Developing a Bivariate AMOEBA Technique

Sang-Il Lee[1], Daeheon Cho[2]

[1]Seoul National University, 1 Gwanak-ro Gwanak-gu, Seoul 151-748, South Korea
Telephone: (82)-2-880-9028
Email: si_lee@snu.ac.kr

[2]Catholic Kwandong University, 579-24 Beomil-ro Gwanak-ro, Gangneung 210-701, South Korea
Telephone: (82)-10-2363-3524
Email: dhncho@gmail.com

1. Introduction

The original AMOEBA technique is directly connected to the spatial cluster analysis which has been a norm for researchers dealing with areal data. The task is to identify areal units associated with certain types of spatial clusters. LISA (local indicators of spatial association) such as local Moran's I_i and Getis-Ord's G_i^* have been extensively utilized. However, people have increasingly recognized that identifying significant areal units is one thing, delineating spatial clusters associated with the areal units is another.

AMOEBA stands for a multidirectional optimal ecotope-based algorithm and has originally been developed by Getis and his associates (Getis and Aldstadt 2004, Aldstadt and Getis 2006, Lee et al 2010). This is basically a way of delineating univariate spatial clusters based on LISA. This paper basically attempts to extend the AMOEBA technique to bivariate situations. Why we need a bivariate AMOEBA? The answer would be that there exists the 'bivariate spatial dependence' in our data. Lee (2001, 2012, 2015) and Lee and Cho (2013) defined the bivariate spatial dependence as "the systematic relationship between the spatial proximity among observational units and the numeric similarity in terms of pairwise bivariate association".

We have four different types of 'pairwise bivariate association' for each areal unit; H=H, H=L, L=H, L=L (the sign of '=' indicates that we are dealing with 'bivariate' situations). A positive bivariate spatial dependence refers to a situation that nearby areal units have better chance to have the same type of pairwise bivariate association; for example, an areal unit of H=H association could be surrounded by ones with the same association. However, the concept of the 'bivariate spatial cluster' is quite complicated. We hardly expect that all the adjacent areal units share the same association type. Then, some crucial questions come out: what if a reference area with H=H is surrounded by 5 neighbors with H=H and one with H=L, or surrounded by 4 with H=H and two with H=L; how to define a bivariate spatial cluster; how to draw its boundary.

Thus, the main objective of the paper is to answer all these questions and to develop a bivariate AMOEBA technique to delineate bivariate spatial clusters.

2. Conceptualizing a Bivariate Spatial Cluster

According to local Moran's I_i, there are four different types of 'univariate spatial association'; H-H, H-L, L-H, L-L (the sign of '-' indicates that we are dealing with

'univariate' situations). However, Getis-Ord's G_i^* reduces those four different types into two values, H^* and L^* by taking a weighted average for a whole local set composed of a reference area and its neighbors. We regard the two values as two different types of 'univariate spatial clusters': H-H and L-L associations are definitely associated respectively with H^* and L^* clusters; however, H-L and L-H associations could go either way.

When one more variable is involved, the situations become much more complicated. For each location, there are 16 different types of 'bivariate spatial association' since two sets of four different types of univariate spatial association should be considered. However, we only have four different types of 'bivariate spatial clusters' when the notions of two univariate spatial cluster types (H^* and L^*) are applied; $H^*=H^*$, $H^*=L^*$, $L^*=H^*$, and $L^*=L^*$. The $H^*=H^*$ type, for example, indicates a situation that a local set shows the H^* cluster type for both variables.

3. A Statistic for a Bivariate AMOEBA

Our statistic for a bivariate AMOEBA is heavily dependent upon Getis-Ord's G_i^*. One of the main advantages of using G_i^* than other LISA is that the statistic can increase when you add more areal units. G_i^* can be written as:

$$G_i^* = \frac{\sum_j w_{ij}^* x_j - w_{i.}^* \bar{x}}{s\sqrt{\left(n w_{i.}^{*(2)} - w_{i.}^{*2}\right)/(n-1)}} \tag{1}$$

If a spatial weights matrix is a row-standardized version of a binary contiguity matrix, equation 1 is simplified into equation 2.

$$G_i^* = \sqrt{\frac{n-1}{n/n_i^* - 1}}\tilde{z}_i^* = \delta\tilde{z}_i^* \tag{2}$$

Now G_i^* is seen as the product of a spatial moving average of z-scores ($\tilde{z}_i^*$) for a local set and a scalar (δ) which is responsible for a larger G_i^* even with a decreased $\tilde{z}_i^*$ due to the addition of areal units.

We utilize a bivariate LISA, L_i^* (Lee 2001, 2004, 2009, 2012, 2015) which can be written as:

$$L_i^* = \frac{n^2}{\sum_i\left(\sum_j w_{ij}^*\right)^2}\frac{\left[\sum_j w_{ij}^*\left(x_j-\bar{x}\right)\right]\left[\sum_j w_{ij}^*\left(y_j-\bar{y}\right)\right]}{\sqrt{\sum_i\left(x_i-\bar{x}\right)^2}\sqrt{\sum_i\left(y_i-\bar{y}\right)^2}} \tag{3}$$

When a spatial weights matrix is row-standardized, equation 3 is reduced to equation 4.

$$L_i^* = \tilde{z}_{X_i}^*\tilde{z}_{Y_i}^* \tag{4}$$

Now, the statistic is seen as the product of spatial moving average of z-scores for *X*-variable and spatial moving average of z-scores for *Y*-variable. By multiplying the δ and the bivariate LISA, the final statistic is presented as:

$$\delta L_i^* = \sqrt{\frac{n-1}{n/n_i^* - 1}} \tilde{z}_{X_i}^* \tilde{z}_{Y_i}^* \tag{5}$$

As more areal units are involved to constitute a region, δ increases but L_i^* usually decreases. If an increase in δ is large enough to compensate for a decrease in L_i^*, the overall statistic increases. If an increase in δ is not large enough, the region stops expanding.

4. An Algorithm

The algorithm is composed of two parts; delineating a maximum boundary for each areal unit and retaining/discarding overlapping boundaries. The first thing you should do is to choose which type of bivariate spatial clusters you are interested in. If you choose the H^*=H^* cluster, then areal units having the H=H type of bivariate association function as seed cells. For only the seed cells, boundaries are drawn.

First, you select a seed cell and calculate $r_i = z_{X_i} z_{Y_i}$. This value is the $\delta L_i^*(0)$, the baseline statistic. Then, consider each of the contiguous neighbors and check which one's addition increases the statistic most. The highest δL_i^* will be $\delta L_i^*(1)$, the maximum value at the stage 1. Now the boundary is the boundary 1. In the second stage, you consider each of the contiguous neighbors and check which one's addition increases the statistic most. The highest δL_i^* will be $\delta L_i^*(2)$, the maximum value at the stage 2. Now the boundary is the boundary 2. If there is no contiguous cell whose addition increases the statistic, then we reach the final stage. The current δL_i^* is $\delta L_i^*(f)$, the final maximum value for cell 1, and the boundary is the maximum δL_i^* boundary.

You can do the same thing for all the other seed cells to obtain their maximum boundaries. Since the resulting boundaries should overlap with each other, we should have an eliminating rule; only the boundary with the highest final statistic survives.

5. An Application

Our example is about Seoul, the capital city of South Korea which is composed of 522 administrative areal units. What we are interested in here is the correlation between average land value, proxy for job centrality, and educational attainment, proxy for income of residents. The global correlation is moderate or somewhat high. In general sense, job centers and good residential areas are spatially separated. But in Seoul, some areas have both. Since we have a positive correlation, it would be reasonable to expect more and larger H^*=H^* and L^*=L^* clusters rather than H^*=L^* and L^*=H^*. We focus H^*=H^* and L^*=L^* spatial cluster boundaries for Seoul. Other situations where a negative overall correlation is dominant will have more and larger H^*=L^* and L^*=H^* rather than H^*=H^* and L^*=L^* clusters.

6. References

Aldstadt J and Getis A, 2006, Using AMOEBA to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis*, 38(4):327-343.

Getis A and Aldstadt J, 2004, Constructing the spatial weights matrix using a local statistic. *Geographical Analysis*, 36(2):90-104.

Lee S-I, 2001, Developing a bivariate spatial association measure: An integration of Pearson's *r* and Moran's *I*. *Journal of Geographical Systems*. 3(4):369-385.

Lee S-I, 2004, A generalized significance testing method for global measures of spatial association. *Environment and Planning A*, 36(9):1687-1703.

Lee S-I, 2009, A generalized randomization approach to local measures of spatial association. *Geographical Analysis*, 41(2):221-248.

Lee S-I, 2012, Exploring bivariate spatial dependence and heterogeneity: A comparison of bivariate measures of spatial association. Annual Meeting of the Association of American Geographers, New York.

Lee S-I, 2015, Conceptualizing and exploring bivariate spatial dependence and heterogeneity: A comparison of bivariate LISA, Prepared for a submission.

Lee S-I and Cho D, 2013, Delineating bivariate spatial clusters: A bivariate AMOEBA Technique. Annual Meeting of the Association of American Geographers, Los Angels.

Lee S-I, Cho D, Sohn H and Chae M, 2010, A GIS-based method for delineating spatial clusters: A modified AMOEBA technique. *Journal of the Korean Geographical Society*, 45(4):502-520 (in Korean).