including the link of point i to itself. Throughout the paper, the d argument is dropped when only a single distance is under consideration. The sum of the weights is written as

$$W_i = \sum_{j \neq i} w_{ij}(d). \tag{3}$$

The numerator of (2) is the sum of all x_j within d of i but not including x_i. The denominator is the sum of all x_j not including x_i. When we set

$$\bar{x}(i) = \frac{\sum_j x_j}{(n-1)} \text{ and } s^2(i) = \frac{\sum_j x_j^2}{(n-1)} - [\bar{x}(i)]^2, \tag{4}$$

it may be shown that

$$\text{Var}(G_i) = \frac{W_i(n-1-W_i)}{(n-1)^2(n-2)} \cdot \left[\frac{s(i)}{\bar{x}(i)}\right]^2. \tag{5}$$

It should be noted that Getis and Ord (1992) used (Y_{i1}, Y_{i2}) in place of $[\bar{x}(i), s^2(i)]$. It was shown (Getis and Ord 1992) that if $E(G_i)$ is bounded away from 0 and from 1, then the permutations distribution of G_i under H_o approaches normality. We now redefine G_i as a standard variate by taking the statistic minus its expectation, $E[G_i] = W_i/(n-1)$, divided by the square root of its variance; at the same time we allow the weights to be nonbinary. The resulting measures are

$$G_i(d) = \frac{\sum_j w_{ij}(d)x_j - W_i\bar{x}(i)}{s(i)\{[(n-1)S_{1i}) - W_i^2]/(n-2)\}^{\frac{1}{2}}}, \quad j \neq i. \tag{6}$$

Similarly, if we include $w_{ii} \neq 0$, the standardized G_i^* statistic is

$$G_i^*(d) = \frac{\sum_j w_{ij}(d)x_j - W_i^*\bar{x}}{s\{[(nS_{1i}^*) - W_i^{*2}]/(n-1)\}^{1/2}}, \quad \text{all } j. \tag{7}$$

In (6) and (7), we have $W_i^* = W_i + w_{ii}$, $S_{1i} = \sum_j w_{ij}^2$, $(j \neq i)$, and $S_{1i}^* = \sum_j w_{ij}^2$ (all j); $\bar{x}$ and s^2 denote the usual sample mean and variance.

Numerical results for (6) and (7) are given in Table 1. As expected, the following patterns emerge for the distributions of G_i^* and similar results hold for G_i:

i. When the underlying distribution is normal, so is that of the test statistics (an exact result);
ii. when the underlying distribution is markedly skew, the distribution of the test statistics is non-normal, but approaches normality as the distance is increased.
iii. the statistics for edge cells approach normality more slowly because they have fewer neighbors; the convergence for corner cells is still slower.

4. PROPERTIES AND ASSOCIATIONS WITH *I*

Moran's statistic can be written as

$$I(d) = \frac{\sum_i (x_i - \bar{x}) \sum_j w_{ij}(x_j - \bar{x})}{Ws^2} - \frac{n}{W}, \tag{8}$$

temporarily dropping the d argument in the weights, for convenience; $W = \sum_i W_i^*$. If we set

$$K_{ri} = \frac{\sum (w_{ij} - \bar{w}_i)^r}{n-1}, r = 2, 3, \ldots \tag{9}$$

where $\bar{w}_i = W_i^*/n$ and $K_{2i} = [nS_{1i}^* - W_i^{*2}]/(n-1)$ and put $K_i^2 = K_{2i}$, again for convenience, we have

$$G_i^* \equiv G_i^*(d) = \frac{\sum_j w_{ij}(x_j - \bar{x})}{sK_i} \tag{10}$$

and

$$I(d) = [\sum z_i K_i G_i^* - n]/W \tag{11}$$

where $z_i = (x_i - \bar{x})/s$, so that $I(d)$ is a weighted average of the local statistics.

TABLE 1

Mean, Standard Deviation, Skewness, and Kurtosis of $G_i^*(d)$ Statistic for Five Thousand Random Permutations of Each of Four Probability Distributions[1] for Five Distances by Type of Cell in a 10 by 10 Matrix

		Central Cell				Edge Cell				Corner Cell			
Dist	(d)	Mean	SD	*Skew*	*Kur*	Mean	SD	*Skew*	*Kur*	Mean	SD	*Skew*	*Kur*
Normal	1.0	−.005	.998	.006	−.217	.004	1.014	.023	.049	−.001	.990	.004	.058
	1.5	.002	1.004	−.021	.036	−.012	1.003	.053	−.019	−.006	1.011	.014	−.086
	2.0	−.005	1.010	−.045	−.053	−.037	.996	−.047	−.113	.005	1.001	−.036	−.041
	2.5	−.000	1.008	−.047	−.079	.001	.991	−.025	.010	−.003	.991	−.060	−.127
	3.0	.009	.989	−.003	−.046	.005	1.002	−.005	−.052	.004	.979	.007	−.067
Binary	1.0	−.000	1.008	−.031	−.434	−.009	1.004	.016	−.482	−.004	.992	.003	−.621
	1.5	−.005	1.010	.065	−.240	−.017	.996	−.027	−.259	−.001	.986	.020	−.446
	2.0	.002	.990	−.033	−.152	−.004	.991	−.008	−.188	.010	.989	−.002	−.221
	2.5	−.001	1.012	.014	.062	−.005	1.012	.026	−.075	−.005	1.006	−.012	−.298
	3.0	−.005	1.017	.040	−.002	.007	1.003	.008	−.016	.001	.992	.038	−.091
Poisson	1.0	.010	1.006	.373	−.101	.016	1.013	.401	−.054	−.007	.989	.550	.340
	1.5	.007	.997	.207	−.094	.003	.995	.358	.064	−.015	.999	.431	.040
	2.0	−.008	.994	.174	.017	.005	1.005	.247	−.024	.002	.994	.322	−.030
	2.5	.011	.998	.173	−.097	.007	1.002	.233	−.123	−.004	1.006	.292	−.006
	3.0	−.002	.992	.082	.002	.001	.996	.097	−.012	.003	1.003	.192	.025
Exponential	1.0	.001	1.004	1.226	1.579	.004	.992	1.383	2.215	.003	.995	1.648	3.138
	1.5	.003	1.012	.934	.880	.006	1.018	1.126	1.442	.009	.993	1.357	2.027
	2.0	.002	1.008	.726	.429	.001	1.001	.867	.814	−.001	.994	1.108	1.362
	2.5	.017	1.004	.437	.001	.018	.996	.651	.207	.021	1.004	.912	.802
	3.0	−.007	.984	.284	−.220	−.003	.998	.480	.001	.003	1.000	.784	.520

[1] Characteristics of the Distributions:

	Mean	St. Deviation	Minimum Value	Maximum Value
Normal	0.075	1.044	−3.0753	2.6186
Binary	0.500	0.503	0.0000	1.0000
Exponential	0.829	1.000	0.0065	6.3343
Poisson	0.970	0.979	0.0000	4.0000

(1) The skewness measure is the standardized third moment, $skew = m_3/m_3^{3/2}$, whereas the kurtosis measure is $kur = (m_4/m_2^2) - 3$, where $m_j = \sum (x - \bar{x})^j/n$. The population values of *skew* and *kur* are zero for the normal.
(2) For samples of size 5,000, the null hypothesis of normality is rejected at the $\alpha = 0.05$ level if |skew| > 0.068, or if kur > 0.14 or < −0.13.

Further, when the permutations approach is used so that $(\bar{x}, s)$ are fixed, the standardized skewness and kurtosis measures (cf. Stuart and Ord 1987, p. 107) reduce to

$$\tau_1(G_i^*) = (K_{3i}/K_i^3)\,\mu_3 \tag{12}$$

$$\tau_2(G_i^*) = (K_{4i}/K_i^4)\,(\mu_4 - 3) \tag{13}$$

where $n\mu_r = \sum_i z_i^r$, so that μ_r represents the moments of the original set of n observations.

For example, suppose location i has m neighbors at distance d or less, and that binary weights are used. It follows from equations (12) and (13) that

$$\tau_1(G_i^*) = (n - 2m)\left[\frac{(n-1)}{nm(n-m)}\right]^{1/2} \cdot \mu_3 \tag{14}$$

and

$$\tau_2(G_i^*) = \frac{[(n-m)^3 + m^3]\,(n-1)}{n^2 m(n-m)}\,(\mu_4 - 3). \tag{15}$$

Generally n will be large relative to m, since the G_i^* are looking at local patterns, so we have, approximately,

$$\tau_1 = m^{-1/2}\mu_3 \quad \text{and} \quad \tau_2 = m^{-1}(\mu_4^{-3}) \tag{16}$$

corresponding to the usual rates of convergence with the Central Limit Theorem. Thus, provided d is not too small and the weights are not too uneven, approximate normality is a reasonable assumption.

EXAMPLE 4.1. Suppose a variable X is spatially distributed as in Figure 1. The numbers in parentheses are the identifying numbers for the observations. Suppose our interest is in the possible clustering of high values in the vicinity of point 5 but not including point 5 itself. We decide to select increments of 10 meters from point 5 to a distance of 30 meters (Figure 2).

First, we use equation (4) to find $\bar{x}(5)$ and $s^2(5)$. These are 0.0986 and 1.4336, respectively. Then, select the $G_i(d)$ statistic since we are excluding point i. In this example, the weights are binary, that is, $w_{ij} = 1$ if point j is within d of point i and zero otherwise. For example, when $d = 10$, $w_{51} = 0$ since the distance between point 5 and point 1 is greater than 10. Using equation (6) we get

$$G_5(10) = \frac{(1.67) - (1)(.0986)}{\{[(1)(8-1-1)(1.4336)]/(8-2)\}^{1/2}} = 1.3125;$$

$$G_5(20) = 2.1562;$$

$$G_5(30) = 1.7692.$$

From these results, it is clear that the clustering of positive values around point 5 reaches a maximum in the neighborhood of twenty meters.

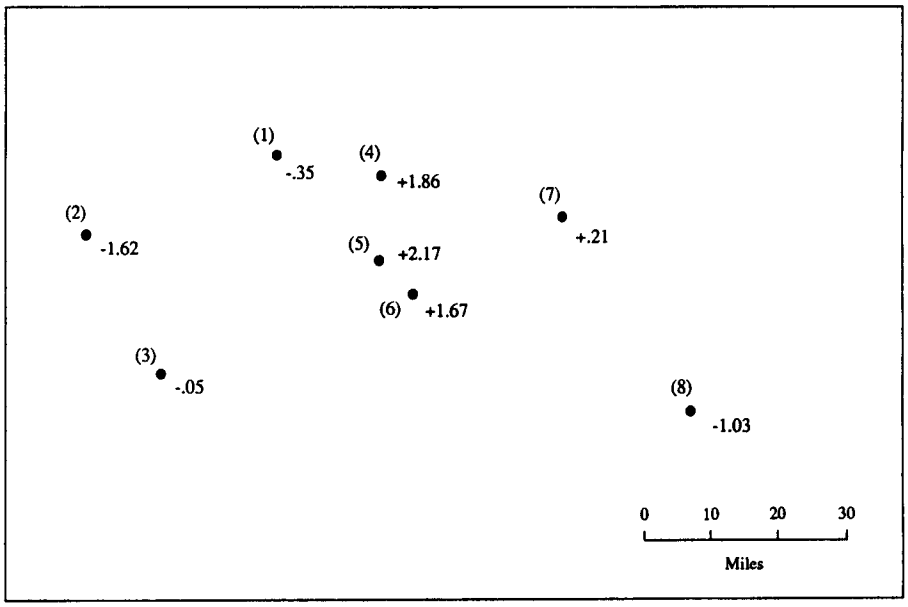

FIG. 1

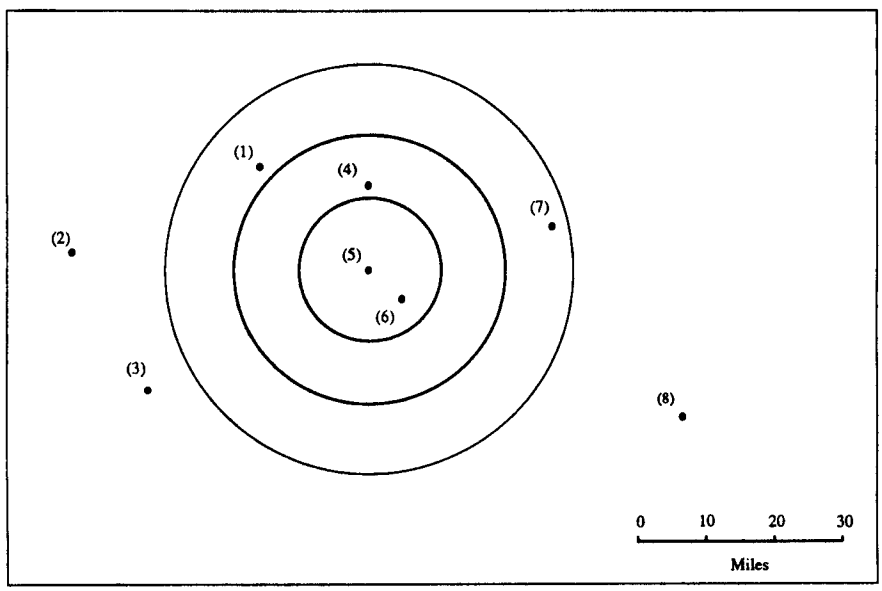

FIG. 2

EXAMPLE 4.2. Now we focus on possible clustering around point 5 including point 5 itself. For this we use the $G_i^*(d)$ statistic. $\bar{x}$ is 0.3575 and s^2 is 1.7237.

$G_5^*(10) = 1.8179;$

$G_5^*(20) = 2.4078;$

$$G_5^*(30) = \frac{(2.17 + 1.67 + 1.86 - .35 + .21) - (5)(.3575)}{\{[(5)(8-5)(1.7237)]/(8-1)\}^{1/2}} = 1.9629.$$

In this example, the clustering of positive values is much more in evidence than in the previous example simply because the value at point 5 is included in the calculations and point 5 happens to be associated with a large positive score.

EXAMPLE 4.3. Now suppose that instead of a binary weighting scheme, we use a standardized approach where the sum of the weights within d of i sum to 1 and each individual weight is $1/W_i$. In this case G_i and G_i^* are homogeneous of order zero in w_{ij} and thus invariant. Thus $G_5(30) = 1.7692$, the same result as in Example 4.1.

EXAMPLE 4.4. In this nonbinary example, we weight each observation by $w_{ij}^\# = (1/d_{5j})/W_i^\#$, where $W_i^\# = \sum w_{ij}^\#$ and $W_i = \sum w_{ij}^\#/W_i^\# = 1$, so that points close to point 5 are given more weight than far points. In this case, we seek only one value of G_i. Values of d for each of the j points are (1) 23, (2) 44, (3) 37, (4) 13, (6) 7, (7) 28, (8) 52. Again, using equation (6) we have $G5(1/d_{5j}) = 1.9893$. This procedure cannot be used for G_i^* simply because $w_{ii}^\# = \infty$; however, modifications to the weights such as $1/(a + d_{ij})$, $a > 0$ could clearly be used.

5. CORRELATION STRUCTURE AND EXPERIMENTAL RESULTS

Clearly the G_i, G_i^* values for various i locations on the same map are not independent, especially if the i locations are within distance d of one another. In the remainder of this section we focus upon G_i^* since the analysis is more straightforward. Similar results hold for G_i. For a particular i, say A, G_i^* is defined in terms of the association of A to all locations j within d of A. If another cell, B, is within d of A, then the G_i^* value associated with B is dependent on a number of the same values on which A is dependent.

In Figure 3, the cells within distance 2.0 (where, as before, a distance of 1.0 is measured from the center of a cell to the center of a contiguous cell) of A are denoted with an a and the cells within 2.0 of B are denoted as b. Of the thirteen

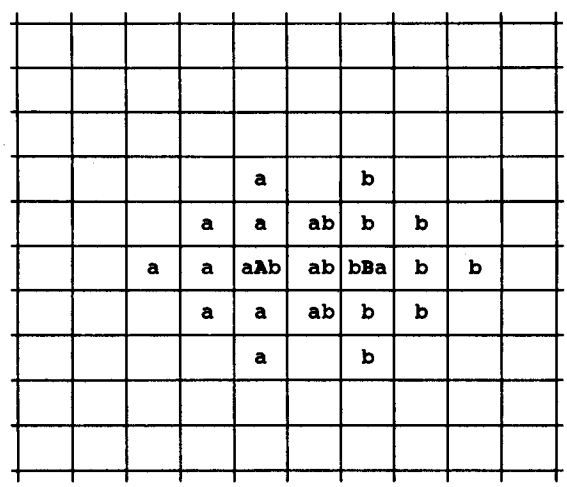

FIG. 3

cells within 2.0 of A, five of them overlap with the cells within 2.0 of B. It is clear, then, that the $G_i^*(d = 2)$ values of A and B correlate to some degree. The degree of correlation not only depends on the overlap, but also the number of regions; see, for example, equation (18) below. In nonlattice situations, the configuration of the units will also bear on the degree of correlation.

When $\{w_{ij}\}$ are binary, the correlations for the G_i^* statistic are found to be

$$r_{ik} = \text{corr}\,[G_i^*, G_k^*] = \frac{n\sum_j w_{ij}w_{kj} - W_i^*W_k^*}{[W_i^*W_k^*(n - W_i^*)(n - W_k^*)]^{1/2}} \tag{17}$$

with the expression for nonbinary weights being different only in that the denominator changes:

$$r_{ik} = \text{corr}\,[G_i^*, G_k^*] = \frac{n\sum_j w_{ij}w_{kj} - W_i^*W_k^*}{\{[nS_{1i} - (W_i^*)^2][nS_{1k}^* - (W_k^*)^2]\}^{1/2}}. \tag{18}$$

Consider a regular lattice, as shown in Figure 3; if we restrict attention to interior cells, such that both i and k have m neighbors, we can write

$$\sum_j w_{ij}w_{kj} = mc_{ik}, \qquad \text{where } 0 \le c_{ik} \le 1 \quad \text{and}$$

$$W_i^* = W_k^* = m.$$

Then

$$r_{ik} = \frac{(nc_{ik} - m)}{(n - m)} \tag{19}$$

so

$$1 \ge r_{ik} \ge -m/(n - m).$$

For example, in the case shown in Figure 3, where $n = 100$, the correlation between the G_i^* of A and the G_i^* of B for a distance of 2.0 is $r_{AB} = .293$, where $m = 13$, $mc_{AB} = 5$, and $c_{AB} = 0.385$.

Figure 4 gives the expected correlations between the statistic for cell A and the statistics based on other cells at the distances shown. For large n, these values approach the limit given by the coefficient of n in the numerator. Thus, for large n, correlations remain positive and decay slowly as m increases; however, even for $n = 100$, the correlation is negative even at $d = 4$.

As a test on the validity of the correlation structure, a series of simulations was carried out. Table 2 gives the results of a battery of correlation experiments under varying conditions. Using equation (19), Table 2 gives the expected correlation between the statistics for two neighboring cells for different distances from the cells' center point. Distance is 1.0 between contiguous cells and the entire matrix is a 10 by 10 with 100 cells total. Correlations were calculated and the mean taken after one thousand simulations. As Table 2 shows, the experimental results approximate the expected results, in all cases. The correla-

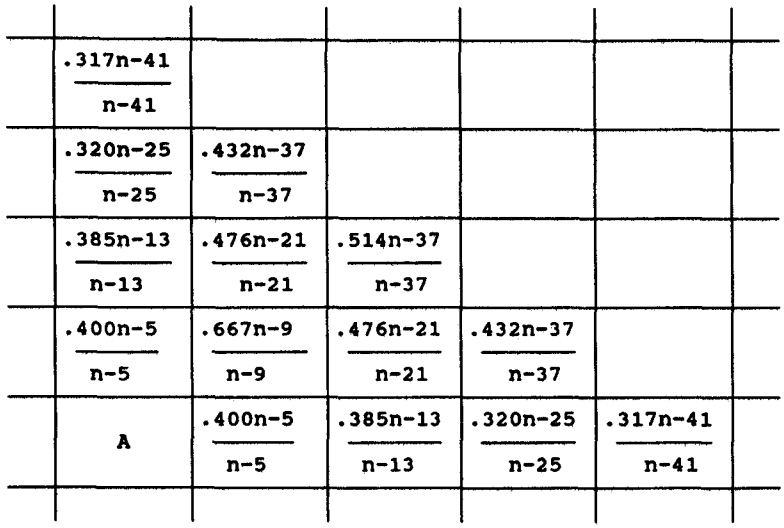

FIG. 4

TABLE 2

Expected and Observed Correlation between Neighboring Cells for Varying Distances (d) for $G_i^*(d)$ Statistic for One Thousand Random Permutations of Each of Four Probability Distributions by Type of Cell in a 10 by 10 Matrix

Distribution	Distance (d)	Central Cell	Edge Cell	Corner Cell
Expected	**1.0**	**0.368**	**0.479**	**0.568**
Normal	1.0	0.401	0.451	0.568
Binary	1.0	0.397	0.460	0.563
Exponential	1.0	0.394	0.471	0.563
Poisson	1.0	0.318	0.454	0.568
Expected	**1.5**	**0.634**	**0.646**	**0.823**
Normal	1.5	0.624	0.646	0.805
Binary	1.5	0.623	0.601	0.802
Exponential	1.5	0.630	0.658	0.810
Poisson	1.5	0.646	0.638	0.812
Expected	**2.0**	**0.557**	**0.634**	**0.707**
Normal	2.0	0.557	0.625	0.703
Binary	2.0	0.548	0.656	0.695
Exponential	2.0	0.551	0.637	0.680
Poisson	2.0	0.543	0.608	0.709
Expected	**2.5**	**0.699**	**0.734**	**0.846**
Normal	2.5	0.714	0.739	0.824
Binary	2.5	0.699	0.730	0.847
Exponential	2.5	0.698	0.742	0.831
Poisson	2.5	0.697	0.734	0.849
Expected	**3.0**	**0.661**	**0.729**	**0.782**
Normal	3.0	0.696	0.734	0.791
Binary	3.0	0.692	0.705	0.781
Exponential	3.0	0.673	0.748	0.794
Poisson	3.0	0.641	0.720	0.806

tions vary only slightly from their expectations and the variation is without pattern.

6. APPROXIMATE TESTS BASED ON EXTREME G STATISTICS

Getis and Ord (1992) showed that the G_i are asymptotically normally distributed and it follows by similar arguments that $\mathbf{G} = (G1, G2, \ldots, G_n)'$ are multivariate normal with means zero, variances one and covariances (correlations) given by (18). In applications, we usually look at the largest and smallest elements of **G**. Here, the extremes are selected from the set of statistics for all the regions; this approach differs from that of Stone (1988) who searches for the maximum value of a statistic with respect to a focused, or preselected, site. In order to make valid inferences, therefore, we need to control the overall Type I error, α, say. Thus, instead of choosing a cutoff value, g, say, such that

$$P(|G_i| < g) = 1 - \alpha, \tag{20}$$

we need to select g so that

$$P(\max|G_1| < g) = 1 - \alpha \tag{21}$$

or equivalently,

$$P(-\mathbf{g} < \mathbf{G} < \mathbf{g}) = 1 - \alpha \tag{22}$$

where **g** is an n-vector with all elements equal to g.

In general, when the elements of **G** are dependent, the distribution of the maximum is intractable. The only cases where tractable expressions exist arise when $r_{ik} = r$ for $i \neq k$, or when $r_{ik} = a_i^* a_k$, $i \neq k$; see Stuart and Ord (1994, pp. 518–19) for details.

One possible approach would be to undertake a simulation study on a case-by-case basis, but this operation would be very time-consuming for large n, and undermines the notion of having the G_i statistics available as a rapid diagnostic device. However, an inequality due to T. W. Anderson (1955) and Sidak (1967) enables us to make some progress. For our present purpose, the inequality may be stated as follows:

> assume that $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are multivariate normally distributed with means zero, variances one. Further, assume that their covariance matrices, $\mathbf{V}_1$ and $\mathbf{V}_2$, are such that $\mathbf{V}_1 \geq \mathbf{V}_2$ meaning that every off-diagonal element of $\mathbf{V}_1$ is equal to or greater than every off-diagonal element of $\mathbf{V}_2$. Then it may be shown that
>
> $$P(-\mathbf{g} \leq \mathbf{Z}_1 \leq \mathbf{g}) \geq P(-\mathbf{g} \leq \mathbf{Z}_2 \leq \mathbf{g}) \tag{23}$$
>
> for any vector of non-negative values, **g**.

As an application of (22), let $\mathbf{V}_1$ have all elements non-negative and set $\mathbf{V}_2 = \mathbf{I}$, the identity matrix, corresponding to independence. If we write $\mathbf{g} = (g, g, \ldots, g)'$, $\mathbf{G} = (G_1, \ldots, G_n)'$ and set

$$P(|G_1| \leq g) = 1 - \alpha_1, \qquad \text{for all } i,$$

it follows from (23) that

$$P(\max|G_i| \leq g) = P(-\mathbf{g} \leq \mathbf{G} \leq \mathbf{g}) \geq P(-\mathbf{g} \leq \mathbf{Z}_2 \leq \mathbf{g}) = (1 - \alpha_1)^n. \tag{24}$$

TABLE 3
Standard Measure for Various Percentiles for the Largest of n Independent and Identically Distributed Normal Random Variables*

n	.90	.95	.99	.999
1	1.2816	1.6450	2.3238	3.0917
2	1.6323	1.9546	2.5759	3.2944
3	1.8183	2.1214	2.7130	3.4100
4	1.9432	2.2342	2.8067	3.4900
5	2.0367	2.3189	2.8769	3.5500
6	2.1107	2.3865	2.9364	3.5875
7	2.1718	2.4425	2.9840	3.6286
8	2.2239	2.4900	3.0225	3.6750
9	2.2692	2.5319	3.0583	3.6927
10	2.3091	2.5683	3.0888	3.7248
11	2.3447	2.6015	3.1189	3.7475
12	2.3767	2.6304	3.1443	3.7655
13	2.4060	2.6575	3.1708	3.7854
14	2.4329	2.6827	3.1893	3.8030
15	2.4575	2.7060	3.2091	3.8212
16	2.4806	2.7274	3.2282	3.8367
17	2.5018	2.7478	3.2455	3.8509
18	2.5219	2.7656	3.2621	3.8659
19	2.5413	2.7825	3.2784	3.8791
20	2.5594	2.7993	3.2931	3.8909
30	2.6964	2.9291	3.4050	3.9885
40	2.7913	3.0175	3.4900	4.0551
50	2.8631	3.0833	3.5488	4.1073
60	2.9218	3.1400	3.5906	4.1488
100	3.0778	3.2889	3.7238	4.2659
500	3.5375	3.7134	4.1075	4.6200
1000	3.7062	3.8855	4.2643	4.7667

* The G_i and G_i^* values are assumed to be independently normal and identically distributed in order to use this table.

Thus, from (21) and (24) we need to choose α_1 such that

$$1 - \alpha = (1 - \alpha_1)^n \tag{25}$$

in order to control the overall probability of Type I error at α or less. Equations (24) and (25) describe the Bonferroni correction for multiple tests, used in Getis and Ord (1992) but not stated explicitly. Table 3 gives the standard measure for various percentiles for the largest G_i or G_i^* of n values. It should be noted that the use of (25) does not absolutely guarantee that the probability of Type I error will be less than or equal to α since, as we saw from Figure 4, the autocorrelations may become negative as the distance between regions increases. However, the risk is very slight unless m/n becomes sizable.

Inequality (23) gives us another way of exploring the distributional issues as follows. From (19) we know that

$$r_{ik} < c_{ik}, < c \quad \text{say, where } c = \max(c_{ik}).$$

Thus, by setting the off-diagonal elements of $\mathbf{V}_1$ equal to c in (23) we may obtain a lower bound for α. If

$$f(x) = (2\pi)^{\frac{1}{2}} \exp(-\frac{1}{2}x^2), \qquad -\infty < x < \infty$$

denotes the p.d.f. for the standard normal distribution, it follows that

$$P(\max G_i \leq g) = \int_{-\infty}^{\infty} f(x) \left[\int_{-\infty}^{v} f(u)du\right]^n dx = 1 - p, \text{ say,} \tag{26}$$

where $v = (g - cx)/(1 - c^2)^n$. It follows that $\alpha > 2p$, so that an appropriate percentage point can be computed. That is, given c, we can find a value $g(c)$ that represents a lower bound for the value of g that should be used in (21), if the overall probability of Type I error is to be controlled. The standard measure for the ninety-fifth percentile when all variates are equally correlated, for $n = 30$, for correlations of 0.0, 0.6, 0.9, and 1.0 are 2.935, 2.845, 2.415, and 1.6454, respectively. These values indicate that the lower bound does not change very much until the correlation becomes quite large, suggesting that the Bonferroni bounds may, in fact, be reasonable. Alternatively, given the observed value, g, the p-value may be computed directly from (26).

Three approximate procedures were also tried in an effort to simplify the calculations, since (26) involves numerical quadrature. The two forms based upon Gumbels' extreme value distributions (cf. Stuart and Ord 1994, pp. 482–87) were poor and are not recommended. That based upon Tukey's lambda distribution (Stuart and Ord 1987, pp. 450–51) fared well except for very large n and represents a reasonable approximation.

In summary, we believe that the limits given by the Bonferroni inequality, possibly with the Tukey approximation, represent a reasonable basis for making inferences until more precise methods are developed.

One of the referees has noted that the Bonferroni limits may be unduly conservative when n becomes large. This issue needs to be explored further.

7. EFFECTS OF GLOBAL AUTOCORRELATION

When n is large, so that $(\bar{x}, s)$ are very close to (μ, σ), we may write (10) as

$$\begin{aligned} G_i^*(d) \equiv G_i^* &= c \cdot \sum w_{ij}(d)\,(x_j - \mu) \\ &= c\mathbf{w}_i'\mathbf{y}, \end{aligned} \tag{27}$$

where $\mathbf{y}' = (y_1, \ldots, y_n)$, $y_j = x_j - \mu$, $\mathbf{w}_i' = \{w_{ij}(d),\ j = 1, \ldots, n\}$ and c is a constant determined directly from (10).

In order to determine the effects of global autocorrelation, we must use a normal model, since the permutations argument is no longer available. Suppose that

$$\mathbf{y} \text{ is } N[\mathbf{0},\ \sigma^2(\mathbf{I} + \mathbf{A})], \tag{28}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{A}$ is a symmetric matrix of weights such that $\mathbf{I} + \mathbf{A}$ is positive definite. The distribution specified in (28) is a general form of spatial moving average (SMA) scheme (cf. Cliff and Ord 1981, pp. 149–51). It follows directly from (27) and (28) that

$$E[G_i^*] = 0,\ \text{Var}\,(G_i^*) = \sigma^2 c^2 \mathbf{w}_i'\,(\mathbf{I} + \mathbf{A})\mathbf{w}_i. \tag{29}$$

In general, the weights are non-negative and the global autocorrelation in the SMA scheme will be positive, so that the variance in (29) will be increased, making individual coefficients less extreme. As a particular case, suppose we

set $w_{ij} = 1$, $0 \leq d(i,j) \leq d$, $w_{ij} = 0$, otherwise, $a_{ij} = \alpha > 0$, $0 < d(i,j) \leq D$, $a_{ij} = 0$, otherwise, where $\mathbf{A} = \{a_{ij}\}$ and $D > d$ to give meaning to the notion of global or larger-scale dependence, relative to the local statistics. If there are n_δ locations within d of i (including i itself), it follows from (28) that

$$\text{Var}\,(G_i^*) = \sigma^2 c^2 n_\delta [1 + \alpha(n_\delta - 1)] \tag{30}$$

so that the variance is increased by the factor $[1 + \alpha(n_\delta - 1)]$.

This effect is as expected; when global autocorrelation exists, local pockets are harder to detect. Conversely, when no global pattern exists, G_i^* helps to monitor local behavior. The question of the interaction between local and global coefficients is an important one and much more remains to be done in this context; Anselin (1995) found that global autocorrelation has a significant impact upon the distribution of the local statistics.

8. AN EXAMPLE: THE OCCURRENCE OF AIDS CASES IN THE SAN FRANCISCO AREA

The epidemiological characteristics of AIDS have recently been studied by Golub, Gorr, and Gould (1993). They find evidence of a spatial hierarchical and expansionary trend since the onset of the disease in Ohio in 1981. In our study we consider the spread of the disease yearly through the counties of California by identifying spatial pattern trends beginning eight and ending twelve years after the first report of a case of AIDS in California. Thus our study relates to a somewhat later stage of the AIDS epidemic than the Ohio study.

Data on the incidence of the AIDS disease are available by county, by month, or quarter for the state of California for the period January 1989 to the present from the California Department of Health Services (see references). Our goal here is to study the pattern each year of the cumulative incidence of cases per 100,000 population in the region surrounding San Francisco County in order to use the G_i statistic to test the notion that AIDS spread spatially from San Francisco to nearby counties during this period. A spatial pattern of the incidence of AIDS had been established by 1989 (see Figure 5). The question here is simply: Have there been recognizable changes in the pattern since that time?

The random permutations approach requires that, under the null hypothesis H_o, all possible assignments of the county values to the counties are equally likely. As noted earlier, Besag and Newell (1991) have pointed out that the clustering of more-urban, and usually more-populous, regions may invalidate this assumption. In most instances, since a standard confidence interval for the proportion is proportional to $1/\{\textit{square root of population}\}$, these errors will not be substantial for sufficiently large population values (t_i). In California, however, county populations vary greatly and so do incidence rates. Comparisons of the raw incidence rates for each of the years 1988 to 1992 with the z scores obtained from (1) produced correlations in the range 0.65 to 0.67 without using San Francisco County data. San Francisco, with its cumulative incidence rate in 1993 of 1664.46 per 100,000 population, has a rate that is six times higher than that of the second-highest county, Marin.

The incidence of AIDS in San Francisco County (which has the same boundaries as the city of San Francisco) is among the highest in the country both absolutely and in terms of the rate per 100,000 population. By December 1992, 12,050 cases of AIDS had been reported since the inception of the disease in 1981. Of those, 8,982 had died. Aldrich et al. (1990) and Smallman-Raynor, Cliff, and Haggett (1992) studied the spatial distribution of AIDS

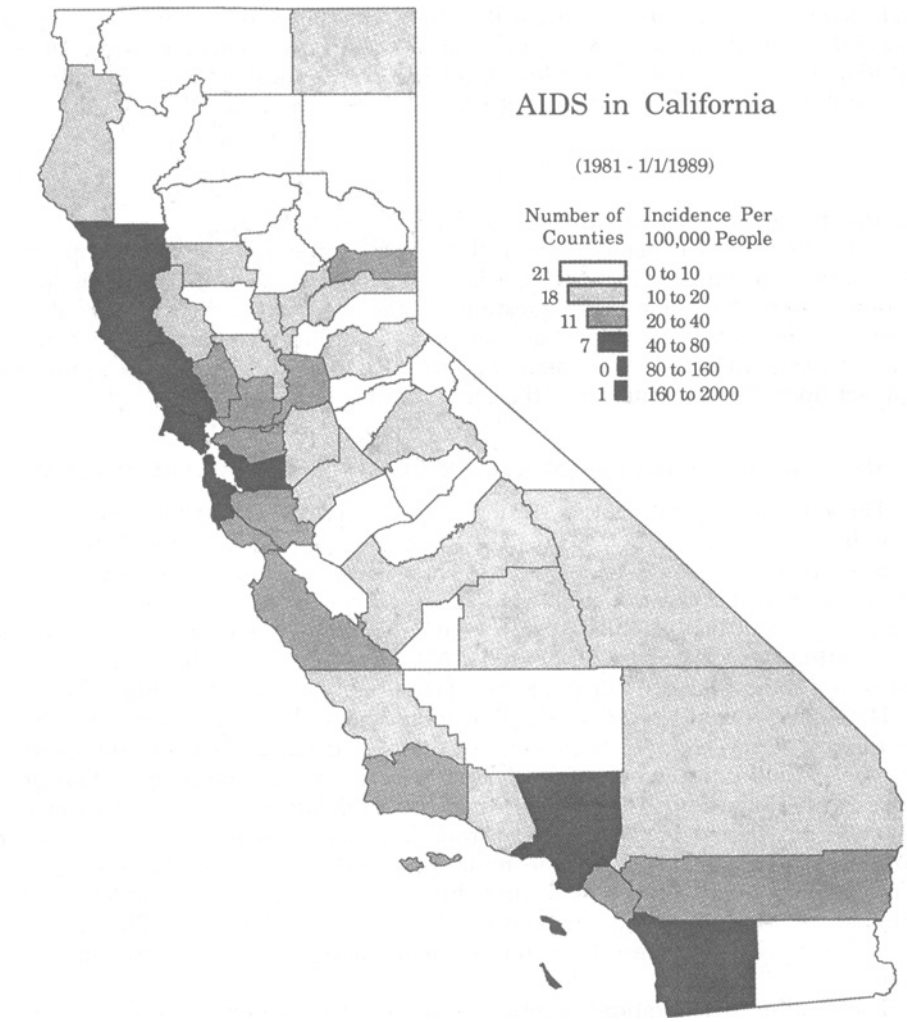

FIG. 5

within San Francisco. They found that the vast majority of cases (87 percent) are among homosexuals and that the cases cluster to a very high degree in central San Francisco, especially the Castro district.

The G_i statistic focuses on clustering around San Francisco since it does not take into account the rate of AIDS in San Francisco itself (recall that $j \neq i$). The statistic allows us to identify clustering of the disease as distance increases from San Francisco without being affected by the large number of cases in San Francisco. The null hypothesis under test is that the observations are not spatially dependent, but we are focusing on how counties relate to San Francisco rather than to all counties of the state. The conservative approach is to apply the Bonferroni adjustment as described earlier, recognizing that San Francisco was selected on the basis of prior knowledge. On the other hand, since we are focusing on spatial pattern rather than the incidence rates themselves, the usual z scores could also be viewed as appropriate. The choice is not clear-cut and a

researcher must weigh the form of the alternative hypothesis carefully before drawing any conclusions.

In our use of the G_i statistic, we employed a number of different distances from San Francisco. In principle, it is possible to choose that value of d for which the standardized statistic is maximized, somewhat in the spirit of the analysis in Stone (1988) and Getis (1995). However, given that the values are assigned to the county population centroids and that some of the counties cover large areas, we felt that multiple $G_i(d)$ values are more informative.

Figures 5 and 6 show the incidence of AIDS at the beginning and the end of the study period. Note that it is difficult to discern visually differences in pattern. Figure 7 shows the circular bands representing six distance increments of twenty miles each that are used in the study. The points representing California

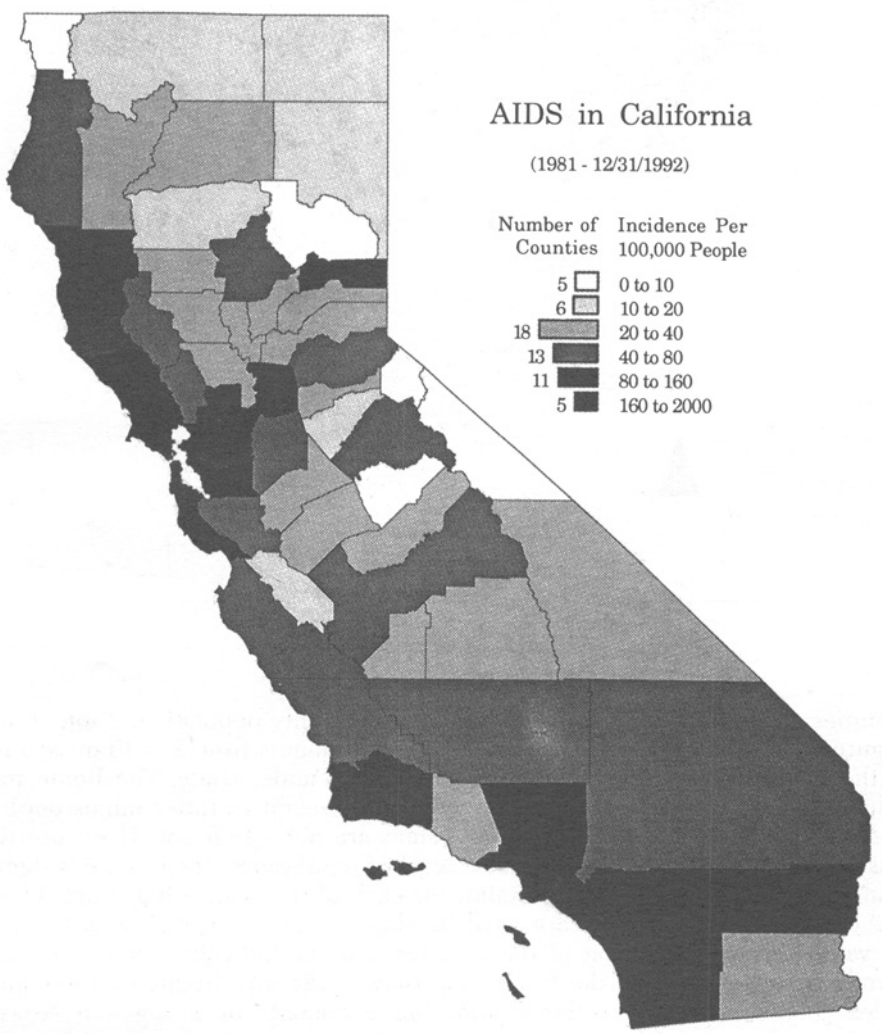

FIG. 6

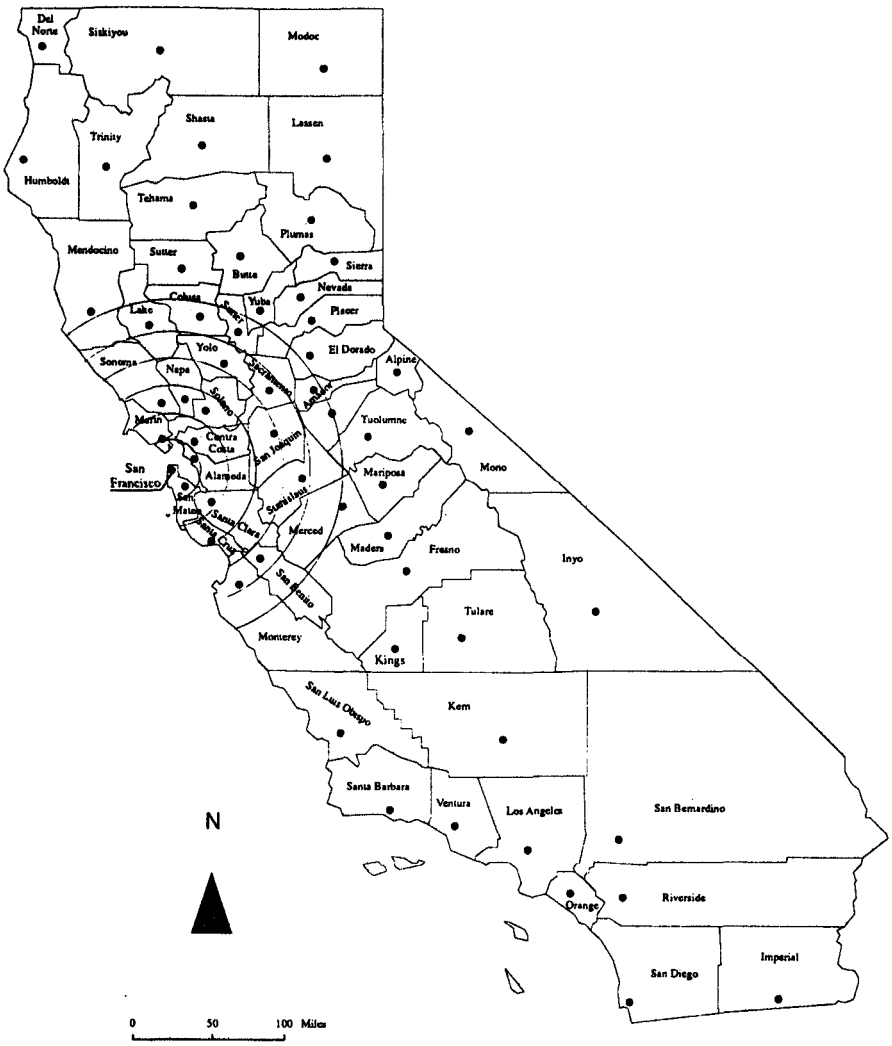

FIG. 7

counties are given by the estimated centers of county population. Table 4 and Figure 8 give the G_i values for twenty-mile increments from San Francisco for z (the corrected rate) for each year of the period under study. The Bonferroni value interpolated from Table 3 for $n = 57$ (fifty-eight counties minus one) at the .05 level is 3.125. That is, the G_i values are not significant. If we use the less conservative approach at the same level of significance, the disease is significantly clustered to at least forty miles for each of the years 1990–1992. When the original incidence rates are used in place of the z scores, the pattern of G_i values is not unlike that of the z scores, but the individual values are uniformly larger and exceed the Bonferroni value 3.125 for distances up to eighty miles (Figure 9), in accordance with the comments of Besag and Newell (1991).

These results are understandable if we suspect that the expected incidence may decrease monotonically with the distance from San Francisco. For example,

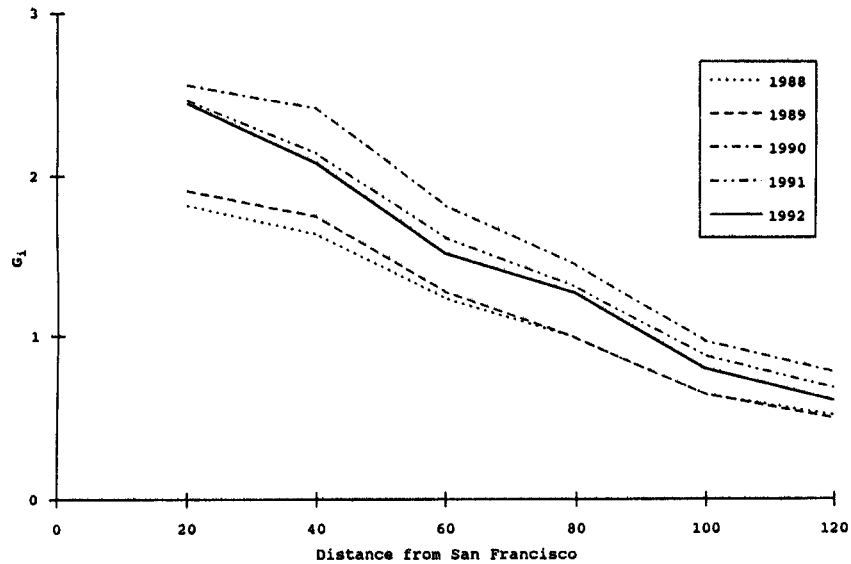

FIG. 8. G_i Values for Twenty-mile Increments from San Francisco for z, the Corrected Cumulative Incidence Rate

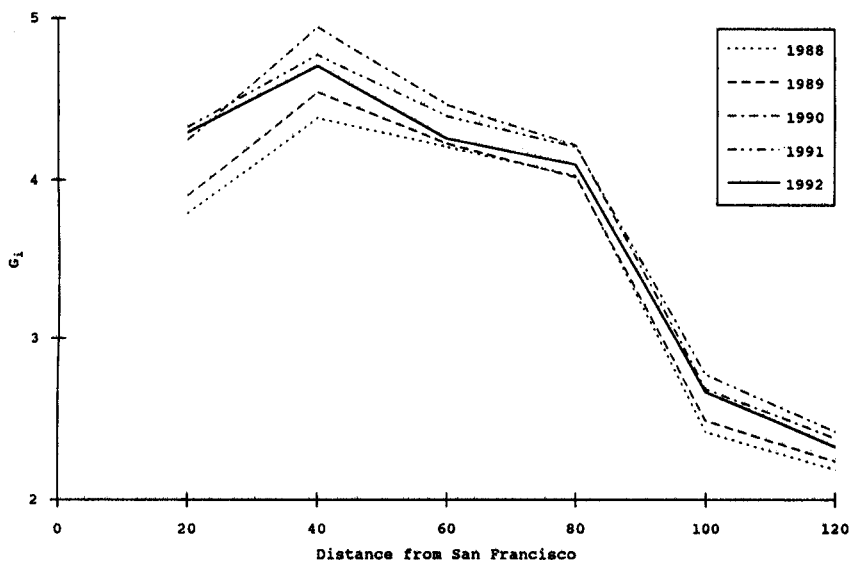

FIG. 9. G_i Values for Twenty-mile Increments from San Francisco for Cumulative Incidence Rates

let incidence p_i in region i be of the form

$$p_i = \alpha - \beta(d_i - D) + \epsilon_i \tag{31}$$

where d_i denotes the distance from the source, D is the mean distance from the source and ϵ_i denotes an error term with mean zero and variance σ^2 for all i. If we use binary weights, we obtain from (1),

TABLE 4
G_i Values of the z_i for the Cumulative Incidence of the AIDS Rate at Twenty-Mile Increments from San Francisco for the Years 1988–1992

Distance	1988	1989	1990	1991	1992
20	1.82	1.91	2.56	2.47	2.45
40	1.64	1.75	2.42	2.14	2.08
60	1.24	1.28	1.82	1.62	1.52
80	0.99	0.99	1.45	1.31	1.27
100	0.64	0.64	0.97	0.88	0.80
120	0.51	0.49	0.78	0.68	0.60

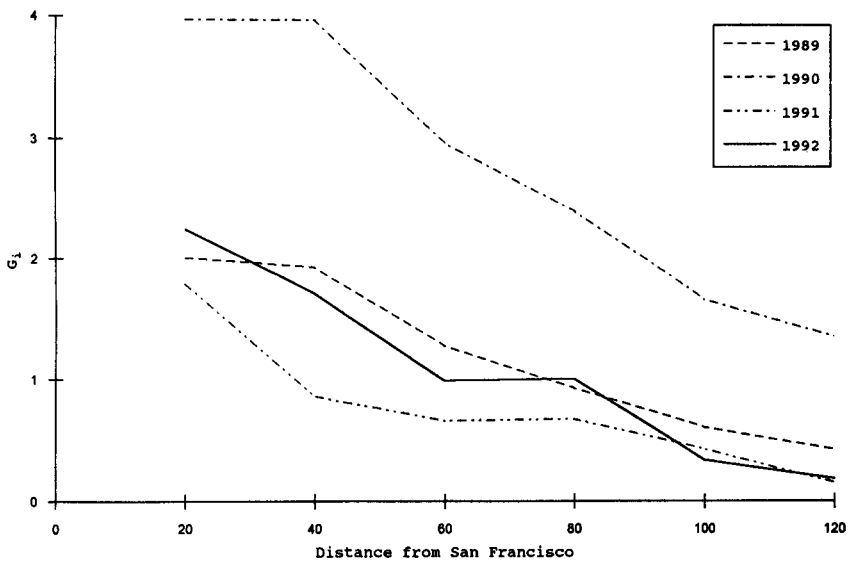

FIG. 10. G_i Values for Twenty-mile Increments from San Francisco for z, the Corrected Incidence Rate for the Given Year Only

$$G_i(d) \propto \sum_j p_j, \qquad \{j : d_j \leq d\}. \tag{32}$$

When (31) holds and the number of counties within d of county i is J (J much less than n), the expected value for (32) is, approximately,

$$-\beta \sum(d_j - D)/\sigma\sqrt{J}. \tag{33}$$

Examination of (33) indicates that the expectation will increase initially and then decline.

Since the distances are fixed, the results in Figure 8 suggest that a similar pattern of declining incidences holds for all time periods. The curves in Figure 8 imply that the rate of AIDS cases increased uniformly over the area of clustering from 1988 to 1990 and declined uniformly after 1990. The detail presented below refines that conclusion.

Figure 10 shows the G_i values for the changes in the rate (z) of AIDS cases (new cases per 100,000 population) for four years beginning in 1989 and ending 1992. All periods displayed roughly similar sequences of G_i values although there was a substantial boost to the already existing pattern in 1990. Such sim-

ilarities are consistent with the simple model (31) if α and β are functions of time but not of location. That is, the increased incidence appears to be following a rather stable pattern over time, with the regions closest to San Francisco exhibiting higher absolute increases. Thus, we obtain from (31) for successive time periods t and $t+1$:

$$E\{p_j(t+1) - p_j(t)\} = [\alpha(t+1) - \alpha(t)] - [\beta(t+1) - \beta(t)](d_j - D) \qquad (34)$$

which decreases as d_j increases. The only evidence of spread of the disease from San Francisco can be seen in the strong forty-mile showing in 1989 and 1990 and the slight increase in values at eighty miles in 1991 and 1992.

This brief spatial analysis allows us to answer the question posed earlier, that is, has there been a change in pattern during the study period? Is there any indication of spread? Our answer is that there has been very little pattern change. If there is any change, it is in the direction of an intensification in all counties, each at about the rate prevailing at the beginning of the period. During the study period, there is only a small indication of greater than local spread of the disease. This interpretation is consistent with studies of more traditional diseases. For example, Cliff et al. (1981, p. 148), in a study of measles epidemics in Iceland, note that "once an epidemic is established in a medical district, within rather than between-area effects dominate."

LITERATURE CITED

Aldrich, M. R., S. F. Payne, S. M. Little, J. Mandel, and H. W. Feldman (1990). "Classic Epidemiological Mapping of AIDS among San Francisco Drug Injectors, 1987–1989." *VI International Conference on AIDS*, San Francisco, 20–24 June, Abstract Th.C. 705.

Anderson, T. W. (1955). "The Integral of a Symmetric Unimodal Function over a Symmetric Convex Set and Some Probability Inequalities." *Proceedings American Mathematical Society* 6, 170–76.

Anselin, L. (1995). "Local Indicators of Spatial Association—LISA." *Geographical Analysis* 27 (April), 93–115.

Besag, J., and J. Newell (1991). "The Detection of Clusters in Rare Diseases." *Journal of the Royal Statistical Society, Series A* 154, 143–55.

California Department of Health Services, Office of AIDS, (1990). California AIDS Update, 1988–1990.

——— (1993). California HIV/AIDS Update, 1990–1993.

Cliff, A. D., P. Haggett, J. K. Ord, and G. R. Versey (1981). *Spatial Diffusion*. London: Cambridge University Press.

Cliff, A. D., and J. K. Ord (1981). *Spatial Processes: Models and Applications*. London: Pion Press.

Cressie, N. (1992). "Smoothing Regional Maps Using Empirical Bayes Predictors." *Geographical Analysis* 24 (January), 75–95.

Cuzik, J., and R. Edwards (1990). "Spatial Clustering for Inhomogeneous Populations." *Journal of the Royal Statistical Society. Series B* 52, 73–104 (with discussion).

Getis, A. (1995). "Spatial Filtering in a Regression Framework: Experiments on Regional Inequality, Government Expenditures, and Urban Crime." In *New Directions in Spatial Econometrics*, edited by L. Anselin and R. Florax. Amsterdam: North Holland.

Getis, A., and J. K. Ord (1992). "The Analysis of Spatial Association by Use of Distance Statistics." *Geographical Analysis* 24 (July), 189–206.

Golub, A., W. L. Gorr, and P. R. Gould (1993). "Spatial Diffusion of the HIV/AIDS Epidemic: Modeling Implications and Case Study of AIDS Incidence in Ohio." *Geographical Analysis* 25 (April), 85–100.

Openshaw, S., M. Charlton, C. Wymer, and A. Craft (1987). "A Mark I Geographical Analysis Machine for the Automated Analysis of Point Data Sets." *International Journal of Geographical Information Systems* 1, 335–58.

Sidak, Z. (1967). "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions."

Journal of the American Statistical Association 62, 626–33.

Smallman-Raynor, M., A. D. Cliff, and P. Haggett (1992). *Atlas of AIDS*. Oxford: Basil Blackwell International.

Stone, R. A. (1988). "Investigations of Excess Environmental Risks around Putative Sources: Statistical Problems and a Proposed Test." *Statistics in Medicine* 7, 649–60.

Stuart, A., and J. K. Ord (1994). *Kendall's Advanced Theory of Statistics*, vol. 1, 6th ed. London: Arnold, and New York: John Wiley & Sons.