Arthur Getis
J. K. Ord

The Analysis of Spatial Association by Use of Distance Statistics

Introduced in this paper is a family of statistics, G, *that can be used as a measure of spatial association in a number of circumstances. The basic statistic is derived, its properties are identified, and its advantages explained. Several of the* G *statistics make it possible to evaluate the spatial association of a variable within a specified distance of a single point. A comparison is made between a general* G *statistic and Moran's* I *for similar hypothetical and empirical conditions. The empirical work includes studies of sudden infant death syndrome by county in North Carolina and dwelling unit prices in metropolitan San Diego by zip-code districts. Results indicate that* G *statistics should be used in conjunction with* I *in order to identify characteristics of patterns not revealed by the* I *statistic alone and, specifically, the* G_i *and* G_i^* *statistics enable us to detect local "pockets" of dependence that may not show up when using global statistics.*

INTRODUCTION

The importance of examining spatial series for spatial correlation and autocorrelation is undeniable. Both Anselin and Griffith (1988) and Arbia (1989) have shown that failure to take necessary steps to account for or avoid spatial autocorrelation can lead to serious errors in model interpretation. In spatial modeling, researchers must not only account for dependence structure and spatial heteroskedasticity, they must also assess the effects of spatial scale. In the last twenty years a number of instruments for testing for and measuring spatial autocorrelation have appeared. To geographers, the best-known statistics are Moran's *I* and, to a lesser extent, Geary's *c* (Cliff and Ord 1973). To geologists and remote sensing analysts, the semi-variance is most popular (Davis 1986). To spatial econometricians, estimating spatial autocorrelation coefficients of regression equations is the usual approach (Anselin 1988).

The authors wish to thank the referees for their perceptive comments on an earlier draft, which led to considerable improvements in the paper.

Arthur Getis is professor of geography at San Diego State University. 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.

Geographical Analysis, Vol. 24, No. 3 (July 1992)
Submitted 9/90. Revised version accepted 4/16/91.

A common feature of these procedures is that they are applied globally, that is, to the complete region under study. However, it is often desirable to examine pattern at a more local scale, particularly if the process is spatially nonstationary. Foster and Gorr (1986) provide an adaptive filtering method for smoothing parameter estimates, and Cressie and Read (1989) present a modeling procedure. The ideas presented in this paper are complementary to these approaches in that we also focus upon local effects, but from the viewpoint of testing rather than smoothing.

This paper introduces a family of measures of spatial association called G statistics. These statistics have a number of attributes that make them attractive for measuring association in a spatially distributed variable. When used in conjunction with a statistic such as Moran's I, they deepen the knowledge of the processes that give rise to spatial association, in that they enable us to detect local "pockets" of dependence that may not show up when using global statistics. In this paper, we first derive the statistics $G_i(d)$ and $G(d)$, then outline their attributes. Next, the $G(d)$ statistic is compared with Moran's I. Finally, there is a discussion of empirical examples. The examples are taken from two different geographic scales of analysis and two different sets of data. They include sudden infant death syndrome by county in North Carolina, and house prices by zip-code district in the San Diego metropolitan area.

THE $G_i(d)$ STATISTIC

This statistic measures the degree of association that results from the concentration of weighted points (or area represented by a weighted point) and all other weighted points included within a radius of distance d from the original weighted point. We are given 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 (a weight) taken from a variable X. The variable has a natural origin and is positive. The $G_i(d)$ statistic developed below allows for tests of hypotheses about the spatial concentration of the sum of x values associated with the j points within d of the ith point.

The statistic is

$$G_i(d) = \frac{\sum_{j=1}^{n} w_{ij}(d)x_j}{\sum_{j=1}^{n} x_j}, \quad j \text{ not equal to } i, \tag{1}$$

where $\{w_{ij}\}$ 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. The numerator 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.

Adopting standard arguments (cf. Cliff and Ord 1973, pp. 32–33), we may fix the value x_i for the ith point and consider the set of $(n - 1)!$ random permutations of the remaining x values at the j points. Under the null hypothesis of spatial independence, these permutations are equally likely. That is, let X_j be the random variable describing the value assigned to point j, then

$$P(X_j = x_r) = 1/(n - 1), \quad r \neq i,$$

and

$$E(X_j) = \sum_{r \neq i} x_r/(n - 1).$$

$$\text{Thus} \qquad E(G_i) = \sum_{j \neq i} w_{ij}(d)\, E(X_j) / \sum_{j \neq i} X_j$$

$$= W_i/(n-1)\,, \tag{2}$$

$$\text{where} \qquad W_i = \Sigma_j\, w_{ij}(d)\,.$$

Similarly,

$$E(G_i^2) = \frac{1}{(\Sigma_j\, x_j)^2}\left[\Sigma_j\, w_{ij}^2(d)\, E(X_j^2) + \sum_{j \neq k}\sum w_{ij}(d) w_{ik}(d) E(X_j X_k)\right]$$

Since $$E(X_j^2) = \sum_{r \neq i} x_r^2/(n-1)$$

and $$E(X_j X_k) = \sum_{r \neq s \neq i}\sum x_r x_s/(n-1)(n-2)$$

$$= \{(\sum_{r \neq i} x_r)^2 - \sum_{r \neq i} x_r^2\}/(n-1)(n-2)\,.$$

Recalling that the weights are binary

$$\sum_{j \neq k}\sum w_{ij} w_{ik} = W_i^2 - W_i$$

and so

$$E(G_i^2) = \frac{1}{(\Sigma_j\, x_j)^2}\left\{\frac{W_i \Sigma_j\, x_j^2}{(n-1)} + \frac{W_i(W_i-1)}{(n-1)(n-2)}\left[(\Sigma_j\, x_j)^2 - \Sigma_j\, x_j^2\right]\right\}.$$

Thus $\mathrm{Var}(G_i) = E(G_i^2) - E^2(G_i)$

$$= \frac{1}{(\Sigma_j\, x_j)^2}\left[\frac{W_i(n-1-W_i)\,\Sigma_j\, x_j^2}{(n-1)(n-2)}\right] + \frac{W_i(W_i-1)}{(n-1)(n-2)} - \frac{W_i^2}{(n-1)^2}\,.$$

If we set $\dfrac{\Sigma_j\, x_j}{(n-1)} = Y_{i1}$ and $\dfrac{\Sigma_j\, x_j^2}{(n-1)} - Y_{i1}^2 = Y_{i2}$,

$$\text{then } \mathrm{Var}(G_i) = \frac{W_i(n-1-W_i)}{(n-1)^2(n-2)}\left(\frac{Y_{i2}}{Y_{i1}^2}\right). \tag{3}$$

As expected, $\mathrm{Var}(G_i) = 0$ when $W_i = 0$ (no neighbors within d), or when $W_i = n-1$ (all $n-1$ observations are within d), or when $Y_{i2} = 0$ (all $n-1$ observations are equal).

Note that W_i, Y_{i1}, and Y_{i2} depend on i. Since G_i is a weighted sum of the variable X_j, and the denominator of G_i is invariant under random permutations of $\{x_j,\ j \neq i\}$, it follows, provided $W_i/(n-1)$ is bounded away from 0 and from 1, that the permutations distribution of G_i *under* H_0 approaches normality as $n \to \infty$; cf. Hoeffding (1951) and Cliff and Ord (1973, p. 36). When d, and thus W_i, is small, normality is lost, and when d is large enough to encompass the whole study

TABLE 1
Characteristics of G_i Statistics

	j not equal to i	j may equal i
Statistic	$G_i(d)$	$G_i^*(d)$
Expression	$\dfrac{\Sigma_j\, w_{ij}(d)x_j}{\Sigma_j\, x_j}$	$\dfrac{\Sigma_j\, w_{ij}(d)x_j}{\Sigma_j\, x_j}$
Definitions	$W_i = \Sigma_j\, w_{ij}(d)$	$W_i^* = \Sigma_j\, w_{ij}(d)$
	$Y_{i1} = \dfrac{\Sigma_j\, x_j}{(n-1)}$	$Y_{i1}^* = \dfrac{\Sigma_j\, x_j}{n}$
	$Y_{i2} = \dfrac{\Sigma_j\, x_j^2}{(n-1)} - Y_{i1}^2$	$Y_{i2}^* = \dfrac{\Sigma\Sigma_{ij}\,(x_i x_j)^2}{n} - (Y_{i1}^*)^2$
Expectation	$W_i/(n-1)$	W_i^*/n
Variance	$\dfrac{W_i\,(n-1-W_i)\,Y_{i2}}{(n-1)^2\,(n-2)\,Y_{i1}^2}$	$\dfrac{W_i^*(n-W_i^*)\,Y_{i2}^*}{n^2(n-1)\,(Y_{i1}^*)^2}$

area, and thus $(n - 1 - W_i)$ is small, normality is also lost. It is important to note that the conditions must be satisfied separately for *each* point if its G_i is to be assessed via the normal approximation.

Table 1 shows the characteristic equations for $G_i(d)$ and the related statistic, $G_i^*(d)$, which measures association in cases where the j equal to i term is included in the statistic. This implies that any concentration of the x values includes the x at i. Note that the distribution of $G_i^*(d)$ is evaluated under the null hypothesis that all $n!$ random permutations are equally likely.

ATTRIBUTES OF G_i STATISTICS

It is important to note that G_i is scale-invariant ($Y_i = bX_i$ yields the same scores as X_i) but not location-invariant ($Y_i = a + X_i$ gives different results than X_i). The statistic is intended for use only for those variables that possess a natural origin. Like all other such statistics, transformations like $Y_i = \log X_i$ will change the results.

$G_i(d)$ measures the concentration or lack of concentration of the sum of values associated with variable X in the region under study. $G_i(d)$ is a proportion of the sum of all x_j values that are within d of i. If, for example, high-value x_js are within d of point i, then $G_i(d)$ is high. Whether the $G_i(d)$ value is statistically significant depends on the statistic's distribution.

Earlier work on a form of the $G_i(d)$ statistic is in Getis (1984), Getis and Franklin (1987), and Getis (1991). Their work is based on the second-order approach to map pattern analysis developed by Ripley (1977).

In typical circumstances, the null hypothesis is that the set of x values within d of location i is a random sample drawn without replacement from the set of all x values. The estimated $G_i(d)$ is computed from equation (1) using the observed x_j values. Assuming that $G_i(d)$ is approximately normally distributed, when

$$Z_i = \{G_i(d) - E[G_i(d)]\}/\sqrt{\text{Var } G_i(d)} \tag{4}$$

is positively or negatively greater than some specified level of significance, then we say that positive or negative spatial association obtains. A large positive Z_i implies that large values of x_j (values above the mean x_j) are within d of point i. A large negative Z_i means that small values of x_j are within d of point i.

A special feature of this statistic is that the pattern of data points is neutralized when the expectation is that all x values are the same. This is illustrated for the case when data point densities are high in the vicinity of point i, and d is just large enough to contain the area of the clustered points. Theoretical $G_i(d)$ values are high because W_i is high. However, only if the observed x_j values in the vicinity of point i differ systematically from the mean is there the opportunity to identify significant spatial concentration of the sum of x_js. That is, as data points become more clustered in the vicinity of point i, the expectation of $G_i(d)$ rises, neutralizing the effect of the dense cluster of j values.

In addition to its above meaning, the value of d can be interpreted as a distance that incorporates specified cells in a lattice. It is to be expected that neighboring G_i will be correlated if d includes neighbors. To examine this issue, consider a regular lattice. When n is large, the denominator of each G_i is almost constant so it follows that corr (G_i, G_j) = proportion of neighbors that i and j have in common.

EXAMPLE 1

Consider the rook's case. Cell i has no common neighbors with its four immediate neighbors, but two with its immediate diagonal neighbors. The numbers of common neighbors are as illustrated below:

	0	1	0	
0	2	0	2	0
1	0	*i*	0	1
0	2	0	2	0
	0	1	0	

All the other cells have no common neighbors with i. Thus, the G-indices for the four diagonal neighbors have correlations of about 0.5 with G_i, four others have correlations of about 0.25 and the rest are virtually uncorrelated.

For more highly connected lattices (such as the queen's case) the array of nonzero correlations stretches further, but the maximum correlation between any pair of G-indices remains about 0.5. ▲

EXAMPLE 2

m	*m*	*m*	*m*	*m*	*m*	*m*	*m*	*m*	*m*
m	*A*	*A*	*A*	*m*	*m*	*B*	*B*	*B*	*m*
m	*A*	*A*	*A*	*m*	*m*	*B*	*B*	*B*	*m*
m	*A*	*A*	*A*	*m*	*m*	*B*	*B*	*B*	*m*
m	*m*	*m*	*m*	*m*	*m*	*m*	*m*	*m*	*m*

Set $A + B = 2m$, therefore $\bar{x} = m$; $n = 50$;
$A \geq 0$;
$B \geq 0$;
put $A = m(1 + c)$, $B = m(1 - c)$, $0 \leq c \leq 1$
Using this example, the G_i and G_i^* statistics are compared in the following table.

G_i and G_i^* Values (queen's case; non-edge cells)

cell	G_i	$Z(G_i)$	G_i^*	$Z(G_i^*)$
A, surrounded by *A*s	$\frac{8 + 8c}{49 - c}$	5.30#	$\frac{9 + 9c}{50}$	5.47
A, adjacent to *m*s	$\frac{8 + 3c}{49 - c}$	2.06#	$\frac{9 + 4c}{50}$	2.43

cell	G_i	$Z(G_i)$	G_i^*	$Z(G_i^*)$
central m, adjacent to As	$\frac{8 + 3c}{49}$	1.89#	$\frac{9 + 3c}{50}$	1.82
other m, adjacent to As	$\frac{8 + 2c}{49}$	1.26#	$\frac{9 + 3c}{50}$	1.21

Values for Bs are the same, with negative signs attached.
These values are lower bounds as $c \rightarrow 1$; they vary only slightly with c.

We note that G_i and G_i^* are similar in this case; if the central A was replaced by a B, $Z(G_i)$ would be unchanged, whereas $Z(G_i^*)$ drops to 4.25. Thus, G_i and G_i^* typically convey much the same information. ▲

EXAMPLE 3

Consider a *large* regular lattice for which we seek the distribution under H_0 for G_i^* with W_i neighbors. Let p = proportion of As = proportion of Bs and $1 - 2p$ = proportion of ms.

Let (k_1, k_2, k_3) denote the number of As, Bs, and ms, respectively so that $k_1 + k_2 + k_3 = n$. For large lattices, in this case, the joint distribution is approximately tri(multi-)nomial with index W and parameters $(p, p, 1 - 2p)$.

Since $$G_i^* = \frac{W_i + (k_1 - k_2)c}{n}$$

clearly $E(G_i^*) = W_i/n$ as expected

and $V(G_i^*) = 2pW_i/n$,

reflecting the large sample approximation. The distribution is symmetric and the standardized fourth moment is

$$3 + \frac{(1 - 6p)}{2pW_i} .$$

This is close to 3 provided pW_i is not too small.

Since we are using G_i and G_i^* primarily in a diagnostic mode, we suggest that $W_i \geq 8$ at least (that is, the queen's case), although further work is clearly necessary to establish cut-off values for the statistics. ▲

A GENERAL G STATISTIC

Following from these arguments, a general statistic, $G(d)$, can be developed. The statistic is general in the sense that it is based on all pairs of values (x_i, x_j) such that i and j are within distance d of each other. No particular location i is fixed in this case. The statistic is

$$G(d) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(d) x_i x_j}{\sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j} , \quad j \text{ not equal to } i . \tag{5}$$

The G-statistic is a member of the class of linear permutation statistics, first introduced by Pitman (1937). Such statistics were first considered in a spatial context by Mantel (1967) and Cliff and Ord (1973), and developed as a general cross-product statistic by Hubert (1977 and 1979), and Hubert, Golledge, and Costanzo (1981).

For equation (5),

$$W = \sum_{i=1} \sum_{j=1} w_{ij}(d) \ , \quad j \text{ not equal to } i$$

so that

$$E[G(d)] = W/[n(n-1)] \ . \tag{6}$$

The variance of G follows from Cliff and Ord (1973, pp. 70–71):

$$E(G^2) = \frac{1}{(m_1^2 - m_2)^2 \, n^{(4)}} [B_0 m_2^2 + B_1 m_4 + B_2 m_1^2 m_2 + B_3 m_1 m_3 + B_4 m_1^4]$$

where $m_j = \sum_{i=1} x_i^j \ , \quad j = 1, 2, 3, 4 \ ,$

and $n^{(r)} = n(n-1)(n-2) \ldots (n-r+1) \ .$

The coefficients, B, are

$$B_0 = (n^2 - 3n + 3)S_1 - nS_2 + 3W^2 \ ;$$

$$B_1 = -[(n^2 - n)S_1 - 2nS_2 + 3W^2] \ ;$$

$$B_2 = -[2nS_1 - (n+3)S_2 + 6W^2] \ ;$$

$$B_3 = 4(n-1)S_1 - 2(n+1)S_2 + 8W^2 \ ;$$

and $B_4 = S_1 - S_2 + W^2$

where $S_1 = \tfrac{1}{2} \sum_i \sum_j (w_{ij} + w_{ji})^2 \ , \quad j \text{ not equal to } i \ ,$

and $S_2 = \sum_i (w_{i.} + w_{.i})^2 \ ; \quad w_{i.} = \Sigma_j w_{ij} \ , \quad j \text{ not equal to } i \ ;$

thus

$$\text{Var}(G) = E(G^2) - \{W/[n(n-1)]\}^2 \ . \tag{7}$$

THE *G(d)* STATISTIC AND MORAN'S *I* COMPARED

The $G(d)$ statistic measures overall concentration or lack of concentration of all pairs of (x_i, x_j) such that i and j are within d of each other. Following equation (5), one finds $G(d)$ by taking the sum of the multiples of each x_i with all x_js within d of all i as a proportion of the sum of all $x_i x_j$. Moran's I, on the other hand, is often used to measure the correlation of each x_i with all x_js within d of i and, therefore,

is based on the degree of covariance within d of all x_i. Consider K_1, K_2 as constants invariant under random permutations. Then using summation shorthand we have

$$G(d) = K_1 \Sigma\Sigma\, w_{ij}\, x_i x_j$$

and

$$I(d) = K_2 \Sigma\Sigma\, w_{ij}\, (x_i - \bar{x})(x_j - \bar{x}) .$$

$$= (K_2/K_1)\, G(d) - K_2 \bar{x}\, \Sigma\, (w_{i.} + w_{.i})x_i + K_2 \bar{x}^2 W$$

where $w_{i.} = \sum_j w_{ij}$ and $w_{.i} = \sum_j w_{ji}$.

Since both $G(d)$ and $I(d)$ can measure the association among the same set of weighted points or areas represented by points, they may be compared. They will differ when the weighted sums $\Sigma w_{i.} x_i$ and $\Sigma w_{.i} x_i$ differ from $W\bar{x}$, that is, when the patterns of weights are unequal. The basic hypothesis is of a random pattern in each case. We may compare the performance of the two measures by using their equivalent Z values of the approximate normal distribution.

EXAMPLE 4

Let us use the lattice of Example 2. As before,
Set $A + B = 2m$, therefore $\bar{x} = m$; $n = 50$;
$A \geq 0$;
$B \geq 0$;
put $A = m(1 + c)$, $B = m(1 - c)$, $0 \leq c \leq 1$.
In addition, put
$a = A - m$;
$B = 2m - A = m - a$;
$B - m = a$;
$m \geq a$;
j not equal to i.

For the rook's case, $W = \Sigma\Sigma\, w_{ij} = 170$.

$$I = \frac{n \Sigma\Sigma\, w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{W \Sigma\, (x_i - \bar{x})^2} = \frac{50 \cdot 24a^2 \cdot 2}{170 \cdot 18a^2} = 0.784$$

for all choices of a, m.

$$\text{Var}\,(I) = 0.010897$$

$$Z(I) = 7.7088 \text{ whenever } A > B .$$

$$G = \frac{\Sigma\Sigma\, w_{ij}\, x_i x_j}{\Sigma\Sigma\, x_i x_j} = \frac{24A^2 + 24B^2 + 24Am + 24Bm + 74m^2}{2500m^2 - 9A^2 - 9B^2 - 32m^2}$$

$$= \frac{170 + 48c^2}{2450 - 18c^2}$$

When $c = 0$, $A = B = m$, and G is a minimum.

$$G_{min} = 170/2450 = 0.0694 \text{ and}$$

$$\text{Var}(G_{min}) = 0.0000 \text{ from equation (7)} .$$

When $c = 1$, $A = 2m$, $B = 0$, and G is a maximum.

$$G_{max} = 218/2432 = 0.0896 .$$

$$\text{Var}(G_{max}) = 0.000011855 .$$

$$Z(G^{max}) = 5.87 \text{ for any } m .$$

G depends on the relative absolute magnitudes of the sample values. Note that I is positive for any A and B, while G values approach a maximum when the ratio of A to B or B to A becomes large. ▲

EXAMPLE 5

m	m	m	m	m	m	m	m	m	m
m	m	m	m	m	m	m	m	m	m
m	m	A	m	m	m	m	B	m	m
m	m	m	m	m	m	m	m	m	m
m	m	m	m	m	m	m	m	m	m

A, B, $\bar{x}$, n, W as in Examples 2 and 4.

$$I = 0, \text{ for any possible } A, B, \text{ or } m .$$

$$Z(I) = 0.1920 \text{ since } E(I) = -1/(n-1), \text{ whenever } A > B .$$

$$G_{min} = G_{max} = 0.0694, \text{ for any possible } A, B, \text{ or } m .$$

$$\text{Var}(G_{min}) = 0, \text{ but } \text{Var}(G_{max}) = 0.00000059 .$$

$$Z(G^{max}) = 0.0739 .$$

Neither statistic can differentiate between a random pattern and one with little spatial variation. Contributions to $G(d)$ are large only when the product $x_i x_j$ is large, whereas contributions to $I(d)$ are large when $(x_i - m)(x_j - m)$ is large. It should be noted that the distribution is nowhere near normal in this case. ▲

EXAMPLE 6

m	m	m	m	m	m	m	m	m	m
m	A	B	A	m	m	B	A	B	m
m	B	A	B	m	m	A	B	A	m
m	A	B	A	m	m	B	A	B	m
m	m	m	m	m	m	m	m	m	m

A, B, $\bar{x}$, n, W as in the above examples.

$$I = -0.7843$$

$$\text{Var}(I) = 0.010897$$

$$Z(I) = -7.3177$$

$$\text{When } A = 2m \text{ and } B = 0 ,$$

$$G = 0.0502$$

$$\text{Var}(G) = 0.00001189$$

$$Z(G) = -5.5760$$

TABLE 2

Standard Normal Variates for $G(d)$ and $I(d)$ under Varying Circumstances for a Specified d Value

Situation	$Z(G)$	$Z(I)$
HH	+ +	+ +
HM	+	+
MM	0	0
Random	0	0
HL	−	− −
ML	−#	−
LL	− −	+ +

Key:
HH = pattern of high values of xs within d of other high x values
M = moderate values
L = low values
Random = no discernible pattern of xs
+ + = strong positive association (high positive Z scores)
+ = moderate positive association
0 = no association

This combination tends to be more negative than HL.

The juxtaposition of high values next to lows provides the high negative covariance needed for the strong negative spatial autocorrelation $Z(I)$, but it is the multiplicative effect of high values near lows that has the negative effect on $Z(G)$. ▲

Table 2 gives some idea of the values of $Z(G)$ and $Z(I)$ under various circumstances. The differences result from each statistic's structure. As shown in the examples above, if high values within d of other high values dominate the pattern, then the summation of the products of neighboring values is high, with resulting high positive $Z(G)$ values. If low values within d of low values dominate, then the sum of the product of the xs is low resulting in strong negative $Z(G)$ values. In the Moran's case, both when high values are within d of other high values and low values are within d of other low values, positive covariance is high, with resulting high $Z(I)$ values.

GENERAL DISCUSSION

Any test for spatial association should use both types of statistics. Sums of products and covariances are two different aspects of pattern. Both reflect the dependence structure in spatial patterns. The $I(d)$ statistic has its peculiar weakness in not being able to discriminate between patterns that have high values dominant within d or low values dominant. Both statistics have difficulty discerning a random pattern from one in which there is little deviation from the mean.

If a study requires that $I(d)$ or $G(d)$ values be traced over time, there are advantages to using both statistics to explore the processes thought to be responsible for changes in association among regions. If data values increase or decrease at the same rate, that is, if they increase or decrease in proportion to their already existing size, Moran's I changes while $G(d)$ remains the same. On the other hand, if all x values increase or decrease by the same amount, $G(d)$ changes but $I(d)$ remains the same.

It must be remembered that $G(d)$ is based on a variable that is positive and has a natural origin. Thus, for example, it is inappropriate to use $G(d)$ to study residuals from regression. Also, for both $I(d)$ and $G(d)$ one must recognize that transformations of the variable X result in different values for the test statistic. As has

been mentioned above, conditions may arise when d is so small or large that tests based on the normal approximation are inappropriate.

EMPIRICAL EXAMPLES

The following examples of the use of G statistics were selected based on size and type of spatial units, size of the x values, and subject matter. The first is a problem concerning the rate of sudden infant death syndrome by county in North Carolina, and the second is a study of the mean price of housing units sold by zip-code district in the San Diego metropolitan region. In both cases the data are explained, hypotheses made clear, and $G(d)$ and $I(d)$ values calculated for comparable circumstances.

1. Sudden Infant Death Syndrome (SIDS) by County in North Carolina

SIDS is the sudden death of an infant one year old or less that is unexpected and inexplicable after a postmortem examination (Cressie and Chan 1989). The data presented by Cressie and Chan were collected from a variety of sources cited in the article. Among other data, the authors give the number of SIDs by county for the period 1979–1984, the number of births for the same period, and the coordinates of the counties. We use as our data the number of SIDs as a proportion of births multiplied by 1000 (see Figure 1). Since no viral or other causes have been given for SIDS, one should not expect any spatial association in the data. To some extent, high or low rates may be dependent on the health care infants receive. The rates may correlate with variables such as income or the availability of physicians' services. In this study we shall not expect any spatial association.

Table 3 gives the values for the standard normal variate of I and G for various distances.

Results using the G statistic verify the hypothesis that there is no discernible association among counties with regard to SIDS rates. The values of $Z(G)$ are less than one. In addition, there seems to be no smooth pattern of Z values as d increases. The $Z(I)$ results are somewhat contradictory, however. Although none are statistically significant at the .05 level, $Z(I)$ values from 30 to 50 miles, about the distance from the center of each county to the center of its contiguous neighboring counties, are well over one. This represents a tendency toward positive spatial autocorrelation at those distances. Taking the two results together, one should be cautious before concluding that a spatial association exists for SIDS

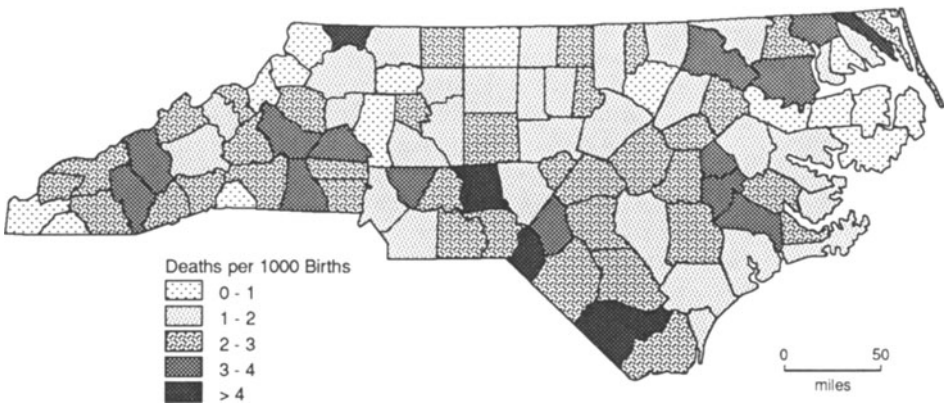

FIG. 1. Sudden Infant Death Rates for Counties of North Carolina, 1979–1984

TABLE 3

Spatial Association among Counties: SIDS Rates by County in North Carolina, 1979–1984

d in miles	Z(G)	Z(I)
10	0.82	−0.55
20	0.29	0.99
30	−0.12	1.68
33*	0.40	1.84
40	−0.04	1.32
50	0.60	1.20
60	−0.36	0.48
70	−0.28	−0.45
80	−0.19	−0.13
90	0.11	−0.19
100	0.30	0.18

*At all distances of this length or longer each county is linked to at least one other county.

TABLE 4

Highest Positive and Negative Standard Normal Variates by County for $ZG_i^*(d)$ and $ZG_i(d)$: SIDS Rates in North Carolina, 1979–1984 (d = 33 miles)

County	$ZG_i^*(d)$	County	$ZG_i(d)$
	Highest Positive		
Richmond	+3.34	Richmond	+3.62
Robeson	+3.12	Robeson	+3.09
Scotland	+2.78	Hoke	+1.78
Hoke	+2.12	Northampton	+1.44
Cleveland	+1.78	Moore	+1.39
	Highest Negative		
Washington	−2.63	Washington	−2.18
Dare	−1.84	Davie	−1.92
Davie	−1.76	Dare	−1.70
Cherokee	−1.55	Bertie	−1.64
Tyrrell	−1.53	Stokes	−1.58

among counties in North Carolina. Perhaps more light can be shed on the issue by using the $G_i(d)$ and $G_i^*(d)$ statistics.

Table 4 and Figure 2 give the results of an analysis based on the $G_i(d)$ and $G_i^*(d)$ statistics for a d of thirty-three miles. This represents the distance to the furthest first-nearest neighbor county of any county.

The $G_i^*(d)$ statistic identifies five of the one hundred counties of North Carolina as significantly positively or negatively associated with their neighboring counties (at the .05 level). Four of these, clustered in the central south portion of the state, display values greater than +1.96, while one county, Washington near Albemarle Sound, has a Z value of less than −1.96 (see Figure 2). Taking into account values greater than +1.15 (the 87.5 percentile), it is clear that several small clusters in addition to the main cluster are widely dispersed in the southern part of the state. The main cluster of values less than −1.15 (the 12.5 percentile) is in the eastern part of the state. It is interesting to note that many of the counties in this cluster are in the sparsely populated swamp lands surrounding the Albemarle and Pamlico Sounds. If *overall* error is fixed at 0.05 and a Bonferroni correction is applied, the cutoff value for each county is raised to about 3.50. However, such a figure is unduly conservative given the small numbers of neighbors.

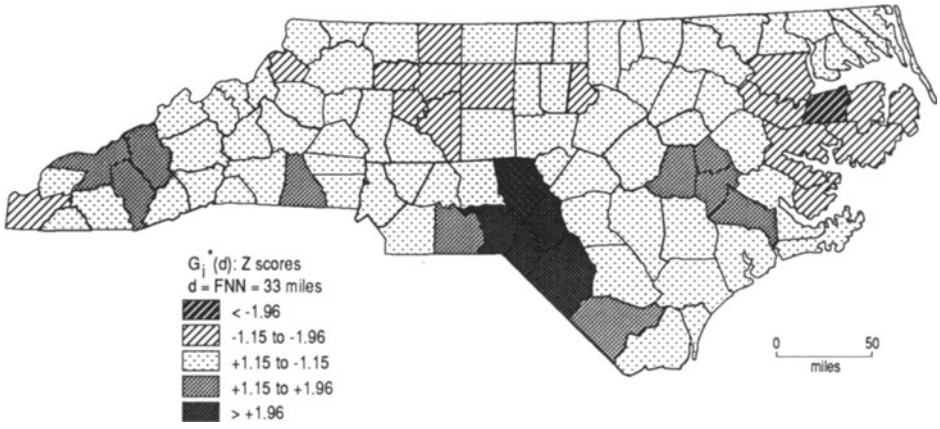

FIG. 2. $Z[G_i^*(d = \text{furthest nearest neighbor} = 33 \text{ miles})]$ for SIDS Rates of Counties of North Carolina, 1979–84.

In this case it becomes clear that an overall measure of association such as $G(d)$ or $I(d)$ can be misleading because it prompts one to dismiss the possibility of significant spatial clustering. The $G_i(d)$ statistics, however, are able to identify the tendency for positive spatial clustering and the location of pockets of high and low spatial association. It remains for the social scientist or epidemiologist to explain the subtle patterns shown in Figure 2.

2. Dwelling Unit Prices in San Diego County by Zip-Code Area, September 1989

Data published in the *Los Angeles Times* on October 29, 1989, give the adjusted average price by zip code for all new and old dwelling units sold by builders, real estate agents, and homeowners during the month of September 1989 in San Diego County (see Appendix). The data are supplied by TRW Real Estate Information Services. One outlier was identified: Rancho Santa Fe, a wealthy suburb of the city of San Diego, had prices of sold dwelling units that were nearly three times higher than the next highest district (La Jolla). Since neither statistic is robust enough to be only marginally affected by such an observation, Rancho Santa Fe was not considered in the analysis.

Although the city of San Diego has a large and active downtown, San Diego County is not a monocentric region. One would not expect housing prices to trend upward from the city center to the suburbs in a uniform way. One would expect, however, that since the data are for reasonably small sections of the metropolitan area, that there would be distinct spatial autocorrelation tendencies (see Figure 3). High positive I values are expected. $G(d)$ values are dependent on the tendencies for high values or low values to group. If the low cost areas dominate, the $G(d)$ value is negative. In this case, $G(d)$ is a refinement of the knowledge gained from I.

Table 5 shows that there are strong positive values for $Z(I)$ for distances of four miles and greater. $Z(G)$ also shows highly significant values at four miles and beyond, but here the association is negative, that is, low values near low values are much more influential than are the high values near high values. Moran's I clearly indicates that there is significant spatial autocorrelation, but, without knowledge of $G(d)$, one might conclude that at this scale of analysis, in general, high income districts are significantly associated with one another.

By looking at the results of the $G_i(d)$ statistics analysis for d equal to five, the individual district pattern is unmistakable. The $Z(G_i^*(5))$ values shown in Table 6

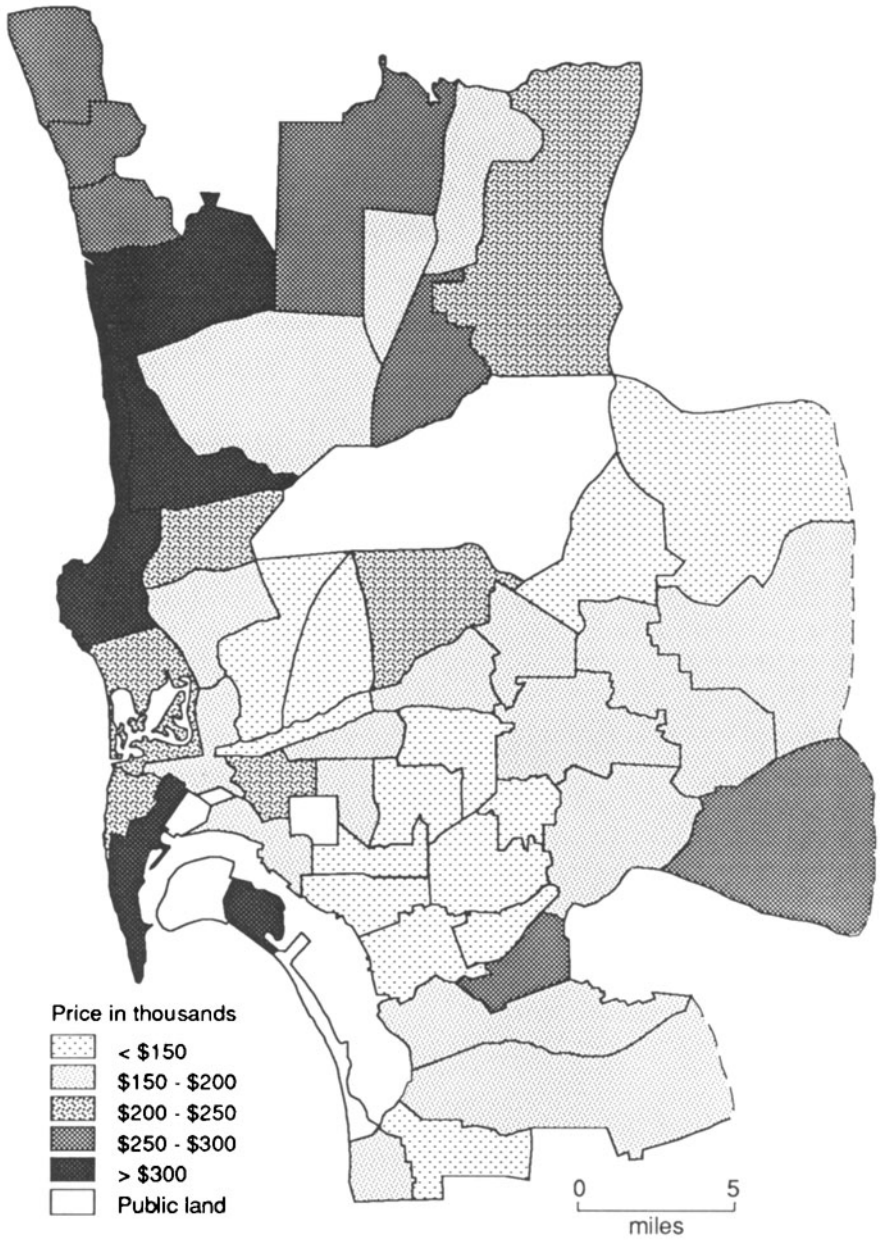

FIG. 3. San Diego House Prices, September 1989.

and Figure 4 provide evidence that two coastal districts are positively associated at the .05 level of significance while eight central and south central districts are negatively associated at the .05 level. There is a strong tendency for the negative values to be higher. It is for this reason that the Z(G) values given above are so decidedly negative. The districts with high values along the coast have fewer near neighbors with similar values than do the central city lower value districts. The

TABLE 5

Spatial Association among Zip Code Districts: Dwelling Unit Prices in San Diego County, September 1989

d in miles	$Z(G)$	$Z(I)$
2	−0.67	0.33
4	−2.36	2.36
5#	−2.32	4.13
6	−2.47	4.16
8	−2.80	3.51
10	−2.66	3.57
12	−2.20	3.53
14	−2.34	3.92
16	−2.54	4.27
18	−2.30	3.57
20	−2.25	2.92

#At all distances of this length or longer each district is connected to at least one other district.

TABLE 6

Highest Positive and Negative Standard Normal Variates by Zip Code District for $G_i^*(d)$ and $G_i(d)$: Dwelling Unit Prices in San Diego County, September 1989 (d = 5 miles)

Neighborhood	$ZG_i^*(d)$	Neighborhood	$ZG_i(d)$
	Highest Positive		
Cardiff	+2.27	Cardiff	+2.08
Solana Beach	+2.02	Solana Beach	+1.81
Point Loma	+1.93	Mira Mesa	+1.56
La Jolla	+1.89	Ocean Beach	+1.37
Del Mar	+1.55	R. Penasquitos	+1.33
	Highest Negative		
East San Diego	−3.22	East San Diego	−2.99
East San Diego	−2.74	East San Diego	−2.54
East San Diego	−2.64	North Park	−2.48
North Park	−2.56	East San Diego	−2.48
Mission Valley	−2.38	College	−2.19

cluster of districts with negative $Z(G_i^*)$ values dominates the pattern. The adjusted Bonferroni cutoff is about 3.27, but again is overly conservative.

CONCLUSIONS

The G statistics provide researchers with a straightforward way to assess the degree of spatial association at various levels of spatial refinement in an entire sample or in relation to a single observation. When used in conjunction with Moran's I or some other measure of spatial autocorrelation, they enable us to deepen our understanding of spatial series. One of the G statistics' useful features, that of neutralizing the spatial distribution of the data points, allows for the development of hypotheses where the pattern of data points will not bias results.

When G statistics are contrasted with Moran's I, it becomes clear that the two statistics measure different things. Fortunately, both statistics are evaluated using normal theory so that a set of standard normal variates taken from tests using each type of statistic are easily compared and evaluated.

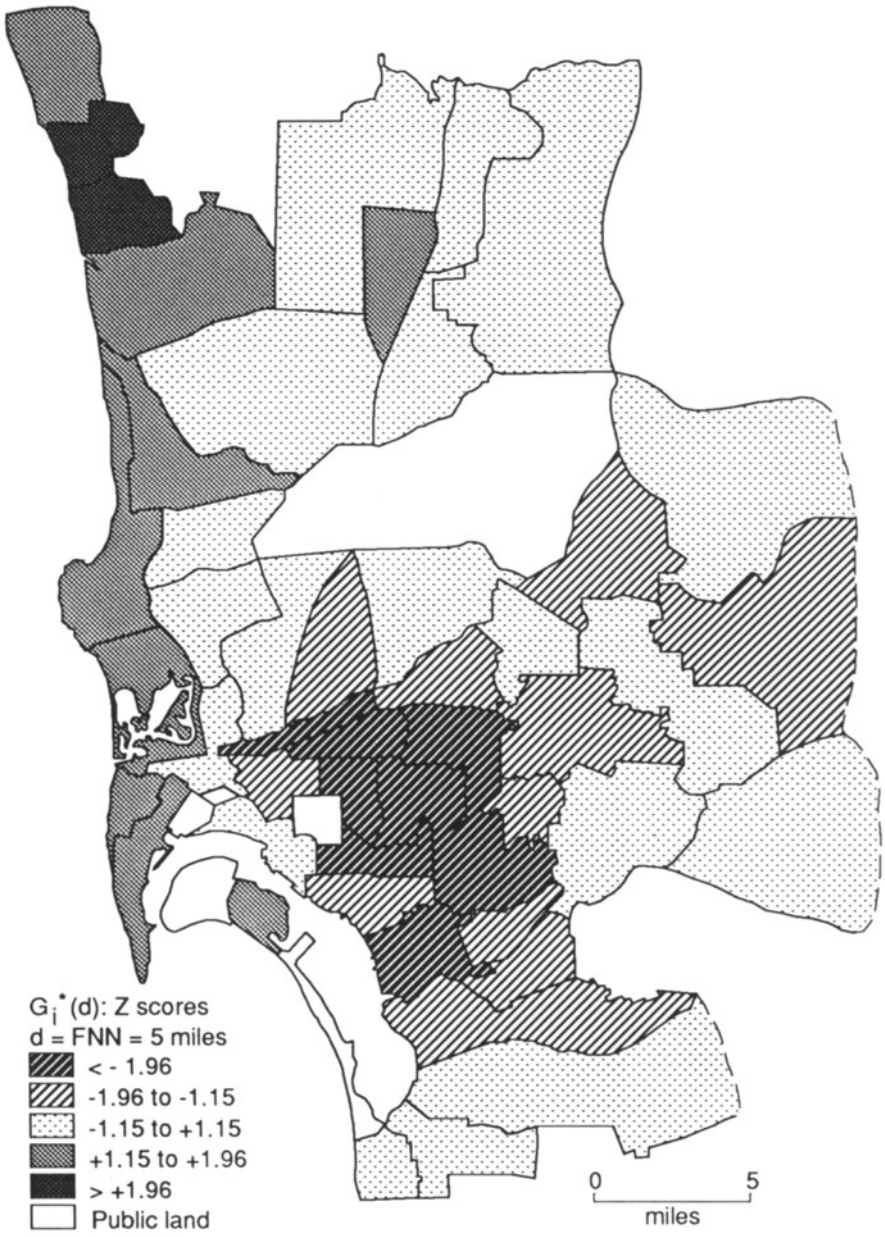

FIG. 4. $Z[G_i^*(d = \text{furthest nearest neighbor} = 5 \text{ miles})]$ for House Prices of San Diego County Zip Code Districts, September 1989.

APPENDIX

San Diego County Average House Prices for September 1989 by Zip-Code District

	Zip Code	Principal Neighborhood Name	Coordinates (miles) x	y	Price (in thousands)
01	92024	Encinitas	1	39	264
02	92007	Cardiff	2	36	260
03	92075	Solana Beach	3	34	261
04	92014	Del Mar	5	32	309
05	92127	Lake Hodges	10	34	265
06	92129	R. Penasquitos	12	32	194
07	92128	R. Bernardo	15	35	191
08	92064	Poway	17	32	236
09	92131	Scripps Ranch	13	29	270
10	92126	Mira Mesa	8	28	162
11	92037	La Jolla	3	22	398
12	92122	University City	6	23	201
13	92117	Clairemont	6	20	192
14	92109	Beaches	4	18	249
15	92110	Bay Park	6	15	152
16	92111	Kearny Mesa	8	19	138
17	92123	Mission Village	10	19	131
18	92124	Tierrasanta	13	20	221
19	92120	Del Cerro	14	18	187
20	92119	San Carlos	17	19	182
21	92071	Santee	20	22	124
22	92040	Lakeside	23	24	147
23	92021	El Cajon	24	19	151
24	92020	El Cajon	22	17	150
25	92041	La Mesa	18	16	169
26	92115	College	14	16	138
27	92116	Kensington	11	16	192
28	92108	Mission Valley	9	16	89
29	92103	Hillcrest	8	14	225
30	92104	North Park	11	14	152
31	92105	East San Diego	13	14	111
32	92045	Lemon Grove	17	13	137
33	92077	Spring Valley	20	13	150
34	92035	Jamul	24	12	291
35	92002	Bonita	17	8	297
36	92139	Paradise Hills	16	9	117
37	92050	National City	13	8	99
38	92113	Logan Heights	11	10	84
39	92102	East San Diego	12	12	88
40	92101	Downtown	8	12	175
41	92107	Ocean Beach	3	14	229
42	92106	Point Loma	3	12	338
43	92118	Coronado	7	10	374
44	92010	Chula Vista	15	6	165
45	92011	Chula Vista	17	4	184
46	92032	Imperial Beach	11	1	164
47	92154	Otay Mesa	15	2	126
48	92114	East San Diego	15	11	126

Source of Data: *Los Angeles Times*, October 29, 1989, page K15.

LITERATURE CITED

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer.

Anselin, L., and D. A. Griffith (1988). "Do Spatial Effects Really Matter in Regression Analysis?" *Papers of the Regional Science Association* 65, 11–34.

Arbia, G. (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Dordrecht: Kluwer.

Cliff, A. D., and J. K. Ord (1973). *Spatial Autocorrelation*. London: Pion.

Cressie, N., and N. H. Chan (1989). "Spatial Modeling of Regional Variables." *Journal of the American Statistical Association* 84, 393–401.

Cressie, N., and T. R. C. Read (1989). "Spatial Data Analysis of Regional Counts." *Biometrical Journal* 31, 699–719.

Davis, J. C. (1986). *Statistics and Data Analysis in Geology*. New York: Wiley.

Foster, S. A., and W. L. Gorr (1986). "An Adaptive Filter for Estimating Spatially Varying Parameters: Application to Modeling Police Hours in Response to Calls for Service." *Management Science* 32, 878–89.

Getis, A. (1984). "Interaction Modeling Using Second-order Analysis." *Environment and Planning A* 16, 173–83.

——— (1985). "A Second-order Approach to Spatial Autocorrelation." *Ontario Geography* 25, 67–73.

——— (1991). "Spatial Interaction and Spatial Autocorrelation: A Cross-Product Approach." *Environment and Planning A* 23, 1269–77.

Getis, A., and J. Franklin (1987). "Second-order Neighborhood Analysis of Mapped Point Patterns." *Ecology* 68, 473–77.

Hoeffding, W. (1951). "A Combinatorial Central Limit Theorem." *Annals of Mathematical Statistics* 22, 558–66.

Hubert, L. J. (1977). "Generalized Proximity Function Comparisons." *British Journal of Mathematical and Statistical Psychology* 31, 179–82.

——— (1979). "Matching Models in the Analysis of Cross-Classifications." *Psychometrika* 44, 21–41.

Hubert, L. J., R. G. Golledge, and C. M. Costanzo (1981). "Generalized Procedures for Evaluating Spatial Autocorrelation." *Geographical Analysis* 13, 224–33.

Mantel, N. (1967). "The Detection of Disease Clustering and a Generalized Regression Approach." *Cancer Research* 27, 209–20.

Pitman, E. J. G. (1937). "The 'Closest' Estimates of Statistical Parameters." *Biometrika* 58, 299–312.

Ripley, B. D. (1977). "Modelling Spatial Patterns." *Journal of the Royal Statistical Society* B39, 172–212.