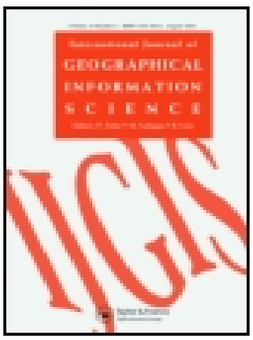

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: http://www.tandfonline.com/loi/tgis20

Spatial association detector (SPADE)

Xuezhi Cang & Wei Luo

To cite this article: Xuezhi Cang & Wei Luo (2018): Spatial association detector (SPADE), International Journal of Geographical Information Science, DOI: 10.1080/13658816.2018.1476693

To link to this article: https://doi.org/10.1080/13658816.2018.1476693

Published online: 24 May 2018.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2018.1476693

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Spatial association detector (SPADE)

Xuezhi Cang and Wei Luo

Department of Geographic and Atmospheric Sciences, Northern Illinois University, DeKalb, IL, USA

ABSTRACT

The geographical detector model can be applied to either spatial or non-spatial data for discovering associations between a dependent variable and potential discrete controlling factors. It can also be applied to continuous factors after they are discretized. However, the power of determinant (PD), measuring data association based on the variance of the dependent variable within zones of a potential controlling factor, does not explicitly consider the spatial characteristics of the data and is also influenced by the number of levels into which each continuous factor is discretized. Here, we propose an improved spatial data association estimator (termed as SPatial Association DEtector, SPADE) to measure the spatial data association by the power of spatial and multilevel discretization determinant (PSMD), which explicitly considers the spatial variance by assigning the weight of the influence based on spatial distribution and also minimizes the influence of the number of levels on PD values by using the multilevel discretization and considering information loss due to discretization. We illustrate our new method by applying it to simulated data with known benchmark association and to dissection density data in the United States to assess its potential controlling factors. Our results show that PSMD is a better measure of association between spatially distributed data than the original PD.

ARTICLE HISTORY
Received 20 June 2017
Accepted 10 May 2018

KEYWORDS
Geographical detector; spatial association; multilevel discretization; spatial variance

1. Introduction

1.1. *Introduction of geographical detector*

The geographical detector (Geo-detector) is a relatively new spatial analysis method (Wang *et al*. 2010) that explores the association between variables. It was first developed in medical geography to estimate the associations between a health outcome, such as mortality rate, and risk factors, such as water pollution and social economic factors, based on their spatial distribution. The premise is that if the dependent variable is controlled by an independent variable (or potential factor), their spatial distribution will be identical or very similar and the similarity can be measured in terms of the variance of the dependent variable within zones of the independent variable (Wang *et al*. 2010). Specifically, if dependent variable Y is controlled by factor X, then the spatial distribution of the two should be identical or very similar, and the similarity in spatial distribution can be measured by dividing the dependent variable into zones (levels) of each independent variable and by comparing the local zonal variance and the global variance. For example, let us assume the dependent variable (such

CONTACT Xuezhi Cang xcang@niu.edu

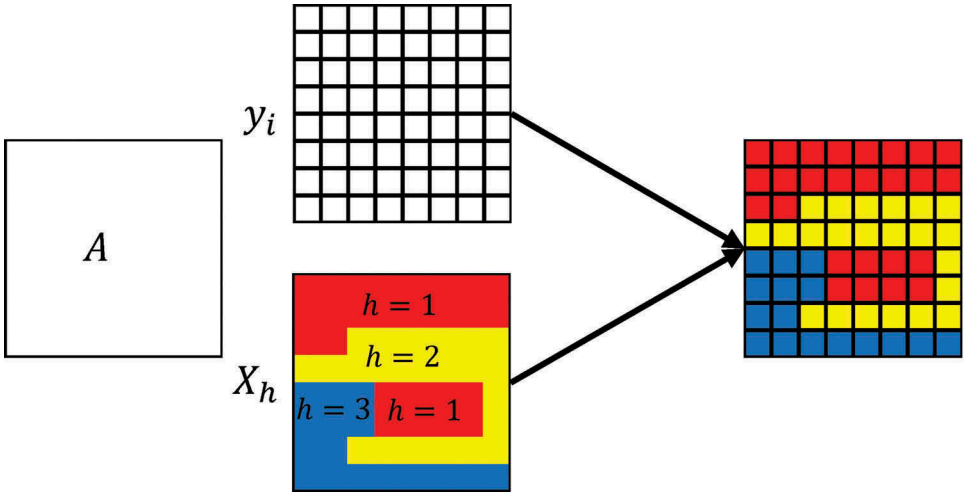

Figure 1. Method of the geographical detector.

as a health outcome or a quantitative geographical phenomenon) is sampled by a grid system $(Y = \{Y_i | i = 1, 2, 3, \ldots n\})$ (see Figure 1). A potential controlling factor (independent variable) is represented by a Geographic Information Science (GIS) data layer *X*; it is either already a categorical variable or can be discretized into a finite number of zones or levels (e.g. in Figure 1, *X* has three zones, $\{X_h | h = 1, 2, 3\}$). The PD, which measures the similarity in spatial distribution between the dependent variable *Y* and a potential factor *X*, is defined as follows,

$$\mathrm{PD} = q = 1 - \frac{\mathrm{SSW}}{\mathrm{SST}} = \frac{\mathrm{SSB}}{\mathrm{SST}} \tag{1}$$

$$\mathrm{SSW} = \sum_{h=1}^{1} \sum_{i=1}^{N_h} \left(Y_{hi} - \overline{Y_h}\right)^2 = N_h \sigma_h^2 \tag{2}$$

$$\mathrm{SSB} = \sum_{h=1}^{L} N_h \left(\overline{Y_h} - \bar{Y}\right)^2 \tag{3}$$

$$\mathrm{SST} = \mathrm{SSW} + \mathrm{SSB} = \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2 = N\sigma^2 \tag{4}$$

where SSW is the sum of squares within zones; SSB is the sum of squares between zones; SST is the total sum of squares; *L* is the total number of zones of *X*; N_h is the count of samples of the *h*th zone; Y_{hi} is the *i*th sample of dependent variable within *h*th zone of *X*; $\overline{Y_h}$ is the mean of *Y* within the *h*th zone of *X*; $\bar{Y}$ is the global mean of *Y* (i.e. mean of *Y* over the entire study area); Y_i is the *i*th sample of the entire study area and *N* is the total count of samples.

1.2. *Analysis of Geo-detector*

The main advantage of the Geo-detector is that it makes fewer assumptions than other methods such as the regression (Wang and Xu 2017). Geo-detector has been applied to many different fields, including physical geography (Du *et al*. 2016, Luo *et al*. 2016) and urban geography (Ren *et al*. 2014, Zhu *et al*. 2015). A complete list of applications and software can be found on the website of the geographical detector (http://geodetector.org/). However, there are a number of drawbacks to the original Geo-detector method. First, PD neglects the characteristics of spatial data by utilizing the sum of squares (within zones or between zones) to describe the similarity of a data set. The sum of squares, although it is used widely in statistics, cannot describe the similarity of spatial data properly, because it cannot represent the most important characteristics of spatial data, that is, spatial dependence. The only thing that is 'spatial' or 'geographic' about the Geo-detector is the implied overlay of GIS data layers in the data preparation, but the method itself never explicitly considers spatial distribution and can be applied to either spatial or non-spatial data; thus, 'Geo-detector' is really a misnomer. Spatial dependence plays two types of roles in spatial data association: within a GIS layer or between GIS layers. Within a GIS layer, spatial dependence refers to the spatial autocorrelation. Spatial autocorrelation, also commonly called spatial association in the literature, describes the association between nearby observations and can be measured by Moran's I, Geary's C or Local Indicators of Spatial Association (Anselin 1995). Between GIS layers, spatial dependence is described by geographically weighted regression (GWR) (Brunsdon *et al*. 1996). The purpose of the Geo-detector and our method is to measure the association between GIS layers. To distinguish the concept of spatial autocorrelation, we will refer to the association between GIS layers as spatial data association or simply spatial association. Spatial data association can be measured by GWR. GWR is suitable for assessing the association between continuous variables or between continuous variable and dummy variable transformed from discrete variable. For the association between variables, GWR takes advantage of the linear model in the assessing and assumes the correlations between variables are linear relationship (Brunsdon *et al*. 1996). However, the association measured by the Geo-detector is not limited by the linear relationship, because the Geo-detector measures the association by stratified heterogeneity (Wang *et al*. 2016, Wang and Xu 2017). The stratified heterogeneity is measured by comparing the difference between strata (zones) of a data layer. Therefore, when estimating association between two continuous variables, one of them needs to be discretized into zones first. The scope of application of the Geo-detector is wider than GWR; however, the Geo-detector neglects the spatial dependence as mentioned above. In a spatial data set, the associations between locations vary with distance. Normally, based on the first law of geography, attribute values at closer locations are more similar and have stronger relations. The drawback of the sum of squares is that it treats every value equally in the same data set. To address the drawback of the sum of squares, we present a solution to this problem by using spatial variance. Besides the problem of similarity measure, the second drawback of the Geo-detector is that the selection of the number of discretization zones of continuous variables is arbitrary and may cause underestimation of the association between two continuous variables. As outlined above, the Geo-detector method works with categorical (or discrete) variables. Continuous variables must first be discretized into a finite number of zones (or levels). How a continuous variable is discretized and how many zones it is discretized into can influence the resultant PD values. Cao *et al*. (2013) discretized continuous

variables into two to eight zones using a number of different discretization methods (e.g. equal interval, natural breaks, quantile, geometrical interval, and standard deviation) and compared the resultant PD values. Their results suggest that the quantile method generally produced the highest PD value and is thus better than other methods. However, they only examined a small number of zones (two to eight) and did not examine in detail the impact of number of zones on PD value. In traditional cartography, the mapping of quantitative variables as different levels of gray is usually limited to less than eight (e.g. Brewer and Pickle 2002, Cauvin *et al*. 2010), because human eyes are only capable of seeing 30 shades of grays (Kreit *et al*. 2013) and a small number of classes is easier for humans to distinguish and understand than a large number of classes. So the rule of thumb in cartography is to use three to seven classes for grayscale representation (Harvey 2015). With the development of geo-computation, advanced spatial analysis methods can take advantage of more classes than the limited number of grayscales designed for easy distinguishing and interpretation by humans (Kwan 2004). Conceptually, if a continuous variable is discretized into a large number of zones, each zone will be small and the variance within each zone will also likely be smaller, leading to a larger PD value. This appears to be the case based on existing publications using the Geo-detector method. For example, Wang *et al*. (2010) used both continuous factors (*distance to fault, distance to buffer, elevation*) and categorical or nominal factors (*soil type, watershed* and *lithozone*). The continuous factors were discretized to five categories. All the categorical variables have more than five categories: *soil type* has nine categories, *watershed* has nine categories and *lithozone* has seven categories. The Geo-detector showed that the top three variables are the three nominal variables. Cao *et al*. (2014) presented a spatial data discretization method which utilized local spatial autocorrelation indices to discretize continuous variables and suggested that the variables need to be discretized into more than 40 classes. They also found that the PD value increases with the increasing number of categories. Luo *et al*. (2016) applied the Geo-detector method to evaluate the factors controlling the surface dissection density in the conterminous United States by physiographic regions. In that study, all the continuous factors are discretized to 6 categories and the categorical variable lithology has 21 categories, much more than 6. The results showed that for four of eight physiographic regions, the dominant factor (with maximum PD value) is lithology. To further confirm this, we discretize the continuous factor elevation into a range of categories (4–22) and calculate their corresponding PD values in Region 3 (Interior Highlands) of the United States. The result is shown in Figure 2. It is clear that there is a general trend of the PD value increasing as the number of discretization zones increases. We will refer to the number of discretization zones as the discretization level hereafter.

Conceptually, one can imagine two extreme cases: if the discretization level is one (i.e. the whole study area is one zone), the zonal variance and global variance are the same, then the PD would be 0; if the discretization level is the same as the total number of samples of a continuous variable (i.e. each sample is its own zone), the zonal variance will be 0 and the PD value will be 1 (see Equation (1)). Thus, the effect of the discretization level on the PD value in the Geo-detector method must be controlled in order for PD values to be more comparable, meaningful and interpretable across different situations. Here we present a solution to this problem using multilevel discretization and considering information loss due to discretization.

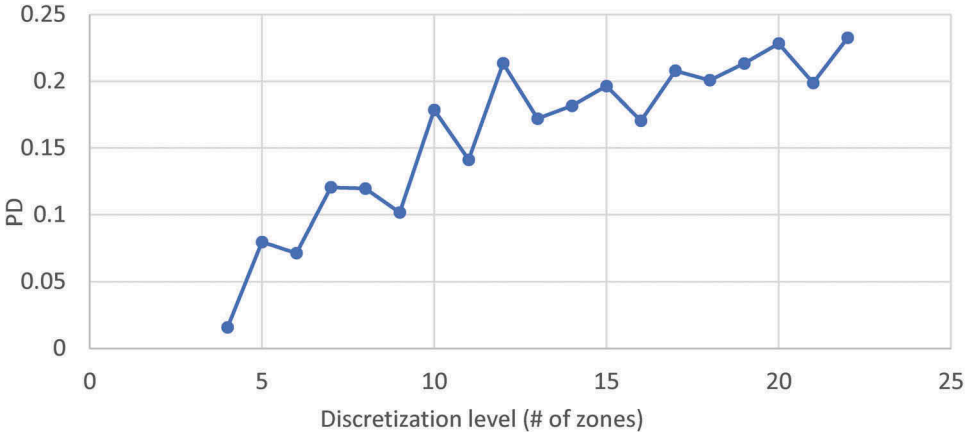

Figure 2. PD values of *elevation* under different discretization level (from 4 to 22).

2. Spatial association detector (SPADE) considering spatial variance, multilevel discretization and information loss

2.1. *Spatial variance*

Spatial dependence, as the most important characteristics of spatial data, is ignored by the Geo-detector (Wang *et al*. 2010). The spatial dependence can be represented as spatial autocorrelation measured by Moran's I, Geary's C, Semivariogram, spatial interaction models, etc. The common point of these models is that they all take advantage of the spatial weighted cross-product statistic (Getis 1991). The general form of the mean of spatial weighted cross-product is shown as below:

$$\Gamma = \frac{\sum_i \sum_{j \neq i} w_{ij} c_{ij}}{\sum_i \sum_{j \neq i} w_{ij}} \tag{5}$$

where w_{ij} is the weight between ith location and jth location. Here, we set the inverse of distance as the weight (see more discussion in Section 5). c_{ij} measures the attribute similarity, such as semi-squared difference $\frac{(y_i - y_j)^2}{2}$ or absolute difference $|y_i - y_j|$. Here we will use semi-squared difference (because doing so will make the original Geo-detector a special case of the new method, as will be shown next):

$$c_{ij} = \frac{(y_i - y_j)^2}{2} \tag{6}$$

In the extreme case that the weight matrix is a matrix of ones, which means that all weights between locations are equal to 1, the mean spatial weighted cross-product becomes the variance equation (Bachmaier and Backes 2008):

$$\Gamma = \frac{\sum_i \sum_{j\neq i} w_{ij} \frac{(y_i - y_j)^2}{2}}{\sum_i \sum_{j\neq i} w_{ij}} = \frac{\sum_i \sum_{j\neq i} \frac{(y_i - y_j)^2}{2}}{N(N-1)} = \frac{1}{2}\frac{1}{N(N-1)} \sum_{all\ i\neq j} (y_i - y_j)^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{N-1} \tag{7}$$

For N samples, there are $N(N\text{-}1)$ pairs of cross-product. As Equation (7) showed, Equation (2) is a special case of Equation (5) when the weights are equal to 1. Based on the equation transformation (Wang *et al*. 2016), the sum of squares can be represented by the product of counts of variable and variance (Equations (2) and (4)). Here, the spatial sum of squares is represented by the product of variable counts and spatial variance. We define the ratio of local spatial sum of squares and the global counterpart as power of spatial determinant (PSD) or q_s:

$$q_s = 1 - \frac{\sum_{h=1}^{L} N_h * \Gamma_h}{N * \Gamma} \tag{8}$$

where N_h is the total count of samples in hth category; Γ_h is the spatial variance within level h; L is the total number of levels; N is the total count of samples and Γ is the total spatial variance. To give a specific example, let us assume X and Y are two spatial variables which cover the same area. Y is a continuous dependent variable. X is an independent categorical variable and has h levels, so we do not need to consider the discretization process in this example. The spatial data association (q_s) or PSD between Y and X is:

$$q_s = 1 - \frac{\sum_{h=1}^{L} N_h * \Gamma_h}{N * \Gamma_t} = 1 - \frac{\sum_{h=1}^{L} N_h \frac{\sum_{h=1}^{N_h} \sum_{j\neq i}^{N_h} w_{hi,hj} \frac{(y_{hi} - y_{hj})^2}{2}}{\sum_{i=1}^{N_h} \sum_{j\neq i}^{N_h} w_{hi,hj}}}{N \frac{\sum_{i=1}^{N} \sum_{j\neq 1}^{N} w_{i,j} \frac{(y_j - y_j)^2}{2}}{\sum_{i=1}^{N} \sum_{j\neq i}^{N} w_{i,j}}} \tag{9}$$

where $w_{i,j}$ is the weight between the ith location and the jth location, which is taken as the inverse of distance in our calculation; subscript $_{hi}$ and $_{hj}$ are the ith and jth samples in the hth category.

2.2. *Multilevel discretization and information loss*

Discretization level corresponds to the minimum perception of measurement. The discretization makes a variable easier to be understood by humans. In essence, discretization keeps the main information of the variable and ignores the finer details. The smaller the discretization level (i.e. fewer categories), the more the information loss. Too few levels cannot represent variation within the original variable well, but too many levels may result in redundant information and could thus hinder the analysis and interpretation. However, it is always difficult to find the best discretization level, because the best discretization level is related to the characteristics of the data and the specific research questions (e.g. Liao *et al*. 2010). Because the study fields are different, the best discretization levels are not even the same for similar research works (Wang *et al*. 2010, Cao *et al*. 2014). Thus, it is desirable to conduct the analysis at multiple discretization levels and compensate information loss due to

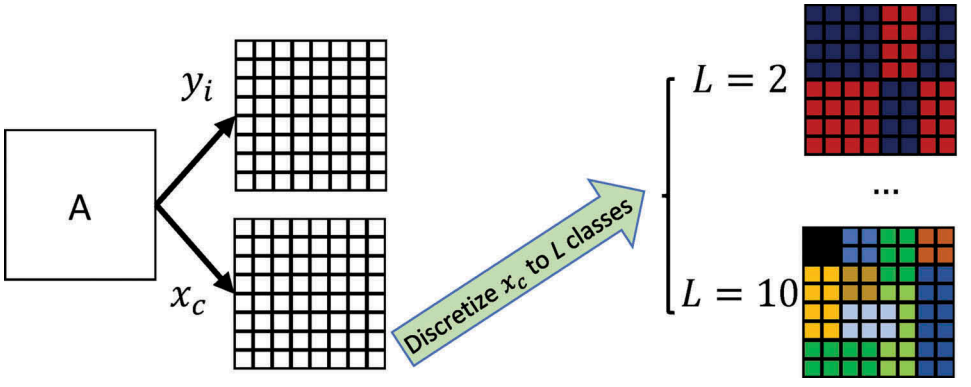

Figure 3. Discretizing continuous independent variable using different discretization levels (zones).

discretization. As shown in Figure 3, we discretize the continuous independent variable into different levels and apply spatial variance to each discretization level.

In this article, we select the quantile discretization for three reasons. First, a previous comparative research suggested that the quantile method is better than others discretization methods (e.g. equal interval, natural breaks, geometrical interval and standard deviation) (Cao *et al*. 2013). Second, the example of this article is from a previous research project (Luo *et al*. 2016). In that article, Luo *et al*. (2016) utilized the quantile discretization. Selecting the same discretization method will greatly facilitate the comparison. Third, the quantile method can minimize the information loss as measured by information entropy, which is defined as Equation (10) (Quinlan 1986, Baez *et al*. 2011):

$$F = -\sum_{i=1}^{N} p(i)\log_2 p(i) - \left(-\sum_{h=1}^{L} p(h)\log_2 p(h)\right) \tag{10}$$

where N is the total count of samples; $p(i)$ is the probability of finding sample i; h is the hth level and $p(h)$ is the probability of a data point or sample belonging to the hth level. The first term of Equation (10) is the information contained in the original un-discretized data; the second term is the information contained after discretization; F is the information lost due to discretization. To minimize Equation (10), the second term should be maximized and this is achieved when all the $p(h)$s are equal (MacKay 2003, p 43–44), that is, quantile discretization. For example, if we have 100 samples (of a continuous variable) and discretize them to four levels ($L = 4$), F will be minimized (second term maximized) if $p(h) = 0.25$. There are two extreme examples, which will not be used in practice but can help understand the concept. One is that if $L = N$ (i.e. there is no discretization), then $F = 0$ which means that there is no information loss; the other is that if $L = 1$ (i.e. all samples are grouped in one level), then F reaches the maximum (first item of this equation), which means that all the information is lost. As described in Figure 3, the SPADE uses levels of the independent variable to stratify (or divide) the dependent variable and to compare the zonal spatial variance and global spatial variance as a way to measure the spatial association between them. It only measures the relationship between the continuous dependent variable and the discretized independent variable, but omits the information loss as a result of the discretization. The information remaining after discretization can be measured by examining the spatial data association

between the original continuous independent variable and its discretized counterpart. If all information is kept (i.e. loss = 0), the spatial data association between the original continuous independent variable and its discretized counterpart should be 1. Following this line of reasoning, we can calculate the spatial data association between the original continuous independent variable and its discretized counterpart as a measure of the information still remaining after discretization (we will call this $q_{s\_infokept}$ to distinguish it from the q_s as defined in Equation (8)). The information kept can be expressed as:

$$q_{s\_infokept} = 1 - \frac{\sum_{h=1}^{L} N_h * \Gamma_{h\_ind}}{N * \Gamma_{total\_ind}} \tag{11}$$

where the subscript $_{h\_ind}$ represents the hth level of independent variable; Γ represents the spatial variance. We define the new compensated power of spatial discretization determinant (Q_s) as the ratio of the above two quantities (i.e. relative to the information kept, thus compensating for the information loss due to discretization):

$$Q_s = q_s / q_{s\_infokept} = \frac{1 - \frac{\sum_{h=1}^{L} N_h * \Gamma_{k\_dep}}{N * \Gamma_{total\_dep}}}{1 - \frac{\sum_{h=1}^{L} N_h * \Gamma_{h\_ind}}{N * \Gamma_{total\_ind}}} \tag{12}$$

For reasons that will become clear later, we define the power of spatial and multilevel discretization determinant as follows:

$$PSMD\_Q_s = MEAN(Q_s) \tag{13}$$

where MEAN represents the mean of Q_s at all discretization levels.

2.3. *Test of significance*

After some transformations (Wang *et al*. 2016), the probability density function (PDF) of PDs is a noncentral F-distribution with the first degree of freedom (d.f.)*L*-1, the second d.f. *N*-*L*, and noncentrality λ. The null hypothesis (two variables have no association) can be tested by critical value test. The PDF of PSMD is unknowable because it is influenced by the spatial pattern of locations which varies in each case. If all the weights are equal to 1, the PD is a noncentral F-distribution, just as the Geo-detector article showed (Wang *et al*. 2016). Although the PDF of PSMD cannot be known, the null hypothesis (two variables have no association) can still be tested. Following the idea of current spatial analysis software (such as ArcGIS, PySAL and GeoDa) (Anselin *et al*. 2006, Rey and Anselin 2010), the null hypothesis can be tested by a *Z*-test, which assumes that the PDF of PSMD under null hypothesis is a normal distribution; or it can be tested by the pseudo *p*-value approach, which has a broader constraint of distribution than traditional significance test. For an observed PSMD, if the *Z*-score (see Equation (14)) is greater than 1.96 or if pseudo *p*-value (see Equation (15)) is smaller than 0.05, the confidence level is above 95%. We select the pseudo *p*-value approach. The reason will be explained in Section 5.

$$Z = \frac{PSMD_{obv} - Mean(PSMD_{random})}{SE(PSMD_{random})} \tag{14}$$

where $PSMD_{obv}$ is the observed PSMD; $Mean(PSMD_{random})$ is the mean of $PSMD_{random}$; $SE(PSMD_{random})$ is the standard error of the mean of $PSMD_{random}$ and $PSMD_{random}$ is an array including $M(99,\ 999, \text{etc.})$ *PSMD*s which are calculated from randomization null hypothesis.

$$\text{pseudo\_p} = \frac{R+1}{M+1} \tag{15}$$

where R is the number of times a computed statistic from the random data sets is equal to or more extreme than the observed PSMD and M is the number of permutations (99, 999, etc.). For a given spatial data set, the critical value can be calculated by the following steps. We calculate the PSMD under the situation that all the dependent and independent values are rearranged randomly. When we select 0.05 as the critical p-value, the R-th ($R = 0.05 * (M+1) - 1$) highest PSMD from null hypothesis is the critical PSMD value. If the observed PSMD value is less than the critical PSMD, the association between dependent and independent variable is not significant.

3. Simulation test of PSMD

3.1. *Simulation scheme*

To compare the result of our new method with Wang *et al*. (2010), we design simulations to test the performances of our method. The simulation includes two parts: spatial data generation and spatial data association assessment. The basic idea is to create simulated variables Y and X with perfect spatial association (=1); we then randomly shuffle the variable X with known shuffling rate and use PSMD to measure the spatial association and compare that with the benchmark association, which is 1 minus the shuffling rate. To generate the variables with perfect spatial association, we first create an area with 30×30 small lattices. Then, we select 50 positions within this area (the locations of points are random, so probably are not at the centers of lattices) and assign 50 random numbers from a random distribution (e.g. Gaussian~$N(0, 10)$) to these points. To obtain a spatial surface data, we select the radial basis function to interpolate the values at the centers of all lattices using the 50 random distributed points. The set of center point values and their locations is the spatial dependent variable (Y). After we created the spatial dependent variable, we produce the independent variable (X) as follows. We generate 900 (30×30) randomly distributed (e.g. Gaussian~$N(0, 10)$) random values, rank them and assign them to the center points of lattices based on the ranking of the dependent variable. After this step, the rankings of dependent and independent variable are matched perfectly. In this case, the ranking of dependent values and their corresponding independent values are the same. The association between two data reaches maximum which is 100% or 1. After we created the maximum association spatial data set, we use the controlled parameter (shuffling rate) to decrease the maximum association to generate benchmark association for testing the measures of spatial association. We switch some independent values with each other randomly (i.e. shuffling) to wane the association between dependent and independent variables. The shuffling rate is the ratio between the counts of rearranged independent values and the total counts of independent values (so the range of shuffling rate is from 0 to 1). If the shuffling rate is 0, all the independent values are not rearranged; the association

between spatial dependent and independent variables is maximum, in other words, it is 1. Conversely, if the shuffling rate is 1, all the independent values are rearranged randomly; the association between the spatial dependent and independent variable is 0 (random).

3.2. *Estimated spatial data association comparison by simulations*

To evaluate our SPADE, we compare the PD (from Wang *et al*. 2010) with PSMD (proposed by this article) under the controlled shuffling rates. First, under some fixed shuffling rates, we compare the PD, compensated PD (PD compensated by the information loss, or CPD), PSD (PD using spatial variance, Equation (9)) and compensated PSD (PSD compensated by the information loss, or CPSD Equation (12)). We also calculate the critical value by the pseudo *p*-value method to test whether the estimated associations are significant. The aims of comparison are as follows: (1) to illustrate which estimator is stable across different discretization levels and (2) to test if the estimated values follow the general trend of different benchmark associations. The results under different shuffling rates (0.0, 0.2, 0.4, 0.6, 0.8, 1.0) are shown in Figure 4. Each point represents the mean of 100 simulations. Figure 5 showed the comparisons between the estimated associations and critical values. Following pseudo *p*-value, we repeated 99 times of the calculation of the spatial association under the full shuffling rate (1.0) and obtained 4^{th} largest value of simulation results ($p = 0.05$) as the critical value ($R = 0.05 * (M + 1) - 1 = 0.05 * (99 + 1) - 1 = 4$).

The results showed that (1) the compensated estimators (CPD and CPSD) are generally stable throughout the discretization levels; (2) the stabilities of CPD and CPSD only have subtle difference; (3) when the shuffling rate is 100%, all the estimators are not significant (shown in Figure 5); (4) when the shuffling rate is 0%, the compensated estimators (CPD and CPSD) reach 1 (shown in Figure 4) and (5) overall, our CPSD more closely follow the trend of different benchmark associations.

4. Application to US surface dissection density data

We applied the new approach to derive PSMD as described above to better capture spatial associations between dissection density and environmental factors and compared them with the results from Luo *et al*. (2016). Dissection density describes the degree of land surface dissection by erosional processes and is defined as the total length of valleys per unit area (Luo *et al*. 2016). On a continental scale, dissection density (a geomorphology concept not requiring identification of channels) is highly correlated with drainage density (a hydrology concept requiring identification of channels). Previous research projects have shown that factors controlling dissection density include climate (Melton 1957, Montgomery and Dietrich 1989, Tucker and Bras 1998), slope and relief (Schumm 1956, Strahler 1964, Oguchi 1997), lithology (Gardiner 1995, Tucker and Slingerland 1996, Xiong *et al*. 2014) and soil properties (Montgomery and Dietrich 1989, Dietrich *et al*. 1992). The understanding of which factors play dominant roles in controlling dissection density is an important theme in geomorphology and hydrology because of its scientific and practical values. The latter relates to assessing the risk of soil loss and to designing measures to reduce such loss. Unlike most previous studies, which were at local scales and lacked an analytical framework designed especially for comparing controlling factors over a regional or continental scale, Luo *et al*. (2016) utilized the Geo-detector as a general framework to

Figure 4. Comparison of PD, compensated PD (CPD), PSD and compensated PSD (CPSD), under different shuffling rates.
The benchmark association equals 1 minus the shuffling rate. The suffix _mean represents the mean of simulation results. *X*-axis is the discretization level which is from 5 to 30 and *Y*-axis is estimated spatial association between two variables.

assess the associations between dissection density and environmental factors in each physiographic region (Fenneman 1928, Figure 6) and tested the hypothesis that the dominant controlling factor, or the interactions between factors, vary from region to region due to differences in each region's local characteristics and geologic history. The dissection density data were derived using geomorphons method (Jasiewicz and Stepinski 2013) and aggregated to the basic units of watersheds based on the 12-digit hydrologic unit boundaries (Federal standards and procedures for the national Watershed Boundary Dataset (WBD) 2013). The 13 controlling factors are shown in Table 1, which include 3 main groups: geology, climate and terrain, and are aggregated to the same basic units of watersheds. The terrain factors are derived from Digital Elevation Model (DEM) data. The geology and climate factors are downloaded from open source database: the precipitation factor is from the website of PRISM climate Group (http://www.prism.oregonstate.edu/), the glaciation and

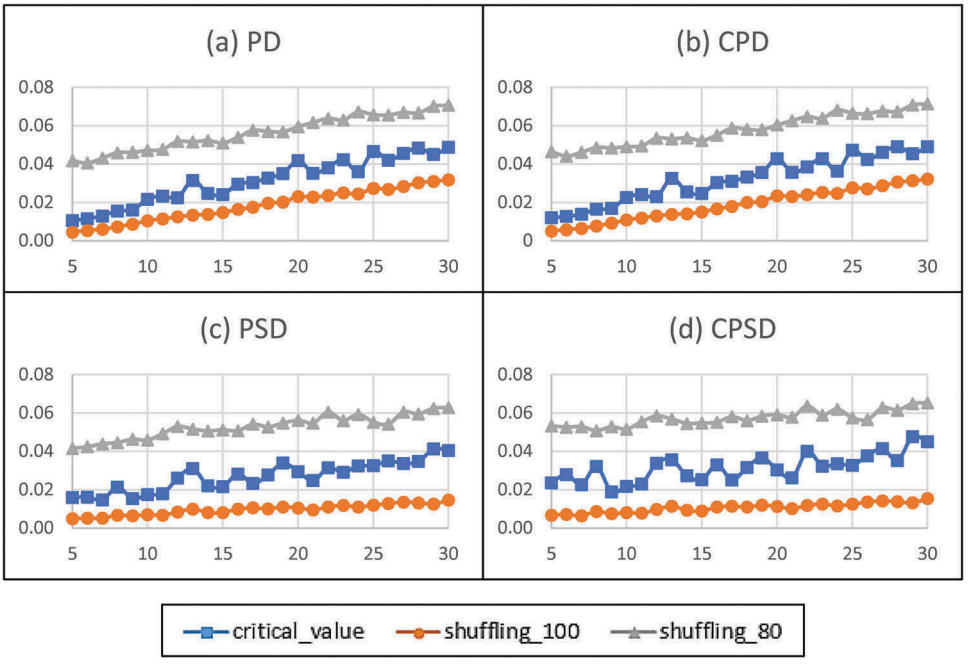

Figure 5. Comparison of estimated spatial association and critical value under different shuffling rates. *X*-axis is the discretization level and *Y*-axis is estimated spatial association between two variables.

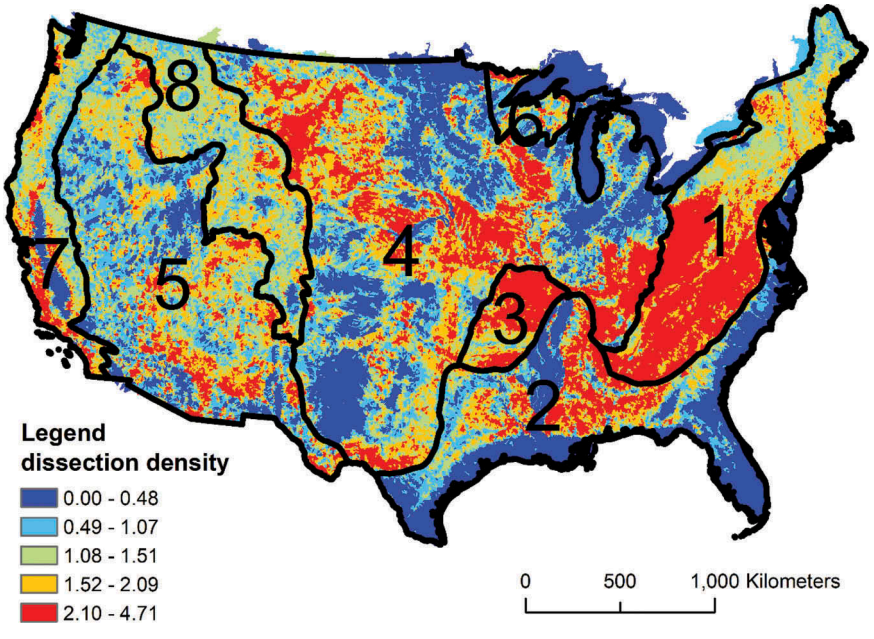

Figure 6. Dissection density in the US.

Table 1. Controlling factors.

Category	Factor	Factor code	Resolution
Geology or soil property	Glaciation	Glaci	Resampled shapefile data to 4 km
	Lithology	Litho	1 km resolution, resampled to 4 km
	Permeability	logK	Resampled shapefile data to 4 km
	Porosity	poro	Resampled shapefile data to 4 km
Climate	Precipitation	precip	4 km resolution
Topography	Elevation	elev	ETOPO1 DEM resampled to 4 km resolution
	Aspect	asp	Derived from DEM, 4 km resolution
	Slope	slp	Derived from DEM, 4 km resolution
	difference in elevation (relief)	difelev	Derived from DEM, 4 km resolution
	distance to erosional base	distb	Derived from DEM, 4 km resolution
	Elevation to erosional base	elevb	Derived from DEM, 4 km resolution
	Planar Curvature	planc	Derived from DEM, 4 km resolution
	Tangential Curvature	tanc	Derived from DEM, 4 km resolution

the lithology factors are from United States Geological Survey (USGS), the permeability factor is from Gleeson *et al*. (2014) (http://crustalpermeability.weebly.com/glhymps.html) and the porosity factor is from the STATSGO2 database.

To illustrate the PSMD, we use Region 3 (Interior Highlands), Region 6 (Laurentian Upland), Region 7 (Pacific Mountain System) and Region 8 (Rocky Mountain System) as examples. In the calculation, the discretization levels in the multilevel discretization are from 4 to 20. Table 2 shows the PSMD values and their rankings in those regions in comparison with those of the original PD. The spatial association for most continuous variables increased. The ranking order of the spatial associations between dependent and independent variables also changed slightly. The ranking of discrete variables, such as the *Litho*, decreased. However, the dominant factors (high ranked factors in PD) still retain their statuses. To illustrate the physical meaning of the PSMD, we map the three controlling factors, which are high, medium and low ranking, to compare the spatial distribution of dissection density (Figure 7). It is clear that the controlling factors with higher PSMD have a more similar spatial distribution with the dependent variable.

The results (Table 2 and Figure 7) show that (1) the associations between dependent and independent continuous variables are underestimated by the original PD in most cases, because the original Geo-detector omits the spatial dependence resulting from dynamical geographical processes and the information loss due to discretization; (2) The information loss from discretizing different continuous independent variables is different, because their distributions are different. Thus, the ranking of the continuous independent variables also changed.

5. Discussion

5.1. *Selection of model parameters*

The major parameters of SPADE include the range of discretization level and the weighting method. The range of discretization level is a robust parameter because the information loss is compensated. From Figure 4, the compensated PSDs are stable across different discretization levels; most of the differences of the compensated PSDs between different discretization levels are smaller than 0.1. Here, we apply the different discretization level ranges on Region 3 as an example to illustrate the effect of discretization level ranges. The results are shown in

Table 2. PSMD results in Regions 3, 6, 7 and 8.

	Region 3 (Interior Highlands) N = 1955				Region 6 (Laurentian Upland) N = 1215			
Ranking	Variable	*PSMD*	Variable	*PD*	Variable	*PSMD*	Variable	*PD*
1	*logk*	0.389	*logk*	0.309	*distb*	0.214	*litho*	0.231
2	*poro*	0.363	*precip*	0.262	*litho*	0.202	*tanc*	0.139
3	*precip*	0.348	*poro*	0.219	*elev*	0.167	*asp*	0.130
4	*distb*	0.336	*litho*	0.185	*poro*	0.164	*planc*	0.128
5	*difelev*	0.280	*elevb*	0.175	*logk*	0.148	*logk*	0.128
6	*elevb*	0.258	*distb*	0.163	*elevb*	0.138	*poro*	0.119
7	*elev*	0.235	*slp*	0.154	*precip*	0.136	*slp*	0.110
8	*litho*	0.235	*difelev*	0.127	*difelev*	0.131	*elevb*	0.109
9	*slp*	0.166	*elev*	0.122	*asp*	0.125	*elev*	0.105
10	*planc*	0.123	*planc*	0.107	*slp*	0.104	*difelev*	0.088
11	*tanc*	0.118	*tanc*	0.094	*planc*	0.100	*distb*	0.081
12	*asp*	0.042	*glaci*	0.057	*tanc*	0.099	*precip*	0.075
13	*glaci*^	0.001	*asp*	0.049	*glaci*	0.017	*glaci*	0.071
	Region 7 (Pacific Mountain System) N = 5134				Region 8 (Rocky Mountain System) N = 6598			
Ranking	Variable	*PSMD*	Variable	*PD*	variable	*PSMD*	Variable	*PD*
1	*elev*	0.469	*elev*	0.280	*precip*	0.150	*litho*	0.104
2	*difelev*	0.431	*litho*	0.243	*difelev*	0.146	*slp*	0.086
3	*logk*	0.386	*difelev*	0.228	*slp*	0.144	*planc*	0.083
4	*slp*	0.383	*logk*	0.215	*elev*	0.131	*tanc*	0.079
5	*tanc*	0.326	*slp*	0.189	*planc*	0.124	*precip*	0.065
6	*planc*	0.323	*tanc*	0.170	*distb*	0.120	*logk*	0.057
7	*litho*	0.299	*planc*	0.153	*tanc*	0.116	*poro*	0.046
8	*poro*	0.266	*poro*	0.121	*litho*	0.104	*difelev*	0.043
9	*distb*	0.203	*precip*	0.098	*logk*	0.093	*elev*	0.042
10	*elevb*	0.194	*elevb*	0.085	*poro*	0.080	*elevb*	0.040
11	*precip*	0.174	*distb*	0.064	*elevb*	0.072	*distb*	0.035
12	*asp*	0.006	*asp*	0.062	*asp*	0.019	*asp*	0.033
13	*glaci*^	0.000	*glaci*	0.048	*glaci*^	0.000	*glaci*	0.031

N is the number of watersheds in the region.

All the PSMD estimated associations are significant, except that the values with the symbol ^, representing that the values are not significant. The PSMD is the mean of PD using spatial variance under different discretization levels. The original PDs are from Luo *et al*. (2016).

Table 3. We apply three groups of discretization level ranges (5–10, 10–15 and 15–20) to calculate PSMDs. The results showed that the rankings of variables are the same and that most of differences between different range are smaller than 0.05. The weighting method includes contiguity-based weights or distance-based weights. The selection of method depends on the nature of the research. In this research, the distance-based weights are more suitable, because the scale of dynamic geography processes (erosion, sedimentation and tectonic processes) is much larger than the statistics unit (watershed), which means that noncontiguous watersheds can be influenced by the same process and the intensity of influence decays through the distance (Taylor and Openshaw 1975). The distance decay can be measured by different functions (summarized by Martínez and Viegas 2013) quantitatively. In this article, we choose the power law because the gravity model, a form of power law, has been used in different geography research fields, both in human geography and in physical geography. The general form of weight based on the power law of distance decay is shown in Equation (16):

$$w = \frac{1}{d^{\beta}} \tag{16}$$

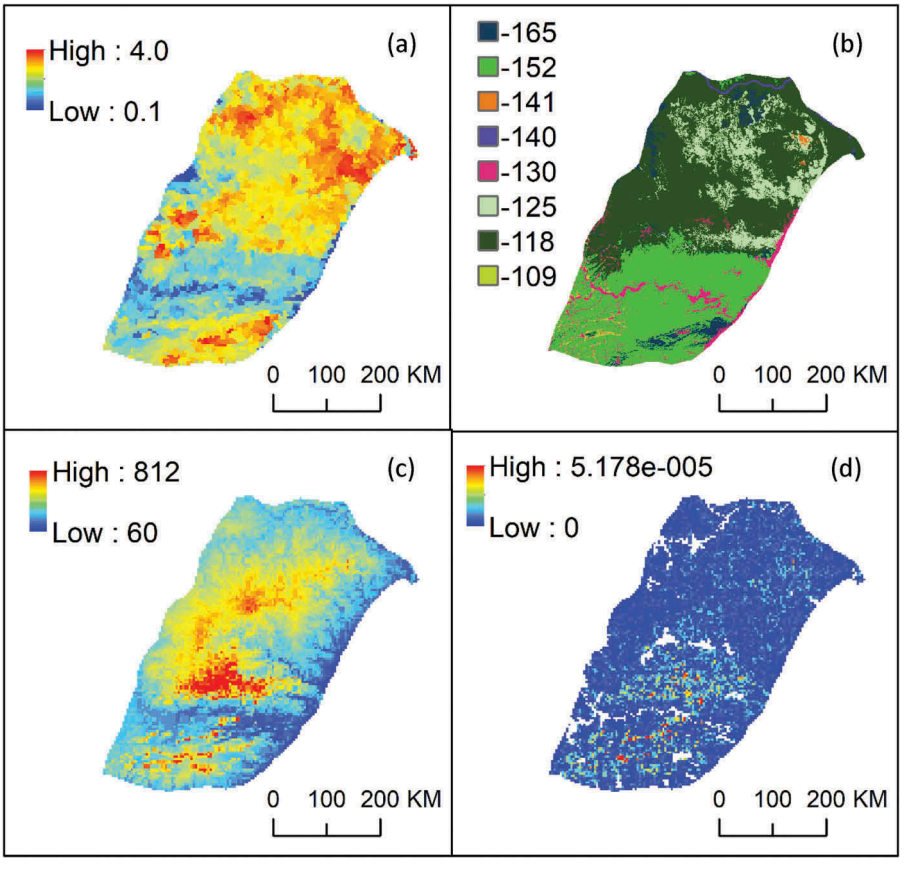

Figure 7. Dissection density and its controlling factors in Region 3. (a) Dissection density (dependent variable); (b) *logk* (highest PSMD); (c) *elevation* (7th controlling factor); (d) *tanc* (11th controlling factor).

Table 3. PSMD of continuous variables under different discretization levels in Region 3.

5_to_10		10_to_15		15_to_20	
logk	0.396	*logk*	0.354	*logk*	0.422
poro	0.367	*poro*	0.335	*poro*	0.389
precip	0.362	*precip*	0.324	*precip*	0.364
distb	0.341	*distb*	0.319	*distb*	0.350
difelev	0.288	*difelev*	0.262	*difelev*	0.295
elevb	0.259	*elevb*	0.251	*elevb*	0.265
elev	0.245	*elev*	0.206	*elev*	0.258
slp	0.168	*slp*	0.158	*slp*	0.174
planc	0.125	*planc*	0.118	*planc*	0.126
tanc	0.121	*tanc*	0.109	*tanc*	0.125

where w is the weight between two locations; d is the Euclidean distance between two locations and β is the exponent of distance, representing the decay rate. The β, with a typical range from 0 to 3 (Chen 2015), is the intensity of influence between neighbors. A larger β in Equation (16) means that the closer values have higher weights in the calculation. The selection of β is usually research dependent and often determined empirically; or can be

Table 4. PSMDs under different distance decay methods.

Region 3				Region 6			
$\beta = 2$		$\beta = 1$		$\beta = 2$		$\beta = 1$	
distb	0.406	*logk*	0.389	*distb*	0.406	*distb*	0.214
logk	0.402	*poro*	0.363	*logk*	0.402	*litho*	0.202
difelev	0.393	*precip*	0.348	*difelev*	0.393	*elev*	0.167
poro	0.391	*distb*	0.336	*poro*	0.391	*poro*	0.164
elevb	0.355	*difelev*	0.280	*elevb*	0.355	*logk*	0.148
precip	0.338	*elevb*	0.258	*precip*	0.338	*elevb*	0.138
elev	0.322	*elev*	0.235	*elev*	0.322	*precip*	0.136
slp	0.225	*litho*	0.235	*slp*	0.225	*difelev*	0.131
tanc	0.187	*slp*	0.166	*tanc*	0.187	*asp*	0.125
planc	0.147	*planc*	0.123	*asp*	0.230	*slp*	0.104
litho	0.133	*tanc*	0.118	*litho*	0.150	*planc*	0.100
asp	0.106	*asp*	0.042	*planc*	0.147	*tanc*	0.099
glaci	0.053	*glaci^*	0.001	*glaci*	0.128	*glaci*	0.017

estimated by the product of the Zipf's exponent of size distributions and the fractal dimension of spatial distributions (Chen and Huang 2018). In producing Table 2, we used inverse distance decay, that is, $\beta = 1$. Here, we test the robustness of β by applying two β values (1, 2) on the same region (Table 4). The dominant factors of Regions 3 and 6 changed subtly, and most of ranking only changed one or two positions. The results showed that the ranking of variables may change with different weighting methods, but they are not very sensitive to different weighting methods.

5.2. *Assumption of the probability distributions*

Normally, the assumption of association estimator is that the PDF of variables are normal distribution; however, in the geography research field, the symmetrical distribution and asymmetrical distribution are both very common. For this reason, we create random variables which follow different PDF to test the influence of different PDFs on SPADE. First, we investigate the PSMD under different distributions. We select the normal distribution to represent the symmetrical distribution and select Pareto distribution to represent the asymmetrical distribution. We create four groups of data sets whose dependent and independent variables are normal distribution ($N(0, 10)$) or Pareto distribution ($P(3)$). Then, we analyze the influence of variables' distribution on the association estimation. Based on simulation test described in Section 3, we average the compensated PSDs from 5 to 30 discretization levels as the result of multilevel associations to compare the association to the shuffling rates. The results are compared with the spatial association benchmark line ($y = -x + 1,\ x \in [0, 1]$) (see Figure 8). The benchmark values of spatial association are from 0 to1, whose interval is 0.2. The results showed that the compensated estimators (PMD_Q and PSMD_Q) can cover the full range of benchmarks and are less sensitive to the changes of variables' distributions than PMD_q and PSMD_q.

5.3. *Selection of significant test*

As mentioned above, the PDF of PSMD is unknowable. The current spatial analysis software (such as ArcGIS, PySAL and GeoDa) (Anselin *et al.* 2006, Rey and Anselin 2010) test the

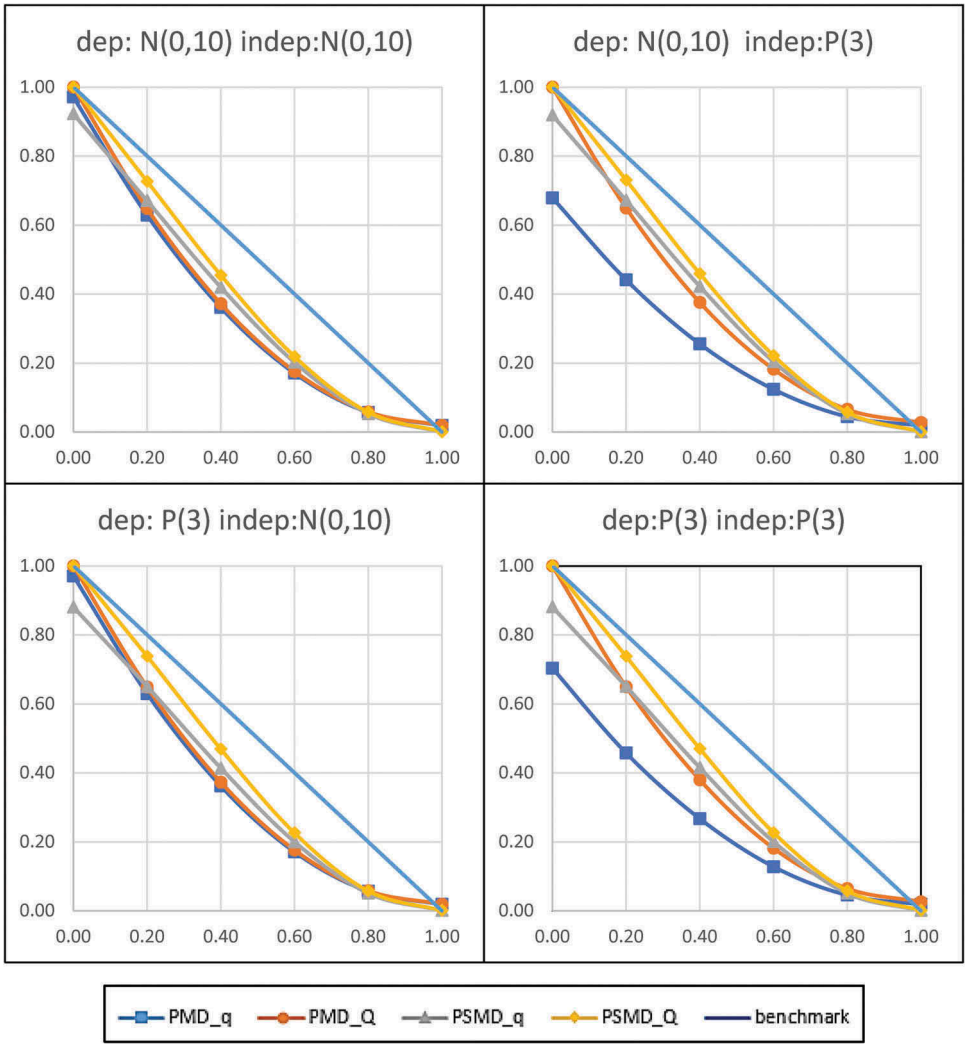

Figure 8. Comparison spatial estimators form different distributed variables.

X-axis is the shuffling rate and *Y*-axis is estimated spatial association between two variables. The suffix _q represents the estimator without compensation and the suffix _Q represents the compensated estimator.

significance of spatial estimator by *Z*-test or pseudo *p*-value approach. We use the simulations to discuss the difference between the *Z*-test and pseudo *p*-value approach and explain why we choose *p*-value approach. We repeat 99 times of the calculation of PSMDs under the 100% shuffling rate. In each simulation, first, we create four pairs of possible combinations from the two PDFs (normal distribution ($N(0, 10)$) or Pareto distribution ($P(3)$)) in one time of simulation; next, shuffle one variable of every pair with 100% shuffling rate; then, calculate the PSMDs. Based on the four groups of PSMDs, the QQ-plot (Figure 9) shows the similarity between null hypothesis distribution and normal distribution. At the right end of each line, the actual values are a little bit greater than the expected value. The shape represents that the tail of distribution from null hypothesis is heavier than the normal distribution. In this situation,

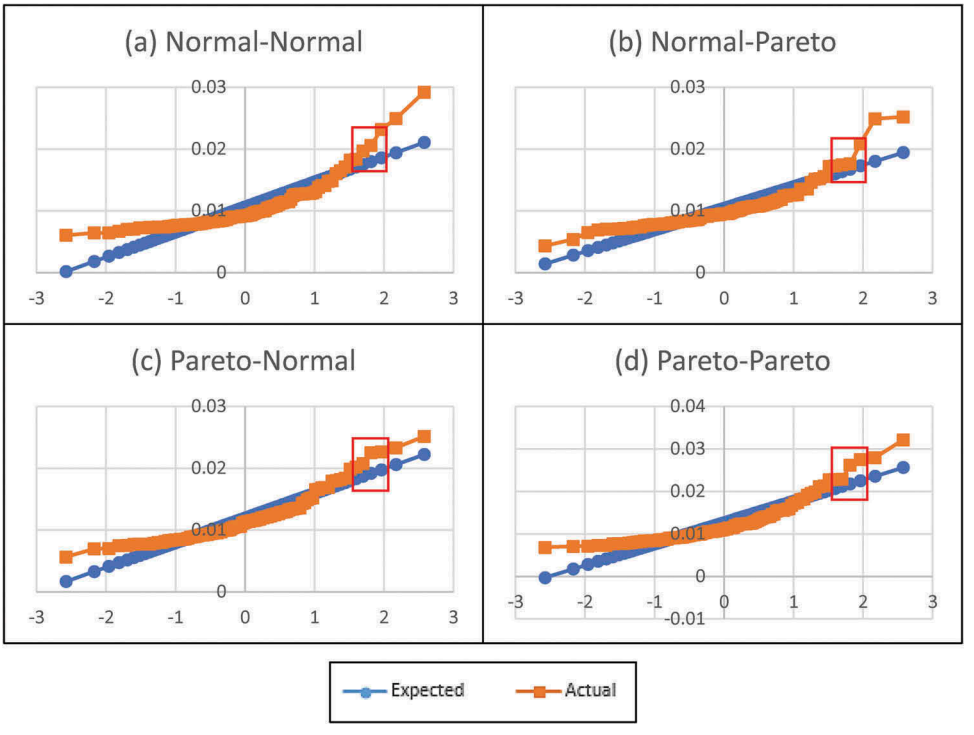

Figure 9. QQ-plot of null hypothesis under varied distributions.
The red rectangle shows the critical values (fourth point).

the pseudo *p*-value approach has a higher critical value than the normal distribution significance test for a given spatial estimator. For this reason, we select pseudo *p*-value approach because it is more conservative.

6. Conclusion

The purpose of this article is to improve the measure of spatial association between dependent variable and potential controlling factors within the Geo-detector framework by explicitly utilizing the spatial information and minimizing the influence of discretization levels. We solved the problem that the original Geo-detector lacks measure of spatial dependence by utilizing spatial variance, which is derived from general spatial weighted cross-product, to replace the traditional variance. We also addressed the problem that the Geo-detector measured association (PD) was influenced by the number of levels into which continuous variables are discretized by compensating the information loss due to discretization. The information kept was measured by PSD value between the continuous variable and its discretized counterpart. Using simulated data with known benchmark association, we demonstrated that (1) the compensated association can cover the whole range of benchmark association; (2) the compensated association is stable across different discretization levels and (3) the significance of null hypothesis (association between variables not significant) can be tested by pseudo *p*-value approach, whose result is more conservative than *Z*-test which assumes that the distribution from null hypothesis is a Gaussian distribution. When applying

the new method to measure the spatial association between dissection density and controlling factors in United States, the ranking of PSMD values of some variables changed, but most dominant factors still remain the same. So the general conclusion of Luo *et al*. (2016) that the dominant factor for each physiographical region reflects that region's geological history and character still holds. The exception happened in Regions 6 and 8, both with low original PD value. The dominant factors of the two regions were *litho*, which means that a higher category number can cause a higher estimated spatial association and that the previous discretization level in Luo *et al*. (2016) underestimated the associations between dependent and independent variables. Through our simulated data with known benchmark association and case study of dissection density in United States, we have demonstrated that the PSMD value is a stable and more accurate measure than original PD because PSMD explicitly considers spatial variation and minimizes the influence of discretization levels. Thus, in practice, we have more confidence in using PSMD to measure the association between spatial data and do not need to discretize the continuous variables into a large number of levels.

Acknowledgments

We thank the anonymous reviewers for their insightful comments and helpful suggestions.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was partially supported by the National Aeronautics and Space Administration [NNX13AK65G].

ORCID

Xuezhi Cang http://orcid.org/0000-0002-8928-2096

References

Anselin, L., 1995. Local indicators of spatial association—LISA. *Geographical Analysis*, 27 (2), 93–115. doi:10.1111/j.1538-4632.1995.tb00338.x

Anselin, L., Syabri, I., and Kho, Y., 2006. GeoDa: an introduction to spatial data analysis. *Geographical Analysis*, 38 (1), 5–22. doi:10.1111/j.0016-7363.2005.00671.x

Bachmaier, M. and Backes, M., 2008. Variogram or semivariogram? Understanding the variances in a variogram. *Precision Agriculture*, 9 (3), 173–175. doi:10.1007/s11119-008-9056-2

Baez, J.C., Fritz, T., and Leinster, T., 2011. A characterization of entropy in terms of information loss. *Entropy*, 13 (11), 1945–1957. doi:10.3390/e13111945

Brewer, C.A. and Pickle, L., 2002. Evaluation of methods for classifying epidemiological data on choropleth maps in series. *Annals of the Association of American Geographers*, 92 (4), 662–681. doi:10.1111/1467-8306.00310

Brunsdon, C., Fotheringham, A.S., and Charlton, M.E., 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28 (4), 281–298. doi:10.1111/j.1538-4632.1996.tb00936.x

Cao, F., Ge, Y., and Wang, J., 2013. Optimal discretization for geographical detectors-based risk assessment. *GIScience & Remote Sensing*, 50 (1). doi:10.1080/15481603.2013.778562

Cao, F., Ge, Y., and Wang, J., 2014. Spatial data discretization methods for geocomputation. *International Journal of Applied Earth Observation and Geoinformation*, 26, 432–440. doi:10.1016/j.jag.2013.09.005

Cauvin, C., Escobar, F., and Serradj, A., 2010. *Cartography and the impact of the quantitative revolution*. London: ISTE.

Chen, Y., 2015. The distance-decay function of geographical gravity model: power law or exponential law? *Chaos Solitons & Fractals*, 77, 174–189. doi:10.1016/j.chaos.2015.05.022

Chen, Y. and Huang, L., 2018. A scaling approach to evaluating the distance exponent of the urban gravity model. *Chaos, Solitons & Fractals*, 109, 303–313. doi:10.1016/j.chaos.2018.02.037

Dietrich, W.E., *et al*., 1992. Erosion thresholds and land surface morphology. *Geology*, 20 (8), 675–679. (1992)020<0675:ETALSM>2.3.CO;2. doi:10.1130/0091-7613

Du, Z., *et al*., 2016. Geographical detector-based identification of the impact of major determinants on aeolian desertification risk. *PloS One*, 11 (3), e0151331. doi:10.1371/journal.pone.0151331

Fenneman, N.M., 1928. Physiographic divisions of the United States. *Annals of the Association of American Geographers*, 18, 261–353. doi:10.2307/2560726

Gardiner, V., 1995. Channel networks: progress in the study of spatial and temporal variations of drainage density. *In*: A. Gurnell, and G.E. Petts, eds. *Changing river channels*. New York: Wiley, 65–85.

Getis, A., 1991. Spatial interaction and spatial autocorrelation: a cross-product approach. *Environment and Planning A*, 23 (9), 1269–1277. doi:10.1068/a231269

Gleeson, T., *et al*., 2014. A glimpse beneath earth's surface: GLobal HYdrogeology MaPS (GLHYMPS) of permeability and porosity. *Geophysical Research Letters*, 41 (11), 3891–3898. doi:10.1002/2014GL059856

Harvey, F., 2015. *A primer of GIS: fundamental geographic and cartographic concepts*. New York: Guilford Publications.

Jasiewicz, J. and Stepinski, T.F., 2013. Geomorphons - a pattern recognition approach to classification and mapping of landforms. *Geomorphology*, 182, 147–156. doi:10.1016/j.geomorph.2012.11.005

Kreit, E., *et al*., 2013. Biological versus electronic adaptive coloration: how can one inform the other? *Journal of the Royal Society Interface*, 10 (78), 20120601. doi:10.1098/rsif.2012.0601

Kwan, M., 2004. GIS methods in time-geographic research: geocomputation and geovisualization of human activity patterns. *Geografiska Annaler, Series B: Human Geography*, 86 (4), 267–280. doi:10.1111/j.0435-3684.2004.00167.x

Liao, Y., *et al*., 2010. Risk assessment of human neural tube defects using a Bayesian belief network. *Stochastic Environmental Research and Risk Assessment*, 24 (1), 93–100. doi:10.1007/s00477-009-0303-5

Luo, W., *et al*., 2016. Spatial association between dissection density and environmental factors over the entire conterminous United States. *Geophysical Research Letters*, 43 (2), 692–700. doi:10.1002/2015GL066941

MacKay, D.J., 2003. *Information theory, inference and learning algorithms*. Cambridge, England: Cambridge University Press, 43–44.

Martínez, L.M. and Viegas, J.M., 2013. A new approach to modelling distance-decay functions for accessibility assessment in transport studies. *Journal of Transport Geography*, 26, 87–96. doi:10.1016/j.jtrangeo.2012.08.018

Melton, M.A., 1957. *An analysis of the relations among elements of climate, surface properties, and geomorphology (No. CU-TR-11)*. New York: Columbia University.

Montgomery, D.R. and Dietrich, W.E., 1989. Source areas, drainage density, and channel initiation. *Water Resources Research*, 25 (8), 1907–1918. doi:10.1029/WR025i008p01907

Oguchi, T., 1997. Drainage density and relative relief in humid steep mountains with frequent slope failure. *Earth Surface Processes and Landforms*, 22 (2), 107–120. (199702)22:2<107::AID-ESP680>3.0.CO;2-U. doi:10.1002/(SICI)1096-9837

Quinlan, J.R., 1986. Induction of decision trees. *Machine Learning*, 1 (1), 81–106. doi:10.1007/BF00116251

Ren, Y., *et al*., 2014. Geographical modeling of spatial interaction between human activity and forest connectivity in an urban landscape of southeast China. *Landscape Ecology*, 29 (10), 1741–1758. doi:10.1007/s10980-014-0094-z

Rey, S.J. and Anselin, L., 2010. PySAL: a Python library of spatial analytical methods. *In*: M.M. Fischer, and G. Arthur, eds. *Handbook of applied spatial analysis*. Berlin, Heidelberg: Springer, 175–193. doi: 10.1007/978-3-642-03647-7_11

Schumm, S.A., 1956. Evolution of drainage systems and slopes in badlands at Perth Amboy, New Jersey. *Geological Society of America Bulletin*, 67 (5), 597–646. (1956)67[597:EODSAS]2.0.CO;2. doi:10.1130/0016-7606

Strahler, A.N., 1964. Quantitative geomorphology of drainage basins and channel networks. *In*: V.T. Chow, ed. *Handbook of applied hydrology*. New York: McGraw-Hill, 439–476.

Taylor, P.J. and Openshaw, S., 1975. Distance decay in spatial interactions. *In Concepts and techniques in modern geography*.

Tucker, G.E. and Bras, R.L., 1998. Hillslope processes, drainage density, and landscape morphology. *Water Resources Research*, 34 (10), 2751–2764. doi:10.1029/98WR01474

Tucker, G.E. and Slingerland, R., 1996. Predicting sediment flux from fold and thrust belts. *Basin Research*, 8 (3), 329–349. doi:10.1046/j.1365-2117.1996.00238.x

US Geological Survey and US Department of Agriculture, Natural Resources Conservation Service, 2013. *Federal standards and procedures for the national Watershed Boundary Dataset (WBD)*. Reston, VA: U.S. Geological Survey, Techniques and methods 11–A3.

Wang, J. and Xu, C., 2017. Geodetector: principle and prospective. *Acta Geographica Sinica*, 72 (1), 116–134. doi:10.11821/dlxb201701010

Wang, J., Zhang, T., and Fu, B., 2016. A measure of spatial stratified heterogeneity. *Ecological Indicators*, 67, 250–256. doi:10.1016/j.ecolind.2016.02.052

Wang, J.F., *et al*., 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. *International Journal of Geographical Information Science*, 24 (1), 107–127. doi:10.1080/13658810802443457

Xiong, L.Y., *et al*., 2014. Modeling the evolution of loess-covered landforms in the Loess Plateau of China using a DEM of underground bedrock surface. *Geomorphology*, 209, 18–26. doi:10.1016/j.geomorph.2013.12.009

Zhu, H., *et al*., 2015. A spatial-temporal analysis of urban recreational business districts: a case study in Beijing, China. *Journal of Geographical Sciences*, 25 (12), 1521–1536. doi:10.1016/j.ecolind.2016.02.052