GIScience & Remote Sensing
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tgrs20

# Optimal discretization for geographical detectors-based risk assessment

Feng Cao [a] , Yong Ge [a] & Jin-Feng Wang [a]
[a] State Key Laboratory of Resources and Environmental Information System , Institute of Geographic Sciences and Nature Resources Research, Chinese Academy of Sciences , A11Datun Road, Beijing , 100101 , China
Published online: 11 Apr 2013.

To cite this article: Feng Cao , Yong Ge & Jin-Feng Wang (2013): Optimal discretization for geographical detectors-based risk assessment, GIScience & Remote Sensing, 50:1, 78-92

To link to this article: http://dx.doi.org/10.1080/15481603.2013.778562

*GIScience & Remote Sensing*, 2013
Vol. 50, No. 1, 78–92, http://dx.doi.org/10.1080/15481603.2013.778562

Taylor & Francis
Taylor & Francis Group



# Optimal discretization for geographical detectors-based risk assessment

Feng Cao, Yong Ge* and Jin-Feng Wang

*State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Nature Resources Research, Chinese Academy of Sciences, A11Datun Road, Beijing 100101, China*

The geographical detectors model is a new spatial analysis method for the assessment of health risks. It is adapted to discrete risk factors. Meanwhile, the geographical detectors model also effectively analyzes the continuous risk factors by discretizing the continuous data into discrete data. The biggest difficulty is in deciding how to discretize continuous risk factors using the most appropriate discretization method. In this paper, we will discuss the selection of an optimal discretization method for geographical detectors-based risk assessment, and exemplify the process using neural tube defects (NTD) from the Heshun County, Shanxi Province, China.

**Keywords:** geographical detectors; discretization; risk assessment; NTD

## Introduction

Spatial analysis is widely used in public health research because of its powerful data analysis and visualization abilities. The earliest application of spatial analysis was in 1854, when John Snow mapped the addresses of cholera cases in London, England and discovered the source of the cholera outbreak. With its increasing development, there is an increase in the use of spatial analysis methods in public health research. Exploratory spatial analysis methods such as Getis G, LISA, and spatial scan statistics can identify hot-spot areas of a disease by comparing the disease rates between regions and their surrounding areas (Getis and Ord 1992; Anselin 1995; Kulldorff 1997; Li et al. 2011). This is very important for the generation of early warning and prevention strategies. Spatial clustering methods can also identify disease hot-spot areas. For example, Wang et al. (2006) used the nearest-neighbor hierarchical clustering method to analyze the spatial clustering of SARS in Beijing, 2003, and found strong associations between the ring roads, the light railway, and SARS. This analysis provided information that was crucial to SARS prevention. Spatial regression analysis methods can reveal potential disease risk factors by analyzing the spatial variability between a health outcome and its associated social, economic, and environmental risk factors (Best, Ickstadt, and Wolpert 2000; Nakaya et al. 2005). Spatial data mining methods such as neural networks, support vector machines, rough sets, and Bayesian networks mine the potential associations between health and its risk factors. They can also provide effective health risk assessments and predictions (Lisboa and Taktak 2006; Magnin 2009; Bai et al. 2010; Liao et al. 2010). The geographical detectors model is a new spatial analysis method used to assess health

*Corresponding author. Email: gey@lreis.ac.cn

and environmental risks. The model can resolve the following four questions: (1) What is the geographical domain of health risk? (2) What are the environmental parameters responsible for the risk? (3) What is the relative importance of each risk factor? (4) Do the risk factors operate independently or are they interconnected? The model has been applied to the risk assessment of neural tube defects (NTD) and under-five earthquake mortality (Wang et al. 2010a; Hu et al. 2011). Wang and Hu (2012) developed a software called GeogDetector to perform the tasks of geographical detectors.

Most spatial analysis methods used in public health primarily focus on continuous risk factors. However, many risk factors are discrete variables, such as land cover, watershed, and soil type. The geographical detectors model is a method that can analyze the relationship between discrete risk factors and health outcomes based on a spatial variance analysis. Moreover, it can still effectively model the continuous risk factors. In the geographical detectors model, continuous risk factors are transformed into discrete intervals before its relationship with health outcomes is analyzed. User-defined discretization is used in the geographical detectors model and the GeogDetector software for continuous risk factors (Wang et al. 2010a; Hu et al. 2011). The user-defined discretization needs to assign the number and values of cut points which is subject to weaknesses such as randomness and subjectivity. In order to overcome these drawbacks, we discuss the selection of an optimal unsupervised discretization method for continuous risk factors. The number of cut points needs to be assigned for unsupervised discretization methods.

The power of determinant (PD) value and the interactive PD value are two novel indicators used to assess the relationship between health outcomes and their environmental risk factors in the geographical detectors model (Wang et al. 2010a; Hu et al. 2011). In this paper, we describe the process of using both the PD and the interactive PD values to select the optimal discretization method. We use the NTD from the Heshun County, Shanxi Province, China to exemplify this process in detail. Five unsupervised discretization methods, namely Equal Interval (EI), Natural Breaks (NB), Quantile (QU), Geometrical Interval (GI), and Standard Deviation (SD), are used to discretize elevation and GDP (continuous risk factors of NTD). The calculated PD and interactive PD values are used to select the optimal method for discretizing elevation and GDP. The relationships between elevation, GDP, and NTD may be more meaningful when the risk factor variables are discretized.

## Methods

### *Geographical detectors*

The geographical detectors model proposed by Wang et al. (2010a) is based on spatial variance analysis. It can be used for assessing the relationship between health outcomes and their environmental risk factors. The geographical detectors model consists of four parts, namely risk detector, factor detector, ecological detector, and interactive detector. For more details about the geographical detectors model please see Wang et al. (2010a). Briefly:

(1) The risk detector compares the differences in average disease rates between sub-regions generated by a risk factor. It uses *t*-tests to identify whether the average disease rates among different sub-regions are significantly different. Greater differences mean greater danger to the health of people within the sub-region.
(2) The factor detector uses the PD to assess the impact of risk factors on the spatial pattern of a disease. Higher PD means the risk factor has a stronger contribution to

the occurrence of the disease. It uses *F*-tests to compare whether the accumulated variance of each sub-region is significantly different from the variance of the entire study region.

(3) The ecological detector assesses whether the impacts of two risk factors on the spatial distribution of a disease are significantly different. It also uses *F*-tests to compare the variance calculated in a sub-region attributed to one risk factor with the variance attributed to another risk factor.

(4) The interaction detector consists of seven parts: Enhance, Enhance-bi, Enhance-nonlinear, Weaken, Weaken-uni, Weaken-nonlinear, and Independent. It compares the combined contribution of two individual risk factors to a disease, as well as their independent contributions. By doing so, it assesses whether the two risk factors weaken or enhance each another, or whether they independently influence the development of the disease.



### *Discretization*

There are four levels of data measurement: nominal, ordinal, interval, and ratio. Nominal and ordinal data are discrete, while interval and ratio data are continuous (Haining 2003). The aim of discretization is to transform the continuous data into discrete data. Compared with continuous data, discrete data are easier to understand, use, and explain and are closer to a knowledge-level representation (Dougherty, Kohavi, and Sahami 1995; Liu et al. 2002).

Data discretization is the process whereby continuous data are divided into a number of intervals with selected cut points, where each interval is mapped to a qualitative symbol. A cut point is a value from the adjacent continuous data that divides it into two intervals. In actual applications, researchers always discretize continuous data with user-defined discretization and select the cut points according to their experience (Liao et al. 2010; Wang et al. 2010a). The user-defined discretization is subject to weaknesses such as randomness and subjectivity. In order to overcome those drawbacks, some discretization methods are generated based on the statistical characteristics of the data (Kerber 1992; Dougherty, Kohavi, and Sahami 1995; Kurgan and Cios 2004; Tsai, Lee, and Yang 2008; Ge, Cao, and Duan 2011; Fischer and Wang 2011). Many researchers have classified the discretization methods with different taxonomies. Supervised and unsupervised are two commonly used method categories (Liu et al. 2002; Yang, Webb, and Wu 2005). Supervised discretization methods relate class information to the selection of cut points. Appropriate cut points are selected so that the data instances have the same class label, and the labels are different across consecutive intervals. Unsupervised discretization methods do not consider class information during the discretization process. Where no class information is available, unsupervised discretization is the only choice for researchers. Unsupervised discretization methods are very simple, and directly divide continuous data with user-defined parameters. The following is a description of the most commonly used unsupervised discretization methods (Jenks 1967; Fischer and Wang 2011):

(1) The equal interval (EI) method equally divides the entire range of data values into specified intervals without taking into account the number of data values in each interval. The choice of cut points only depends on the data range, ignoring the

distribution of the data. The method is very simple and performs better when the data has a normal distribution.

(2) The quantile (QU) method classifies data into a specified number of intervals with an equal number of units in each interval. The algorithm may lead to objects with the same values being assigned to different discrete intervals. In addition, the values within the same discrete interval may be very different. This method is suitable for data with a linear distribution.

(3) The natural breaks (NB) method is designed to determine the best arrangement of values into different intervals. This is done by seeking to minimize each interval's average deviation from the interval mean, while maximizing each interval's deviation from the means of the other groups. In other words, the method seeks to reduce the variance within intervals and maximize the variance between intervals.

(4) The geometrical interval (GI) method creates geometrical intervals by minimizing the square sum of the elements per interval. This ensures that each interval has approximately the same number of values and that the change between intervals is consistent. This method was designed to work on data that are heavily skewed by a preponderance of duplicate values. The specific benefit of the geometrical interval method is that it works reasonably well on data that are not normally distributed.

(5) The standard deviation (SD) method calculates mean values and standard deviations. Cut points are then created using these values. The method shows how much a feature's attribute value varies from the mean. A large standard deviation indicates that the data points are far from the mean and a small standard deviation indicates that they are clustered closely around the mean.

### *Flowchart of discretization method assessment*

The transformation of continuous into discrete risk factors using appropriate unsupervised discretization methods is an important and difficult issue. However, there are no clear ways to assess the effectiveness of unsupervised discretization methods. Here, the PD and interactive PD values calculated by the geographical detectors model are used to assess the effectiveness of different discretization methods. The PD value is an indicator of the influence of risk factors on the spatial pattern of a disease. The interactive PD value is an indicator to assess whether two risk factors when taken together weaken or enhance each other, or if they are independent in developing the disease. The higher PD and interactive PD values can better reflect the relationship between health outcomes and their environmental risk factors (Wang et al. 2010a, Hu et al. 2011). The higher the PD and interactive PD values, the better the discretization method. In this paper, we describe the process of using PD and interactive PD values to assess unsupervised discretization methods. Figure 1 displays a flowchart of the assessment of the discretization methods. The three assessment steps are as follows: (1) select the discretization methods and range of intervals representing different discretization levels, and then use them to discretize the continuous risk factors; (2) calculate the PD values of the discretized continuous risk factors using the factor detector, and the interactive PD values with the interaction detector; and (3) compare the PD and interactive PD values at each level of the discretization method, and output the optimal discretization method that has the highest PD

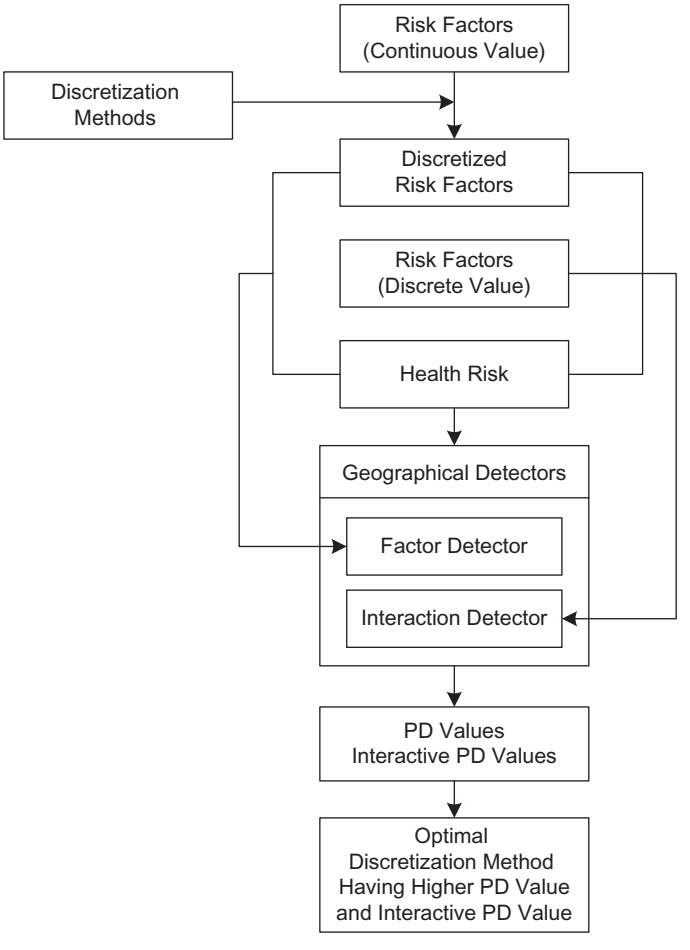


Figure 1. Flow chart of optimal discretization method assessment.

and interactive PD values. The PD value is the main indicator for the assessment of the discretization method; the interactive PD value is the auxiliary indicator of when the PD values are nearly the same for different discretization methods. We will exemplify the above process with NTD data.

## Experiments

### *Data description*

The study area is the Heshun County, which includes 326 administrative villages located in the northern region of Shanxi Province, China (Figure 2). One hundred and eighty-seven cases of neural tube defects (NTD) were reported during the period 1998–2005, and its incidence rate was very high. A great deal of NTD research has been conducted in this area. Both physical and man-made environmental exposures are thought to contribute to the prevalence of NTD (Wu et al. 2004). Liao et al. (2010) used spatial filtering to detect potential annual clusters, and Kruskal–Wallis tests and Multivariate regression analysis

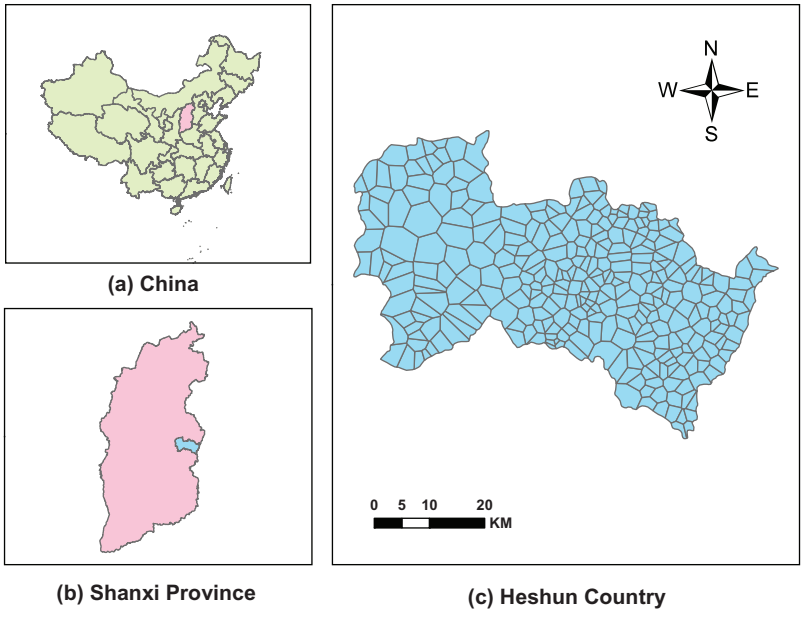


Figure 2. Study area: (a) China, (b) Shanxi Province, and (c) Heshun Country.

were used to identify the environmental risk factors for NTD. Bai et al. (2010) used rough set theory to mine the decision rules of environmental risk factors for NTD. Some continuous risk factors were discretized using the MDL method before applying to the risk prediction of NTD using rough set theory. Wang et al. (2010a) used the geographical detectors model to study the relationship between environmental risk factors and NTD incidence rates. The continuous risk factors were discretized with user-defined discretization. The efficacy of the discretization used was not discussed. In this paper, we discuss this problem in terms of the PD and interactive PD values calculated by the geographical detectors model.

In previous studies, watershed type, soil type, elevation, and GDP were considered risk factors of NTD (Bai et al. 2010; Liao et al. 2010; Wang et al. 2010b). There probably exists an underlying relationship between the direct cause of NTD and the four risk factors. Therefore, in our paper, the four risk factors are used to discuss the suitability of unsupervised discretization methods. Watershed type and soil type are the discrete risk factors, while elevation and GDP are continuous risk factors. These data can be acquired from the sample data attached to the GeogDetector software. Figure 3 displays maps of these factors. We will compare the PD and interactive PD values generated by the geographical detectors model with elevation and GDP discretized using different unsupervised discretization methods, and describe the selection of the optimal method.

### *Discretization of elevation and GDP*

In this paper, we use the EI, QU, GI, NB, and SD methods to discretize elevation and GDP. The primary number of intervals needs to be set for all of these discretization methods. Selecting the optimum number of intervals is a very difficult task. Here, we set

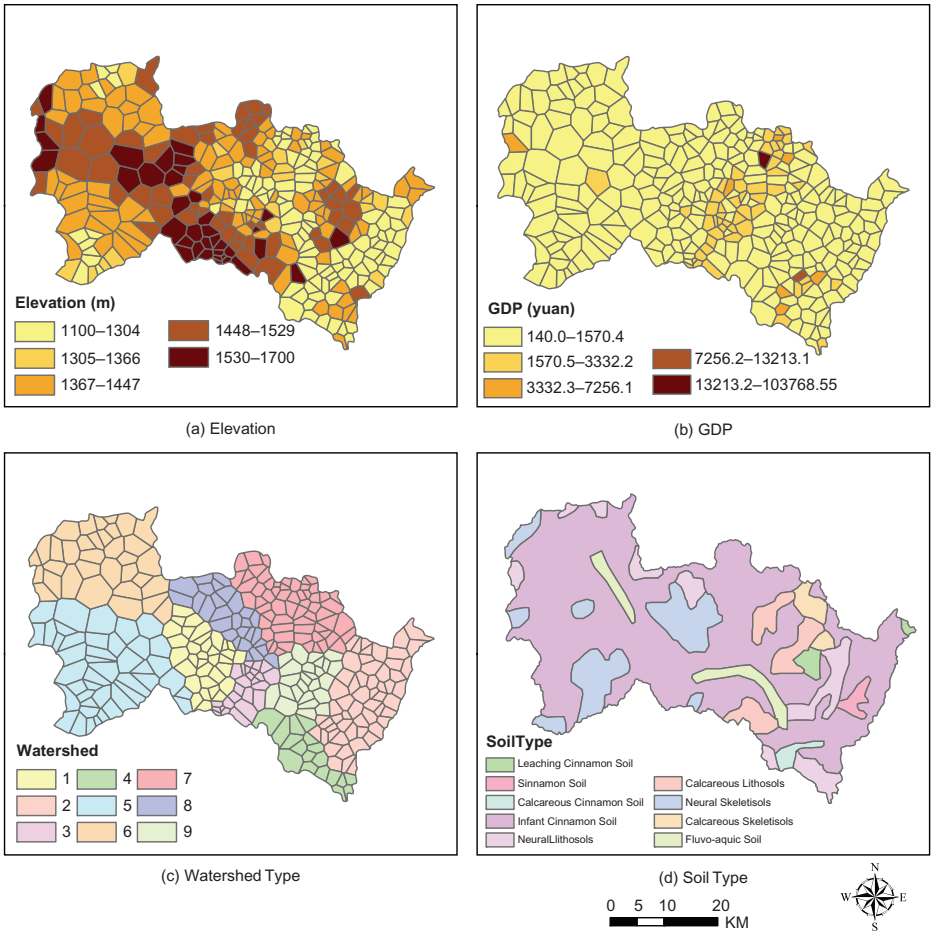


Figure 3. Maps of (a) elevation, (b) GDP, (c) watershed type, and (d) soil type in Heshun Country.

the number of discretization intervals as 2–8, representing different discretization levels. When the number of discretization intervals is determined, we can get the cut points of elevation and GDP according to the five discretization methods. The values lower than the cut points belong to the former interval. We then compare the PD values of the five discretization methods at each level. The optimal number of intervals and discretization method are determined when the PD and interactive PD values are relatively high.

## Results

The PD values of elevation and GDP divided into a range of 2–8 intervals using the five discretization methods were calculated using the factor detector (Table 1). The interactive PD values for watershed, elevation, and GDP and the interactive values for soil type, elevation, and GDP were calculated using the interaction detector (Table 2). To effectively analyze the PD and interactive PD values in Tables 1 and 2, six maps were drawn (Figures 4–9), which give direct and visual expressions of the PD and interactive PD values and reflect the changes among different discretization levels very well.

Table 1. The PD values for elevation and GDP divided into a range of 2–8 intervals using the five discretization methods.

| Discretization methods | Risk factors | PD values | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| EI | Elevation | 0.171 | 0.206 | 0.137 | 0.184 | 0.202 | 0.177 | 0.173 |
| QU | | 0.171 | 0.206 | 0.194 | 0.226 | 0.230 | 0.226 | 0.226 |
| NB | | 0.171 | 0.189 | 0.194 | 0.184 | 0.188 | 0.183 | 0.183 |
| GI | | 0.171 | 0.186 | 0.190 | 0.185 | 0.198 | 0.189 | 0.181 |
| SD | | – | – | – | – | – | 0.184 | – |
| EI | GDP | 0.000 | 0.000 | 0.000 | 0.000 | 0.00 | 0.000 | 0.012 |
| QU | | 0.040 | 0.028 | 0.096 | 0.185 | 0.147 | 0.120 | 0.117 |
| NB | | 0.000 | 0.047 | 0.049 | 0.066 | 0.121 | 0.226 | 0.220 |
| GI | | 0.042 | 0.046 | 0.045 | 0.049 | 0.039 | 0.182 | 0.120 |
| SD | | – | – | 0.000 | – | – | 0.000 | – |

## Discussion

### *Optimal discretization of elevation*

Figure 4 displays a map of the PD values for elevation divided into a range of 2–8 intervals using the five discretization methods. For the EI method, the PD value was highest when elevation was divided into three intervals. For the QU method, the PD value was highest when the elevation was divided into six intervals, and had an increasing trend

Table 2. Interactive PD values for watershed type, soil type and elevation, GDP divided into a range of 2–8 intervals using the five discretization methods.

| Discretization methods | Risk factors | Interactive PD values | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| EI | Watershed type and elevation | 0.721 | 0.720 | 0.718 | 0.718 | 0.729 | 0.715 | 0.722 |
| QU | | 0.71 | 0.715 | 0.724 | 0.733 | 0.739 | 0.738 | 0.738 |
| NB | | 0.71 | 0.717 | 0.715 | 0.723 | 0.723 | 0.724 | 0.724 |
| GI | | 0.71 | 0.714 | 0.711 | 0.733 | 0.725 | 0.725 | 0.716 |
| SD | | – | – | – | – | – | 0.717 | – |
| EI | Soil type and elevation | 0.327 | 0.327 | 0.276 | 0.318 | 0.337 | 0.327 | 0.315 |
| QU | | 0.318 | 0.333 | 0.327 | 0.351 | 0.368 | 0.369 | 0.369 |
| NB | | 0.318 | 0.321 | 0.338 | 0.333 | 0.333 | 0.333 | 0.337 |
| GI | | 0.318 | 0.318 | 0.327 | 0.331 | 0.331 | 0.334 | 0.337 |
| SD | | – | – | – | – | – | 0.318 | – |
| EI | Watershed type and GDP | 0.693 | 0.693 | 0.693 | 0.693 | 0.693 | 0.693 | 0.698 |
| QU | | 0.718 | 0.728 | 0.785 | 0.81 | 0.806 | 0.804 | 0.826 |
| NB | | 0.694 | 0.711 | 0.708 | 0.716 | 0.771 | 0.787 | 0.786 |
| GI | | 0.728 | 0.724 | 0.718 | 0.721 | 0.737 | 0.801 | 0.798 |
| SD | | – | – | 0.694 | – | – | 0.694 | – |
| EI | Soil type and GDP | 0.172 | 0.172 | 0.172 | 0.172 | 0.172 | 0.172 | 0.186 |
| QU | | 0.217 | 0.211 | 0.282 | 0.356 | 0.328 | 0.315 | 0.326 |
| NB | | 0.173 | 0.205 | 0.215 | 0.239 | 0.281 | 0.390 | 0.390 |
| GI | | 0.225 | 0.223 | 0.224 | 0.224 | 0.228 | 0.351 | 0.303 |
| SD | | – | – | 0.173 | – | – | 0.173 | – |

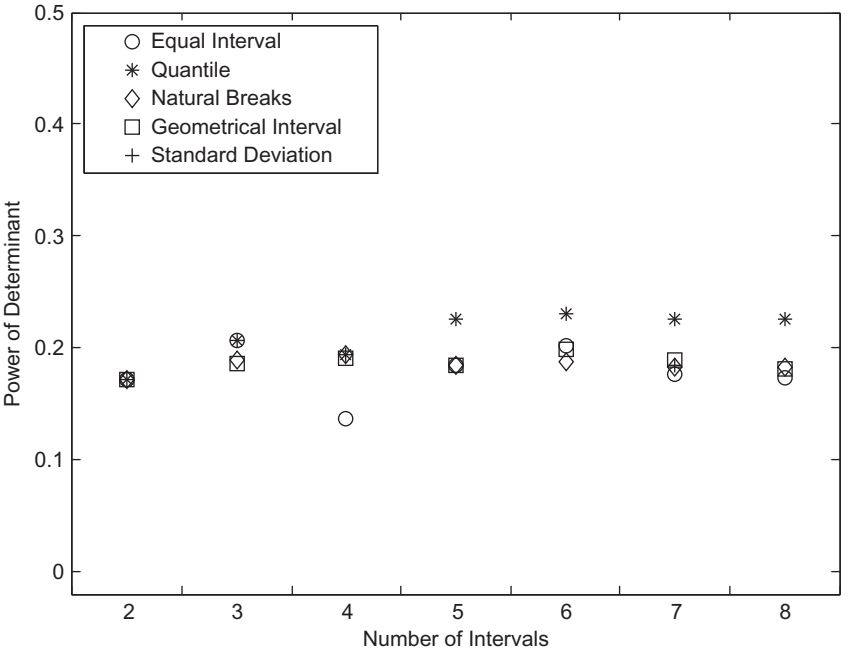


Figure 4. Map of PD values for elevation divided into a range of 2–8 intervals using the five discretization methods.

when the number of intervals was lesser than six and a decreasing trend when the number of intervals was greater than six. The PD values were very close when elevation was divided using the NB or GI methods. Using the SD method, elevation was divided into four intervals and the range was lower than one standard deviation for each interval when the required number of intervals was less than eight. Following a comparison of all PD values, the PD value was highest when elevation was divided into six intervals using the QU method. Figure 5 displays a map of the interactive PD values for watershed type and elevation divided into a range of 2–8 intervals using the five discretization methods. Figure 6 displays a map of the interactive PD values for soil type and elevation divided into a range of 2–8 intervals using the five discretization methods. From these two figures, it can be seen that the interactive PD values between elevation and watershed or soil type were highest when elevation was divided into 6–8 intervals using the QU method. Following a comparison of the PD and interactive PD values, the QU method with six intervals was regarded as the optimal discretization method for elevation.

The PD value is a quantitative assessment of the impact of elevation on the spatial pattern of NTD. It was 0.23 when the optimal discretization method was used, which is higher than the PD value of 0.1 calculated by Wang et al. (2010a) with a user-defined discretization. Therefore, our discretization method is thought to be more suitable for the assessment of the impact of elevation on NTD incidence rates. The PD value for watershed type was 0.669 and the PD value for soil type was 0.132. From Table 2, we know that the lowest interactive PD value between elevation and watershed type was 0.71, which is higher than the independent watershed type PD value of 0.669. The lowest interactive PD value for elevation and soil type was 0.276, which is also higher than the independent soil PD value of 0.139. This means that elevation and watershed or soil type

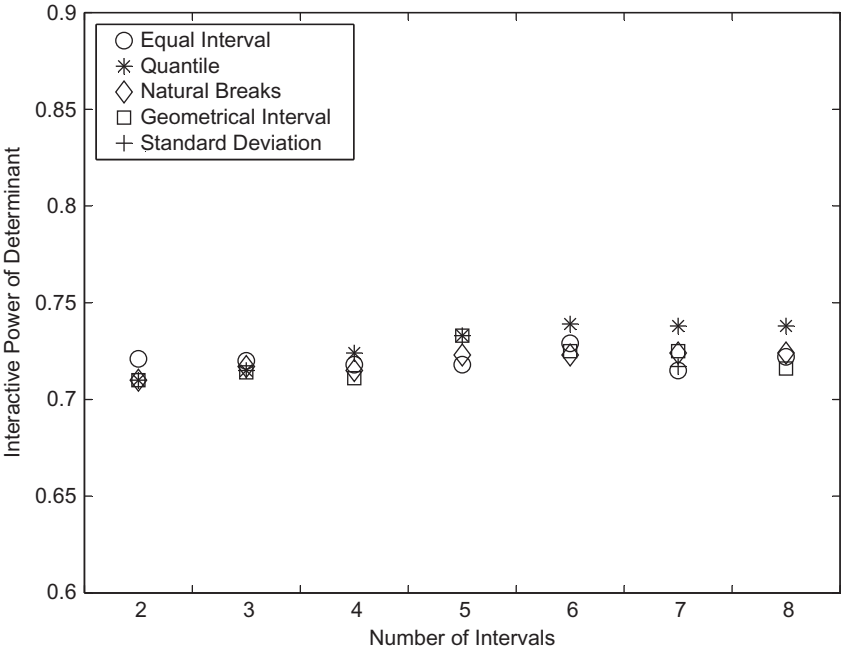


Figure 5. Map of interactive PD values for watershed type and elevation divided into a range of 2–8 intervals using the five discretization methods.

taken together will enhance the independent risks of watershed or soil type. In addition, the interactive PD value of 0.739 for watershed and elevation discretized using the optimal method is higher than the 0.734, which was calculated using the user-defined discretization. The interactive PD value of 0.368 for soil type and elevation discretized using the optimal method was also higher than the 0.279, which was calculated using the user-defined discretization.

We also used the risk factors to find the average NTD incidence rate for each interval, and the highest two rates were 6.77 and 6.53 per hundred persons when elevation was discretized using the optimal discretization method. The corresponding ranges of elevation were 1298–1300 m, and lower than 1298 m, respectively, and were the two lowest intervals of elevation. Higher NTD incidence rates are closely related to lower elevation.

### *Optimal discretization method of GDP*

We analyzed the impact of discretization methods on the PD values for GDP and interactive PD values for GDP and watershed type or soil type. The optimal discretization method for GDP was found. Figure 7 displays a map of the PD values for GDP divided into a range of 2–8 intervals using the five discretization methods. From Figure 7, it can be seen that the PD values were very low and nearly equivalent when GDP was discretized using the EI and SD methods. This indicates that the EI and SD methods performed badly in the discretization of GDP. The reason for this is that GDP has special data that are much larger than other values. Because the EI method equally divided the entire range of data values into specified intervals, some intervals had no values. Therefore, the geographical strata and PD values generated for these intervals were almost

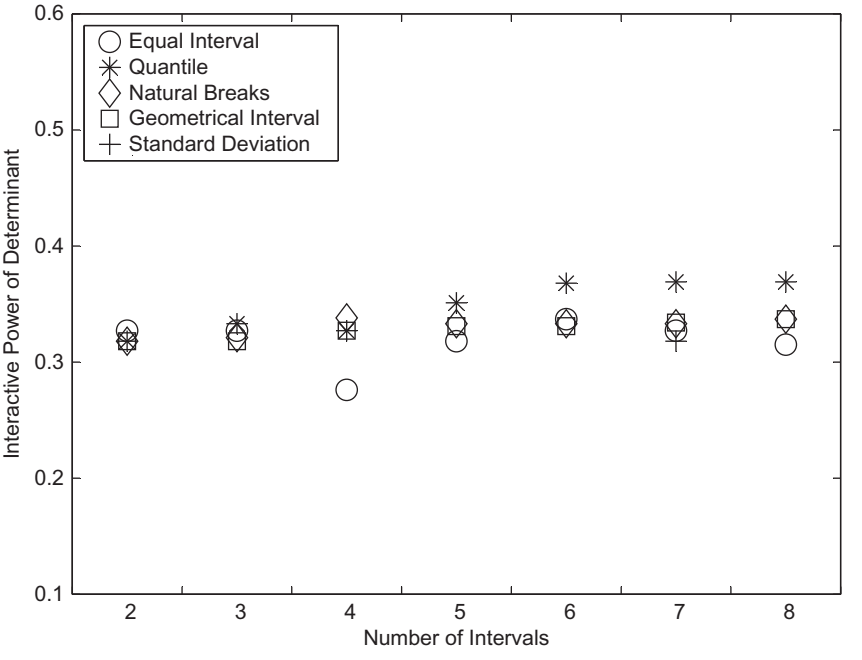


Figure 6. Map of interactive PD values for soil type and elevation divided into a range of 2–8 intervals using the five discretization methods.

the same. Similar to the EI method, when applying the SD method some GDP intervals had no values, and the geographical strata generated were almost the same. Meanwhile, the PD values were lowest. For the QU method, the PD value was highest when GDP was divided into five intervals. It had an increasing trend when the number of intervals was lesser than five and a decreasing trend when the number of intervals was greater than five. For the NB method, the PD value was highest with an initial increase followed by a drop when GDP was divided into seven intervals. Similar to the NB method, when applying the GI method the PD value was highest with an initial increase followed by a drop when GDP was divided into seven intervals. For the SD method, GDP was divided into four and seven intervals when the number of required intervals was lesser than eight, and the PD values were lowest. The PD value was highest among when GDP was divided into seven intervals using the NB method.

Following the analysis of Table 2, we know that watershed type and GDP will together enhance the independent risk of watershed type, and that soil type and GDP will together enhance the independent risk of soil type when GDP is discretized with any method. Figure 8 displays a map of the interactive PD values for watershed type and GDP divided into 2–8 intervals using the five discretization methods. Similar to the PD values, the interactive PD values were very low and almost the same when GDP was discretized using the EI and SD methods. For the QU method, the interactive PD value was highest when GDP was divided into eight intervals. For the NB and GI methods, the interactive PD values had similar trends of an initial increase followed by a drop, and were highest when GDP was divided into seven intervals. Figure 9 displays a map of the interactive PD values for soil type and GDP divided into 2–8 intervals using the five discretization

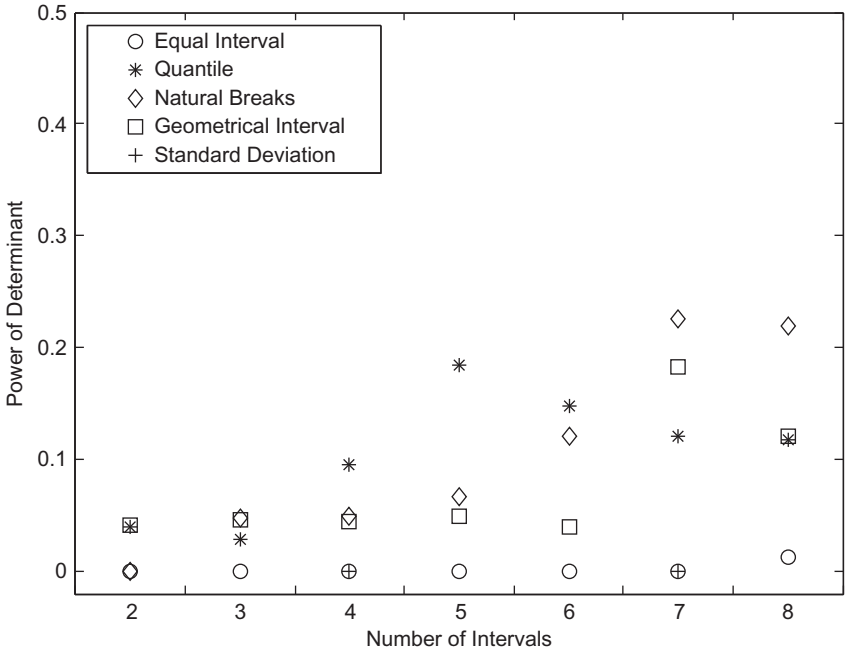


Figure 7. Map of PD values for GDP divided into a range of 2–8 intervals using the five discretization methods.

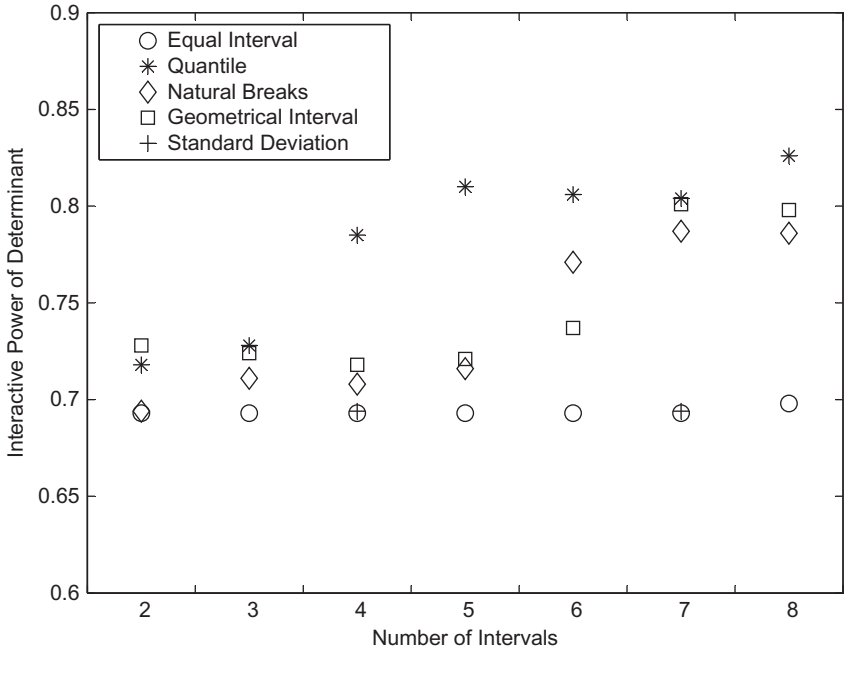


Figure 8. Map of interactive PD values for watershed type and GDP divided into a range of 2–8 intervals using the five discretization methods.

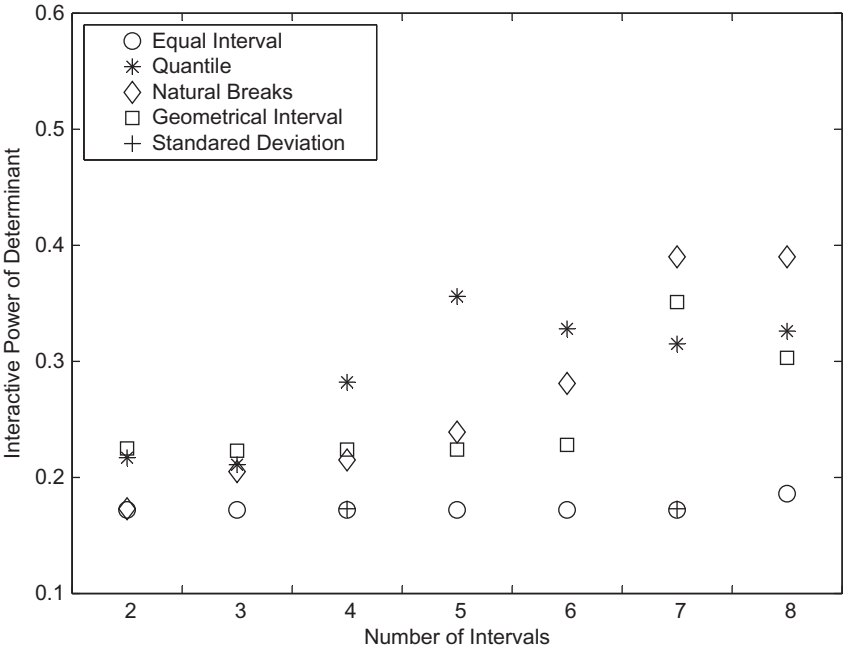


Figure 9. Map of interactive PD values for soil type and GDP divided into a range of 2–8 intervals using the five discretization methods.

methods. The interactive PD values were still very low and almost the same when GDP was discretized using the EI and SD methods. For the QU method, the interactive PD value was highest when GDP was divided into five intervals. For NB and GI methods, the interactive PD values were highest when GDP was divided into seven intervals. Among the values displayed in Figure 9, the interactive PD value was also highest when GDP was divided into seven intervals using the NB method.

Following the comparison of PD and interactive PD values, we regarded the NB method with seven intervals as the optimal GDP discretization method. The GDP PD value was 0.226 when GDP was discretized using the optimal method. The interactive PD values for GDP, watershed, and soil type were 0.787 and 0.390, respectively. They are suitable for assessing the impact of GDP on NTD incidence rates as well as the joint impacts of GDP, watershed, and soil type. The average NTD incidence rate was calculated for each interval and the two highest were 7.27 and 6.97 per hundred persons when GDP was divided using the optimal discretization method. The ranges of these two GDP intervals were 7256.10–13,213.08 yuan and lower than 642.60 yuan, respectively. The NTD incidence rates were high when GDP was in the lowest and second highest intervals. This indicates that countries with a low GDP lack nutrition, which may be related to their high NTD rates. The countries with a relatively higher GDP (but not the highest) have better nutrition, larger population, and a higher natality. Higher natality may be the reason behind the higher NTD incidence rates. The higher NTD incidence rates may also be significantly related to some other risk factors.

## Conclusions

The geographical detectors model is a new spatial analysis method used to assess health risk based on spatial variance analysis. It is suitable for both discrete and continuous environmental risk factors, although the continuous risk factors should be made discrete before analysis. Because the method used for discretizing continuous risk factors is a very important issue in the application of the geographical detectors model, in this paper we discussed the selection of the optimal method. A selection process based on PD values is introduced. This methodology will also assist in the selection of the optimal discretization method for geographical detectors-based risk assessment research. However, there are still some issues to be resolved. For example, the five discretization methods used in our example are very commonly used unsupervised discretization methods. They are simple, easy to implement, and more effective when the data satisfies a distributional assumption. However, the number of intervals must be set before use. Because more intervals do not help us mine the information of continuous risk factors, we set the number of intervals to 2–8, depending on our experience. Therefore, the optimal discretization method chosen with our strategy is only locally optimal. However, when the range of intervals has been determined, our strategy is useful. Meanwhile, the five discretization methods do not take into account the spatial distribution of the data. Disease risk factors always have spatial distribution features such as spatial association and heterogeneity. These distribution features include important spatial risk factor information that is closely related to health. If we only use traditional discretization methods we lose useful spatial information, reducing the effectiveness of the analysis. In future research, we will explore discretization methods the spatial distribution of the continuous data into account. The methods will then be used in the process of discretizing continuous disease risk factors and their analysis based on the geographical detectors model. Finally, the PD and interactive PD values can be used as indicators of the effectiveness of discretization methods.

## Acknowledgements

This work is supported in part by the NSFC (40971222), NSFC (41023010), and MOST (2012CB955503).

## References

Anselin, L. 1995. “Local Indicators of Spatial Association-LISA.” *Geographical Analysis* 27: 93–115.

Bai, H. X., Y. Ge, J. F. Wang, and Y. L. Liao. 2010. “Using Rough Set Theory to Identify Villages Affected by Birth Defects: The Example of Heshun, Shanxi, China.” *International Journal of Geographical Information Science* 24 (4): 559–576.

Best, N. G., K. Ickstadt, and R. L. Wolpert. 2000. “Spatial Poisson Regression for Health and Exposure Data Measured at Disparate Resolutions.” *Journal of the American Statistical Association* 95 (452): 1076–1088.

Dougherty, J., R. Kohavi, and M. Sahami. 1995. “Supervised and Unsupervised Discretization of Continuous Features,” In *Proceedings of the Twelfth International Conference on Machine Learning*, Tahoe City, CA, July 9–12, 194–202. San Francisco, CA: Morgan Kaufmann Publisher.

Fischer, M. M., and J. F. Wang. 2011. *Spatial Data Analysis: Models, Methods and Techniques*, 82. Berlin: Springer.

Ge, Y., F. Cao, and R. F. Duan. 2011. “Impact of Discretization Methods on the Rough Set-Based Classification of Remotely Sensed Images.” *International Journal of Digital Earth* 4 (4): 330–346.

Getis, A., and J. K. Ord. 1992. "The Analysis of Spatial Association by Use of Distance Statistics." *Geographical Analysis* 24 (3): 189–206.

Haining, R. 2003. *Spatial Data Analysis: Theory and Practice*, 432. Cambridge: Cambridge University Press.

Hu, Y., J. F. Wang, X. H. Li, D. Ren, and J. Zhu. 2011. "Geographical Detector-Based Risk Assessment of the Under-Five Mortality in the 2008 Wenchuan Earthquake, China." *PLoS ONE* 6 (6): e21427.

Jenks, G. F. 1967. "The Data Model Concept in Statistical Mapping." *International Yearbook Cartography* 7: 186–190.

Kerber, R. 1992. "Chimerge: Discretization of Numeric Attributes." In *Proceedings of the Tenth National Conference on Artificial Intelligence*, San Jose, CA, July 12–16, 123–128. Menlo Park, CA: AAAI Press.

Kulldorff, M. 1997. "A Spatial Scan Statistic." *Communications in Statistics: Theory and Methods* 26 (6): 1481–1496.

Kurgan, L. A., and K. J. Cios. 2004. "Discretization Algorithm That Uses Class-Attribute Interdependence Maximization." *IEEE Transactions on Knowledge and Data Engineering* 16 (2): 145–153.

Li, X. Z., J. F. Wang, W. Z. Yang, Z. J. Li, and S. J. Lai. 2011. "A Spatial Scan Statistic for Multiple Clusters." *Mathematical Biosciences* 233: 135–142.

Liao, Y. L., J. F. Wang, Y. Q. Guo, and X. Y. Zheng. 2010. "Risk Assessment of Human Neural Tube Defects Using a Bayesian Belief Network." *Environmental Research and Risk Assessment* 24 (1): 93–100.

Lisboa, P. J., and A. F. G. Taktak. 2006. "The Use of Artificial Neural Networks in Decision Support in Cancer: A Systematic Review." *Neural Networks* 19: 408–415.

Liu, H., F. Hussain, C. L. Tam, and M. Dash. 2002. "Discretization: An Enabling Technique." *Data Mining and Knowledge Discovery* 6: 393–423.

Magnin, B., L. Mesrob, S. Kinkingnéhum, M. Pélégrini-Issac, O. Colliot, M. Sarazin, B. Dubois, S. Lehéricy, and H. Benali. 2009. "Support Vector Machine-Based Classification of Alzheimer's Disease from Whole-Brain Anatomical MRI." *Neuroradiology* 51: 73–83.

Nakaya, T., A. S. Fotheringham, C. Brunsdon, and M. Charlton. 2005. "Geographically Weighted Poisson Regression for Disease Association Mapping." *Statistics in Medicine* 24: 2695–2717.

Tsai, C. J., C. I. Lee, and W. P. Yang. 2008. "A Discretization Algorithm Based on Class-Attribute Contingency Coefficient." *Information Sciences* 178: 714–731.

Wang, J. F., and Y. Hu. 2012. "Evironmental Health Risk Detection with GeogDetector." *Environmental Modelling & Software* 33: 114–115.

Wang, J. F., X. H. Li, G. Chirstakos, Y. L. Liao, T. Zhang, X. Gu, and X. Y. Zheng. 2010a. "Geographcial Detectors-Based Health Risk Assessment and Its Application in the Neural Tube Defects Study of the Heshun Region, China." *Journal of Geographical Information Science* 24 (1): 107–127.

Wang, J. F., X. Liu, G. Christakos, Y. L. Liao, X. Gu, and X. Y. Zheng. 2010b. "Assessing Local Determinants of Neural Tube Defects in the Heshun Region, Shanxi Province, China." *BMC Public Health* 10: 52.

Wang, J. F., A. J. McMichael, B. Meng, N. G. Becker, W. G. Han, K. Glass, J. L. Wu, et al. 2006. "Spatial Dynamics of an Epidemic of Severe Acute Respiratory Syndrome in an Urban Area." *Bulletin of the World Health Organization* 84 (12): 965–968.

Wu, J. L., J. F. Wang, B. Meng, G. Chen, L. H. Pang, X. M. Song, K. L. Zhang, T. Zhang, and X. Y. Zheng. 2004. "Exploratory Spatial Data Analysis for the Identification of Risk Factors to Birth Defects." *BMC Public Health* 4: 23.

Yang, Y., G. I. Webb, and X. D. Wu. 2005. "Discretization Methods." In *Data Mining and Knowledge Discovery Handbook*, edited by O. Maimon and L. Rokach, 113–130. New York: Springer.