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 Gi and G* 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.

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 $i$th point.

The statistic is

$$G_i(d) = \frac{\sum_{j=1, j \neq i}^{n} w_{ij}(d)x_j}{\sum_{j=1}^{n} x_j}$$

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 $i$th 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

$$E(X_j) = \frac{\sum_{r \neq i} x_r}{(n - 1)} .$$
Thus
\[
E(G_i) = \sum_{j \neq i} w_{ij}(d) \frac{E(X_j)}{\sum_{j \neq i} X_j} = W_i/(n - 1),
\]

(2)

where
\[W_i = \sum_j w_{ij}(d).\]

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

Since
\[E(X_j^2) = \sum_{r \neq i} x_j^2/(n - 1)\]

and
\[E(X_j X_k) = \sum_{r \neq i} x_j x_k/(n - 1)(n - 2)\]

\[= \left\{ \left( \sum_{r \neq i} x_j \right)^2 - \sum_{r \neq i} x_k^2 \right\}/(n - 1)(n - 2).\]

Recalling that the weights are binary
\[\sum_{j \neq k} w_{ij} w_{ik} = W_i^2 - W_i\]

and so
\[E(G_i^2) = \frac{1}{(\sum_j x_j)^2} \left( W_i \sum_j x_j^2/(n - 1) + W_i(W_i - 1)/(n - 1)(n - 2) \right) \left[ \left( \sum_j x_j \right)^2 - \sum_j x_j^2 \right].\]

Thus
\[\text{Var}(G_i) = E(G_i^2) - E^2(G_i)\]

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

If we set
\[\frac{\sum_j x_j}{(n - 1)} = Y_{i1}\]

and
\[\frac{\sum_j x_j^2}{(n - 1)} - Y_{i1}^2 = Y_{i2},\]

then
\[\text{Var}(G_i) = \frac{W_i(n - 1 - W_i)}{(n - 1)^2(n - 2)} \left( \frac{Y_{i2}}{Y_{i1}} \right).
\]

As expected, \(\text{Var}(G_i) = 0\) 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

<table>
<thead>
<tr>
<th>Statistic</th>
<th>( j ) not equal to ( i )</th>
<th>( j ) may equal ( i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression</td>
<td>( \frac{\sum_j w_{ij}(d) x_j}{\sum_j x_j} )</td>
<td>( \frac{\sum_j w_{ij}(d) x_j}{\sum_j x_j} )</td>
</tr>
<tr>
<td>Definitions</td>
<td>( \frac{\sum_j x_j}{n(n-1)} )</td>
<td>( \frac{\sum_j x_j}{n} )</td>
</tr>
<tr>
<td>Expectation</td>
<td>( \frac{W_i}{(n-1)} )</td>
<td>( \frac{W_i}{n} )</td>
</tr>
<tr>
<td>Variance</td>
<td>( \frac{W_i(n-1-W_i) Y_{ii}^2}{(n-1)^2(n-2) Y_{ii}^2} )</td>
<td>( \frac{W_i(n-W_i)}{n^2(n-1) Y_{ii}^2} )</td>
</tr>
</tbody>
</table>

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 \( x_i \). 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 \( x \)'s 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

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

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 \( x_j \)s. 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 \( \text{corr}(G_i, G_j) = \text{proportion of neighbors that } i \text{ and } j \text{ have in common} \).

**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:

\[
\begin{array}{cccccc}
0 & 1 & 0 & & & \\
0 & 2 & 0 & 2 & 0 & \\
1 & 0 & i & 0 & 1 & \\
0 & 2 & 0 & 2 & 0 & \\
0 & 1 & 0 & & & \\
\end{array}
\]

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**

\[
\begin{array}{cccccccc}
m & m & m & m & m & m & m & m & m \\
m & A & A & A & m & m & B & B & B & m \\
m & A & A & A & m & m & B & B & B & m \\
m & A & A & A & m & m & B & B & B & m \\
m & m & m & m & m & m & m & m & m & m \\
\end{array}
\]

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) | \( G_i \) | \( Z(G_i) \) | \( G^*_i \) | \( Z(G^*_i) \) |
|---|---|---|---|
| \( A \), surrounded by As | \( 8 + 8c \) | 5.30* | \( 9 + 9c \) | 5.47 |
| | \( 49 - c \) | | \( 50 \) | |
| | \( 8 + 3c \) | | \( 50 \) | |
| \( A \), adjacent to \( ms \) | \( 2.06* \) | | \( 9 + 4c \) | 2.43 |
| | \( 49 - c \) | | 50 | |
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

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

$$W = \sum_{i=1}^{n} \sum_{j \neq i} w_{ij}(d),$$

so that

$$E[G(d)] = W/[n(n - 1)].$$

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

$$E(G^2) = \frac{1}{(m_1^2 - m_2)^2 n^{(4)}} [B_0m_2^2 + B_1m_4 + B_2m_1m_2 + B_3m_1m_3 + B_4m_1^4]$$

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

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

The coefficients, $B$, are

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

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

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

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

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

where $S_1 = \frac{1}{2} \sum_{i} \sum_{j \neq i} (w_{ij} + w_{ji})^2, j \neq i,

and $S_2 = \sum_{i} (w_{i} + w_{.i})^2; w_{i} = \sum_{j \neq i} w_{ij}, j \neq i;

thus

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

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 $x_j$s 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 $x_j$s 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

\[
G(d) = K_1 \sum_i w_{ij} x_i x_j
\]

and

\[
I(d) = K_2 \sum_i w_{ij} (x_i - \bar{x})(x_j - \bar{x})
= (K_2/K_1) G(d) - K_2 \bar{x} \sum_i (w_i + w_j) x_i + K_2 \bar{x}^2 W
\]

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 \( \sum w_i x_i \) and \( \sum w_j x_j \) differ from \( \bar{x} \), 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 \( A = m(1 + c), B = m(1 - c), \quad 0 \leq c \leq 1. \)

In addition, put

\[
a = A - m; \\
b = 2m - A = m - a; \\
c = B - m = a; \\
m \geq a; \\
J \text{ not equal to } i.
\]

For the rook’s case, \( W = \sum w_{ij} = 170. \)

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

for all choices of \( a, m. \)

\[
\text{Var}(I) = 0.010897 \\
Z(I) = 7.7088 \text{ whenever } A > B.
\]

\[
G = \frac{\sum w_{ij} x_i x_j}{\sum x_i x_j} = \frac{24A^2 + 24B^2 + 24Am + 24Bm + 74m^2}{2500m^2 - 9A^2 - 9B^2 - 32m^2}
= \frac{170 + 48c^2}{2450 - 18c^2}
\]

When \( c = 0, A = B = m, \) and \( G \) is a minimum.

\[
G_{\text{min}} = \frac{170}{2450} = 0.0694 \text{ and} \\
\text{Var}(G_{\text{min}}) = 0.0000 \text{ from equation (7).}
\]
When \( c = 1, A = 2m, B = 0 \), and \( G \) is a maximum.

\[
G_{\text{max}} = \frac{218}{2432} = 0.0896.
\]

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

\[
Z(G_{\text{max}}) = 5.87 \text{ for any } m.
\]

\( G \) 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**

\[
\begin{array}{cccccccccc}
m & m & m & m & m & m & m & m & m & m \\
m & m & m & m & m & m & m & m & m & m \\
m & m & A & m & m & m & B & m & m & m \\
m & m & m & m & m & m & m & m & m & m \\
m & m & m & m & m & m & m & m & m & m
\end{array}
\]

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

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

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

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

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

\[
Z(G_{\text{max}}) = 0.0739.
\]

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**

\[
\begin{array}{cccccccccc}
m & m & m & m & m & m & m & m & m & m \\
m & A & B & A & m & m & B & A & B & m \\
m & B & A & B & m & m & A & B & A & m \\
m & A & B & A & m & m & B & A & B & m \\
m & m & m & m & m & m & m & m & m & m
\end{array}
\]

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

\[
I = -0.7843
\]

\[
\text{Var}(I) = 0.010897
\]

\[
Z(I) = -7.3177
\]

When \( A = 2m \) and \( B = 0 \),

\[
G = 0.0502
\]

\[
\text{Var}(G) = 0.000001189
\]

\[
Z(G) = -5.5760
\]
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 \(x\)s 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.
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.
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. Dwellings 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...
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

<table>
<thead>
<tr>
<th>$d$ in miles</th>
<th>$Z(G)$</th>
<th>$Z(I)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-0.67</td>
<td>0.33</td>
</tr>
<tr>
<td>4</td>
<td>-2.36</td>
<td>2.36</td>
</tr>
<tr>
<td>5*</td>
<td>-2.32</td>
<td>4.13</td>
</tr>
<tr>
<td>6</td>
<td>-2.47</td>
<td>4.16</td>
</tr>
<tr>
<td>8</td>
<td>-2.80</td>
<td>3.51</td>
</tr>
<tr>
<td>10</td>
<td>-2.66</td>
<td>3.57</td>
</tr>
<tr>
<td>12</td>
<td>-2.20</td>
<td>3.53</td>
</tr>
<tr>
<td>14</td>
<td>-2.34</td>
<td>3.92</td>
</tr>
<tr>
<td>16</td>
<td>-2.54</td>
<td>4.27</td>
</tr>
<tr>
<td>18</td>
<td>-2.30</td>
<td>3.57</td>
</tr>
<tr>
<td>20</td>
<td>-2.25</td>
<td>2.92</td>
</tr>
</tbody>
</table>

*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 $G^*_h(d)$ and $G_t(d)$: Dwelling Unit Prices in San Diego County, September 1989 ($d = 5$ miles)

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

A cluster of districts with negative $Z(G^*_h)$ 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. $Z[G_*(d = \text{furthest nearest neighbor} = 5 \text{ miles})]$ 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

<table>
<thead>
<tr>
<th>Zip Code</th>
<th>Principal Neighborhood Name</th>
<th>Coordinates (miles)</th>
<th>Price (in thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>92024 Encinitas</td>
<td>1 39</td>
<td>264</td>
</tr>
<tr>
<td>02</td>
<td>92007 Cardiff</td>
<td>2 36</td>
<td>260</td>
</tr>
<tr>
<td>03</td>
<td>92075 Solana Beach</td>
<td>3 34</td>
<td>261</td>
</tr>
<tr>
<td>04</td>
<td>92014 Del Mar</td>
<td>5 32</td>
<td>309</td>
</tr>
<tr>
<td>05</td>
<td>92127 Lake Hodges</td>
<td>10 34</td>
<td>265</td>
</tr>
<tr>
<td>06</td>
<td>92129 R. Penasquitos</td>
<td>12 32</td>
<td>194</td>
</tr>
<tr>
<td>07</td>
<td>92128 R. Bernardo</td>
<td>15 35</td>
<td>191</td>
</tr>
<tr>
<td>08</td>
<td>92064 Poway</td>
<td>17 32</td>
<td>236</td>
</tr>
<tr>
<td>09</td>
<td>92131 Scripps Ranch</td>
<td>13 29</td>
<td>270</td>
</tr>
<tr>
<td>10</td>
<td>92126 Mira Mesa</td>
<td>8 28</td>
<td>162</td>
</tr>
<tr>
<td>11</td>
<td>92037 La Jolla</td>
<td>3 22</td>
<td>396</td>
</tr>
<tr>
<td>12</td>
<td>92122 University City</td>
<td>6 23</td>
<td>201</td>
</tr>
<tr>
<td>13</td>
<td>92117 Clairemont</td>
<td>6 20</td>
<td>192</td>
</tr>
<tr>
<td>14</td>
<td>92109 Beaches</td>
<td>4 18</td>
<td>249</td>
</tr>
<tr>
<td>15</td>
<td>92110 Bay Park</td>
<td>6 15</td>
<td>152</td>
</tr>
<tr>
<td>16</td>
<td>92111 Kearny Mesa</td>
<td>8 19</td>
<td>135</td>
</tr>
<tr>
<td>17</td>
<td>92123 Mission Village</td>
<td>10 19</td>
<td>131</td>
</tr>
<tr>
<td>18</td>
<td>92124 Tierrasanta</td>
<td>13 20</td>
<td>221</td>
</tr>
<tr>
<td>19</td>
<td>92120 Del Cerro</td>
<td>14 18</td>
<td>187</td>
</tr>
<tr>
<td>20</td>
<td>92119 San Carlos</td>
<td>17 19</td>
<td>182</td>
</tr>
<tr>
<td>21</td>
<td>92071 Santee</td>
<td>20 22</td>
<td>124</td>
</tr>
<tr>
<td>22</td>
<td>92040 Lakeside</td>
<td>23 24</td>
<td>147</td>
</tr>
<tr>
<td>23</td>
<td>92021 El Cajon</td>
<td>24 19</td>
<td>151</td>
</tr>
<tr>
<td>24</td>
<td>92020 El Cajon</td>
<td>22 17</td>
<td>150</td>
</tr>
<tr>
<td>25</td>
<td>92041 La Mesa</td>
<td>18 16</td>
<td>169</td>
</tr>
<tr>
<td>26</td>
<td>92115 College</td>
<td>14 16</td>
<td>138</td>
</tr>
<tr>
<td>27</td>
<td>92116 Kensington</td>
<td>11 16</td>
<td>192</td>
</tr>
<tr>
<td>28</td>
<td>92108 Mission Valley</td>
<td>9 16</td>
<td>89</td>
</tr>
<tr>
<td>29</td>
<td>92103 Hillcrest</td>
<td>8 14</td>
<td>225</td>
</tr>
<tr>
<td>30</td>
<td>92104 North Park</td>
<td>11 14</td>
<td>152</td>
</tr>
<tr>
<td>31</td>
<td>92105 East San Diego</td>
<td>13 14</td>
<td>111</td>
</tr>
<tr>
<td>32</td>
<td>92045 Lemon Grove</td>
<td>17 13</td>
<td>137</td>
</tr>
<tr>
<td>33</td>
<td>92077 Spring Valley</td>
<td>20 13</td>
<td>150</td>
</tr>
<tr>
<td>34</td>
<td>92035 Jamul</td>
<td>24 12</td>
<td>291</td>
</tr>
<tr>
<td>35</td>
<td>92002 Bonita</td>
<td>17 8</td>
<td>297</td>
</tr>
<tr>
<td>36</td>
<td>92139 Paradise Hills</td>
<td>16 9</td>
<td>117</td>
</tr>
<tr>
<td>37</td>
<td>92050 National City</td>
<td>13 8</td>
<td>99</td>
</tr>
<tr>
<td>38</td>
<td>92113 Logan Heights</td>
<td>11 10</td>
<td>84</td>
</tr>
<tr>
<td>39</td>
<td>92102 East San Diego</td>
<td>12 12</td>
<td>88</td>
</tr>
<tr>
<td>40</td>
<td>92101 Downtown</td>
<td>8 12</td>
<td>175</td>
</tr>
<tr>
<td>41</td>
<td>92107 Ocean Beach</td>
<td>3 14</td>
<td>229</td>
</tr>
<tr>
<td>42</td>
<td>92106 Point Loma</td>
<td>3 12</td>
<td>338</td>
</tr>
<tr>
<td>43</td>
<td>92118 Coronado</td>
<td>7 10</td>
<td>374</td>
</tr>
<tr>
<td>44</td>
<td>92010 Chula Vista</td>
<td>15 6</td>
<td>165</td>
</tr>
<tr>
<td>45</td>
<td>92011 Chula Vista</td>
<td>17 4</td>
<td>184</td>
</tr>
<tr>
<td>46</td>
<td>92032 Imperial Beach</td>
<td>11 1</td>
<td>164</td>
</tr>
<tr>
<td>47</td>
<td>92154 Otay Mesa</td>
<td>15 2</td>
<td>126</td>
</tr>
<tr>
<td>48</td>
<td>92114 East San Diego</td>
<td>15 11</td>
<td>126</td>
</tr>
</tbody>
</table>

LITERATURE CITED