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

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 the global per-class and inter-class spatial association can be calculated by the new method. One illustrative example and two real life examples were given to validate the new method. Comparisons between the proposed method and other three methods including JCS, $Q(m)$ statistics and CLQ were also made in the experiments. The results demonstrate that the new method is consistent with other methods and can effectively measure the degrees of the higher order spatial associations.

The rest of this paper is organized as follows. Section 2 presents some notations and terminologies used in this paper. Section 3 describes the new spatial association detection method. Section 4 describes two experiments that illustrate and verify the new method. Finally, there is a discussion of the results of the two examples and the relation between the proposed method and other methods in Section 5. The last section presents concluding remarks.

2. Notation and terminology

The method proposed in this paper is suitable for lattice data with one nominal attribute. Accordingly, we restrict our discussion to the lattices with only one nominal attribute. For convenience, we review some notations in this section.

Definition 2.1. A one nominal attribute lattice is a pair (U, C), where $U = \{u_1, u_2, \ldots, u_N\}$ is the set of surface objects that represent different regions in a study area and C is a nominal attribute with domain $V_C = \{c_1, c_2, \ldots, c_K\}(K \geq 2)$.

In this paper, a lattice has one nominal attribute, for simplicity. Meanwhile, our model requires that any surface objects must have exactly one label, and each $c_i \in V_C$ is observed in the lattice. In Definition 2.1, a surface object u_α is represented by the corresponding region's location. A value c_i in the domain V_C indicates a category associated with attribute C, and it can be regarded as a mapping from U to $\{0, 1\}$. If u_i belongs to c_i, then $c_i(u_\alpha) = 1$; otherwise $c_i(u_\alpha) = 0$. Meanwhile, attribute C can be regarded as a mapping from U to V_C. If a surface object u_α is labeled c_i, we denote $C(u_\alpha) = c_i$.

Definition 2.2. Let $L = (U, C)$ be a lattice with $|U| = N$. The adjacency matrix of L is defined as $M^1 = [m^1_{ij}]_{N\times N}$, where $m^1_{ij} = m^1_{ji} = 1$ if u_i and u_j are 1-adjacent; otherwise $m^1_{ij} = m^1_{ji} = 0$.

In Definition 2.2, the 1-adjacency of two surface objects can be established using any connectivity algorithm [17]. For example, two surface objects u_i and u_j in the study area are called 1-adjacent if they share a border. The adjacency matrix of a lattice represents all the 1-adjacency relationships between all the surface objects in the lattice. Different lattice data can be represented by a well-defined adjacency matrix and the attributes of surface objects. For example, each polygon in an irregular lattice represents a spatial object and the polygons' adjacency is calculated through judging if two polygons touch each other. For a regular lattice, each grid in the lattice can be a surface object, and a surface object u_α is adjacent to its four neighbors in the rook directions (immediately above, below, left and right).

The kth order adjacency matrices M^k can be recursively derived using the concept of relation composition. That is, $M^k = M^{(k)} - \cup_{i=1}^{k-1} M^{(i)} - E$, where $M^{(i)}$ is the relation composition of i Ms. The kth order adjacency matrix of a lattice is defined as $M^k = [m^k_{ij}]_{N\times N}$, where $m^k_{ij} = 1$ only if two spatial objects u_i and u_j are adjacent to each other via other $k-1$ surface objects and they are not adjacent to each other via any $k' < k-1$ surface objects; otherwise $m^k_{ij} = 0$. Two surface objects are kth order neighbors to each other if and only if the corresponding element in M^k equals 1.

Let u_α be a surface object and u^{+k}_α be its kth order neighbor. The pair $(u_\alpha, u^{+k}_\alpha)$ denotes the path of length k from u_α to u^{+k}_α, ignoring the intermediate surface objects. u_α is called the tail and u^{+k}_α is called the head.

We randomly select a surface object u_α in U, and then observe if it belongs to category c_i. By Definition 2.1 and its explanation, $c_i(u_\alpha)$ can be thought of as an indicator random variable of category c_i at u_α. A random function can be defined on U as $\{c_i(\alpha), \alpha \in U\}$, and can be characterized using an N-variate cumulative probability distribution function (cdf). In this paper, the analysis of spatial association only uses univariate and bivariate cdfs of the random function and their corresponding moments. The model proposed requires the assumption of stationary, i.e., any N-dimensional probability distribution function of the random variable is invariant with respect to its position. According to [26], univariate and bivariate cdfs of the random function and their moments are location-independent under the stationary assumption, and α can be dropped from expressions.

The stationary assumption has been generally accepted when modeling spatial data. As noted by Anselin [4], "In most instances of analyzing spatial data, the proper perspective is not to consider spatial data as a random sample with many observations, but instead as a single realization of a stochastic process. Provided that the underlying stochastic process is sufficiently stationary, the observed pattern will yield information on the characteristics of that process." Many spatial association detecting and measuring methods for lattice data are based on the stationary assumption. That is, they need the assumption that "every location is assumed to have the same chance of receiving any particular value [9,30,33,41]". Our model also uses this generally accepted stationary assumption.

For a surface object pair $(u_\alpha, u^{+k}_\alpha)$, $P(c_i)$ represents the probability that the tail is labeled c_i, i.e., $P(c_i) = P(c_i(u_\alpha) = 1)$ and $P_k(c_i)$ represents the probability that the head is labeled c_i, i.e., $P_k(c_i) = P(c_i(u^{+k}_\alpha) = 1)$. Under the stationary assumption, $P_k(c_i) = P(c_i)$ [26]. $P_k(c_j|c_i)$ represents the conditional probability of the event $c_j(u^{+k}_\alpha) = 1$ when the event $c_i(u_\alpha) = 1$ occurs, i.e., $P_k(c_j|c_i) = P(c_j(u^{+k}_\alpha) = 1|c_i(u_\alpha) = 1)$. $P_k(c_j|c_i)$ can also be regarded as the probability that the surface object category transforms from c_i into c_j along the path $(u_\alpha, u^{+k}_\alpha)$ [33]. $P_k(c_j \cap c_i)$ represents the co-occurrence probability of two events $c_j(u^{+k}_\alpha) = 1$ and $c_i(u_\alpha) = 1$. That is, the probability of a pair with a tail labeled c_i and a head labeled c_j. $P\{C(u_\alpha) = C(u^{+k}_\alpha)\}$ represents the probability of a pair with a tail and a head labeled with the same category. By $E_k(c_i)$ and $E(c_i)$, we denote the expectations of the random variables "the head of a k-th pair is labeled with c_i" and "the tail of a kth pair is labeled with c_i", respectively. That is, $E_k(c_i) = E(c_i(u_\alpha^{+k}))$ and $E(c_i) = E(c_i(u_\alpha))$. By $E_k(c_jc_i)$, we denote the expectation of the product of $c_j(u_\alpha^{+k})$ and $c_i(u_\alpha)$, i.e., $E_k(c_jc_i) = E(c_j(u_\alpha^{+k})c_i(u_\alpha))$. By $Cov_k(c_j, c_i)$, we denote the cross covariance of $c_j(u_\alpha^{+k})$ and $c_i(u_\alpha)$, i.e., $Cov_k(c_j, c_i) = E((c_j(u_\alpha^{+k}) - E(c_j(u_\alpha^{+k})))(c_i(u_\alpha) - E(c_i(u_\alpha))))$.

3. Measures and algorithms

Intuitively, if category c_i attracts category c_j in space, then c_j will be observed in the neighbors of category c_i more frequently than randomly expected. On the contrary, if category c_i repels category c_j in space, then c_j will be observed in the neighbors of c_i less frequently than randomly expected. Therefore, one can measure the spatial attraction of c_i to c_j using $P_k(c_j|c_i)$. Additionally, if category c_j is randomly distributed in the neighbors of category c_i, the occurrence of c_i in one location will not affect the probability of the occurrence of c_j in its neighbor. Consequently, under the stationary assumption, if $P_k(c_j|c_i)$ is larger (less) than $P(c_j)$, then c_i tends to attract (repel) c_j. When $c_i = c_j$, the attraction and repulsion are known as the positive and negative spatial auto-correlations.

3.1. Measuring the spatial association between categories

Let $CP_k(c_j|c_i) = P_k(c_j|c_i) - P(c_j)$. The above discussion suggests that $CP_k(c_j|c_i)$ can depict the spatial association between category c_i and category c_j in the kth order neighbors. However, the following example shows that $CP_k(c_j|c_i)$ cannot perfectly achieve our requirement.

Example 3.1. Suppose that there are two lattices $L1 = \{U_1, C\}$ and $L2 = \{U_2, C\}$, where $V_C = \{c_1, c_2, \ldots, c_K\}$ $(K \geq 2)$. In $L1$, $P(c_i) = 0.3$, $P(c_j) = 0.3$ and $P_1(c_j|c_i) = 1.0$. In $L2$, $P(c_i) = 0.1$, $P(c_j) = 0.1$ and $P_1(c_j|c_i) = 0.85$. Then $CP_1(c_j|c_i)$ for the two lattices $L1$ and $L2$ are 0.7 and 0.75, respectively. This suggests that category c_i attracts category c_j in both lattices. Although $CP_1(c_j|c_i)$ of $L2$ is larger than that of $L1$, the degree that c_i attracts c_j in $L2$ is smaller than in $L1$. Apparently, all the neighbors of the category c_i must belong to the category c_j in $L1$. However, approximately 15% of the neighbors of the category c_i do not belong to category c_j in $L2$.

Because $CP_k(c_j|c_i)$ is an absolute quantity but not a relative quantity, it is unsuitable to the relative comparison of spatial association degree. In other words, we cannot use $CP_k(c_{j_1}|c_i) = CP_k(c_{j_2}|c_i)$ to determine which pair, (c_{j_1}, c_i) or (c_{j_2}, c_i), has a larger kth order spatial association. Take, for example, the positive spatial association, i.e., $P_k(c_j|c_i) > P(c_j)$, we hope to know that the enlarged degree of the probability of c_j under the condition "c_i occurs" with respect to the largest potential enlarged degree $1 - P(c_j)$, i.e., $CP_k(c_j|c_i)/[1 - P(c_j)]$. For negative spatial associations, i.e., $P_k(c_j|c_i) < P(c_j)$, the largest potential change in the probability of c_j under "c_i occurs" is $P(c_j)$, so we can use $CP_k(c_j|c_i)/P(c_j)$ to depict the negative spatial association.

By the above discussion we have the following definition.

Definition 3.1. Let $L = (U, C)$ be a lattice with $U = \{u_1, u_2, \ldots, u_N\}$ $V_C = \{c_1, c_2, \ldots, c_K\}$ $(K \geq 2)$, and $P(c_j) > 0$ for all $c_j \in V_C$. For a kth order pair $(u_\alpha, u_\alpha^{+k})$ with $C(u_\alpha) = c_i$ and $C(u_\alpha^{+k}) = c_j$, we define the index of the kth order inter-class spatial association of c_j with respect to c_i as

$$NCP_k(c_j|c_i) = \begin{cases} \dfrac{CP_k(c_j|c_i)}{1 - P(c_j)}, & CP_k(c_j|c_i) \geq 0 \\ \dfrac{CP_k(c_j|c_i)}{P(c_j)}, & CP_k(c_j|c_i) < 0 \end{cases}. \tag{1}$$

In Definition 3.1, if $c_i = c_j$, $NCP_k(c_j|c_i)$ is called the index of the kth order spatial auto-correlation (per-class spatial association) of c_i, and is denoted as $NCP_k(c_i)$. Note that our model requires $P(c_j) > 0, \forall c_j \in V_C$. If $P(c_j) = 1$ for category c_j, then $P_k(c_j) = 1$ under the stationary assumption. This means that the whole study area has only one category. This is of little use when measuring spatial associations for lattice data, because the whole area is filled with one category. Some measures can be applied to the spatial associations of spatial data with one category, for example the Ripley's K function [47]. However, these measures are designed for point based data rather than lattice data. The spatial associations of point-based data are beyond the scope of this paper.

Definition 3.1 implies that $-1 \leq NCP_k(c_j|c_i) \leq 1$. $NCP_k(c_j|c_i) = 0$ means that c_j is neither attracted nor repelled by c_i at its kth order neighbors. $NCP_k(c_j|c_i) > 0$ means that c_j is attracted by c_i at its kth order neighbors. $NCP_k(c_j|c_i) < 0$ means that c_i repels c_j at its kth order neighbors.

Property 3.1. The index of spatial association $NCP_k(c_j|c_i)$ can be rewritten as

$$NCP_k(c_j|c_i) = \begin{cases} \dfrac{1}{P(c_i)(1-P(c_j))}Cov_k(c_j, c_i), & CP_k(c_j|c_i) \geq 0 \\ \dfrac{1}{P(c_i)P(c_j)}Cov_k(c_j, c_i), & CP_k(c_j|c_i) < 0 \end{cases}. \tag{2}$$

Proof. From the stationary assumption it follows that $P_k(c_j) = P(c_j)$. By the conditional probability formula, it is obvious that

$$CP_k(c_j|c_i) = \frac{P_k(c_j|c_i)P(c_i) - P(c_j)P(c_i)}{P(c_i)} = \frac{P_k(c_j \cap c_i) - P_k(c_j)P(c_i)}{P(c_i)}.$$

Let ξ be an indicator random variable, i.e., ξ only takes two values 0 or 1. Then the expectation $E(\xi)$ is the probability of the event $\xi = 1$, i.e., $E(\xi) = P(\xi = 1)$.

It should be noticed that $c_j(u_\alpha^k)c_i(u_\alpha)$ is also an indicator random variable when both $c_j(u_\alpha^k)$ and $c_i(u_\alpha)$ are indicator random variables. Then, $c_j(u_\alpha^k)c_i(u_\alpha) = 1$ if and only if $c_j(u_\alpha^k) = 1$ and $c_i(u_\alpha) = 1$. So $P_k(c_j(u_\alpha^k)c_i(u_\alpha) = 1) = P_k(c_j \cap c_i)$. This means that $P_k(c_j \cap c_i) = E(c_jc_i)$.

We have that

$$CP_k(c_j|c_i) = \frac{E_k(c_jc_i) - E_k(c_j)E(c_i)}{P(c_i)} = \frac{Cov_k(c_j, c_i)}{P(c_i)}.$$

We can immediately obtain Formula (2) by plug this formula into Formula (1). This completes the proof of the property. □

Formula (1) shows that each central surface object u_α with label c_i, as the tails of some kth order pairs, may contribute to $NCP_k(c_j|c_i)$. Some make $NCP_k(c_j|c_i)$ biased towards a positive value (positive contribution), and the others make $NCP_k(c_j|c_i)$ biased towards a negative value (negative contribution). If u_α is labeled c_i and the proportion of all its neighboring surface objects that are labeled c_j is larger than $P(c_j)$, it will have a positive contribution to $NCP_k(c_j|c_i)$. Otherwise, it will have a negative contribution to $NCP_k(c_j|c_i)$. A positive contribution of u_α to $NCP_k(c_j|c_i)$ means that the surface objects labeled c_j tend to congregate in the neighbors of u_α.

Remark. How to evaluate if a central surface object u_α with category c_i significantly contributes to $NCP_k(c_j|c_i)$? Suppose there are q neighbors that are labeled c_j in all s kth order neighbors of u_α. One can evaluate the contribution significance of u_α to $NCP_k(c_j|c_i)$ by inspecting if the event "q neighbors are labeled c_j in all s kth order neighbors of u_α" is a small probability event under the marginal probability $P(c_j)$ of the category c_j. It is easy to compute this probability using the binomial distribution (see also [54] and [9]).

$$RP_s^q(u_\alpha) = \binom{s}{q} P^q(c_j)(1 - P(c_j))^{s-q}. \tag{3}$$

A vary small $RP_s^q(u_\alpha)$, for example $RP_s^q(u_\alpha) \leq 0.05$, means that the event is a small probability event. According to the principle of a small probability event, i.e., "a small probability event cannot actually happen in once test", we have reason to believe that the observed fact "q neighbors are labeled c_j in all s kth order neighbors of u_α" cannot happen by chance. In other words, occurring q neighbors with category c_j in all s kth order neighbors of u_α significantly depends on u_α belonging to c_i. In the significance test, the values 0.05 or 0.01 are usually known as the significance level.

3.2. Measuring the spatial association of an attribute

Let $L = (U, C)$ be a lattice with $U = \{u_1, u_2, \ldots, u_N\}$, $V_C = \{c_1, c_2, \ldots, c_K\}$ $(K \geq 2)$, and $P(c_j) > 0$ for all $c_j \in V_C$. To determine if attribute C is kth order spatially auto-correlated, we must consider every category $c_i \in V_C$. The neighbors of surface objects tend to belong to the same category as the central surface object, if there is a positive spatial association. Therefore, the kth order spatial association of an attribute can be measured via comparing the probability of observing pairs of surface objects with the same category with the theoretical value under the assumption of no spatial association. The assumption of no spatial association means that "every location is assumed to have the same chance of receiving any particular value and the chance of receiving that value at any location is assumed to be independent of values at other locations [41]". Under the assumptions of no spatial association and stationary, the following lemma holds.

Lemma 3.1. *Let $L = (U, C)$ be a lattice with $U = \{u_1, u_2, \ldots, u_N\}$, $V_C = \{c_1, c_2, \ldots, c_K\}$ $(K \geq 2)$ and $P(c_j) > 0$ for all $c_j \in V_C$. If any surface objects have exactly one label, the theoretical value of $P\{C(u_\alpha) = C(u_\alpha^{+k})\}$ under the assumptions of no spatial association and stationary is*

$$P_E^k = P\{C(u_\alpha) = C(u_\alpha^{+k})\} = \sum_{c_i \in V_C} P^2(c_i).$$

Proof. $P\{C(u_\alpha) = C(u_\alpha^{+k})\}$ is the probability that pairs have tails and heads from the same category. There are $|V_C|$ categories in the study area, so there are at most $|V_C|$ possible outcomes for $C(u_\alpha) = C(u_\alpha^{+k})$, i.e., $C(u_\alpha) = C(u_\alpha^{+k}) = c_i$, $i \in \{1, \ldots, |V_C|\}$.

Algorithm 1 Calculation of $NCP_k(c_j|c_i)$.

```
Input:
  $L(U, C)$: the lattice containing $N$ surface objects to be analyzed;
  $c_i, c_j$: the categories of the tail and head respectively;
  $M^k$: the $k$-th order adjacency matrix;
Output:
  $NCP_k(c_j|c_i)$;
  function INTERNCP($L(U, C), c_i, c_j, M^k$)
      $P(c_i) = P(c_j) = 0$;
      for each $u_\alpha \in U$ do
          $T_{u_\alpha} = 0; I_{u_\alpha} = 0$;
          if $c_j(u_\alpha) == 1$ then
              $P(c_j) = P(c_j) + 1$;
          end if
          if $c_i(u_\alpha) == 1$ then
              $P(c_i) = P(c_i) + 1$;
              for each $u_\alpha^{+k}$ in terms of $M^k$ do
                  $T_{u_\alpha}$=$T_{u_\alpha}$+1;
                  if $c_j(u_\alpha^{+k}) == 1$ then
                      $I_{u_\alpha}$=$I_{u_\alpha}$+1;
                  end if
              end for
          end if
          Set $T = T + T_{u_\alpha}$ and $I = I + I_{u_\alpha}$;
      end for
      $P(c_i) = P(c_i)/N; P(c_j) = P(c_j)/N$;
      $P_k(c_j|c_i) = I/T$;
      Calculate $NCP_k(c_j|c_i)$ in terms of Equation (1).
      return $NCP_k(c_j|c_i)$;
  end function
```

Accordingly, $P\{C(u_\alpha) = C(u_\alpha^{+k})\} = P\{\cup_{c_i \in V_C}(C(u_\alpha) = C(u_\alpha^{+k}) = c_i)\}$. Meanwhile, because the surface objects have exactly one label, then events $C(u_\alpha) = C(u_\alpha^{+k}) = c_i$ and $C(u_\alpha) = C(u_\alpha^{+k}) = c_i$ are disjoint if c_i and c_j are different. Therefore, $P\{\cup_{c_i \in V_C}(C(u_\alpha) = C(u_\alpha^{+k}) = c_i)\} = \sum_{c_i \in V_C} P_k\{C(u_\alpha) = C(u_\alpha^{+k}) = c_i\} = \sum_{c_i \in V_C} P_k(c_i \cap c_i)$. Under the assumption of no spatial association and stationary, $P_k(c_i|c_i) = P_k(c_i) = P(c_i)$. Accordingly, $P_k(c_i|c_i) = P_k(c_i \bigcap c_i)/P_k(c_i) = P(c_i)$. Then, $P_k(c_i \bigcap c_i) = P_K(c_i)P(c_i) = P^2(c_i)$. Therefore, $P\{\cup_{c_i \in V_C}(C(u_\alpha) = C(u_\alpha^{+k}) = c_i)\} = \sum_{c_i \in V_C} P_k(c_i \bigcap c_i) = \sum_{c_i \in V_C} P^2(c_i)$. □

Let $CP_k^O(C) = P_k\{C(u_\alpha) = C(u_\alpha^{+k})\} - P_E^k$. Similarly to measuring the spatial association between categories, we divide $CP_k^O(C)$ by the largest possible change, $1 - P_E^k$ and P_E^k under $CP_k^O(C) \geq 0$ and $CP_k^O(C) < 0$, respectively. That is,

$$NCP_k^O(C) = \begin{cases} \dfrac{CP_k^O(C)}{1 - P_E^k}, & CP_k^O(C) \geq 0 \\ \dfrac{CP_k^O(C)}{P_E^k}, & CP_k^O(C) < 0 \end{cases}. \tag{4}$$

We call $NCP_k^O(C)$ as the index of the kth order overall spatial association of an attribute. In practical applications, $P\{C(u_\alpha) = C(u_\alpha^{+k})\}$ can be approximated using the ratio of joins with the same category to all joins in a map, which is similar to counting the *rr* joins in JCS [13].

It is easy to show that $-1 \leq NCP_k^O(C) \leq 1$. And $NCP_k^O(C) = 0$ means that the attribute has no kth order auto-correlation. $NCP_k^O(C) > 0$ means the kth order positive auto-correlation of the attribute, and $NCP_k^O(C) < 0$ means the kth order negative auto-correlation of the attribute.

3.3. Spatial association detection algorithm

Based on the measures proposed in the previous sections, we developed a new algorithm for detecting spatial associations using NCP values. This algorithm calculates different order NCPs. The orders of adjacency are determined in terms of the applications. $NCP_k(c_j|c_i)$ is approximated by estimating of $P(c_i)$ and $P_k(c_j|c_i)$, and is the key ingredient of the algorithm. $NCP_k(c_i)$ is a special case of $NCP_k(c_j|c_i)$ when $c_i = c_j$. Here, we suppose that there are N surface objects in the study area, and that M^k is the kth order adjacency matrix used. Algorithm (1) shows how to calculate $NCP_k(c_j|c_i)$.

To calculate the overall NCP, we must approximate $P_k\{C(u_\alpha) = C(u_\alpha^{+k})\}$ and P_E^k. $P_k\{C(u_\alpha) = C(u_\alpha^{+k})\}$ can be approximated using the quotient of the total number of joins between surface objects divided by the number of *rr* joins. P_E^k can be approximated using the marginal probability of each category in the study area. The calculation of the overall NCP is shown in Algorithm (2).

Algorithm 2 Calculation of $NCP_k^O(C)$.

```
Input:
  $L(U, C)$: the lattice containing $N$ surface objects to be analyzed;
  $M^k$: the $k$-th order adjacency matrix;
Output:
  $NCP_k^O(C)$;
  function OverallNCP($L(U, C), M^k$)
      for each $c_i$ in $V_C$ do
          $P(c_i) = 0$;
      end for
      $J = J_{rr} = 0$;
      for each $u_\alpha$ in $U$ do
          for each $c_i$ in $V_C$ do
              $P(c_i) = P(c_i) + c_i(u_\alpha)$;
          end for
          for each $u_\alpha^{+k}$ in terms of $M^k$ do
              $J = J + 1$;
              if $C(u_\alpha) = C(u_\alpha^{+k})$ then
                  $J_{rr} = J_{rr} + 1$;
              end if
          end for
      end for
      for each $c_i$ in $V_C$ do
          $P(c_i) = P(c_i)/N$;
      end for
      $P_E^k = \sum_{c_i \in V_C} (P(c_i))^2$;
      Calculate $NCP_k^O(C)$ using Equation (4);
      return $NCP_k^O(C)$;
  end function
```

The statistical significances of $NCP_k(c_j|c_i)$ and NCP_k^O are tested using permutation test [39], because it does not require any assumptions regarding the expected distribution and has been widely used by other methods, such as JCS, Moran's I, CLQ and $Q(m)$ statistics. The permutation test proceeds as follows. First, $n \geq 1000$ random permutations are generated from the original data. Each permutation is a random reshuffle of the original data over space. The values of NCPs for all the permutations are recomputed and then all the NCPs of all permutations result a reference distribution. Then, the reference distribution is compared with the NCPs of the original data to determine the probability that the observed NCPs come from a random distribution. For example, if the original NCPs are more extreme than all but n' of the calculated NCPs from permutations, the permutation's *p*-value would be $(n' + 1)/(n + 1)$. If the *p*-value is smaller than a given threshold (significance level, e.g., 0.01), then the calculated NCP is statistically significant.

3.4. Illustrative examples

We used three simulated data sets with known spatial patterns to illustrate and validate the algorithm. These data sets were: (a) a negative auto-correlated lattice; (b) a positive auto-correlated lattice; and (c) a randomly distributed lattice. For simplicity, there were only two categories, denoted as B (black) and W (white). The three simulated data sets are shown in Fig. 1. In this example, there were 100 units in each simulated data set and four nearest neighbors in the rook directions were the first order neighbors. The marginal probability of each category in all the data sets was 0.5. In this paper, the significant levels were all set to 0.01.

We calculated the first order NCPs for all three data sets to illustrate the new method. Consider the example in Fig. 1(b). There are 180 neighbors of white units, and 170 of these are white. Therefore, $P_1(W|W) = 17/18$ and $NCP_1(W) = (17/18 - 0.5)/(1 - -0.5) \approx 0.889$, according to Eq. 1. There are ten black neighboring units. Therefore, $P_1(B|W) = 1/18$ and $NCP_1(B|W) = (1/18 - 0.5)/0.5 = -0.889$. The $NCP_1(B)$ and $NCP_1(W|B)$ can also be calculated in the same way. Among all neighbors of white units, there are 180 joins between units, and 170 of them connect two units with same category. Therefore, $P_1\{C(u_\alpha) = C(u_\alpha^{+k})\} = 17/18$. Given that $P_E^1 = 0.5$, $NCP_1^O(C) \approx 0.889$.

To validate the effectiveness of the new method, we calculate the first order JCS, Q(m) statistics, and CLQ for all three data sets. The first order per-class and overall NCPs together with the first order per-class and overall JCSs for all three data sets are shown

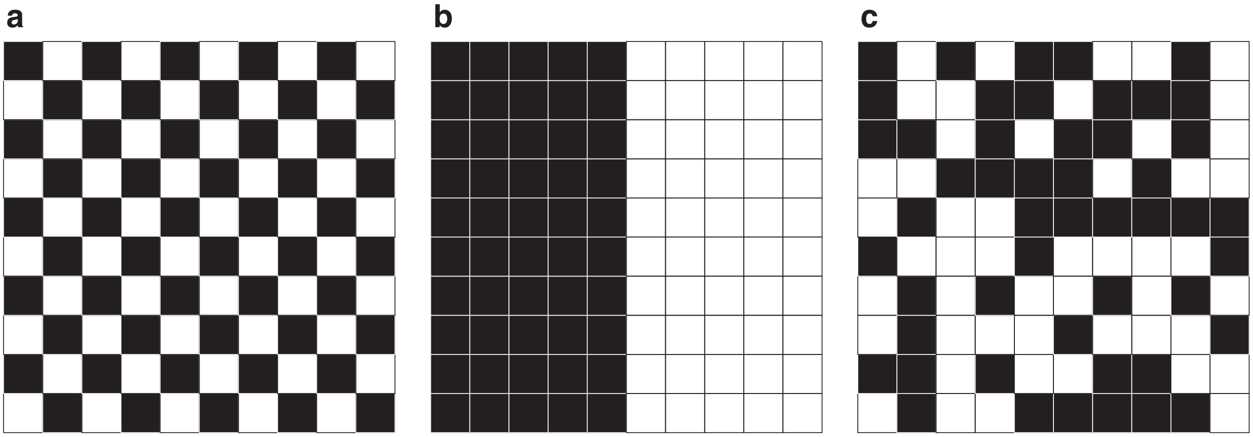

Fig. 1. Three simulated data sets: (a) negatively auto-correlated simulated data; (b) positively auto-correlated simulated data; and (c) randomly distributed simulated data.

Table 1
The first order adjacency per-class and overall NCPs, JCSs and Q statistics for data sets (a), (b) and (c). W represents white, B represents black and C ={W,B}.

	B_a	W_a	C_a	B_b	W_b	C_b	B_c	W_c	C_c
NCP	−1	−1	−1	0.889	0.889	0.889	−0.151[a]	−0.138[a]	−0.144[a]
JCS	90	−45	−45	−80	40	40	13[a]	−6[a]	−7[a]
Q(5)	NA	NA	398.718	NA	NA	422.046	NA	NA	31.2214[a]

[a] Failed to pass the permutation test.

Table 2
The inter-class NCPs for data sets (a), (b), and (c). W represents white and B represents black.

	$W\|B$	$B\|W$
Data (a)	1	1
Data (b)	−0.889	−0.889
Data (c)	0.150[a]	0.138[a]

[a] Failed to pass the permutation test.

Table 3
The CLQs of data sets (a), (b), and (c). W represents white and B represents black.

	$W\|W$	$B\|W$	$B\|B$	$W\|B$	Overall
Data (a)	0	2	0	2	0
Data (b)	1.93	0.11	1.93	0.11	1.91
Data (c)	0.86[a]	1.16[a]	0.88[a]	1.13[a]	0.86[a]

[a] Failed to pass the permutation test.

in Table 1. The first order inter-class NCP is shown in Table 2. The $Q(m)$ statistics for the first order neighbors $Q(5)$ are shown in Table 1 for the three data sets. The CLQs for the three data sets are shown in Table 3. For convenience, the original symbol '$c_i \rightarrow c_j$' used in [32] has been replaced by '$c_j|c_i$'.

For the spatially negatively auto-correlated data, the observed number of *rs* joins was larger than the expected number and the number of *rr* joins for each category was less than the corresponding expected value. The per-class and overall JCSs passed the permutation test. This means that Fig. 1(a) is not from a random distribution. The overall and per-class NCPs were −1, and both passed the permutation test. This means there were negative auto-correlations in the data and NCP was consistent with JCS. The inter-class measures $NCP_1(W|B)$ and $NCP_1(B|W)$ were both equal 1, which means that surface objects with different categories tended to be neighbors. Similarly, the NCPs and JCSs were consistent with each other for the positively correlated data. For the randomly distributed data, the JCSs and NCPs failed to pass the permutation test, which supports that Fig. 1(c) is from a random distribution. The CLQ result was also consistent with the NCPs. For example, the overall CLQ for data sets (a) and (b) were 0 and 1.91, respectively, and the overall CLQ for data set (c) failed to pass the permutation test. Accordingly, data sets (a) and (b)

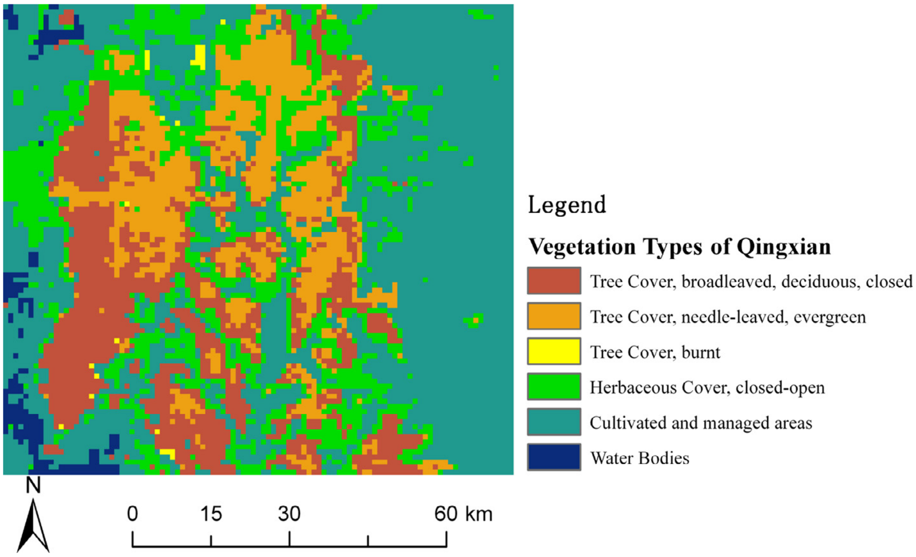
Legend
Vegetation Types of Qingxian
Tree Cover, broadleaved, deciduous, closed
Tree Cover, needle-leaved, evergreen
Tree Cover, burnt
Herbaceous Cover, closed-open
Cultivated and managed areas
Water Bodies
N
0
15
30
60 km

Fig. 2. Map of vegetation types in Qingxian, Shanxi, China.

have negative and positive spatial associations, and data set (c) has no significant spatial associations. This is consistent with the overall NCPs.

Compared with other indices, the $Q(m)$ statistics could not discern the positive and negative spatial associations without the help of the distribution of different configurations of the m-surrounding patterns. For example, the $Q(m)$ statistics of data sets (a) and (b) both passed the permutation test. However, it is hard to judge the difference between the spatial associations of these two data sets using this measure alone. For (a), the most frequent configurations were {B,W,W,W,W} and {W,B,B,B,B}. Therefore, the units in (a) tend to have neighbors from different categories. Similarly, units in (b) tend to have neighbors from the same category. This is consistent with the results from the NCPs.

4. Experiments

We ran two experiments to further validate the new method. The first experiment analyzed spatial associations of vegetation types in Qingxian, Shanxi, China. The vegetation types were stored in a regular lattice. We calculated the first to ninth order NCPs. The second experiment analyzed the spatial distribution of the Neural Tube Birth Defects (NTD) in Heshun, Shanxi, China. We calculated the first to fourth order NCPs, and the contribution of each village. In both experiments, We compared the NCPs with other indices, such as JCS, $Q(m)$ statistics and CLQ, to validate the effectiveness of the new method.

4.1. Vegetation types in Qingxian

In the first experiment, we analyzed the spatial association trends of vegetation types in Qingxian, Shanxi, China over distance. The data (Fig. 2) is from the Global Land Cover 2000 Project [19], and has a spatial resolution of 1km at Equator. The upper-left latitude and longitude are 111°47′53.87′′E and 37°6′26.28′′N, and the lower-right latitude and longitude are 112°48′26.31′′E and 36°12′22.66′′N. There are six different types of vegetation in the study area: (1) broadleaved, deciduous and closed tree cover; (2) needle-leaved and evergreen tree cover; (3) burnt tree cover; (4) closed-open herbaceous cover; (5) cultivated and managed areas; and (6) water bodies. For convenience, we identified the six categories by 'Veg' plus the corresponding code number. For example, the category of broadleaved, deciduous and closed tree cover was identified using 'Veg1'.

To inspect the trends of the spatial associations of different types of vegetation, we calculated the first to ninth order NCPs using the new method. Two units of the lattices are first order neighbors if they are neighbors in the rook directions. Higher order adjacency matrices can be established recursively. The first to ninth order inter-class and per-class NCPs are shown in Fig. 3, which contains 36 trellises. In each trellis, the horizontal axis of the chart represents the order of adjacency and the vertical axis represents the corresponding NCP. The strip on each trellis denotes its content. For example, the trellis corresponding to 'Veg2|Veg3' contains the first to ninth order inter-class NCP of 'Veg2' with respect to 'Veg3'. The gray line in the middle of each trellis represents zero NCP. All NCPs that fail to pass the permutation test were set to zero.

Different order JCSs were also calculated to validate the new measure. Tables 4 and 5 contain the different order per-class and overall NCPs and JCSs, respectively. In these two tables, the first row contains the order of adjacency, and the first column indicates the vegetation type or overall spatial association. Table 4 contains the per-class or overall NCPs. Table 5 contains the differences between the observed and expected numbers of joins.

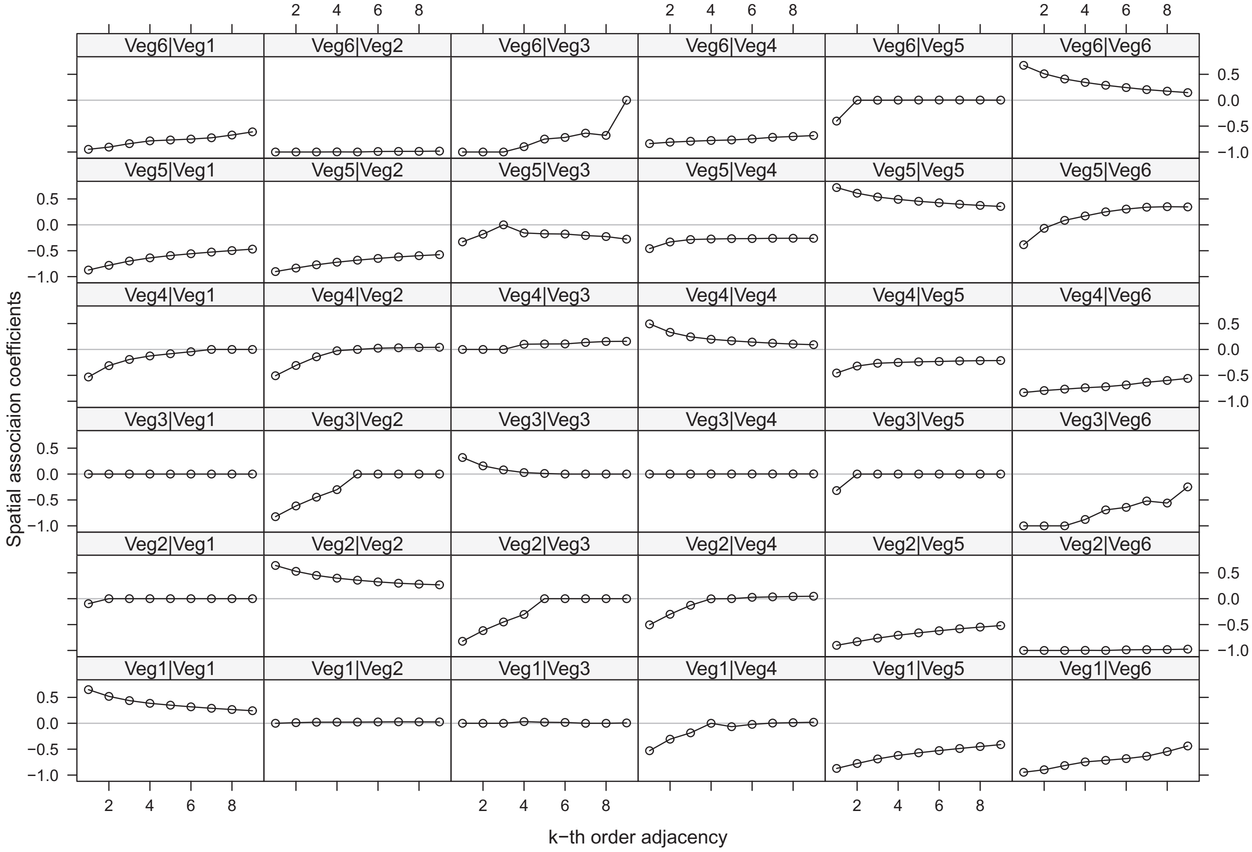

Fig. 3. The per-class and inter-class NCP trends of the vegetation types in Qingxian, Shanxi, China.

Table 4
The first to ninth order per-class and overall NCPs of the vegetation types in Qingxian, Shanxi, China.

	1	2	3	4	5	6	7	8	9
Overall	0.64	0.51	0.43	0.38	0.34	0.31	0.28	0.26	0.24
Veg1	0.65	0.52	0.44	0.39	0.35	0.32	0.29	0.27	0.24
Veg2	0.64	0.53	0.45	0.40	0.36	0.32	0.30	0.28	0.27
Veg3	0.32	0.16	0.08	0.03	0.01	0.00[a]	−0.11[a]	−0.22[a]	0.00[a]
Veg4	0.49	0.33	0.24	0.20	0.17	0.14	0.12	0.10	0.09
Veg5	0.72	0.61	0.54	0.49	0.45	0.43	0.40	0.37	0.36
Veg6	0.67	0.51	0.41	0.34	0.29	0.24	0.20	0.17	0.15

[a] Failed to pass the permutation test.

Table 5
The first to ninth order JCSs of the vegetation types in Qingxian, Shanxi, China.

	1	2	3	4	5	6	7	8	9
Overall	−8342	−13196	−16485	−19202	−21438	−23174	−24431	−25519	−26386.9
Veg1	5064	8604	11446	14012	16434	18554	20446	22202	23688
Veg2	3952	6800	9034	11000	12786	14376	15890	17452	19024
Veg3	36	36	28	14	8	4[a]	2[a]	2[a]	4[a]
Veg4	3982	6114	7660	9114	10462	11678	12706	13636	14618
Veg5	14166	25928	36356	46108	55146	63622	71444	78886	85894
Veg6	532	776	902	974	994	980	930	882	812

[a] Failed to pass the permutation test.

Table 6
$Q(m)$ statistics of the vegetation types in Qingxian, Shanxi, China. ‘ $\geq$20’ is the number of configurations that occurred at least 20 times.

m	5	13	25	41	61	85	113	145	181
Q(m)	88049.8	308307	696453	1227720	1896900	2701320	3640840	4715510	5925240
$\geq$ 20	72	12	6	2	2	2	1	1	1

*Failed to pass the permutation test.

Table 7
CLQs of the vegetation types in Qingxian, Shanxi, China.

	Veg1	Veg2	Veg3	Veg4	Veg5	Veg6
Veg1	3.79	0.90	1.18[a]	0.47	0.13	0.05
Veg2	0.90	4.59	0.18	0.49	0.10	0
Veg3	1.18[a]	0.18	112	0.69[a]	0.67	0
Veg4	0.47	0.49	0.68[a]	3.22	0.54	0.16
Veg5	0.13	0.10	0.67	0.54	1.87	0.61
Veg6	0.05	0	0	0.17	0.60	31
Overall	2.51					

[a] Failed to pass the permutation test.

Table 8
The first to forth order NCPs and JCSs for Heshun, Shanxi, China. Zero represents “Have no NTD” and one represents “Have NTD”. 00 represents joins between two surface objects both labeled “Have no NTD”. 11 represents joins between two surface objects both labeled “Have NTD”. 01 represents joins between two surface objects labeled “Have no NTD” and “Have NTD”, respectively.

k	$NCP_k^O(C)$	$NCP_k(0)$	$NCP_k(1)$	$O_{01}-E_{01}$	$O_{00}-E_{00}$	$O_{11}-E_{11}$
1	0.13	0.12	0.15	−49	477[a]	139
2	0.12	0.11	0.13	−77	833[a]	241
3	0.05[a]	0.03[a]	0.09	−57[a]	1240[a]	355
4	0.00[a]	0.01[a]	−0.06[a]	−2[a]	1664[a]	484[a]

[a] Failed to pass the permutation test.

The $Q(m)$ statistics and CLQ were also calculated for comparison. Table 6 contains the $Q(m)$ statistics for the vegetation types. The first row contains m, which corresponds to the order of adjacency. For example, if the second order adjacency is taken into account, then m is the sum of the central grid, the first order neighbors, and the second order neighbors. The second row contains the $Q(m)$ statistics and the last row contains the number of configurations that occurred at least 20 times. The neighboring grids were sorted according to the rules introduced in [48]. Table 7 contains the CLQs for the vegetation types. The first column of the table corresponds the tail values of the surface objects' pairs, and the first row corresponds to the head. The last row of the table contains the overall CLQs for all categories. Each entry of the table represents a corresponding $CLQ_{c_j|c_i}$. For example, 0.90 in the first row is the value of $CLQ_{Veg2|Veg1}$.

4.2. Neural tube birth defects of Heshun

The NTD data set has been investigated in many previous studies [6,7,34,35,55,57]. Most inhabitants of Heshun are farmers whose living environment seldom changes. There has not been any significant wide-range migration in this district in the past. People here have similar inherited and congenital causes of birth defects. This only explains a few NTD cases. In the study area, there are 322 villages and one town. The locations of these villages were determined by the Geographical Information System for spatial analysis (see Figure 4). All the data were collected by our own field survey. This research project was approved by the Ministry of Science and Technology of the People's Republic of China. The study used only local statistical data. There is no experimental work or ethical issue. As there are no boundaries defined for the villages, we drew them for each village using Voronoi polygons. In [7,57], the spatial auto-correlations of the occurrence rate of NTDs were carried out to detect hot-spots using Getis' G and Moran's I, respectively. In our experiment, we used zero to represent “Have no NTD” and one to represent “Have NTD”.

We used four methods, NCPs, JCS, $Q(m)$ statistics and CLQ, to detect spatial associations of the occurrence of NTDs in the village. If a village had NTD instances, then the attribute was one; otherwise it was zero. Two villages are adjacent to each other in the first order adjacency matrix if they share borders. High order adjacency matrices can be recursively derived in terms of the first order adjacency matrix.

The first to forth order per-class and overall JCSs and NCPs are shown in Table 8. The first to forth order inter-class NCPs are shown in Table 9. we also calculated the $Q(m)$ statistics and CLQ using the point based NTD data that was used to generate the

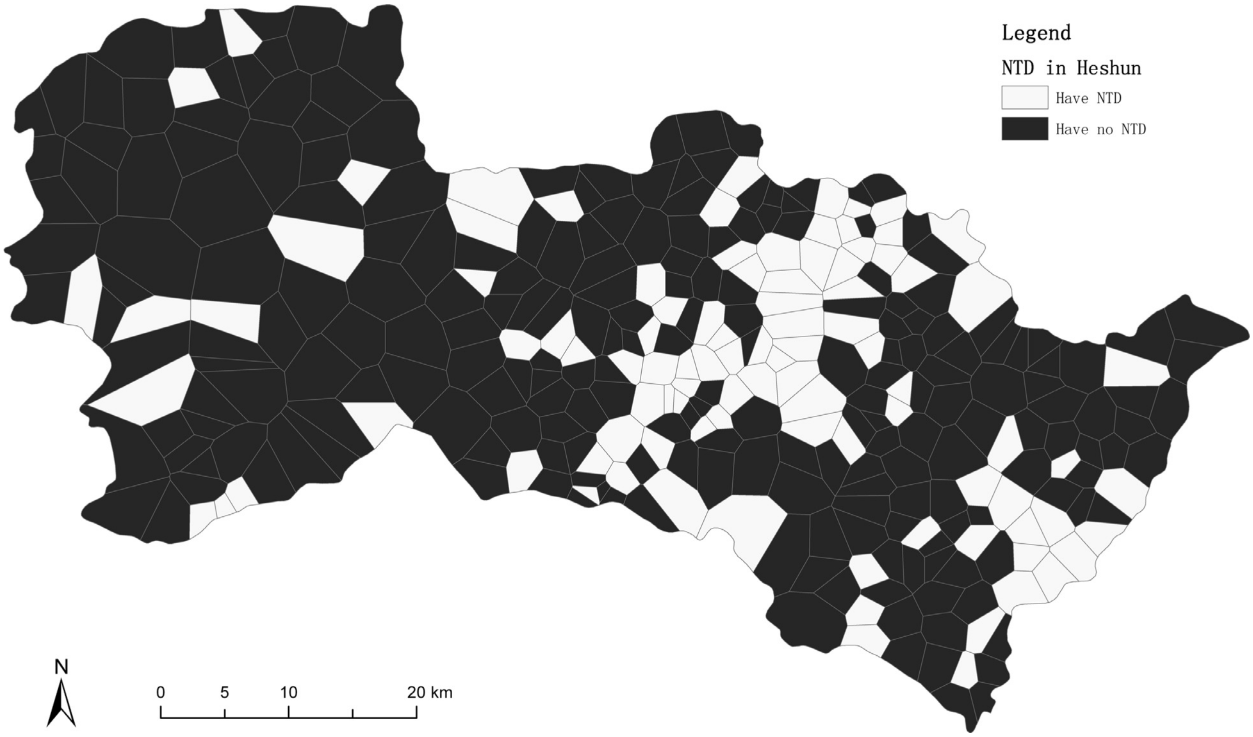

Fig. 4. Map of NTD instances or not for Heshun, Shanxi, China.

Table 9
The first to fourth order inter-class NCPs for the NTD data.

k	$NCP_k(1\|0)$	$NCP_k(0\|1)$
1	−0.12	−0.15
2	−0.11	−0.13
3	−0.03[a]	−0.09
4	−0.01[a]	−0.06[a]

[a] Failed to pass the permutation test.

Table 10
Standard symbol based and equivalent symbol based $Q(m)$ statistics for the NTD data. STD represents standard symbol based Q(m) statistics, and EQU represents equivalent symbol based Q(m) statistics. ‘ ≥20’ is the number of configurations that occurred at least 20 times.

m	5	6	7	8
STD	348[a]	437[a]	591[a]	800
≥ 20	6	3	1	1
EQU	1274	1631	2014	2393
≥ 20	3	3	2	2

[a] Failed to pass the permutation test.

Voronoi polygons. Both standard symbol based and equivalent symbol based $Q(m)$ statistics were calculated. The parameter m of $Q(m)$ was set to five to eight which represents the numbers of the first order neighbors for most villages. Table 10 shows the results of the $Q(m)$ statistics and the number of configurations that occurred at least 20 times. Table 11 shows the results from the CLQ using the point based NTD data. All the per-class, inter-class and overall CLQs were calculated for the NTD data.

Table 11
CLQ for the NTD data.

$CLQ_{1\|1}$	$CLQ_{0\|1}$	$CLQ_{0\|0}$	$CLQ_{1\|0}$	CLQ_{Global}
1.44[a]	0.82	1.03[a]	0.95[a]	1.09[a]

[a] Failed to pass the permutation test.

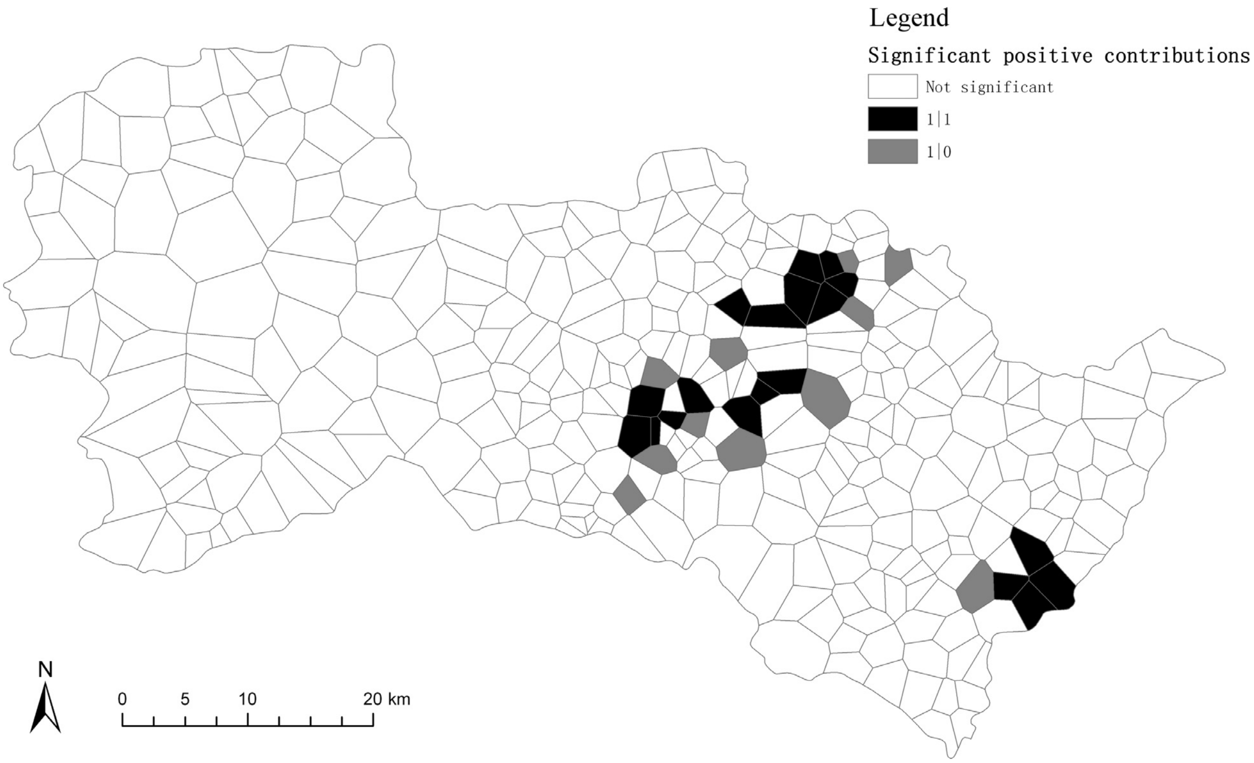

Fig. 5. Significant positive contributions of each village.

The positive contribution of each village to $\{NCP_1(c_j|c_i)|c_i, c_j \in \{0, 1\}\}$ is shown in Fig. 5. Villages where $RP_s^q(u_\alpha) > 0.05$ are colored white in this Figure. Villages colored black significantly positively contribute to $NCP_1(1|1)$, and villages colored gray significantly positively contribute to $NCP_1(1|0)$, given the significance level 0.05.

5. Discussion

5.1. Comparison with other methods

We compared the proposed method with three existing methods, JCS, $Q(m)$ statistics and CLQ, to show its effectiveness and advantages. The comparison shows that the NCP results are consistent with other methods. Additionally, the proposed method can more effectively detect the spatial associations in some cases. For example, NCP can detect inter-class associations but JCS cannot. Compared with $Q(m)$ statistics, NCP does not need to reanalyze the distribution of the configurations of the m neighbors. Although CLQ and NCP both provide inter-class, per-class, and overall spatial association measures, NCP is not confined to the nearest neighbors and can measure higher order spatial associations.

5.1.1. Comparison with JCS

Both experiments demonstrated that the NCP results are consistent with the JCS results. In the first experiment, the overall and per-class NCPs agreed with JCS. Consider the first order spatial associations. The overall NCP was 0.64, while the observed number of *rs* joins was less than the expected count (a difference of −8342). Both the overall NCP and JCS passed the permutation test. The overall JCS supports that the vegetation types are statistically significantly positively associated in space. Because the overall NCP was larger than zero, the new method also showed that there were statistically significantly positive associations. For each vegetation type ‘Vegi’, $NCP_1(\text{Vegi}) > 0$ and the corresponding observed number of *rr* joins was greater than the expected number under the assumption of randomness. the per-class NCPs and JCSs both passed the permutations test. There were only four entries in Tables 4 and 5 that failed to pass the permutation tests, i.e., the sixth to ninth order NCPs and JCSs for category 'Veg3' in Table 4. This indicates that they are from a random distribution.

The situation in the second experiment is complicated. For the third and fourth order adjacency, the NCP and JCS were consistent. The two measures were consistent when measuring the overall spatial association and the association for "Have NTD". However, they were not consistent when measuring the first and second order spatial associations for "Have no NTD". The NCP passed the permutation test but the JCS did not. Although the JCSs of these two special cases did not pass the permutation test, their p-values were very close to 0.01 (0.014 and 0.043).

As an absolute count of the difference between the actual and expected number of joins, JCS is sensitive to the shape and arrangement of the lattices [50]. In the NTD data, the average neighbor number of the villages labeled "Have no NTD" is less than the average neighbor number of all villages from the first to forth order. Compared with the configurations where the villages labeled "Have no NTD" have an average number of neighbors, there were less joins with "Have no NTD" tail in current situation. Accordingly, the number of *rr* joins for "Have no NTD" was also smaller than that of the situations that villages labeled "Have no NTD" have average number of neighbors, when the degrees of spatial association were the same. This leads to smaller JCSs for "Have no NTD", and may lead to large p values in the permutation test when the spatial association is weak. However, NCP takes the number of neighbors into account via the probability. It is robust with respect to the number of neighbors. This may lead to the difference in the first and second order JCSs and NCPs in the second experiment.

5.1.2. Comparison with $Q(m)$ statistics

We used a simple illustrative example to compare our method with $Q(m)$ statistics. Although $Q(m)$ statistics can mine complex spatial patterns, its explanation depends on the probability distributions of different configurations of the m-surrounding no matter whether equivalent symbols are used or not. Therefore, the distributions of different configurations are as important as testing the significance of the $Q(m)$ statistics. Meanwhile, because $Q(m)$ statistics do not test which configuration is significant, the frequency of a configuration is also important in detecting spatial associations.

The $Q(m)$ statistics were also consistent with the NCP. Consider the first experiment. All the $Q(m)$ statistics passed the permutation test. Accordingly, there were significant configurations of the m-surrounding pattern. When m was small, there were many configurations that occurred at least 20 times. When m increased, the number of configurations that occurred at least 20 times decreased. In these configurations, the neighboring surface objects tended to have the same category as the central object. This is consistent with the meaning of the overall NCP, i.e., there are positive first to ninth order spatial associations in the study area. When m was larger than 85, there was only one configuration. in this configuration, the grids were all labeled 'Veg5'. This is also consistent with the per-class NCPs. The per-class NCP of Veg5 was the highest of all the categories from the first to ninth order, in the first experiment.

NCP was more effective than $Q(m)$ statistics in some cases. Consider the second experiment. When m was less than eight, the standard symbol based $Q(m)$ statistics could not pass the permutation test. When $m = 8$, although the $Q(m)$ statistics were significant, there was only one m-surrounding configuration that occurred at least 20 times. This configuration contained objects labeled "Have no NTD". Accordingly, the $Q(m)$ statistics only revealed the auto-correlation of the "Have no NTD" category. All the equivalent based $Q(m)$ statistics for m from five to eight passed the permutation test, and there were at least two configurations that occurred at least 20 times. However, the equivalent symbol configurations neglected the sequence of the symbols. Accordingly, we cannot distinguish between positive or negative spatial associations for a frequent equivalent based m-surrounding pattern. For example, the configuration of four villages with no NTD instances and one village with NTD instances occurs 129 times when $m = 5$. This configuration included at least two situations: the central village having no NTD instances, and the central village having NTD instances. These two situations lead to completely contradict conclusions. The former configuration means that villages with no NTD instances tends to have neighbors with no NTD instances, whereas the later indicates that villages with NTD instances tend to have neighbors with no NTD instances. When using NCP, it is clear that there are first order per-class and overall positive spatial associations.

5.1.3. Comparison with CLQ

In the first experiment, the first order NCP and CLQ results were consistent. The difference between these two indices is in their physical meanings. NCP is the degree that the probability of one category conditional to another category deviates from its expected value, whereas CLQ is the ratio of the observed to expected proportions of one category among another category's nearest neighbors. However, CLQ does not take higher order neighbors into account, so it cannot detect relationships between categories at higher orders. For example, $CLQ_{\mathrm{Veg5}|\mathrm{Veg6}}$ is less than one, which represents that water bodies repel cultivated and managed areas. However, according to the $NCP_k(\mathrm{Veg5}|\mathrm{Veg6})$ when $k > 2$, managed areas tend to congregated in the neighbors of water bodies when the order of adjacency is larger than two. This information was not revealed by the CLQ.

We also compared the NCP and CLQ in the second experiment. Table 11 shows that only one inter-class index $CLQ_{0|1}$ [32] was less than one and statistically significant. Similarly, $NCP_1(0|1)$ was less than zero and passed the permutation test. Both NCP and CLQ support that category "Have NTD" repels category "Have no NTD". However, other CLQ indices did not pass the permutation test, and there were significantly positive overall spatial associations and significant positive spatial associations for the category "Have NTD", in terms of JCS and NCP results. Moreover, the NCP result shows that different categories significantly repel each other in the first order. Although other CLQ indices have same meanings as JCS and NCP, these indices did not pass the permutation tests. This is because the nearest neighbors are only an approximation of near things. In the second experiment, although there are some sort of spatial relationships, either attraction of repulsion, the nearest neighbor cannot sufficiently model near things.

CLQ only takes the nearest neighbors into account [37]. If the relationships between categories cannot be seen in the nearest neighbors as in the second experiment, CLQ may ignore some attractions or repulsions between categories. Compared with CLQ, NCP takes more neighbors into account when detecting the spatial associations. Therefore, NCP is not sensitive to the nearest neighbors and may lower the odds of neglecting relationships between categories.

5.2. Trends of spatial association

NCP provides summary indices of the spatial associations of nominal attributes analogous to Moran's I index, so these indices can be used to mine spatial associations trends in the same way as Moran's I or Getis G. For example, one can use different order adjacency matrices to mine spatial association trends over distances using NCPs. Another type of trend is the contribution of each surface object in the study area. This can be studied using Local Indicators of Spatial Association (LISA) [5] or Local Indicators for Categorical Data (LICD) [9], which are broadly used in hot-spot detection. We analyzed these two types of trends using two experiments.

The trend of the spatial association with respect to distance was analyzed in both experiments. In the first experiment, the trend of spatial association was analyzed via nine different order adjacency matrices that represented different distances between surface objects. The per-class and overall NCPs are summarized in Table 4. The trend of the per-class NCPs are shown in the diagonal of Fig. 3. The overall NCP decreased and tended to become stable when the distance increased. The same situation occurred in the per-class NCPs. This is similar to people's intuition that near things are more related than distant things. For the "Veg3" category, $NCP_k(\text{Veg3})$ cannot pass the permutation test after the fifth order adjacency, which means that there were no significant differences with a random distribution. From Fig. 2, it can be seen that the "Veg3" category mainly distributed in three clusters and the grids in each cluster were almost within four to five grids of each other. Consequently, $NCP_k(\text{Veg3})$ tended to zero after the fifth order.

The inter-class spatial association trends between two vegetation types are shown in Fig. 3, for the first experiment. Most of the inter-class NCPs were less than or equal to zero, which means that most categories repels the other categories or have no significant relationships with other categories. In addition, when the order of adjacency increased, most of the inter-class NCPs approached to zero and became stable. However, there are also some attractions between categories. For example, the inter-class NCP of 'Veg5' with respect to 'Veg6' was greater than zero since the third order. This means that cultivated and managed areas are significantly attracted by water bodies after the third order. Meanwhile, the inter-class NCP for 'Veg6' with respect to 'Veg5' was equal to zero after the first order, which means that water bodies have no correlation to cultivated and managed areas. This is consistent with the distribution of these two categories. 'Veg6' is mainly distributed in the left of the map and 'Veg5' is the most common category besides water bodies. However, there are many grids with vegetation types other than 'Veg6' in the vicinity of 'Veg5'.

In the second experiment, the attribute was randomly distributed after the second order. This means that "Have NTD" in one village is only correlated with the first and second order neighbors, i.e., approximately 5–10 km. This is consistent with the conclusion from [57] that there are grouped distributions of NTDs at a distance scale of 6.2–9.3 km, which are caused by social-economic activities. According to the inter-class NCPs, villages labeled "Have NTD" repel villages labeled "Have no NTD" from the first to third order, and villages labeled "Have no NTD" repel villages labeled "Have NTD" from the first and second order. This also describes the spread of NTDs from the perspective of spatial associations.

We considered the contribution of each surface object in the second experiment. A map of significant positive contributions is given in Figm 5. The villages colored black positively contributed to $NCP_1(1|1)$, and the villages colored gray positively contributed to $NCP_1(1|0)$. There were no villages that significantly positively contributed to $NCP_1(0|0)$ or $NCP(0|1)$. The contribution result of NCP may be used as a starting point for further investigations into the direct courses of NTDs. For example, according to the contribution distribution of the villages, further studies and investigations should be taken on the villages colored gray and its neighbors to attempt to determine the factors that protect the central villages from NTDs. Additionally, it is also important to inspect why the villages colored black tend to have neighbors labeled "Have NTD".

5.3. Relationship with join count statistics

Although the NCP results were consistent with the JCS results, the two methods are based on different principles. NCP uses conditional probability to measure the spatial associations whereas JCS uses the number of *rs* or *rr* joins. The per-class NCP uses the conditional probabilities of different categories of neighboring surface objects, whereas JCS counts the number of *rr* joins for a category. The overall NCP is more closely related to JCS than the per-class NCP. JCS compares the observed number of *rs* joins with the theoretically expected number of *rs* joins. The overall NCP compares the observed probability of the number of *rr* joins with the theoretically expected probability of the occurrence of *rr* joins for all categories. The ranges of these two measures are different. JCS does not provide a summary index. The value of JCS depends on the number of adjacent edges in a map and can be any integer. NCPs are summary indices that range between $[-1, 1]$. In some sense, NCP can be regarded as a generalized version of JCS.

Compared with JCS, NCP can measure inter-class spatial associations, provides summary values, and can more conveniently compare the degree of spatial associations. JCS results are the differences between the number of observed and theoretical joins. The explanation of the number depends on the configuration of the surface objects. For example, $JCS1 = JCS2 = 20$ for two different experiments may have different meanings. The total number of joins and the probability distribution of different categories both influence the explanation of JCS. Another example of the explanation of JCS is that although the second order JCS for "Have NTD" is larger than the first order JCS for "Have NTD" in Table 8, the fourth order JCS is not significantly different from a random distribution while the corresponding first order JCS is. Therefore, it is not safe to conclude that a larger or smaller JCS value corresponds to a stronger or weaker spatial association. As a generalized version of JCS, NCP can determine if there are spatial associations and provides the relative degree of the association. A larger $|NCP|$ corresponds to a conditional probability that has a larger deviation from the theoretical probability and, therefore a stronger positively or negatively spatial association.

5.4. Relationships with Transiogram

Transiogram is a spatial relationship measure proposed by Li [33]. It is based on the bivariate conditional probability function $p_{ij}(h)$, for two different categories of an attribute over a distance lag h. Li [33] interprets $p_{ii}(h)$ and $p_{ij}(h)$, $i \neq j$ as the auto-Transiogram and cross-Transiogram, respectively. Although the Transiogram inherits characteristics, such as sill and range, from geostatistics and can be used as an input in Markov chain models [33], it cannot be directly applied to measure spatial associations because it lacks a baseline for comparison.

NCP can also be regarded as an extension of Transiogram. Obviously, $P_k(c_j|c_i)$ is similar to $p_{ij}(k)$ if the spatial lag is represented using the order of adjacency instead of a distance. NCP compares the conditional probability with the theoretical value from a random distribution, to measure the spatial associations. A permutation test is used to judge if the measure is statistically significant. Compared with Transiogram, the NCP can measure the overall spatial association (that is, if any two neighboring surface objects tend to belong to the same category), as well as the inter-class and per-class spatial associations.

6. Conclusions

In this paper, we proposed a new method for measuring the degree of spatial associations of nominal variables, based on the conditional probability distributions of the categories of neighboring surface objects. We can measure the spatial associations of a nominal attribute by comparing the observed conditional probability and corresponding theoretical value from a random spatial distribution. Not only the per-class and overall spatial association, but also the attractions and repulsions between categories can be measured using the new measure NCP. A positive NCP represents positive spatial associations (attraction), and a negative NCP represents negative spatial associations (repulsion). Meanwhile, a larger absolute NCP value corresponds to a larger divergence between the observed and expected values, and a stronger positive (attraction) or negative (repulsion) spatial association.

To show the effectiveness of the new methods, we compared the results with other commonly used methods using one illustrative and two real-life examples. In all experiments, the new method was consistent with the existing methods. Compared with JCS, NCP provides comparable indices and can detect inter-class spatial associations. Compared with $Q(m)$ statistics, NCP provides per-class and inter-class measures and does not need to reanalyze of different m-surrounding patterns. Compared with CLQ, NCP are not confined to the nearest neighbors and can detect higher order spatial associations. In summary, as an extension of JCS and Transiogram, NCP provides comparable indices that can measure higher order per-class, inter-class, and overall spatial associations.

The new method proposed in this paper uses the permutation test to judge if the spatial association is statistically significant. The permutation test is computationally expensive for large data. In the future, we will study the statistical characteristics of NCP and search for other significance test methods. Additionally, the local contribution of the new method does not completely satisfy two standards of local indicators for spatial associations proposed by Anselin [5]. We should further investigate this issue and develop local indices. Finally, NCP can be applied to many aspects of spatial analysis. For example, the multiple point simulation relies heavily on the selection of the sufficient nearby grids to acquire acceptable quality simulation result. NCP can be used as a tool to find the relationship between the simulation quality and the extent to which the simulation should be considered. Meanwhile, we also plan using information theory to inspect the spatial associations. For example, conditional entropy and mutual information can potentially be used to explain inter-class spatial associations from different and interesting perspectives.

Acknowledgments

The work is supported by National Natural Science Foundation of China (Nos. 41101440 and 61272095), the Natural Science Foundation of Shanxi, China (Nos. 2012021015–1 and 2013011066–4), Project Supported by National Science and Technology (2012BAH33B01), Shanxi Scholarship Council of China (2013–014) and China Postdoctoral Science Foundation (2013M530891).

References

[1] N. Ahuja, Mosaic models for images–III. Spatial correlation in mosaics, Inf. Sci. 24 (1) (1981) 43–69.
[2] P.V. Amaral, L. Anselin, Finite sample properties of Moran's I test for spatial autocorrelation in tobit models, Pap. Reg. Sci. 93 (4) (2014) 773–781, doi:10.1111/pirs.12034.
[3] L. Anselin, Spatial Econometrics: Methods and Models, Studies in Operational Regional Science, Kluwer Academic Publishers, Dordrecht, Netherlands, 1988.
[4] L. Anselin, What is special about spatial data? Alternative perspectives on spatial data analysis, in: D.A. Griffith (Ed.), Spatial statistics, past, present and future, Institute of Mathematical Geography, Ann Arbor, Michigan, 1989, pp. 63–77.
[5] L. Anselin, Local indicators of spatial association–LISA, Geogr. Anal. 27 (2) (1995) 93–115, doi:10.1111/j.1538-4632.1995.tb00338.x.

[6] H. Bai, Y. Ge, J. Wang, D. Li, Y. Liao, X. Zheng, A method for extracting rules from spatial data based on rough fuzzy sets, Knowl. Based Syst. 57 (0) (2014) 28–40, doi:10.1016/j.knosys.2013.12.008.

[7] H. Bai, Y. Ge, J. Wang, Y. Liao, Using rough set theory to identify villages affected by birth defects: the example of Heshun, Shanxi, China., Int. J. Geogr. Inf. Sci. 24 (4) (2010) 559–576, doi:10.1080/13658810902960079.

[8] T.G. Barbounis, J.B. Theocharis, Locally recurrent neural networks for wind speed prediction using spatial correlation, Inf. Sci. 177 (24) (2007) 5775–5797, doi:10.1016/j.ins.2007.05.024.

[9] B. Boots, Developing local measures of spatial association for categorical data, J. Geogr. Syst. 5 (2) (2003) 139–160, doi:10.1007/s10109-003-0110-3.

[10] B. Boots, Local configuration measures for categorical spatial data: binary regular lattices, J. Geogr. Syst. 8 (1) (2006) 1–24, doi:10.1007/s10109-005-0010-9.

[11] Y. Byun, A texture-based fusion scheme to integrate high-resolution satellite SAR and optical images, Remote Sens. Lett. 5 (2) (2014) 103–111, doi:10.1080/2150704X.2014.888817.

[12] P.J. Clark, F.C. Evans, Distance to nearest neighbor as a measure of spatial relationships in populations, Ecology 35 (4) (1954) 445–453, doi:10.2307/1931034.

[13] A.D. Cliff, Spatial Processes: Models and Applications, Pion Ltd, London, United Kindom, 1973.

[14] A.D. Cliff, J.K. Ord, Spatial autocorrelation: a review of existing and new measures with applications, Econ. Geogr. 46 (1970) 269–292.

[15] A.D. Cliff, J.K. Ord, Spatial Processes: Models and Applications, Pion Ltd, London, United Kindom, 1981.

[16] P. Congdon, Estimating life expectancies for US small areas: a regression framework, J. Geogr. Syst. 16 (1) (2014) 1–18, doi:10.1007/s10109-013-0177-4.

[17] M.R.T. Dale, M.-J. Fortin, Spatial Analysis: A Guide for Ecologists, Cambridge University Press, Cambridge, United Kingdom, 2014.

[18] J.A.F. Diniz-Filho, T. Siqueira, A.A. Padial, T.F. Rangel, V.L. Landeiro, L.M. Bini, Spatial autocorrelation analysis allows disentangling the balance between neutral and niche processes in metacommunities, Oikos 121 (2) (2012) 201–210, doi:10.1111/j.1600-0706.2011.19563.x.

[19] J. R. C.European Commission, Global Land Cover 2000 Database, 2003.

[20] S. Farber, M.R. Marin, A. Páez, Testing for spatial independence using similarity relations, Geograph. Anal. 47 (2) (2015) 97–120, doi:10.1111/gean.12044.

[21] M.M. Fuller, B.J. Enquist, Accounting for spatial autocorrelation in null models of tree species association, Ecography 35 (6) (2012) 510–518, doi:10.1111/j.1600-0587.2011.06772.x.

[22] E.F. Galiano, The use of conditional probability spectra in the detection of segregation between plant species, Oikos 46 (2) (1986) 132–138, doi:10.2307/3565459.

[23] R.C. Geary, The contiguity ratio and statistical mapping, Inc. Stat. 5 (3) (1954) 115–127+129-146, doi:10.2307/2986645.

[24] A. Getis, J.K. Ord, The analysis of spatial association by use of distance statistics, Geogr. Anal. 24 (3) (1992) 189–206, doi:10.1111/j.1538-4632.1992.tb00261.x.

[25] M.F. Goodchild, Spatial Autocorrelation, Concepts and Techniques in Modern Geography, Geo Books, Norwich, United Kingdom, 1986.

[26] P. Goovaerts, Geostatistics for Natural Resources Evaluation, Oxford University Press, Oxford, United Kingdom, 1997.

[27] L. Guo, S. Du, R. Haining, L. Zhang, Global and local indicators of spatial association between points and polygons: a study of land use change, Int. J. Appl. Earth Observation Geoinf. 21 (0) (2013) 384–396, doi:10.1016/j.jag.2011.11.003.

[28] R.P. Haining, Spatial Data Analysis in the Social and Environmental Sciences, Cambridge University Press, Cambridge, United Kingdom, 1990.

[29] F. Jin, L.-f. Lee, On the bootstrap for Moran's I test for spatial dependence, J. Econom. 184 (2) (2015) 295–314, doi:10.1016/j.jeconom.2014.09.005.

[30] S. Kabos, F. Csillag, The analysis of spatial association on a regular lattice by join-count statistics without the assumption of first-order homogeneity, Comput. Geosci. 28 (8) (2002) 901–910, doi:10.1016/S0098-3004(02)00007-9.

[31] N.S.-N. Lam, M. Fan, K.-b. Liu, Spatial-temporal spread of the aids epidemic, 1982-1990: a correlogram analysis of four regions of the united states, Geogr. Anal. 28 (2) (1996) 93–107, doi:10.1111/j.1538-4632.1996.tb00923.x.

[32] T.F. Leslie, B.J. Kronenfeld, The colocation quotient: a new measure of spatial association between categorical subsets of points, Geogr. Anal. 43 (3) (2011) 306–326, doi:10.1111/j.1538-4632.2011.00821.x.

[33] W. Li, Transiogram: a spatial relationship measure for categorical data, Int. J. Geogr. Inf. Sci. 20 (6) (2006) 693–699, doi:10.1080/13658810600607816.

[34] Y. Liao, J. Wang, Y. Guo, X. Zheng, Risk assessment of human neural tube defects using a bayesian belief network, Stoch. Environ. Res. Risk Assess. 24 (1) (2009a) 93–100, doi:10.1007/s00477-009-0303-5.

[35] Y. Liao, J. Wang, X. Li, Y. Guo, X. Zheng, Identifying environmental risk factors for human neural tube defects before and after folic acid supplementation, BMC Public Health 9 (2009b) 391, doi:10.1186/1471-2458-9-391.

[36] F. López, M. Matilla-García, J. Mur, M.R. Marín, A non-parametric spatial independence test using symbolic entropy, Reg. Sci. Urban Econ. 40 (2-3) (2010) 106–115, doi:10.1016/j.regsciurbeco.2009.11.003.

[37] F.A. López, A. Páez, Distribution-free inference for q(m) based on permutational bootstrapping: an application to the spatial co-location pattern of firms in madrid, Estadística Española 54 (177) (2012) 135–156.

[38] Y. Meng, C. Lin, W. Cui, J. Yao, Scale selection based on Moran's I for segmentation of high resolution remotely sensed images, in: Proceedings of IEEE 2014 International Geoscience and Remote Sensing Symposium (IGARSS), Québec City, Canada, 2014, pp. 4895–4898, doi:10.1109/IGARSS.2014.6947592.

[39] D.S. Moore, G.P. McCabe, Introduction to the Practice of Statistics, second ed., W. H. Freeman and Company, New York, 1993.

[40] P.A.P. Moran, The interpretation of statistical maps, J. R. Stat. Soc. Ser. B (Methodol.) 10 (2) (1948) 243–251, doi:10.2307/2983777.

[41] J. Odland, Spatial Autocorrelation, Sage Publications, Thousand Oaks, California, United States, 1987.

[42] A. Okabe, K.-I. Okunuki, S. Shiode, The SANET toolbox: new methods for network spatial analysis, Trans. GIS 10 (4) (2006) 535–550, doi:10.1111/j.1467-9671.2006.01011.x.

[43] A. Okabe, I. Yamada, The k-function method on a network and its computational implementation, Geogr. Anal. 33 (3) (2001) 271–290, doi:10.1111/j.1538-4632.2001.tb00448.x.

[44] K. Overmars, G. de Koning, A. Veldkamp, Spatial autocorrelation in multi-scale land use models, Ecol. Model. 164 (2-3) (2003) 257–270, doi:10.1016/S0304-3800(03)00070-X.

[45] A. Páez, M. Ruiz, F. López, J. Logan, Measuring ethnic clustering and exposure with the Q statistic: an exploratory analysis of Irish, Germans, and Yankees in 1880 Newark, Ann. Assoc. Am. Geogr. 102 (1) (2012) 84–102, doi:10.1080/00045608.2011.620502.

[46] M.B. Pietrzak, J. Wilk, T. Kossowski, R. Bivand, The Identification of Spatial Dependence in the Analysis of Regional Economic Development Join-Count Test Application, Technical Report, Institute of Economic Research, Toruń Poland, 2014.

[47] B.D. Ripley, The second-order analysis of stationary point processes, J. Appl. Probab. 13 (2) (1976) 255–266, doi:10.2307/3212829.

[48] M. Ruiz, F. López, A. Páez, Testing for spatial association of qualitative data using symbolic dynamics, J. Geogr. Syst. 12 (3) (2010) 281–309, doi:10.1007/s10109-009-0100-1.

[49] M. Ruiz, F. López, A. Páez, Comparison of thematic maps using symbolic entropy, Int. J. Geogr. Inf. Sci. 26 (3) (2012) 413–439, doi:10.1080/13658816.2011.586327.

[50] M.J. de Smith, M.F. Goodchild, P.A. Longley, Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools, Troubador Publishing Ltd, Leicester, United Kindom, 2009.

[51] M.-D. Su, M.-C. Lin, C.-H. Lin, S.-F. Wang, T.-H. Wen, H.-I. Hsieh, A spatial aggregation index for effective fallow decision in paddy irrigation demand planning, Paddy Water Environ. 10 (2012) 31–39, doi:10.1007/s10333-011-0258-2.

[52] T.-Q. Thach, Q. Zheng, P.-C. Lai, P.P.-Y. Wong, P.Y.-K. Chau, H.J. Jahn, D. Plass, L. Katzschner, A. Kraemer, C.-M. Wong, Assessing spatial associations between thermal stress and mortality in Hong Kong: a small-area ecological study, Sci. Total Environ. 502 (2015) 666–672, doi:10.1016/j.scitotenv.2014.09.057.

[53] R. Thomas, Introduction to Quadrat Analysis, Concepts and Techniques in Modern Geography, Geo Abstracts, Norwich, United Kingdom, 1977.

[54] M.G. Turner, R.V. O'Neill, R.H. Gardner, B.T. Milne, Effects of changing spatial scale on the analysis of landscape pattern, Landsc. Ecol. 3 (3-4) (1989) 153–162, doi:10.1007/BF00131534.

[55] J. Wang, X. Li, G. Christakos, Y. Liao, T. Zhang, X. Gu, X. Zheng, Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China, Int. J. Geogr. Inf. Sci. 24 (1) (2010) 107–127, doi:10.1080/13658810802443457.

[56] J. Wang, J. Liu, D. Zhuan, L. Li, Y. Ge, Spatial sampling design for monitoring the area of cultivated land, Int. J. Remote Sens. 23 (2) (2002) 263–284, doi:10.1080/01431160010025998.

[57] J. Wu, J. Wang, B. Meng, G. Chen, L. Pang, X. Song, K. Zhang, T. Zhang, X. Zheng, Exploratory spatial data analysis for the identification of risk factors to birth defects, BMC Public Health 4 (2004) 23, doi:10.1186/1471-2458-4-23.

[58] M. Yang, J. Ma, P. Jia, Y. Pu, G. Chen, The use of spatial autocorrelation to analyze changes in spatial distribution patterns of population density in Jiangsu province, China, in: X. Li (Ed.), Proceedings of the 19th International Conference on Geoinformatics, 2011, Blackwell Publishing Ltd, Shanghai, China, 2011, pp. 1–6, doi:10.1109/GeoInformatics.2011.5980909.