Comparing multiple testing methods to detect statistically more accurate spatial clusters

Monghyeon Lee
U. Texas at Dallas

Abstract

Spatial cluster is a notable geographical phenomenon, where mass of events with similar characteristics are closer to each other. Many different methods have been developed and used to detect spatial clusters for various purposes. However, hypothesis testing for spatial cluster detection has not yet been fully explored. Spatial cluster detection methods involve multiple tests. Hence, an alternate level of p-value using multiple testing techniques needs to be set, in order to avoid Type I error. This research compares different multiple testing methods in different spatial situations; different level of spatial autocorrelation and sizes.

Keywords: multiple testing, spatial cluster, spatial autocorrelation.

1. Introduction

Geographical events are frequently related to each other in a spatial context. When geographical events are located nearby with statistical significance of similarity, they are known as spatial clusters. Various studies have attempted to detect spatial clusters using different techniques such as K-function (Ripley, 1977), kernel intensity function (Kelsall and Diggle, 1995), SaTScan (Kulldorff, 1995), local Moran's I_i, local Geary's C_i (Anselin, 1995) and Getis-Ord's G_i^* (Ord and Getis, 1995). However, statistical significance test for detected spatial clusters has been rarely provided in the methods mentioned above. In spatial cluster detection, chances are that Type I error might occur but goes unnoticed, if multiple testing is not used for significance test. Greater the number of times the testing is performed higher is the chance that Type I error is detected. Multiple testing leads to a larger chance of finding a rare event, hence, the chance of detecting Type I error increases, and consequently α level is exaggerated. To avoid this inflation, several methods have been utilized for multiple testing; Bonferroni adjustment (Miller, 1977), Šidák correction (Šidák, 1967), sharper Bonferroni correction based on a stepwise procedure (Paez et al. 2002) and False Discovery Rate (FDR) controlling procedure (Benjamini & Hochberg, 1995). Bonferroni adjustment is one of the most frequently used techniques which provides alternative significance level for multiple testing and, so does Šidák correction which is more computationally intensive. These two methods are generally very close to each other but comparatively, Bonferroni adjustment is more pessimistic than Šidák correction's because as Šidák correction's α is always less than Bonferroni adjustment's. Benjamini & Hochberg (1995) argue that these two methods are conservative and propose a new criterion, FDR controlling procedure. It is philosophically different from the classical approaches, because it controls the rate of rejection.

Several attempts have been made to utilize multiple testing for spatial analysis (Lagendra and Fortin,1989; Paez et al., 2002; Castro and Singer, 2006). Spatial clusters in the literature have been tested by different multiple testing techniques that account for both spatial autocorrelation and multiplicity. However, these multiple testing methods have not yet been fully explored in the spatial context with different spatial

situations such as different level of spatial autocorrelation, which leads to redundant information that affects degree of freedom, and different size of study areas that differentiate the number of tests.

In this study, multiple testing methods are compared in spatial context to detect spatial clusters more accurately. The comparison is based on a set of simulations. Detail steps of simulations are as below:

Step1: Generate 10,000 random numbers from the standard normal distribution with virtually no spatial autocorrelation for the 10-by-10 hexagonal tessellation.

Step2: Add weighted 3rd eigenvector of 10-by-10 hexagonal tessellation, $\beta \cdot E_3$, to each random number to generate artificial clusters.

Step3: Count the number of detected clusters using local Moran's I_i with different multiple testing methods: Bonferroni adjustment, Šidák correction, and FDR.

Step4: Repeat steps 2–3 with different β to have different spatial autocorrelation.

Step5: Repeat steps 1–4 for the other 2 tessellations. (i.e., 25-by-25, and 50-by-50).

In this simulation design, 3rd eigenvector can be replaced by the other eigenvectors to have different number of clusters (i.e. 6th). It could also be replaced with different sizes of the tessellation. Different βs are not determined yet. After the simulation, the numbers of detected clusters with different multiple testing methods will be comparable to each other. The numbers should be close to the number of clusters in the eigenvector that has been chosen for the data generation.

References

Anselin, L., (1995), Local indicators of spatial association-LISA, *Geog. Analysis*, 27, 93-115.

Benjamini & Hochberg, (1995), Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Serie B*, 57, 289–300.

Caldas de Castro, M., and Singer, B. H., (2006), Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association. *Geographical Analysis,* 38(2), 180-208.

Kulldorff, M., (1997), A spatial scan statistic, *Communications in Statistics- Theory and Methods*, 26(6), 1481–1496.

Miller Jr, R.G., 1977, Developments in multiple comparisons, *J. Amer. Stat. Ass*, 72, 779-788.

Ord, J. and Getis, A., (1995), Local Spatial Autocorrelation Statistics: Distributional Issues and an Application, *Geographical Analysis*, 27(4), 286-306.

Paez, A., T. Uchida, and K., Miyamoto, (2002), ''A General Framework for Estimation and Inference of Geographically Weighted Regression Models: 1. Location-Specific Kernel Bandwidths and a Test for Locational Heterogeneity.'' *Env. & Planning A*, 34, 733–54.

Ripley, B. D., (1977), Modeling spatial patterns (with discussion), *Journal of the Royal Statistical Society*, 39, 172-212.

Šidák, Z. K., (1967), "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions". *Journal of the American Statistical Association*, 62 (318): 626–633.