advances in geographic information systems, Bremen, pp 250–259
Litman T (2006) Lessons from Katrina and Rita: what major disasters can teach transportation planners. J Trans Eng 132(1):11–18
Metropolitan Washington Council of Governments (2005) 2004 performance of regional high-occupancy vehicle facilities on freeways in the Washington region. Analysis of person and vehicle volumes
Theodoulou G, Wolshon B (2004) Alternative methods to increase the effectiveness of freeway contraflow evacuation. J Trans Res Board Transp Res Rec 1865:48–56
Tuydes H, Ziliaskopoulos A (2004) Network re-design to optimize evacuation contraflow. Technical report 04-4715, Presented at 83rd Annual Meeting of the Transportation Research Board
Wolshon B (2001) One-way-out: contraflow freeway operation for hurricane evacuation. Natl Hazard Rev 2(3):105–112
Wolshon B, Urbina E, Levitan M (2002) National review of hurricane evacuation plans and policies. Technical report, Hurricane Center, Louisiana State University, Baton Rouge

Contraint Relations

► Constraint Databases and Data Interpolation

Convergence of GIS and CAD

► Computer Environments for GIS and CAD

Converging

► Movement Patterns in Spatio-Temporal Data

Co-occurrence

► Co-location Pattern Discovery
► Co-location Patterns, Algorithms
► Statistically Significant Co-location Pattern Mining

Coordinate Systems

► Reference Frames

Coregistration

► Registration

Co-registration

► Registration

Correlated

► Patterns, Complex

Correlated Frailty Models

► Spatial Survival Analysis

Correlated Walk

► CrimeStat: A Spatial Statistical Program for the Analysis of Crime Incidents

Correlation and Spatial Autocorrelation

Sang-Il Lee
Department of Geography Education, Seoul National University, Seoul, South Korea

Definition

Spatial autocorrelation or spatial dependence can be defined as a particular relationship between the spatial proximity among observational units and the numeric similarity among their values; positive spatial autocorrelation refers to situations in which the nearer the observational units, the more similar their values (and vice versa for its negative counterpart). The presence of spatial autocorrelation or dependence means that a certain amount of information is shared and duplicated among neighboring locations, and thus, an entire data set possesses a certain amount of redundant information. This feature violates the assumption of independent observations upon which many standard statistical treatments are predicated. This entry revolves around what happens to the nature and statistical significance of correlation coefficients (e.g., Pearson's r) when spatial autocorrelation is present in both or either of the two variables under investigation.

Historical Background

A lack of independence results in reduced degrees of freedom or effective sample size; the greater the level of spatial autocorrelation, the smaller the number of degrees of freedom or effective sample size. This means that any type of statistical test based on an original sample size could be flawed in the presence of spatial autocorrelation, thus heightening the probability of committing a Type I error. Suppose that $n!$ different map patterns are generated from n observations. Because the $n!$ different map patterns are identical in terms of sample mean and variance, any statistical inferences based on these values are identical. However, all of the map patterns possess different degrees of freedom or effective sample size, and thus $n!$ different statistical estimations should be obtained.

This type of problem occurs in situations dealing with the correlation between two variables, which has long been known (Bivand 1980; Griffith 1980; Haining 1980; Richardson and Hémon 1981). The presence of spatial autocorrelation in both or either of two variables under investigation (i.e., bivariate spatial dependence) means that when the nature of a bivariate association at a location is known, one can guess the nature of bivariate associations at nearby locations. For example, if a location has a pair of higher-than-average values for two variables, there is a more-than-random chance to observe similar pairs in nearby locations. This feature again violates the assumption of independent observations and reduces the number of degrees of freedom or effective sample size. In this context, standard inferential tests tend to underestimate the true sampling variance of the Pearson's correlation coefficient when positive spatial autocorrelation is present in two variables under investigation, resulting in a heightened chance of committing a Type I error. One can generate $n!$ different pairs of spatial patterns from the original variables; all of the pairs are identical in terms of Pearson's correlation coefficient, but they are different in terms of the number of degrees of freedom or effective sample size (Clifford and Richardson 1985; Clifford et al. 1989; Haining 1991; Dutilleul 1993). These notions can extend to situations dealing with a pair of regression residuals (Tiefelsdorf 2001).

Two different approaches exist, addressing the problem of spatial autocorrelation in bivariate correlation. One is to seek to remedy the problem by providing modified hypothesis testing procedures taking the degree of spatial autocorrelation into account (for a comprehensive review and discussion, see Griffith and Paelinck 2011). The other is to develop bivariate spatial autocorrelation statistics to capture the degree of spatial *co-patterning* between two map patterns and, further, to propose some techniques for exploratory spatial data analysis (ESDA) that allow the detecting of bivariate spatial clusters (among others, Lee 2001; Anselin et al. 2002; Lee 2012).

Scientific Fundamentals

For this section, I seek to conceptualize and illustrate the concept of *bivariate spatial dependence* with which the problems of correlation in the presence of spatial autocorrelation are better captured and tackled. For simplicity, subsequent discussions about spatial autocorrelation tend to refer to its positive component.

Nearly all studies about spatial autocorrelation focus on univariate cases, i.e., on the similarity/dissimilarity in nearby locations in a single map pattern in terms of their values. However, *correlation* could be a legitimate statistical concept endemic to bivariate situations. A correlation coefficient should gauge the nature (direction and magnitude) of the relationship between two variables under investigation. Interestingly, spatial autocorrelation can be viewed as a particular case of correlation, although only a single variable is involved, which is why it is known as *auto*correlation. Because any type of correlation should entail two vectors, another vector should be *spatially* derived for spatial autocorrelation to be a type of correlation. One of the most commonly used concepts for this case is a *spatial lag* vector, each element of which represents a weighted mean of a location's neighbors. In this sense, spatial autocorrelation could be rephrased as the correlation between one variable and its spatial lag vector (Lee 2001).

But, what kinds of issues can arise when we combine the two concepts, correlation and spatial autocorrelation? This question might be better captured by a rather new concept known as *bivariate spatial dependence*, which is a simple extension of the general concept of spatial dependence, and can be defined as "a particular relationship between the spatial proximity among observational units and the numeric similarity of their bivariate associations" (Lee 2001, 2012). In a bivariate situation, each observational unit contains a pair of values, and the nature of the bivariate association is assumed to be conceptually defined and numerically evaluated. If the distribution of bivariate associations is not spatially random, then we might legitimately state that bivariate spatial dependence exists.

Before attempting to illustrate the concept of bivariate spatial dependence, we begin with univariate spatial dependence. Any local set composed of a reference observational unit and its neighbors takes on one of the following four types of *univariate spatial association*:

$$\mathbf{H}-\tilde{\mathbf{H}} \qquad \mathbf{H}-\tilde{\mathbf{L}} \qquad \mathbf{L}-\tilde{\mathbf{H}} \qquad \mathbf{L}-\tilde{\mathbf{L}} \tag{1}$$

Here, **H** denotes a value at a reference unit that is greater than or equal to a threshold value (usually the average) or a positive z-score (original values having the mean subtracted and then divided by the standard deviation), and **L** denotes the opposite. $\tilde{\mathbf{H}}$ denotes a *spatial lag* that is greater than or equal to the global average, and $\tilde{\mathbf{L}}$ denotes the opposite. The symbol '–' denotes a univariate horizontal relationship. The symbol "∼" is introduced here to make a clear distinction between an original value at a location and a derived value from a set of locations. This conceptualization is the basis for the Moran's I statistic.

If another concept (i.e., the *spatial moving average*) is introduced, the situation changes substantively. Unlike the spatial lag, this concept treats the reference unit itself as one of its neighbors. Consider

$$\tilde{\mathbf{H}}^{*} \quad \tilde{\mathbf{L}}^{*} \tag{2}$$

Here, $\tilde{\mathbf{H}}^{*}$ and $\tilde{\mathbf{L}}^{*}$ denote the spatial moving averages at each location. This conceptualization forms the foundation for the Getis-Ord's G_i^* statistic. The four types of univariate spatial association listed in (1) reduce to the two values in (2); $\mathbf{H}-\tilde{\mathbf{H}}$ and $\mathbf{L}-\tilde{\mathbf{L}}$ respectively are linked to $\tilde{\mathbf{H}}^{*}$ and $\tilde{\mathbf{L}}^{*}$, but $\mathbf{H}-\tilde{\mathbf{L}}$ and $\mathbf{L}-\tilde{\mathbf{H}}$ can point either way, depending on the differences in values and/or spatial weights. These two values can be conceptualized as two different types of *univariate spatial clusters* (Lee and Cho 2013). This distinction between spatial association types and spatial cluster types is critical because it can represent the two contrasting perspectives of *spatial modeling* and *spatial exploration*. This distinction plays a pivotal role in addressing various issues about multivariate spatial dependence, a particular case of which is bivariate spatial dependence.

We now move to bivariate situations in which two variables, denoted by X and Y, are under investigation. Each observational unit should take on one of the following four types of *bivariate association* (Lee 2012):

$$\begin{matrix} \mathbf{H} & \mathbf{H} & \mathbf{L} & \mathbf{L} \\ | & | & | & | \\ \mathbf{H} & \mathbf{L} & \mathbf{H} & \mathbf{L} \end{matrix} \tag{3}$$

In this work, the symbol '|' denotes a bivariate vertical relationship at a location. Pearson's correlation coefficient is predicated upon this conceptualization and is *aspatial* in nature in the sense that it does not consider the spatial distribution of the pair-wise local bivariate associations.

Suppose that a location has only one neighbor at which the four different types of bivariate association are possible, resulting in the following 16 different types of bivariate spatial association:

$$
\begin{array}{cccc}
\begin{array}{ccc}\mathbf{H} & - & \mathbf{H} \\ | & & | \\ \mathbf{H} & - & \mathbf{H}\end{array} &
\begin{array}{ccc}\mathbf{H} & - & \mathbf{H} \\ | & & | \\ \mathbf{H} & - & \mathbf{L}\end{array} &
\begin{array}{ccc}\mathbf{H} & - & \mathbf{L} \\ | & & | \\ \mathbf{H} & - & \mathbf{H}\end{array} &
\begin{array}{ccc}\mathbf{H} & - & \mathbf{L} \\ | & & | \\ \mathbf{H} & - & \mathbf{L}\end{array} \\
\\
\begin{array}{ccc}\mathbf{H} & - & \mathbf{H} \\ | & & | \\ \mathbf{L} & - & \mathbf{H}\end{array} &
\begin{array}{ccc}\mathbf{H} & - & \mathbf{H} \\ | & & | \\ \mathbf{L} & - & \mathbf{L}\end{array} &
\begin{array}{ccc}\mathbf{H} & - & \mathbf{L} \\ | & & | \\ \mathbf{L} & - & \mathbf{H}\end{array} &
\begin{array}{ccc}\mathbf{H} & - & \mathbf{L} \\ | & & | \\ \mathbf{L} & - & \mathbf{L}\end{array} \\
\\
\begin{array}{ccc}\mathbf{L} & - & \mathbf{H} \\ | & & | \\ \mathbf{H} & - & \mathbf{H}\end{array} &
\begin{array}{ccc}\mathbf{L} & - & \mathbf{H} \\ | & & | \\ \mathbf{H} & - & \mathbf{L}\end{array} &
\begin{array}{ccc}\mathbf{L} & - & \mathbf{L} \\ | & & | \\ \mathbf{H} & - & \mathbf{H}\end{array} &
\begin{array}{ccc}\mathbf{L} & - & \mathbf{L} \\ | & & | \\ \mathbf{H} & - & \mathbf{L}\end{array} \\
\\
\begin{array}{ccc}\mathbf{L} & - & \mathbf{H} \\ | & & | \\ \mathbf{L} & - & \mathbf{H}\end{array} &
\begin{array}{ccc}\mathbf{L} & - & \mathbf{H} \\ | & & | \\ \mathbf{L} & - & \mathbf{L}\end{array} &
\begin{array}{ccc}\mathbf{L} & - & \mathbf{L} \\ | & & | \\ \mathbf{L} & - & \mathbf{H}\end{array} &
\begin{array}{ccc}\mathbf{L} & - & \mathbf{L} \\ | & & | \\ \mathbf{L} & - & \mathbf{L}\end{array}
\end{array}
\qquad (4)
$$

This association is both bivariate *and* spatial because two pairs (bivariate) in adjacent locations (spatial) are compared. The four main diagonal elements clearly show positive bivariate spatial dependence because exactly the same types of pairs of bivariate association are connected. In contrast, the four anti-diagonal elements can be viewed as examples of negative bivariate spatial dependence because rather different types of bivariate association are placed next to each other.

We next consider additional neighbors. If one more neighbor is added, then we have $4^3 = 64$ different types of bivariate spatial association for each local set. The situation becomes more complicated, although a decent chance still exists for observational units to show perfect and positive bivariate spatial dependence. Because areal units in the real world (i.e., administrative units, school districts, and other types of functional regions) have been reported to have approximately six contiguous neighbors on average, we consider $4^7 = 16{,}384$ different types of bivariate spatial associations at each location, and little chance exists of identifying those showing typical positive bivariate spatial dependence.

We might be able to simplify the situation by applying the notion of spatial lag as seen in (1). Because each variable has four different types of univariate spatial association at a location, we always have only 16 different types of *bivariate spatial association* (Lee 2012; Lee and Cho 2013), no matter how many neighbors are involved. Consider

$$
\begin{array}{cccc}
\begin{array}{ccc}\mathbf{H} & - & \tilde{\mathbf{H}} \\ | & & | \\ \mathbf{H} & - & \tilde{\mathbf{H}}\end{array} &
\begin{array}{ccc}\mathbf{H} & - & \tilde{\mathbf{H}} \\ | & & | \\ \mathbf{H} & - & \tilde{\mathbf{L}}\end{array} &
\begin{array}{ccc}\mathbf{H} & - & \tilde{\mathbf{H}} \\ | & & | \\ \mathbf{L} & - & \tilde{\mathbf{H}}\end{array} &
\begin{array}{ccc}\mathbf{H} & - & \tilde{\mathbf{H}} \\ | & & | \\ \mathbf{L} & - & \tilde{\mathbf{L}}\end{array} \\
\\
\begin{array}{ccc}\mathbf{H} & - & \tilde{\mathbf{L}} \\ | & & | \\ \mathbf{H} & - & \tilde{\mathbf{H}}\end{array} &
\begin{array}{ccc}\mathbf{H} & - & \tilde{\mathbf{L}} \\ | & & | \\ \mathbf{H} & - & \tilde{\mathbf{L}}\end{array} &
\begin{array}{ccc}\mathbf{H} & - & \tilde{\mathbf{L}} \\ | & & | \\ \mathbf{L} & - & \tilde{\mathbf{H}}\end{array} &
\begin{array}{ccc}\mathbf{H} & - & \tilde{\mathbf{L}} \\ | & & | \\ \mathbf{L} & - & \tilde{\mathbf{L}}\end{array} \\
\\
\begin{array}{ccc}\mathbf{L} & - & \tilde{\mathbf{H}} \\ | & & | \\ \mathbf{H} & - & \tilde{\mathbf{H}}\end{array} &
\begin{array}{ccc}\mathbf{L} & - & \tilde{\mathbf{H}} \\ | & & | \\ \mathbf{H} & - & \tilde{\mathbf{L}}\end{array} &
\begin{array}{ccc}\mathbf{L} & - & \tilde{\mathbf{H}} \\ | & & | \\ \mathbf{L} & - & \tilde{\mathbf{H}}\end{array} &
\begin{array}{ccc}\mathbf{L} & - & \tilde{\mathbf{H}} \\ | & & | \\ \mathbf{L} & - & \tilde{\mathbf{L}}\end{array} \\
\\
\begin{array}{ccc}\mathbf{L} & - & \tilde{\mathbf{L}} \\ | & & | \\ \mathbf{H} & - & \tilde{\mathbf{H}}\end{array} &
\begin{array}{ccc}\mathbf{L} & - & \tilde{\mathbf{L}} \\ | & & | \\ \mathbf{H} & - & \tilde{\mathbf{L}}\end{array} &
\begin{array}{ccc}\mathbf{L} & - & \tilde{\mathbf{L}} \\ | & & | \\ \mathbf{L} & - & \tilde{\mathbf{H}}\end{array} &
\begin{array}{ccc}\mathbf{L} & - & \tilde{\mathbf{L}} \\ | & & | \\ \mathbf{L} & - & \tilde{\mathbf{L}}\end{array}
\end{array}
\qquad (5)
$$

Each observational unit is assigned to one of these 16 types in terms of *local* bivariate spatial dependence. Certain interesting findings are drawn from this illustration. First, the four cases of perfect positive spatial dependence are observed in the four corners, and their four negative counterparts are observed in the middle. Second, the four cases in the main diagonal are more closely associated with a positive aspatial correlation, measured by Pearson's correlation coefficient, and the four cases in the anti-diagonal are more strongly associated with a negative aspatial correlation. This notion is not confined to Pearson's r, but can extend to other linear correlation coefficients (see Griffith and Amrhein 1991).

By combining these two aspects, we can make certain general statements. First, with a decent level of positive Pearson's r, the main diagonal cases are expected to be more observable than the anti-diagonal cases. If the first and last cases prevail for a local set, a positive bivariate spatial dependence can be said to exist; if the second and third cases prevail, a negative bivariate spatial dependence can be said to exist. Second, with a decent level of negative Pearson's r, the anti-diagonal cases are expected to be more observable than the main diagonal counterparts. If the first and last cases prevail for a local set, a positive bivariate spatial dependence can be said to exist; if the second and third cases prevail, a negative bivariate spatial dependence can be said to exist. In an overall sense, if no bivariate spatial autocorrelation exists, the 16 different types of bivariate spatial association (occurrences of which are subordinate to the nature of the global aspatial correlation) must be randomly distributed; otherwise, they should show a certain degree of spatial clustering.

These situations are further simplified by incorporating the notion of spatial moving average. Because the four different types of univariate spatial association defined in (1) reduce to the two different values seen in (2), the 16 different types of bivariate spatial association defined in (5) can reduce to the following four:

$$\begin{array}{cccc} \tilde{\mathbf{H}}^* & \tilde{\mathbf{H}}^* & \tilde{\mathbf{L}}^* & \tilde{\mathbf{L}}^* \\ | & | & | & | \\ \tilde{\mathbf{H}}^* & \tilde{\mathbf{L}}^* & \tilde{\mathbf{H}}^* & \tilde{\mathbf{L}}^* \end{array} \tag{6}$$

These classifications can be referred to as four different types of *bivariate spatial clusters* (Lee and Cho 2013). The cases in the four corners in (5) represent typical examples of the four types; the others are classified into one of the four cases, depending on differences in values and/or spatial weights.

Key Applications

For this section, I focus on two strands of endeavors that have been undertaken in this particular field: one is to develop a means to remedy the problem of correlation in the presence of bivariate spatial dependence; the other is to devise bivariate spatial autocorrelation statistics for the bivariate counterparts of Moran's I and Getis-Ord's G_i^* statistics.

The test statistic for Pearson's r is given by

$$t = r\sqrt{n-2}\Big/\sqrt{1-r^2} \tag{7}$$

with $n-2$ degrees of freedom when the following two assumptions are satisfied: pairs of observations are drawn from the same, approximately bivariate normal, distribution with constant expectation and finite variance (Haining 1991) and observations of each variable are mutually independent. This standard hypothesis testing procedure for the correlation coefficient might not hold for spatial data. The first assumption of a constant mean structure cannot be assumed because of the potential presence of a global trend. More importantly, the second assumption cannot be sustained because of the usual presence of univariate spatial autocorrelation for both or either of the variables under investigation, which alludes to bivariate spatial dependence.

The standard error of Pearson's r, which is also a part of (7), is given by

$$\hat{\sigma}_r = \sqrt{\frac{1-r^2}{n-2}}, \tag{8}$$

where the denominator is associated with the number of degrees of freedom. This standard error should be adjusted according to the degree of spatial autocorrelation in the variables; it should be larger when positive spatial autocorrelation prevails (and vice versa for negative spatial autocorrelation) (Haining 1991). This outcome can be shown in (8); the lack of independence among pairs of observations due to positive bivariate spatial dependence reduces the number of degree of freedom or effective sample size, thus making the standard error larger.

Several approaches have been proposed in order to remedy or at least alleviate the problem of underestimation of the true sampling variance that the standard inferential test commits (Clifford and Richardson 1985; Dutilleul 1993). In this entry, we focus solely on the Clifford-Richardson's solution (for a more comprehensive treatment, see Griffith and Paelinck 2011). They redefine the equation for the standard error by replacing n in (8) with n', their "effective sample size," which arguably refers to the number of equivalent, independent samples:

$$\hat{\sigma}_r = \sqrt{\frac{1 - r^2}{n' - 2}}. \tag{9}$$

They also provide the equation for computing the effective sample size as

$$n' = 1 + n^2 \left[trace \left(\hat{\mathbf{R}}_X \hat{\mathbf{R}}_Y \right) \right]^{-1}, \tag{10}$$

where $\hat{\mathbf{R}}_X$ and $\hat{\mathbf{R}}_Y$ are the estimated $n \times n$ spatial autocorrelation matrices for the two variables and the *trace* is a matrix operation which is the sum of the diagonal elements. Because each diagonal element of matrix $\hat{\mathbf{R}}_X \hat{\mathbf{R}}_Y$ can be seen as the relative degree of spatial autocorrelation at each location (1 for no spatial autocorrelation, more than 1 for positive spatial autocorrelation), $trace\left(\hat{\mathbf{R}}_X \hat{\mathbf{R}}_Y\right)$ captures the overall degree of bivariate spatial dependence. If no spatial autocorrelation is present for either of the two variables across all locations, each diagonal element of matrix $\hat{\mathbf{R}}_X \hat{\mathbf{R}}_Y$ is 1, $trace\left(\hat{\mathbf{R}}_X \hat{\mathbf{R}}_Y\right) = n$, and thus $n' \cong n$ (Haining 1991). If a positive bivariate spatial dependence prevails, n' is less than n, resulting in a reduced effective sample size or a lesser number of degrees of freedom.

Suppose, for example, that we have 50 pairs of observations and a Pearson's r of 0.3. The test statistic and the number of degrees of freedom according to the standard hypothesis testing method as shown in (7) are, respectively, 2.179 and 48, which implies that r is statistically significant ($p = 0.0343$). If we have a positive bivariate spatial autocorrelation of 2.0 on average across locations, then we have $t = 1.541$ with the effective sample size of 26 ($1 + 50^2/100$) according to (10), which is not statistically significant ($p = 0.1365$). This Clifford-Richardson's solution is implemented in an *R* package named SpatialPack (Vallejos et al. 2013).

Any bivariate spatial autocorrelation statistic should capture the degree of *spatial co-patterning* by measuring both pair-wise covariance and spatial clustering (Lee 2001). One of the most important considerations in determining how to measure bivariate spatial dependence might be the fact that both Pearson's r and Moran's I are cross-product statistics (Getis 1991), which take the form of an average of the sum of products of two vectors. Pearson's r is defined as an average of the cross-product of two standardized vectors, $\mathbf{z}_X$ and $\mathbf{z}_Y$; similarly, Moran's I can be defined as an average of the cross-product of two standardized vectors, $\mathbf{z}_X$ and $\tilde{\mathbf{z}}_X$ (a standardized spatial lag vector), when a spatial weights matrix is row standardized (Lee 2001):

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \tag{11}$$

$$= \frac{1}{n}\sum_i z_{X_i} z_{Y_i}, \text{ and}$$

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

$$= \frac{\sum_i (x_i - \bar{x}) \sum_j w_{ij} (x_j - \bar{x})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (x_i - \bar{x})^2}}$$

$$= \frac{1}{n}\sum_i z_{X_i} \tilde{z}_{X_i}. \tag{12}$$

Predicated upon all of the discussions about univariate statistics for spatial autocorrelation, Lee (2012) identifies six vectors that may play roles in defining bivariate spatial autocorrelation statistics conforming to the general form of cross-product statistic: $\mathbf{z}_X$, $\tilde{\mathbf{z}}_X$, and $\tilde{\mathbf{z}}_X^*$ (a standardized spatial moving average vector) for the X variable and $\mathbf{z}_Y$, $\tilde{\mathbf{z}}_Y$, and $\tilde{\mathbf{z}}_Y^*$ for the Y variable. Using these two sets of vectors, one can obtain various types of bivariate spatial autocorrelation statistics. In this entry, only the following two are discussed (i.e., the cross-Moran or bivariate Moran statistic denoted by CM and Lee's L^* statistic):

$$CM = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x}) (y_j - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} = \frac{1}{n} \sum_i z_{X_i} \tilde{z}_{Y_i}, \text{ and}$$

$$L^* = \frac{n}{\sum_i \left(\sum_j w^*_{ij}\right)^2} \frac{\sum_i \left[\left(\sum_j w^*_{ij} (x_j - \bar{x})\right) \left(\sum_j w^*_{ij} (y_j - \bar{y})\right)\right]}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} = \frac{1}{n} \sum_i \tilde{z}^*_{X_i} \tilde{z}^*_{Y_i}. \quad (13)$$

Here, w_{ij} and w^*_{ij} are elements from a zero diagonal and nonzero diagonal spatial weights matrix, respectively. The former statistic is one derived from a multivariate spatial correlation matrix proposed by Wartenberg (1985) and is a simple extension of univariate Moran's I, thus gauging the correlation between one variable at original locations and the other variable at the neighboring locations (a spatial lag vector). In contrast, the latter, which was proposed by Lee (2001, 2004, 2009), is defined as the correlation between one variable and the other variable's spatial moving average vectors. In comparison, cross-Moran is more congruent with the concept of *cross-correlation*, whereas Lee's L^* deals more directly with the concept of *co-patterning* by considering not only bivariate association at the original locations but also their spatial association with neighboring locations.

In examining the different advantages and weaknesses, one can conclude that the bivariate Moran's statistic is more congruent with the spatial modeling perspective, whereas Lee's statistic is more strongly associated with the spatial exploration perspective. For example, many situations might exist in which one should postulate that a dependent variable at a given set of locations is influenced by independent variables in the neighboring locations. However, if the main interest lies in measuring the spatial similarity between the two map patterns, and exploring and detecting possible bivariate spatial clusters, L^* might be the better option. In addition, L^* is much more congruent with what is documented in (6). The higher the Pearson's aspatial correlation coefficient, and at the same time the higher the level of spatial clustering of bivariate association, the higher the L^* statistic. Certain exploratory spatial data analysis (ESDA) techniques using Lee's local L^*_i (see Eq. 14) can be developed like ones using cross-Moran (Anselin et al. 2002), which is beyond the scope of this entry (see Lee 2012; Lee and Cho 2013):

$$L^*_i = \frac{n^2}{\sum_i \left(\sum_j w^*_{ij}\right)^2} \frac{\left(\sum_j w^*_{ij} (x_j - \bar{x})\right) \left(\sum_j w^*_{ij} (y_j - \bar{y})\right)}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} = \tilde{z}^*_{X_i} \tilde{z}^*_{Y_i} \quad (14)$$

The distributional properties for all bivariate spatial autocorrelation statistics have been established with the randomization assumption (Lee 2004, 2009), which might be crucial to develop certain kinds of ESDA techniques, such as bivariate cluster maps.

Future Directions

Bivariate spatial dependence points to situations in which nearby observational units carry shared information in terms of bivariate association, thus violating the assumption of independent sampling, and the shared information spuriously strengthens (or weakens) the nature of correlation between two variables under investigation, making any conventional statistical inferences or judgments considerably questionable.

The notion and procedure of correlation coefficient decomposition based on the eigenvector spatial filtering (ESF) technique (Griffith and Paelinck 2011; Chun and Griffith 2013) provides an invaluable insight into our understanding of correlation with spatial autocorrelation. It allows an aspatial correlation coefficient to be decomposed into five *sub*-correlations between spatially filtered variables, common spatial autocorrelation components, unique spatial autocorrelation components, one's spatially filtered variable and the other's unique spatial autocorrelation component, and one's unique spatial autocorrelation component and the other's spatially filtered variable.

Bivariate spatial dependence or autocorrelation is a special case of multivariate spatial dependence or autocorrelation (Wartenberg 1985). For example, "trivariate" spatial dependence is simply defined as "a particular relationship between the spatial proximity among observational units and the numeric similarity of their trivariate associations." Thus, we have $4^3 = 64$ different types of *trivariate spatial association*, similar to (5), and $2^3 = 8$ different types of *trivariate spatial clusters*, similar to (6).

Because each pair of variables in a multivariate data set can be viewed as a building block for statistical treatments, the notion of bivariate spatial dependence should have certain implications in spatializing any form of multivariate statistical techniques, e.g., spatial principal components analysis (e.g., Griffith 1988; Dray et al. 2008; Lee and Cho 2014; Lee 2015) and spatial canonical correlation analysis.

Cross-References

- ▸ Spatial Autocorrelation and Spatial Interaction
- ▸ Spatial Autocorrelation Measures
- ▸ Spatial Filtering
- ▸ Spatial Statistics and Geostatistics: Basic Concepts

References

Anselin L, Syabri I, Smirnov O (2002) Visualizing multivariate spatial correlation with dynamically linked windows. In: Anselin L, Rey S (eds) New tools for spatial data analysis: proceedings of the specialist meeting, Center for Spatially Integrated Social Science (CSISS), University of California, Santa Barbara

Bivand R (1980) A Monte Carlo study of correlation coefficient estimation with spatially autocorrelated observations. Quaest Geogr 6:5–10

Clifford P, Richardson S (1985) Testing the association between two spatial processes. Stat Decis Suppl Issue 2:155–160

Clifford P, Richardson S, Hémon D (1989) Assessing the significance of the correlation between two spatial processes. Biometrics 45:123–134

Chun Y, Griffith DA (2013) Spatial statistics & geostatistics: theory and applications for geographic information science & technology. Sage, Los Angeles

Dray S, Sonia S, François D (2008) Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. J Veg Sci 19:45–56

Dutilleul P (1993) Modifying the t test for assessing the correlation between two spatial processes. Biometrics 49:305–314

Getis A (1991) Spatial interaction and spatial autocorrelation: a cross-product approach. Environ Plan A 23:1269–1277

Griffith DA (1980) Towards a theory of spatial statistics. Geogr Anal 12:325–339

Griffith DA (1988) Advanced spatial statistics: special topics in the exploration of quantitative spatial data series. Kluwer, Dordrecht

Griffith DA, Amrhein CG (1991) Statistical analysis for geographers. Prentice-Hall, Englewood Cliffs

Griffith DA, Paelinck, JH (2011) Non-standard spatial statistics and spatial econometrics. Springer, New York

Haining RP (1980) Spatial autocorrelation problems. In: Herbert DT, Johnston RJ (eds) Geography and the urban environment, vol 3. Wiley, New York, pp 1–44

Haining RP (1991) Bivariate correlation with spatial data. Geogr Anal 23:210–227

Lee S-I (2001) Developing a bivariate spatial association measure: an integration of Pearson's r and Moran's I. J Geogr Syst 3:369–385

Lee S-I (2004) A generalized significance testing method for global measures of spatial association: an extension of the Mantel test. Environ Plan A 36:1687–1703

Lee S-I (2009) A generalized randomization approach to local measures of spatial association. Geogr Anal 41:221–248

Lee S-I (2012) Exploring bivariate spatial dependence and heterogeneity: a comparison of bivariate measures of spatial association. Paper presented at the annual meeting of the association of American geographers, New York, 24–28 Feb

Lee S-I (2015) Some elaborations on spatial principal components analysis. Paper presented at the annual meeting of the association of American geographers, Chicago, 21–25 Apr

Lee S-I, Cho D (2013) Delineating the bivariate spatial clusters: a bivariate AMOEBA technique. Paper presented at the annual meeting of the association of American geographers, Los Angeles, 9–13 Apr

Lee S-I, Cho D (2014) Developing a spatial principal components analysis. Paper presented at the annual meeting of the association of American geographers, Tampa, 8–12 Apr

Richardson S, Hémon D (1981) On the variance of the sample correlation between two independent lattice processes. J Appl Probab 18:943–948

Tiefelsdorf M (2001) Specification and distributional properties of the spatial cross-correlation coefficient $C_{\varepsilon_1,\varepsilon_2}$. Paper presented at the Western Regional Science Conference, Palm Springs, 26 Feb

Vallejos R, Osorio F, Cuevas F (2013) SpatialPack – an R package for computing spatial association between two stochastic processes defined on the plane. Available via DIALOG. http://rvallejos.mat.utfsm.cl/Time%20Series%20I%202013/paper3.pdf. Accessed 12 Feb 2016

Wartenberg D (1985) Multivariate spatial correlation: a method for exploratory geographical analysis. Geogr Anal 17:263–283

Correlation Queries

► Correlation Queries in Spatial Time Series Data

Correlation Queries in Spatial Time Series Data

Pusheng Zhang
Microsoft Corporation, Redmond, WA, USA

Synonyms

Correlation Queries; Spatial Cone Tree; Spatial Time Series

Definition

A **spatial framework** consists of a collection of locations and a neighbor relationship. A **time series** is a sequence of observations taken sequentially in time. A **spatial time series dataset** is a collection of time series, each referencing a location in a common spatial framework. For example, the collection of global daily temperature measurements for the last 10 years is a spatial time series dataset over a degree-by-degree latitude-longitude grid spatial framework on the surface of the Earth.

Correlation queries are the queries used for finding collections, e.g. pairs, of highly correlated time series in spatial time series data, which might lead to find potential interactions and patterns. A strongly correlated pair of time series indicates potential movement in one series when the other time series moves.

Historical Background

The massive amounts of data generated by advanced data collecting tools, such as satellites, sensors, mobile devices, and medical instruments, offer an unprecedented opportunity for researchers to discover these potential nuggets of valuable information. However, correlation queries are computationally expensive due to large spatio-temporal frameworks containing many locations and long time sequences. Therefore, the development of efficient query processing techniques is crucial for exploring these datasets.

Previous work on query processing for time series data has focused on dimensionality reduction followed by the use of low dimensional indexing techniques in the transformed space. Unfortunately, the efficiency of these approaches deteriorates substantially when a small set of dimensions cannot represent enough information in the time series data. Many spatial time series datasets fall in this category. For example, finding anomalies is more desirable than finding well-known seasonal patterns in many applications. Therefore, the data used in anomaly detection is usually data whose seasonality has been removed. However, after transformations (e.g., Fourier transformation) are applied to deseasonalize the data, the power spectrum spreads out over almost all dimensions. Furthermore, in most spatial time series datasets, the number of spatial locations is much greater than the length of the time series. This makes it possible to improve the performance of query processing of spatial time series data by exploiting spatial proximity in the design of access methods.

In this chapter, the spatial cone tree, an spatial data structure for spatial time series data, is dis-