Information Sciences 329 (2016) 701–718

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Detecting nominal variables' spatial associations using conditional probabilities of neighboring surface objects' categories

Hexiang Bai [a,*], Deyu Li [a,b], Yong Ge [c], Jinfeng Wang [c]

[a] School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China
[b] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan 030006, China
[c] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

ARTICLE INFO

Article history:
Received 17 November 2013
Revised 30 April 2015
Accepted 2 October 2015
Available online 8 October 2015

Keywords:
Spatial associations
Nominal variable
Conditional probability distribution
Adjacency matrix

ABSTRACT

How to automatically mining the spatial association patterns in spatial data is a challenging task in spatial data mining. In this paper, we propose three indices that represent the per-class, inter-class, and overall spatial associations of a nominal variable, which are based on the conditional probabilities of surface object categories. These indices represent relative quantities and are normalized to the region $[-1, 1]$, which more accord with the intuitive cognition of people. We present some algorithms for detecting spatial associations that are based on these indices. The proposed method can be regarded as an extension of join count statistics and Transiogram. Several constructive examples were used to illustrate the advantages of the new method. Using two real data sets, vegetation types in Qingxian, Shanxi, China and neural tube birth defects in Heshun, Shanxi, China, we ran comparative experiments with other commonly used methods, including join count statistics, co-location quotient, and $Q(m)$ statistics. The experimental results show that the proposed method can detect more subtle spatial associations, and is not sensitive to the sequence of neighbors.

1. Introduction

Spatial associations play an active role in spatial analysis. As an important source of information, they can assist scientists to make more accurate decisions. This is a fundamental issue in spatial analysis, and has been extensively researched. Ahuja [1] used spatial association as the second order image statistics to fit models to a given ensemble of images. Lam et al. [31] applied spatial association analysis to county-level Acquired Immune Deficiency Syndrome (AIDS) data of four regions of the United States for the period 1982–1990 to characterize the spatial-temporal spread of the AIDS epidemic. Overmars et al. [44] demonstrated the presence of the spatial associations in the land use data of Ecuador at different spatial scales. Barbounis and Theocharis [8] used spatial auto-correlation to predict the wind speeds in wind farms. Yang et al. [58] used spatial auto-correlation to analyze the changes in the spatial distribution patterns of population density. Fuller and Enquist [21] used Moran's I to take spatial associations into account in the null models of tree species' association. Diniz-Filho et al. [18] analyzed the spatial associations in the abundance of 28 terrestrially breeding anuran species from Central Amazonia. Meng et al. [38] used Moran's I to select

* Corresponding author. Tel.: +86 351 7010566; fax:+86 351 7018176.
E-mail address: baihx@sxu.edu.cn, crafly@gmail.com (H. Bai).

http://dx.doi.org/10.1016/j.ins.2015.10.003
0020-0255/

an optimal segmentation scale for high resolution remotely sensed imagery. Thach et al. [52] studied the relationship between thermal stress and mortality in Hong Kong using global and local spatial association measures.

Spatial associations are particular important in geosciences. Spatial associations describe the patterns that the similar objects or activities tend to agglomerate in space. These patterns lead to non-Gaussian distribution of the regression residuals of spatial data when using the ordinary least squares regression [3,28]. They produce redundant information in samples of spatial objects which leads to probability reasoning with low accuracy when using traditional statistical inference methods [26,56]. The existence of spatial associations in data will greatly influence the analysis of spatial data. Therefore, the analysis of spatial associations is a necessary step in analyzing spatial data.

Due to that spatial associations commonly exist in spatial data and greatly influences spatial analysis in many aspects, how to effectively detect and measure the spatial associations has attracted many researchers' attentions in the last few decades. Spatial associations can be measured for different types of spatial data. Measures for the point based data include quadrat analysis [53], nearest neighbor analysis [12], Ripley's K-function [47], network K-function [42,43], etc. Measures for the area-based data include Moran's I [15,40], Geary's C [23], Getis' G [24], join count statistics (JCS) [14], etc. Some researchers have extended point based method, for example the Ripley's K-function, to measure spatial associations among points, lines and polygons [27].

The spatial associations of lattice data can be measured for two different types of variables: continuous and interval variables, and nominal variables. There are three commonly used measures for continuous or interval variables, Moran's I [2,15,29,40], Geary's C [23] and Getis' G [24]. These measures depict the spatial association from different perspectives. Moran's I is based on the covariance of a regionalized attribute, and measures the similarity of two surface objects; Geary's C is based on the variance of the attribute [31]; and Getis' G is based on the distance statistics [24].

For nominal variables, JCS [13–15,40] is an effective tool for detecting spatial associations. This method has been extensively applied in ecology [17], remote sensing [11], economics [46], and sociology [16]. JCS compares the observed number of joins that connect objects with the same category (*rr* join) or different categories (*rs* join) with the corresponding expected join number from the random distribution to judge whether there are spatial associations in spatial data or not. Some extensions and modifications have been proposed. For example, Kabos and Csillag [30] proposed a JCS model that did not assume the first order homogeneity on regular lattices. [9,10] proposed local indicators for nominal attributes based on JCS. [51] proposed a modified JCS to take into account the influences of the underlining irrigation systems on the spatial aggregation. Farber et al. [20] used a similarity count to construct new statistical tests based on both random permutation simulations and derived asymptotic distributions for detecting nonlinear dependencies.

The other objective when considering spatial associations is to measure the degree of the dependence between different categories for nominal variables [32,33]. However, [25] noted that, "join count statistics do not lead to a simple summary index or indices analogous to the Geary or Moran measures". The interpretation of JCS depends on the shape and configuration of surface objects. As an index for testing the significance of spatial associations, JCS is a relative quantity associated with the observed join number and the expected join number. This means it is not appropriate for measuring the degree of spatial association. In addition, JCS cannot detect whether one category attracts or repels another category [32].

Many researchers have attempted to solve the problems of JCS. An easily interpretive measure, the co-location quotient (CLQ) [32], was designed for detecting and measuring spatial associations for point-based data. This measure can detect the attraction and resistance between two categories. Nonetheless, CLQ only uses the nearest neighbors of surface objects, and can hardly detect higher order spatial associations [37]. Furthermore, selecting the nearest neighbors rather than all the necessary neighbors of the surface objects means that CLQ can overlook the existence of spatial associations in some situations. Additionally, CLQ is not suitable for irregular lattice data.

$Q(m)$ statistics [36,37,45,48,49] utilized the symbolic entropy to inspect whether the m-surrounding pattern is significantly different from that of a random distribution or not. This measure can detect the existence of complex patterns of $m-1$ nearest neighbors. However, $Q(m)$ cannot lead to a spatial association index for a category. The probability distribution of different configurations is also needed besides $Q(m)$ to find which patterns are in the m-surrounding. By $Q(m)$, one may not judge which kind of spatial association, positive or negative, exists among the surface object and its neighbors when using the equivalent based m-surrounding. An example of this situation is presented in Section 5.1.2. In addition, if there are many possible configuration patterns in the m-surrounding, $Q(m)$ is computationally expensive.

Spatial association can be detected through the conditional probability of observing surface objects from one category with neighbors from another category. This idea has been used by Galiano [22] to detect the segregation between plant species for point based data. However, Galiano's method only checked if the conditional probability is larger than the marginal probability, and could not give an explicit metric for detecting spatial associations. Same idea was also used by Transiogram [33] in describing spatial variabilities of nominal variables. Unlike $Q(m)$ statistics and CLQ, Transiogram can detect the higher order spatial variabilities of nominal variables and is insensitive to the sequence of neighboring surface objects. However, Transiogram did not provide an overall measure of spatial association with respect to all categories, and a baseline of the random spatial distribution for comparison, which make it hard to detect attraction or repulsion between categories.

This paper combines the merits of Transiogram and JCS to develop some new measures for detecting the degrees of the spatial associations of a nominal variable. This new method inherits some advantages of Transiogram and JCS. For example, compared with CLQ and $Q(m)$ statistics, this method can detect higher order spatial associations and is not sensitive to the sequence of neighboring surface objects. Meanwhile, it extends Transiogram and JCS to measure inter-class, per-class and overall spatial associations for a nominal variable. This method quantifies and normalizes the per-class, inter-class, and overall spatial associations of a study area using several indices that range between $[-1, 1]$. Furthermore, each surface object's contribution to