Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-274
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 30 November 2018

Geoscientific Model Development
Discussions
Open Access

SBDM v1.0: A scaling-based discretization method for the Geographical Detector Model

Xiaoyu Meng[1,2], Xin Gao[1], Shengyu Li[1], Wenjing Huang[3], and Jiaqiang Lei[1]

[1]State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, 818 South Beijing Road, Urumqi 830011, Xinjiang, China.
[2]University of Chinese Academy of Science, Beijing 100049, China.
[3]College of Resources, Environment and Tourism, The Capital Normal University, Beijing 100048, China.

Correspondence: Xin Gao (gaoxin@ms.xjb.ac.cn)

Abstract. Geographical Detector Model (GDM) can be used to assess the affinity between potential environmental factors and the response variables. If environmental factors entered are continuous, the first step for application of GDM is to discretize the continuous variable into category strata with an appropriate discretization method. Many one-dimensional discretization methods have been arbitrarily applied to GDM but failed to obtain the optimal strata of environmental factors, resulting in an inaccurate model output. In this paper, we present the Scaling-Based Discretization Method (SBDM) as a novel discretization method that can be used to obtain the optimal strata for GDM. The SBDM takes the power of determinant as a criterion function through upscaling and downscaling processes to obtain the optimal discretization. The software was tested with two case studies: (1) The distance to river was discretized with SBDM to reveal the effect of rivers on the sand cover ratio in the Maowusu (Mu Us) Sandy Land, northern China. The SBDM obtained more accurate information for the influence of rivers on the sand cover ratio than the results from Priori Knowledge discretization method. (2) Seven environmental factors were discretized using SBDM to detect potential associations between these factors and NDVI spatial pattern in Xinjiang, north-western China. Then we compared the q values from SBDM with the values from four commonly used one-dimensional discretization methods, demonstrating that for all considered factors, SBDM gets a larger q value than other methods. Collectively, SBDM offers a new way for data discretization that accurately reveals the relationship between controlling factors and response variables.

1 Introduction

Spatial data analyses are useful for analysing continuous data, such as temperature or altitude. Because of the complexity of continuous data, it is often less efficient to completely understand all the information contained in the data (Catlett, 1991; Kerber, 1992; Richeldi and Rossotto, 1995; Frank and Witten, 1999; Xu and Ii, 2005; Cheng and Lu, 2017). Discretization is a data processing procedure that partitions continuous data into a set of strata, which can be used to understand the geographical relationships among environmental processes and support decision-making.

1.1 Discretization methods for continuous data

There are one-dimensional and multi-dimensional discretization methods used in geoscientific research. Here we focus on one-dimensional methods and briefly summarize the frequently used discretization methods.

(1) Equal Interval (EI) equally divides data into specified ranges, but without taking data distribution into account (Ren et al., 2014; Hu et al., 2015). (2) Natural Breaks (NB) (Jenks) and one-dimensional K-means (MacQueen, 1967) both seek intervals that have the smallest in-class variance and highest variance between classes (Wu et al., 2016; Hong et al., 2017; Shrestha and Luo, 2017; Du et al., 2017). (3) Quantile (QU) classifies data into specified classes that contain an equal number of elements or units. Each quantile places similar elements in adjacent classes and the elements demonstrate large differences for the same class. This method is suitable for data with a linear distribution (Luo et al., 2016; Liang and Yang, 2016). (4) Geometrical Interval (GI) method is based on a principle that the sums of the squares of the number of elements for each class are minimal. The biggest advantage of this method is that it can handle non-normal data (Cao et al., 2013; Tian et al., 2017). The four methods above are integrated into spatial analysis software (e.g., ArcGIS®) and are the most commonly used methods for univariate discretization. (5) Systematical Clustering method sorts the data by the distance criteria (e.g., Euclidean distance, Mahalanobis distance). The categories are mostly reduced until the appropriate classification requirements are met (Liao et al., 2016). Some studies also use prior knowledge to categorize the continuous data (Xu et al., 2017; Zhou et al., 2018).

Although there are several discretization methods for data mining, two fundamental principles for discretization should be satisfied (Cao et al., 2013). Firstly, a good discretization method always attempts to minimize the information losses (maximizing the within-group similarity and minimizing the between-group similarity when considering only the univariate), such as Natural Breaks and K-means methods. Secondly, the results of discretization should be as suitable as possible for the subsequent models. If we arbitrarily discretize continuous data before modelling, the subsequent analysis and application may miss valuable information or lead to misspecification of models, e.g., Boolean networks (Kauffman, 1969), generalized logical networks (Song et al., 2009), or Sandwich Interpolation model (Wang et al., 2002, 2013; Liu et al., 2018). Therefore, the effectiveness of the discretization algorithm is not only related to discrete data distribution, but also related to the subsequent application models. The degree of information loss is the best criteria to evaluate the discretization methods if there are no subsequent models. Quinlan (1986) proposed a concept of information entropy measuring information loss before and after discretization of continuous data. Cang and Luo (2018) defined a concept to assess the effect of discretization on data loss based on the spatial data association, and suggested that the Quantile discretization method can minimize information loss. However, the unified criteria for evaluating the discretization effect is to maximize the accuracy of the model results as much as possible when discretized data is used as an input data to subsequent model.

1.2 Introduction of the Geographical Detector Model

In geographical research, most purposes of discretization are restricted to divide continuous data for efficient cognition without establishing a geographical data mining model. Based on the spatial variation theory, Wang et al. (2010) presented a new technique: Geographical Detector Model (GDM), which was developed from medical geography and has now been applied to many geographical fields. GDM is designed to measure spatially stratified heterogeneity of a response variable and reveal the impact of driving factors. For example, the geographic elements (X and Y) of the area A can be represented using a grid in Geographic Information Systems (GIS) (Fig. 1). In this case, X is an environmental factor and Y is a response variable and they have potential spatial relationships . The hypothesis of GDM is that if an environmental factor (X) contributes to a response variable (Y), there may be a similar spatial distribution for the two, and this similarity or spatial association can be measured by the power of determinant (PD) value (Wang et al., 2010) or q-statistics (Wang et al., 2016). If the variable X is the continuous data, the first step of GDM is to discretize it into categorical data or strata, i.e., X to X_h (blue arrow in Fig. 1), and then based on the spatial pattern of strata of X, GDM calculates the ratio between the sum of spatial variance of Y within each stratum and population variance ($X_h \sim Y$, red arrows in Fig. 1). The PD value (q value) can be expressed by Eq. (1) (Wang et al., 2010),

$$q = 1 - \frac{\sum_{h=1}^{L} N_h \sigma^2}{N\sigma^2}, \tag{1}$$

where N denotes the units of study area (N = 32 in Fig. 1) stratified into h = 1, 2,..., L strata (L = 5 in Fig.1). σ^2 means the stratum variance. The value of q is required to be within [0, 1] (q = 0 when there is no stratified heterogeneity in Y, and q = 1 when Y is completely stratified). In fact, if X and Y are two different variables, the q value ($X_h \sim Y$) is defined as the driving force of environmental factors (X) to a response variable (Y). However, if we replace the Y with X, the q value ($X_h \sim X$, green arrow in Fig. 1) can be a measure of spatial stratified heterogeneity of X. Spatial stratified heterogeneity refers to the degree of within-strata variance less than the between strata variance. It is a common phenomenon and essence of nature in extensive geoscience (Wang et al., 2016). Based on q-statistic, four detectors have been derived from GDM: Factor detector, Ecological detector, Risk detector, and Interaction detector.

1.3 Imperfections of the Geographical Detector Model

GDM is based on spatial variance analysis of strata, hence, the q value (Eq. (1) is dependent upon the way continuous variables are discretized. Therefore, the key factor affecting the accuracy of GDM is the discretization methods (X to X_h, the blue arrow step in Fig. 1).

For $X_h \sim X$, the two most common algorithms for discretizing data with the highest spatial stratified heterogeneity, namely Natural Breaks (Fisher-Jenks algorithms) (D. Fisher, 1958) and K-means (K-means algorithm) (Hartigan and Wong, 1979). These two methods have equivalent criteria that function as the q-statistic in the geographical detector. In Natural Breaks, the criteria function is the goodness of variance fit (GVF) (Dent et al., 1999), which has the same meaning and formula as q-statistic. In K-means, the criteria function is to minimize the within-cluster sum of square (i.e., variance), which is equivalent to the numerator ($\sum_{h=1}^{L} N_h \sigma^2$) in q-statistic. A faster algorithm, referred to as Dynamic Programming, emerged for optimizing one-dimensional K-means and is available as “Ckmeans.1d.dp“ in the R Package (Wang and Song, 2011).

For $X_h \sim Y$, the power of determinant q value can be used as a quantitative evaluation index to the effective degree of continuous data discretization. A larger q value indicates that the feature has a higher contribution to the discretization degree (Liao et al., 2016). When we use one discretization method to divide the spatial data into different strata, the q value is relatively higher, suggesting that the spatial distribution pattern of X_h has a stronger influence on Y. In this application, discretization is the key step that has great impact on the model results.

In the geographical detector application, different discretization methods may lead to disparate q values (Cao et al., 2013; Huang, 2014; Ju et al., 2016; Zhao et al., 2017). A number of studies (Cao et al., 2013; Ju et al., 2016; Zhao et al., 2017) tried to look for a larger q value using the aforementioned discretization methods . When assessing the driving force of environmental factors (X) to the response variable (Y), those traditional discretization methods are mainly based on the distribution of data X to be stratified. However, q is a comprehensive value which is not only related to the distribution of X itself, but also associated with the distribution of Y. Therefore, the conventional discretization methods might be not suitable for GDM. It is imperative to propose an appropriate discretization method to maximize the q value and ensure optimal stratified continuous data.

1.4 Aims of this study

In GDM, the q value as a criterion can assess the spatial stratified heterogeneity and measure the power of determinant of X to Y (Wang et al., 2016). Following this line of reasoning, we provide a new discretization method based on the scale transformation theory that takes the magnitude of q-statistic of GDM as the criteria function to discretize the continuous environmental factors into categorical strata for subsequent application in GDM. Details of our method to obtain the q value for each spatial pattern and to speed up the calculation are put forward in Sect. 2. The case studies and software instructions of this method are presented in Sect. 3, which is followed by the results, discussion, and conclusion in Sects. 4 and 5.

2 SBDM software development

Based on the principle of different amounts of information at various scales (Burt and Adelson, 1987; Portilla et al., 2003), we propose an efficient discretization algorithm: Scaling-Based Discretization Method (SBDM). Taking the q-statistic of GDM as a criterion function, the design of the SBDM algorithm includes three parts: (1) listing all cut points which are used to classify continuous data into discrete strata and upscaling the data to reduce the number of the cut points, (2) based on the current cut points, making an exhaustive search to find the optimal results (the maximum q value and the corresponding cut points in the current scale), and (3) setting the buffer scale for each optimal cut point in the last scale to get the next calculated cut points and then making an exhaustive search algorithm again to find the most accurate results in the current scale in a short time, this step is called downscaling. After that, if the current cut points precision reaches the highest, the algorithm ends, if not, we then loop steps 2 and 3. Figure 2 shows the conceptual architecture of our algorithm.

2.1 Upscaling

Exhaustive search that enumerates all possible combinations of cut points and select the one that most closely meets the condition is an effective method for obtaining the global optimal strata. For the optimal discretization, it needs to take all the stratified patterns (combinations of different cut points) into account and select the one that corresponds to the maximum q value. Here, we illustrate the algorithm briefly with an example (Fig. 3). Based on a raw dataset consisting of 9 × 9 cells (Fig.

3a), we firstly transform the data into a unidimensional array and remove the duplicates. We then obtain 16 cut points (i.e., represented by red lines between different numbers) and 16 combinations by discretizing the raw dataset into two categories (strata). Using exhaustive search algorithm, the quantity Q of all combinations is determined by the Combination Formula (Eq. (2)).

$$Q = C_{N-1}^{P-1} = \frac{(N-1)!}{(P-1)!(N-P)!}, \tag{2}$$

where N and P refer to the number of different values of the response variables and categories appointed by users, respectively. Supposing that we divide the raw dataset into two strata (i.e., selecting one cut point), there are 16 combinations of cut points using Eq. (2) (i.e., Q = 16).

The main disadvantage of exhaustive search is that numerous combinations exist and increases exponentially with the number of cut points. Most geographic elements have properties with a wide range of values (e.g., in complex topography, elevation may range over thousands of meters). The calculation has high time complexity and cannot be accomplished by a common computer. Therefore, we designed an upscaling method to divide the raw data by a pre-specified scale number to decrease the quantity of the cut points. Figure 3b is obtained by dividing the raw data (Fig. 3a) by three. Then, Fig. 2b has five cut points and five combinations when two strata are assumed. Processing that occurs during upscaling dulls the boundaries of strata (fewer red lines in Fig. 3b). The optimal cut points in the raw data and the upscaled data are adjacent in spatial position (e.g., when the optimal cut point is 19 in Fig. 3a, the optimal one could be 6 in Fig. 3b, the positions of 19 and 6 are close). The upscaling processing seems powerful for data compression in efficaciously reducing the amount of calculations, but it generates two defects:

1. Cut points drift. After upscaling processing, each of the five cut points in Fig. 2b arises from merging the raw data (e.g., cut point 6 in Fig. 2b corresponds to [18, 19, 20] in Fig. 2a). This merging may cause cut points to drift. The drift will be nearby the original optimal cut points. Supposing 19 is the optimal cut point in Fig. 2a, for such a case, 19 divides Fig. 2a into two strata: [10, 19) and [19, 26], which correspond to the maximum q value. 19 corresponds to 6 or 7 in Fig. 2b. Exhaustive search lists all cut points in Fig. 2b, there are two possibilities: (1) If 6 is the optimal cut point (the cut point 18 in the raw data), it reduces q value comparing with that when 19 is the optimal cut point in the previous hypothesis. (2) If 7 is the optimal cut point, the subsequent 21 is the optimal one in the raw data, it also reduces q value. Finally, depending on the corresponding larger q value, the optimal cut point is either 6 or 7 in Fig. 2b.

2. Low accuracy. Upscaling processing may cause loss of precision so that we cannot find the exact location of the optimal cut points in the raw data. For instance, the cut points in Fig. 2b might either be 6 or 7, the corresponding optimal one might be one of the values [18, 19, 20, 21, 22, 23] in the raw data so that we cannot find the most accurate cut points under the current upscaling scale number, three.

2.2 Downscaling

We designed a downscaling method that addresses the above problems, while extracting the exact cut points in a short time. The procedure of downscaling method is to set multiple buffer scales for the raw data. We can obtain the data in Fig. 2c by dividing the raw data in Fig. 2a by the number two and rounding the quotient down. Instead of calculating all combinations in Fig. 2c, we set a rule for the cut point using the following:

$$[M_i, N_i] = [(P_i - \Delta\sigma) \times S_{last}, (P_i + \Delta\sigma) \times S_{last}] / S_{next}, \ (i = 1, 2, 3 \ldots). \tag{3}$$

In the results of the last scale S_{last} (i.e., S_{last} = 3 for upscaling), the optimal cut points are $P_i\,(i = 1, 2, \ldots)$. Based on $P_i\,(i = 1, 2, \ldots)$ and taking $\Delta\sigma$ as the neighbourhood (sliding number in SBDM), we obtain the range of cut points $[M_i, N_i]$ in the next scale (S_{next}). In Fig. 2b, when the optimal cut point is 6, the sliding number $\Delta\sigma$ is designed as 1, then the range [5, 7] in Fig. 2b corresponds to [15, 21] in Fig.2a. Next, we divide the set [15, 21] by the number two (next scale number), the subsequent data set [7, 10] in Fig. 2c are the all cut points to be calculated. Other cut points [5, 6] and [11, 13] do not be considered. Supposing the optimal result is 10 in Fig. 2c, the next calculation range in the raw data could be [18, 22] using Eq. (3). Finally, there are only five combinations needed to be calculated to get the maximum q value. During the downscaling, the q value will increase, and the dissimilar data will decline in each stratum, because there are more highly accurate cut points to suit the spatial distribution of Y.

While for cut points with a small interval, all cut points need to be treated as a group instead of individuals to be listed through exhaustive search. For example, assuming that the raw data in Fig. 2a need to be divided into three strata and the data in Fig. 2b have two cut points, 5 and 6 ([3, 5), [5, 6), [6, 8]), the set of the first and second cut points in Fig. 2c are [6, 7, 8, 9] and [7, 8, 9, 10], respectively. Now, we need to decide the category of cut points [7, 8, 9]. In this case, we treat the two sets of cut points as one category for exhaustive search, for example the data sets [6, 7, 8, 9] and [7, 8, 9, 10] can be merged into one category [6, 7, 8, 9, 10]. We conclude that when the range of cut points shows narrow distributions, it may cause conflict of cut points allocated during downscaling. The solution is to merge all the cut points into exhaustive search. In SBDM, the option to implement this feature is demonstrated as Set Cut Points.

3 Case study

We designed two application modes, Line and Surface, to apply SBDM for different data types. The Line mode is designed for the situation when the environmental factor X is a linear element including highways, rivers, and pipelines. Users can detect different amounts of impact to the response variable with the Line mode in SBDM. The Surface mode is used to quantify the power of determinant of surface data to the response variable. For these two application modes, we implement two case studies. (1) Using the Line mode, we detect the authentic influence power of rivers on Sand Cover Ratio (SCR) in Maowusu (Mu Us) Sandy Land, northern China. Then we compare the results from SBDM with the results from Liang and Yang (2016). (2) Using the Surface mode, we quantify the Normalized Difference Vegetation Index (NDVI) dominant controlling factors and the significance of the rank orders for its environmental factors. Seven environmental factors include Elevation (ELEV), Slope Degree (SD), Slope Aspect (SA), Precipitation (PREC), Wind Velocity (WV), Temperature (TEMP), and Specific Humidity (SH). We compare NDVI to the environmental factors in Xinjiang Uygur Autonomous Region, north-western China and compare the results from SBDM with the results from four conventional discretization methods (EI, NB, QU, GI) as described in Sect.1. We then test the widely used four detectors (Risk, Factor, Ecological, and Interaction) of GDM with SBDM.

3.1 Datasets

Mu Us Sandy Land (34,500 km^2) is in the center of Loess Plateau, China, and surrounded by the Yellow River (Fig. 4a). The study area (6,201 km^2) in our case is a part of Mu Us Sandy Land, which contains a portion of tributaries of the Yellow River (perennial river) and some intermittent streams (Fig. 4c). In this case study, we set the closest distance (Euclidean Distance) to the river of each non-river points in the study area as the X variable and SCR as the response variable, Y. The river location is determined based on the interpretation of multi-period remote sensing images and corrected according to high resolution Google Earth images. The SCR is a quantization value of landscape spatial patterns obtained via Eq. (4) after sand dune and vegetation zone landscape types were converted to binary values. We applied the Landsat 8 OLI images for supervised classification with maximum likelihood classification method provided by ENVI®5.3 to obtain the landscape types in Mu Us Sandy Land with the classification result shown in Fig. 4c. Then, the SCR can be obtained by Eq. (4) (Liang and Yang, 2016),

$$SCR = \frac{\text{Sand dune area}\,(\text{km}^2)}{1(\,\text{km}^2)} \times 100\,\%, \tag{4}$$

where the denominator is the sampling area and the numerator is the area of sand dunes in each 1 km^2. Figure. 4d shows the distribution of SCR. For more details of SCR, readers can refer to Liang and Yang (2016).

The second case is undertaken in the Xinjiang Uygur Autonomous Region (1,660,000 km^2) located in the border of north-western China. Figure. 4b shows the NDVI distribution in 2013 after processing from the maximum value composite method (Holben, 1986). The NDVI dataset is MOD13 A2 (MODIS/Terra Vegetation Indices 16-Day L3 Global 1 km SIN Grid) acquired from LAADS DAAC (https://ladsweb.modaps.eosdis.nasa.gov/). The distributions of seven environmental factors are shown in the first column of Fig. 7. The PREC, WV, TEMP, and SH of 1960-2010 are from the Global Meteorological Forcing Dataset (http://hydrology.princeton.edu/data.pgf.php) (Sheffield et al., 2006). Digital Elevation Models (DEMs) from SRTM (Shuttle Radar Topography Mission/90 m resolution Digital Elevation Database) are used as the Elevation datasets, and are downloaded from CGIAR-CSI (http://srtm.csi.cgiar.org/SELECTION/inputCoord.asp). The SD and SA are derived from DEM. Seven environment factors, X_i (i=1,2,...7), are resampled to 1 km × 1 km to fit the Y ($NDVI$) variable resolution.

3.2 Steps of SBDM operation

SBDM can take Image (*.tif) and Text (*.txt) files as the input data formats to support the analysis of spatial and non-spatial data using GDM. Here, we take image formats as an example. All of the above data are made into the GeoTIFF format (*.tif) file, which is one of the widely used raster file format in GIS (Bernard and Ostlander, 2008), we can demonstrate the steps of using the software are as follows:

Step 1. We constrain each image to be the same size and spatial extent, which is the basic condition for the raster calculation. Then, all digital numbers for environmental factors (X) are converted into integer format, which is required for input into the SBDM software. For example, the range of WV in the second case study [2.2, 5.9] is multiplied by 10 to convert into [22, 59] and avoid loss of precision.

Step 2. Mode selection. Firstly, we define the types of environmental factors, X. Then, we select an appropriate mode: Line mode (investigating the strata of influence of a line type data such as a river) or Surface mode (detecting the power of environmental factors of surface data type to the response variables).

Generally, the order of the computing procedure must experience one-time upscaling and several times downscaling. In this part, we first show an example of Line mode with a general desktop PC (i7-2600 CPU, 4-core, 3.4GHz, 8G RAM). Before using the Line mode, there are two pre-treatments in GDM: (1) If the line element data is a shapefile format (i.e., Polyline [*.shp]), we calculate the minimum distances between non-river points and river points in raster format. This step can be completed using the Euclidean Distance tool in toolbox of ArcGIS®10.3 software (ESRI, 2015). The size of pixels of the resultant raster is the same size as the data Y (SCR). (2) In geographic science, the degree of influence of linear geographical elements is always reduced as the distance increases and will be not be affected beyond a certain distance. Therefore, we set a parameter in SBDM to reduce the cost of calculation, whereby Extremum Number refers to the greatest distance of the linear geographical element to Y. For example, Liang and Yang (2016) show that an Extremum Distance of 21 km indicates that beyond 21 km the rivers have few or no influence on SCR. Here, we also take 21 km as the extremum number.

Step 3. Upscale processing. We set the upscaling number according to the range and size of X, and the strata number is user specified. Here, the distance range of the river is [0, 27730] m, and the size is 117 × 53 units. The scale number of upscaling is set as 900, meaning that there are approximately 31 cut points (27730/900 $\approx$ 31) and 593,775 combinations when the strata number is defined as seven. After pushing the run button, it takes four seconds to finish the calculation (Fig. 5). The q and cut points shown in the Upscaling Results box are the optimal values at the current scale (i.e., 900).

Step 4. Downscale processing. The purpose of this step is to obtain more accurate q values and cut points. In step 3, the optimal cut points are under the scale of 900 (defined as the Last Scale number in this step). The downscaling path could be 900-1 directly or step by step (e.g., 900-300-1) to get the most accurate results. In this example, we compare the downscaling path 900-1 and 900-300-1. The results show that the path 900-300-1 (19 seconds) was less time consuming but produced the same results as 900-1 (603 seconds), shown in Fig. 5. Finally, the results shown in Downscaling Results box are the maximum q value (0.1545) and the corresponding optimal cut points (0, 1000, 1414, 2236, 3000, 4242, 10049, 27730) that are most accurate.

The above steps show the instructions of the Line mode in SBDM that are used to find the best strata for the influence of line element (river) to the response variables (e.g., Sand Cover Ratio). For the Surface mode, the only difference is that there an extremum number is not required as input. The detailed meanings of each button and I/O boxes in the software and how to use the Text files (*.txt) as an input data, readers can refer to the $SBDM\_Software\_Instruction\_Manual.docx$ in the Supplementary files.

4 Results and discussion

4.1 River influence on SCR in the Maowusu (Mu Us) Sandy Land, northern China (Line Mode)

Based on the field investigation or prior knowledge, Liang and Yang (2016) stratified the buffer region of the river to seven strata. Figures. 6a and 6b display the discretization results of the river distance from the SBDM and Liang and Yang (2016), respectively. The discretization results (Strata), the average value SCR in each stratum and q values from the SBDM and Prior Knowledge are shown in Table 1. It can be seen that two discretization methods yield very different results. In SBDM, the strata [1414, 2236) have the highest value of SCR, suggesting that this range of distance from river is more likely to form the sandy land. While in the results of Prior Knowledge, we only know that the strata [1000, 3000) have the highest contribution rate to sandy land formation. Furthermore, the strata [3000, 4242) also have a higher driving force to the sandy land forming but we cannot find this information in the Prior Knowledge method. In addition, the q value is 0.154 in SBDM which is higher than the value 0.136 in the Prior Knowledge. That is because the field investigation has certain limitations and the mechanism of the effects of rivers on vegetation is complicated. Different levels of influence distance defined by Prior Knowledge are commonly not accurate enough, which may result in imprecise (lower) q value in the GDM. Therefore, according to the response variable itself to stratify the environmental factors, SBDM can reveal a more precise biotope range of the geographical variables.

4.2 Impacts of seven environmental factors on NDVI spatial pattern in Xinjiang, north-western China (Surface Mode)

Figures 7 and 8 show the results of the factor detector for the impacts of seven environmental factors on the spatial pattern of NDVI in Xinjiang, north-western China, with different discretization methods. For each factor, comparing with other methods, the q value is the maximum when SBDM is applied. The q value can be used to measure the degree of the impact factor on the response variable. For larger q values, the environmental factor will have a greater impact on the geophysical variable. The impact power can be quantified as $q \times 100\ \%$ (Wang and Xu, 2017). SBDM results show that the order of determinant power of seven factors to the NDVI are roughly the same for all discretization methods (i.e., the q value of PREC > WV > TEMP > SD > SH > SA, except for TEMP > DEM < SD in the GI method, see $Factor\_Detector.xlsx$ in the Supporting Information). Note that q values of DEM with different discretization methods are quite distinct. It might be caused by the large ranges of DEM from -157 m to 7913 m, which can generate thousands of cut points and result in obviously different q values. In addition, the results of factor detector reveal that the TEMP (q = 0.175) and DEM (q = 0.174) factors have roughly the same impact power, while in the GI method, TEMP (q = 0.159) is more dominant than DEM (q = 0.046).

The ecological detector assesses the relative importance of two factors on the spatial distribution of determinant variables (Wang et al., 2010). The ecological detector also shows that TEMP and DEM have no significant difference with SBDM (p = 0.00, see $Ecological\_Detector.xlsx$ of the Supporting Information). Therefore, the SBDM software allows us to obtain a realistic discretization and confirm the ratio of the power of determinant between different environmental factors.

The risk detector calculates the mean value of each stratum for one factor and tests whether the mean value is statistically significantly different with that of another stratum of the factor, revealing where the risk areas are. Bigger differences suggest more risks to the response variable within the stratum (Cao et al., 2013). One more time, the results are closely associated with the cut points determined by the discretization methods. In the case study, the risk detector results include the average value of each stratum and p value of student t test using five different discretization methods (see $Risk\_Detector.xlsx$ of the Supporting Information). Using SA as an example, we find that the average value of first stratum using SBDM is about two times higher than other discretization methods, indicating that the first level of SA has a strong influence on vegetation formation. Therefore, by finding out the differentiation between strata and improving the consistency of data within each stratum, SBDM can effectively detect the relationship between different level strata of environmental factors and response variables.

The interaction detector quantifies the interactive effect of two factors to the response variable. For example, comparing with the effects of an independent factor, we examine whether the temperature and elevation interact or independently contribute to the spatial pattern of NDVI. The interaction detector can solve the collinearity problem of independent variables. We expand the interaction quantities from the traditional two to seven factors to make a comprehensive comparison (Wu et al., 2017; Shrestha and Luo, 2017; Zou et al., 2017). Our results show that the interactive effect combination with seven factors can only account for 65.9 % of the influence from all underlying factors (see $Interaction\_Detector.xlsx$ of the Supporting Information). It suggests that about 34.1% of the influence, because of other factors, has been neglected when selecting the current seven environmental factors. However, when QU method is applied, the interactive q values of seven factors are 0.67, which is greater than the q value from SBDM, implying that single factor optimal discretization does not represent the optimal one when the interaction detector is applied. Because there are more factors considered by interaction, a complete evaluation may result in a more complicated combination. Hence, the optimal discretization of a single factor with the maximum q value might not produce the maximum interacted q. The q values derived from a single factor and the interacted one might be independent. At present, we cannot judge the effect of discretization methods by the interactive q value (Cao et al., 2013). Future work should explore mechanisms that explain the interactive effects between factors.

4.3 Report of calculation

The SBDM software is able to run on a desktop PC with no special hardware (e.g., Graphics Processing Unit). In this paper, we have made a test report for the results of the SBDM software based on the second case study. The report is shown in $Report.xlsx$ in the Supporting Information. Using SH as an example (Table 2), it takes the SBDM software about 1.5 minutes to process the data, which is much improved compared to 7.6 days when using exhaustive search directly to acquire the most accurate cut points. Moreover, when an exhaustive search method processes a large range of factors, such as DEM in this study, computing time could be 1.4 million years to get the better cut points without downscaling processing on a desktop PC environment, but the SBDM software only spends 1.5 hours. Collectively, we demonstrate that using exhaustive search to list all possible combinations and using SBDM have the same consequence of q value and cut points, but the SBDM software takes an extremely short time to obtain the optimal discretization results. In addition, according to the time costing report, more computing time is required when there is a larger interval between successive scales, particularly when downscaling is close to the highest resolution (i.e., scale 1).

4.4 Parameter selection in the software

The SBDM software is written in Python3.6, an interpreted high-level programming language for general-purpose programming. In the software, there are four important parameters that need to be specified manually: the degree of upscaling (Upscaling Number), downscaling intervals (distance between the Last Scale number and the Next Scale number), the Sliding Number, and the Strata Number. Based on a set of experiments and theoretical analysis, there are some principles for the parameter setting. For the first upscaling, more precise q and cut points are possible with smaller upscaling scale. The range of data and computer performance should also be accounted for. However, in our test, the results derived from SBDM with smaller Upscaling Number (4 in Table 2) are the same as the results obtained with one time for bigger Upscaling Number (6 in Table 2) and one time for Downscaling Number (4 in Table 2), but a one-time upscaling process takes much longer (7.6 minutes) than one-time upscaling (1 minute) and downscaling (23 seconds) processes. For downscaling intervals, users need to keep the linear relationship between the last scale number and the next scale number to reduce the drifting extent of results between different scales. More computation time can be saved when longer times of buffer scale or smaller downscaling intervals are utilized. For the sliding number, most of the optimal results in the next scale appear within one neighbourhood of the cut points in the last scale. When it comes to the variables with a large range, the sliding number should be set to two or three (e.g., when downscaling was used in the factor of DEM, the optimal results were obtained with the sliding number two and three, see $Report.xlsx$ file in the Supporting Information). It may be caused by the cut points drift during the division operation. Normally, users only need to set sliding number as one or two with a small range of factor data, and three when the range is large. In that case, the q value and cut points would have a great probability to reach the optimal. For strata number, when the number of strata is too small, we may miss important discoveries that occur at a lower resolution, but when it is very large, it is not conducive for human understanding of data results, and we may not have enough data to infer the distribution patterns (Bay, 2001). In the SBDM software, according to different situations, users should select an appropriate strata number.

5 Conclusions

GDM is an effective model for measuring the power of determinant of environmental factors to the response variables, but model results depend on the discretization methods of the continuous variables. Inappropriate discretization methods might lead to an inexact model result. In this study, we introduced a novel one-dimensional discretization method: SBDM and an ancillary software. Based on scale transformation theory, SBDM takes q value as a criterion function through upscaling and downscaling processes to obtain the optimal strata of environmental factors. When we compare SBDM with different one-dimensional discretization methods by testing the Line mode and Surface mode in the software with two case studies, the major conclusions are: (1) The SBDM can be used to obtain a larger or accurate q value for GDM. Through SBDM, the power of determinant of each environmental factor to the response variable could be ascertained and not affected by the discretization methods. For example, in the second case study, the TEMP (q = 0.175) and ELEV (q = 0.174) factors have roughly the same impact power from SBDM, but in the GI method, TEMP (q = 0.159) is dominant relative to ELEV (q = 0.046). (2) The SBDM can get the optimal cut points (strata) of the environmental factors. When the strata are obtained by an improper discretization method, useful information may be lost about the relationship between X and Y. For example, in the case study of Maowusu (Mu Us) Sandy Land, SBDM infers that the range [1414, 2236) of the distance from the river is more likely to form the sandy land, while from Prior Knowledge method, we only know that the [1000, 3000) is the risk strata without specific information. (3) The SBDM software leads to a dramatic reduction of the number of strata combinations and computation time. A large amount of data can be processed in a short time. Meanwhile, unlike the conventional discretization methods that discretize data based on data itself, taking q value as the criteria, SBDM combines the response variable (Y) with the factor determinants (X) jointly to discretize the continuous data X for GDM. Collectively, the SBDM software is an effective and theoretically well-founded tool that can more realistically discern the "biotope" of the observed response variables and make GDM more accurate.

Code and data availability. The SBDM code and software are archived at:http://doi.org/10.5281/zenodo.1475892. The MIT license governs the distribution and use of the code and associated documentation files. Permission is granted, free of charge, to any person to deal with the software without restriction.

Author contributions. Xiaoyu Meng was the primary developer of the SBDM software. Xiaoyu Meng and Xin Gao contributed to the formulation, analysis and writing of this paper. Other authors contributed equally to the discussion.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We acknowledge the financial support from the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDA20030102), the National Natural Science Foundation of China (No. 41571011), the National Thousand Talents Program (No. Y772091001), the High-level Talent Recruitment Program of Xinjiang Uygur Autonomous Region, China (No. Y742071001) and the China-initiated "the Belt and Road" Special Project (No. 131965KYSB20170038).

References

Bay, S. D.: Multivariate Discretization for Set Mining, Knowledge & Information Systems, 3, 491–512, https://doi.org/10.1007/PL00011680, 2001.

Bernard, L. and Ostlander, N.: Assessing climate change vulnerability in the arctic using geographic information services in spatial data infrastructures, Climatic Change, 87, 263–281, https://doi.org/10.1007/s10584-007-9346-0, 2008.

Burt, P. J. and Adelson, E. H.: The Laplacian Pyramid as a Compact Image Code, Readings in Computer Vision, 31, 671–679, https://doi.org/10.1109/TCOM.1983.1095851, 1987.

Cang, X. and Luo, W.: Spatial association detector (SPADE), International Journal of Geographical Information Science, 32, 2055–2075, https://doi.org/10.1080/13658816.2018.1476693, https://doi.org/10.1080/13658816.2018.1476693, 2018.

Cao, F., Ge, Y., and Wang, J.: Optimal discretization for geographical detectors-based risk assessment, Gisci Remote Sens, 50, 78–92, https://doi.org/10.1080/15481603.2013.778562, 2013.

Catlett, J.: On Changing Continuous Attributes into Ordered Discrete Attributes, in: European Working Session on Machine Learning, pp. 164–178, https://doi.org/10.1007/BFb0017012, 1991.

Cheng, S. and Lu, F.: A Two-Step Method for Missing Spatio-Temporal Data Reconstruction, International Journal of Geo-Information, 6, 187, https://doi.org/10.3390/ijgi6070187, 2017.

D. Fisher, W.: On Grouping for Maximum Homogeneity, Journal of The American Statistical Association - J AMER STATIST ASSN, 53, 789–798, https://doi.org/10.1080/01621459.1958.10501479, 1958.

Dent, B. D., Torguson, J. S., and Hodler, T. W.: Cartography: Thematic map design, vol. 5, WCB/McGraw-Hill Boston, 1999.

Du, Z., Zhang, X., Xu, X., Zhang, H., Wu, Z., and Pang, J.: Quantifying influences of physiographic factors on temperate dryland vegetation, Northwest China, Scientific Reports, 7, 40 092, https://doi.org/10.1038/srep40092, 2017.

ESRI: ArcMap for Desktop: Release 10.3. Environmental Systems Research Institute, Redlands, CA, USA, https://www.esri.com/, 2015.

Frank, E. and Witten, I. H.: Making Better Use of Global Discretization, in: Proc of the Sixteenth International Conference on Machine Learning, pp. 115–123, https://doi.org/10.1007/978-0-387-30164-8-221, 1999.

Hartigan, J. A. and Wong, M. A.: Algorithm AS 136: A k-means clustering algorithm, Journal of the Royal Statistical Society. Series C (Applied Statistics), 28, 100–108, https://doi.org/10.2307/2346830, 1979.

Holben, B.: Characteristics of maximum-value composite images from temporal (AVHRR) data, International Journal of Remote Sensing, 7, 1417–1434, https://doi.org/10.1080/01431168608948945, 1986.

Hong, Y. E., Xinyue, H. U., Ren, Q., Lin, T., Xinhu, L. I., Zhang, G., and Shi, L.: Effect of Urban Micro-climatic Regulation Ability on Public Building Energy Usage Carbon Emission, Energ. Buildings, 154, https://doi.org/10.1016/j.enbuild.2017.08.047, 2017.

Hu, Y., Li, R., Bergquist, R., Lynn, H., Gao, F., Wang, Q., Zhang, S., Sun, L., Zhang, Z., and Jiang, Q.: Spatio-temporal transmission and environmental determinants of Schistosomiasis Japonica in Anhui Province, China, PLoS Neglect. Trop. D., 9, e0003 470, https://doi.org/10.1371/journal.pntd.0003470, 2015.

Huang, J.and Wang, J. B. Y. X. C. H. M. H. D.: Identification of health risks of hand, foot and mouth disease in China using the geographical detector technique., International Journal of Environmental Research & Public Health, 11, 3407, https://doi.org/10.3390/ijerph110303407, 2014.

Jenks, G. F.: The data model concept in statistical mapping, in: International Yearbook of Cartography, vol. 7, pp. 186–190, https://ci.nii.ac.jp/naid/10021899676/en/.

Ju, H., Zhang, Z., Zuo, L., Wang, J., Zhang, S., Wang, X., and Zhao, X.: Driving forces and their interactions of built-up land expansion based on the geographical detector a case study of Beijing, China, International Journal of Geographical Information Science, 30, 2188–2207, https://doi.org/10.1080/13658816.2016.1165228, 2016.

Kauffman, S. A.: Metabolic stability and epigenesis in randomly constructed genetic nets, Journal of theoretical biology, 22, 437–467, https://doi.org/10.1016/0022-5193(69)90015-0, 1969.

Kerber, R.: ChiMerge: discretization of numeric attributes, in: Tenth National Conference on Artificial Intelligence, pp. 123–128, http://dl.acm.org/citation.cfm?id=1867135.1867154, 1992.

Liang, P. and Yang, X.: Landscape spatial patterns in the Maowusu (Mu Us) Sandy Land, northern China and their impact factors, Catena, 145, 321–333, https://doi.org/10.1016/j.catena.2016.06.023, 2016.

Liao, Y., Wang, J., Du, W., Gao, B., Liu, X., Chen, G., Song, X., and Zheng, X.: Using spatial analysis to understand the spatial heterogeneity of disability employment in China, TRANSACTIONS IN GIS, https://doi.org/10.1111/tgis.12217, 2016.

Liu, T., Wang, J., Xu, C., Ma, J., Zhang, H., and Xu, C.: Sandwich mapping of rodent density in Jilin Province, China, Journal of Geographical Sciences, 28, 445–458, https://doi.org/10.1007/s11442-018-1483-z, 2018.

Luo, W., Jasiewicz, J., Stepinski, T., Wang, J., Xu, C., and Cang, X.: Spatial association between dissection density and environmental factors over the entire conterminous United States, Geophys. Res. Lett., 43, n/a–n/a, https://doi.org/10.1002/2015GL066941, 2016.

MacQueen, J.: Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pp. 281–297, University of California Press, Berkeley, Calif., https://projecteuclid.org/euclid.bsmsp/1200512992, 1967.

Portilla, J., Strela, V., Wainwright, M. J., and Simoncelli, E. P.: Image denoising using scale mixtures of Gaussians in the wavelet domain, IEEE Transactions on Image Processing, 12, 1338, https://doi.org/10.1109/TIP.2003.818640, 2003.

Quinlan, J. R.: Induction of Decision Trees, Machine Learning, 1, 81–106, https://doi.org/10.1007/BF00116251, 1986.

Ren, Y., Deng, L., Zuo, S., Luo, Y., Shao, G., Wei, X., Hua, L., and Yang, Y.: Geographical modeling of spatial interaction between human activity and forest connectivity in an urban landscape of southeast China, Landscape Ecol., 29, 1741–1758, https://doi.org/10.1007/s10980-014-0094-z, 2014.

Richeldi, M. and Rossotto, M.: Class-driven statistical discretization of continuous attributes (Extended abstract), Lecture Notes in Computer Science, 912, 335–338, https://doi.org/10.1007/3-540-59286-5-81, 1995.

Sheffield, J., Goteti, G., and Wood, E. F.: Development of a 50-Year High-Resolution Global Dataset of Meteorological Forcings for Land Surface Modeling, J. Climate, 19, 3088–3111, https://doi.org/10.1175/JCLI3790.1, 2006.

Shrestha, A. and Luo, W.: An assessment of groundwater contamination in Central Valley aquifer, California using geodetector method, Annals of Gis, 23, 1–18, https://doi.org/10.1080/19475683.2017.1346707, 2017.

Song, M., Lewis, C. K., Lance, E. R., Chesler, E. J., Yordanova, R. K., Langston, M. A., Lodowski, K. H., and Bergeson, S. E.: Reconstructing Generalized Logical Networks of Transcriptional Regulation in Mouse Brain from Temporal Gene Expression Data, EURASIP J. Bioinformatics Syst. Biol., 2009, 5:1–5:13, https://doi.org/10.1155/2009/545176, 2009.

Tian, L., Li, Y., Yan, Y., and Wang, B.: Measuring urban sprawl and exploring the role planning plays: A shanghai case study, Land Use Policy the International Journal Covering All Aspects of Land Use, 67, 426–435, https://doi.org/10.1016/j.landusepol.2017.06.002, 2017.

Wang, H. and Song, M.: Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming, The R journal, 3, 29, https://www.ncbi.nlm.nih.gov/pubmed/27942416, 2011.

Wang, J. and Xu, C.: Geodetector: Principle and prospective, Acta Geographica Sinica, 72, 116–134, https://doi.org/10.11821/dlxb201701010, 2017.

Wang, J. F., Liu, J. Y., Zhuan, D. F., Li, L. F., and Ge, Y.: Spatial sampling design for monitoring cultivated land, International Journal of Remote Sensing, 23, 263–284, https://doi.org/10.1080/01431160010025998, 2002.

Wang, J. F., Li, X. H., Christakos, G., Gu, X., Gu, X., Gu, X., and Zheng, X. Y.: 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, 107–127, https://doi.org/10.1080/13658810802443457, 2010.

Wang, J. F., Haining, R., Liu, T. J., Li, L. F., and Jiang, C. S.: Sandwich estimation for multi-unit reporting on a stratified heterogeneous surface, Environment & Planning A, 45, 2515–2534, https://doi.org/10.1068/a44710, 2013.

Wang, J. F., Zhang, T. L., and Fu, B. J.: A measure of spatial stratified heterogeneity, Ecol. Indic., 67, 250–256, https://doi.org/10.1016/j.ecolind.2016.02.052, 2016.

Wu, C., Ye, X., Du, Q., and Luo, P.: Spatial effects of accessibility to parks on housing prices in Shenzhen, China, Habitat International, 63, 45–54, https://doi.org/10.1016/j.habitatint.2017.03.010, 2017.

Wu, R., Zhang, J., Bao, Y., and Zhang, F.: Geographical Detector Model for Influencing Factors of Industrial Sector Carbon Dioxide Emissions in Inner Mongolia, China, Sustainability-Basel, 8, 149, https://doi.org/10.3390/su8020149, 2016.

Xu, Q., Dong, Y., and Yang, R.: Influence of different geographical factors on carbon sink functions in the Pearl River Delta:, Sci. Rep.-UK, 7, 110, https://doi.org/10.1038/s41598-017-04856-6, 2017.

Xu, R. and Ii, D. C. W.: IEEE, Survey of clustering algorithms, IEEE Transactions on Neural Networks, 16, 645–678, https://doi.org/10.1109/TNN.2005.845141, 2005.

Zhao, Y., Deng, Q., Lin, Q., and Cai, C.: Quantitative analysis of the impacts of terrestrial environmental factors on precipitation variation over the Beibu Gulf Economic Zone in Coastal Southwest China., Scientific Repots, 7, 44 412, https://doi.org/10.1038/srep44412, 2017.

Zhou, C., Chen, J., and Wang, S.: Examining the effects of socioeconomic development on fine particulate matter (PM2.5) in China's cities using spatial regression and the geographical detector technique, Science of the Total Environment, 619-620, 436–445, https://doi.org/10.1016/j.scitotenv.2017.11.124, 2018.

Zou, B., Jiang, X., Duan, X., Zhao, X., Jing, Z., Tang, J., and Sun, G.: An Integrated H-G Scheme Identifying Areas for Soil Remediation and Primary Heavy Metal Contributors: A Risk Perspective, Scientific Reports, 7, 341, https://doi.org/10.1038/s41598-017-00468-2, 2017.

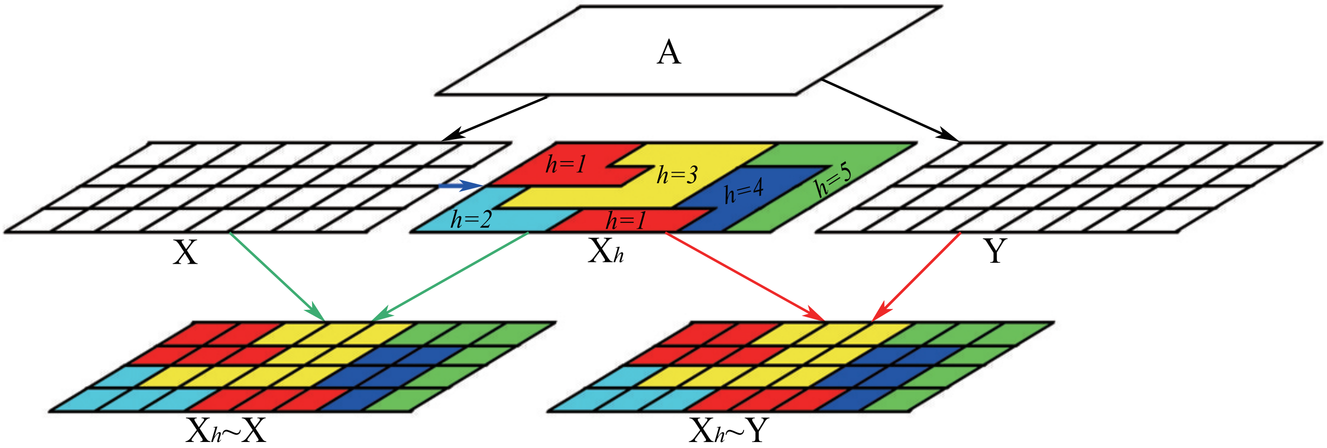


Figure 1. Diagrammatic sketch of the Geographical Detector Model (GDM).

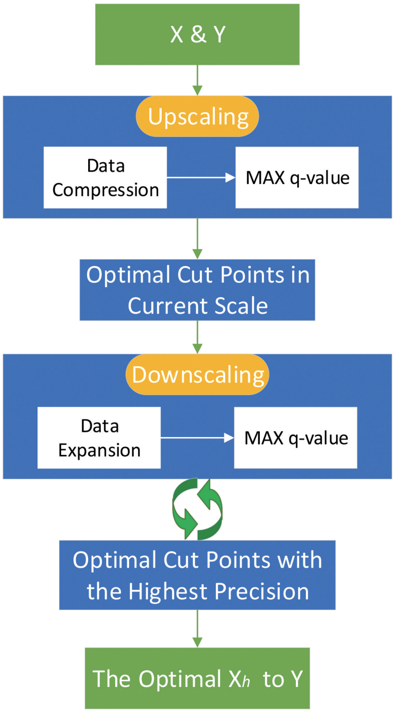
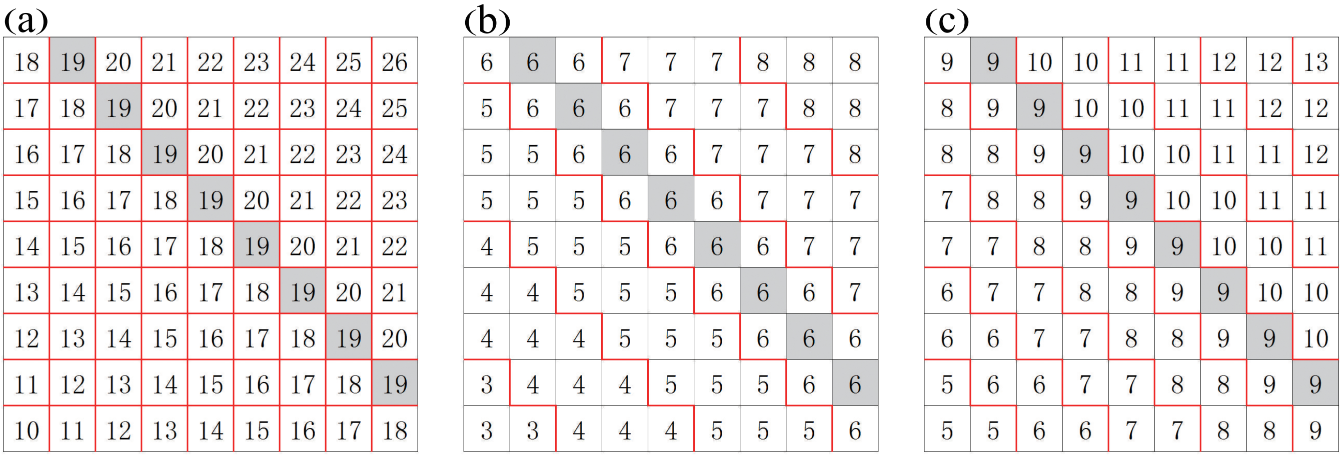


Figure 2. Conceptual architecture of the algorithm with the upscaling and downscaling processing steps in the SBDM software.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-274
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 30 November 2018

Geoscientific Model Development
Discussions
Open Access



(a)

18	19	20	21	22	23	24	25	26
17	18	19	20	21	22	23	24	25
16	17	18	19	20	21	22	23	24
15	16	17	18	19	20	21	22	23
14	15	16	17	18	19	20	21	22
13	14	15	16	17	18	19	20	21
12	13	14	15	16	17	18	19	20
11	12	13	14	15	16	17	18	19
10	11	12	13	14	15	16	17	18

(b)

6	6	6	7	7	7	8	8	8
5	6	6	6	7	7	7	8	8
5	5	6	6	6	7	7	7	8
5	5	5	6	6	6	7	7	7
4	5	5	5	6	6	6	7	7
4	4	5	5	5	6	6	6	7
4	4	4	5	5	5	6	6	6
3	4	4	4	5	5	5	6	6
3	3	4	4	4	5	5	5	6

(c)

9	9	10	10	11	11	12	12	13
8	9	9	10	10	11	11	12	12
8	8	9	9	10	10	11	11	12
7	8	8	9	9	10	10	11	11
7	7	8	8	9	9	10	10	11
6	7	7	8	8	9	9	10	10
6	6	7	7	8	8	9	9	10
5	6	6	7	7	8	8	9	9
5	5	6	6	7	7	8	8	9

Figure 3. Diagram illustrating the steps of upscaling and downscaling processing for the key algorithm of the SBDM software. (a) Raw data consisted of 9 × 9 cells. (b) Upscaling result from the raw data divided by the scale number three. (c) Downscaling result from the raw data divided by the buffer scale number two. All data rounded down as the integer format. The red lines represent the cut points which divide the data into discrete strata.

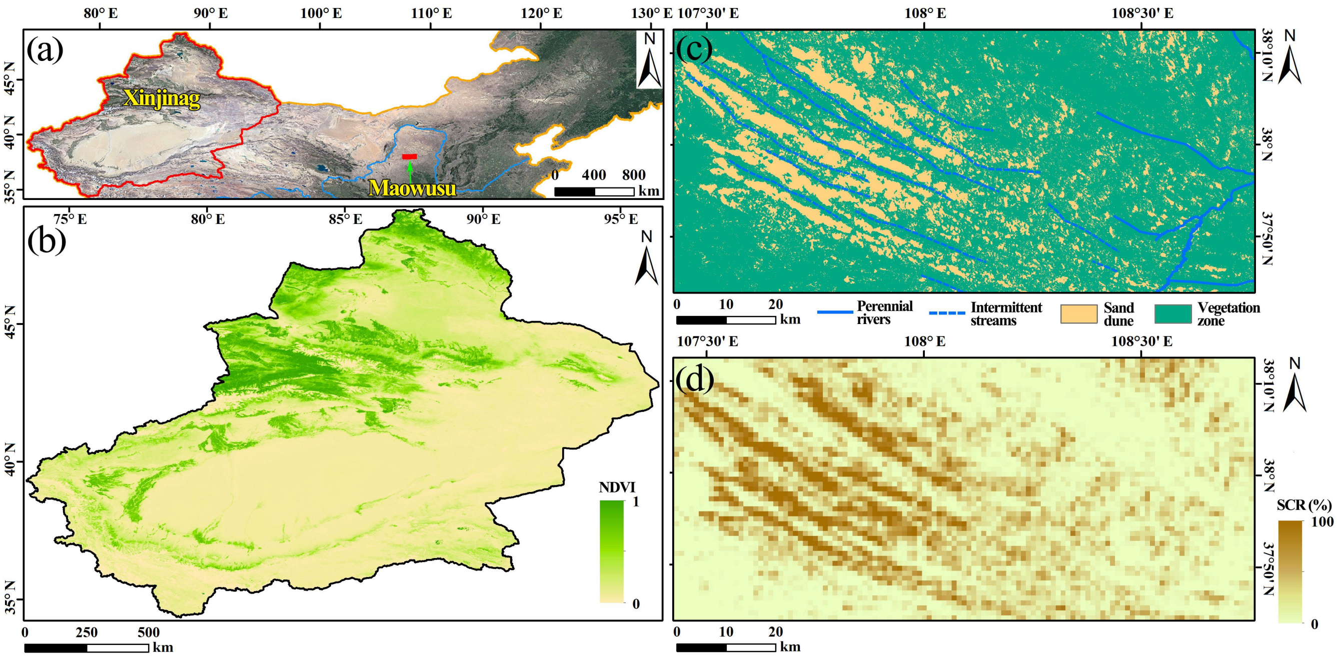
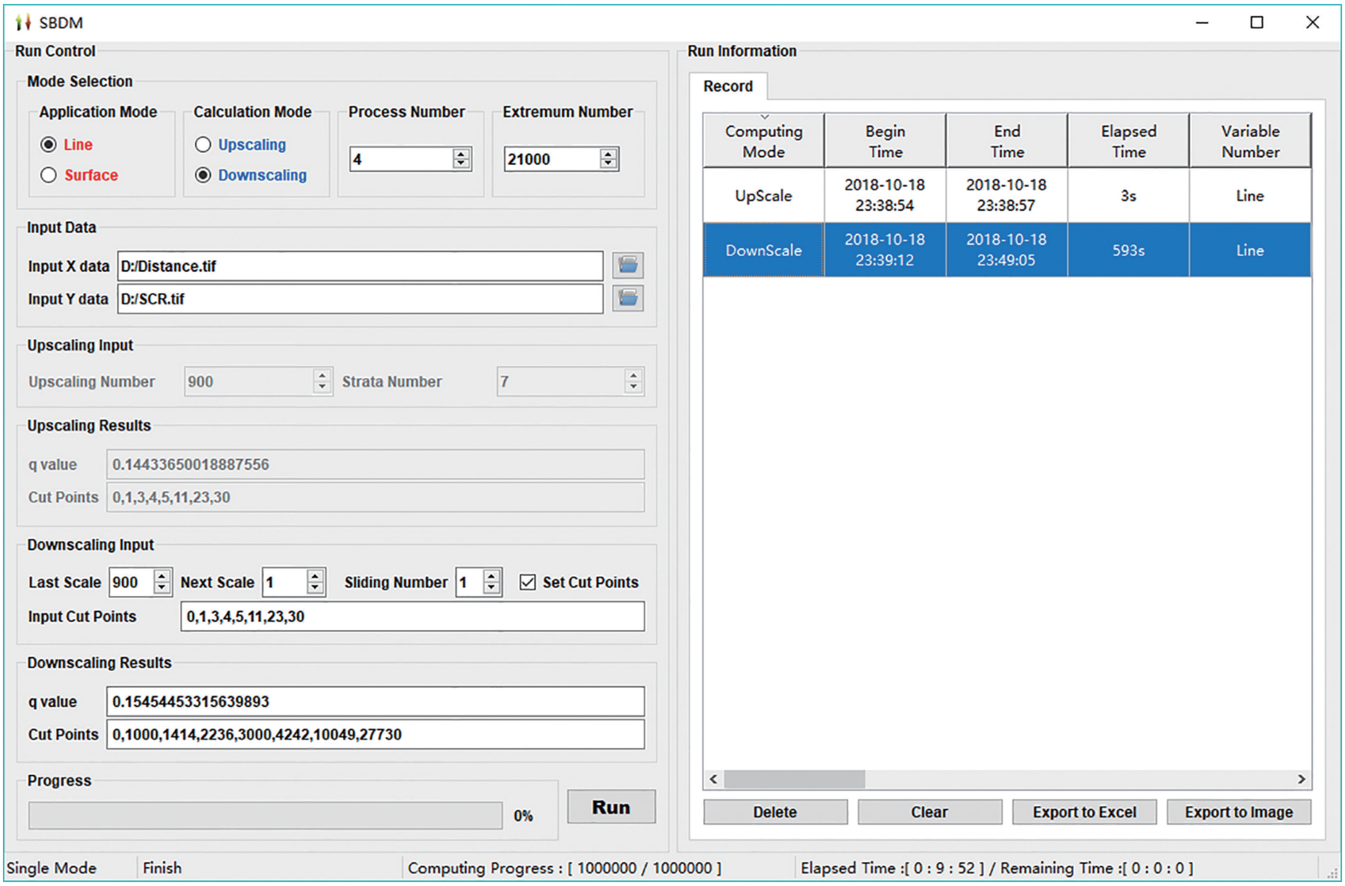


Figure 4. Two case studies. (a) Location of the two study areas (Maowusu (Mu Us) Sandy Land and Xinjiang) in China. (b) The spatial pattern of $NDVI$ (1 km × 1 km resolution) in Xinjiang, north-western China in 2013 processed by the MVC method. (c) Binary map (sand dune and vegetation zone) of the landscape and river distribution in the Maowusu (Mu Us) Sandy Land, northern China. (d) Sand Cover Ratio map derived from Fig. 4c via zonal statistics method (Liang and Yang, 2016).



SBDM

Run Control

Mode Selection

Application Mode: ● Line ○ Surface

Calculation Mode: ○ Upscaling ● Downscaling

Process Number: 4

Extremum Number: 21000

Input Data

Input X data: D:/Distance.tif

Input Y data: D:/SCR.tif

Upscaling Input

Upscaling Number: 900 Strata Number: 7

Upscaling Results

q value: 0.14433650018887556

Cut Points: 0,1,3,4,5,11,23,30

Downscaling Input

Last Scale: 900 Next Scale: 1 Sliding Number: 1 ☑ Set Cut Points

Input Cut Points: 0,1,3,4,5,11,23,30

Downscaling Results

q value: 0.15454453315639893

Cut Points: 0,1000,1414,2236,3000,4242,10049,27730

Progress: 0%

Run Information

Record

Computing Mode	Begin Time	End Time	Elapsed Time	Variable Number
UpScale	2018-10-18 23:38:54	2018-10-18 23:38:57	3s	Line
DownScale	2018-10-18 23:39:12	2018-10-18 23:49:05	593s	Line

Single Mode | Finish | Computing Progress : [1000000 / 1000000] | Elapsed Time :[0 : 9 : 52] / Remaining Time :[0 : 0 : 0]

Figure 5. The processing window of the SBDM software for obtaining the optimal strata of river distance to SCR in the Maowusu (Mu Us) Sandy Land, northern China.

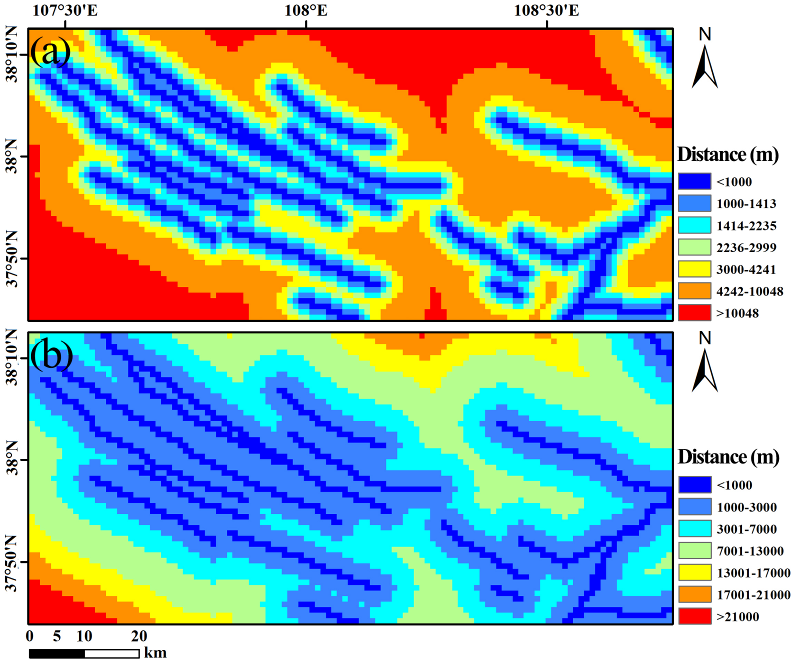


Figure 6. Stratified map of river distance map in the Maowusu (Mu Us) Sandy Land, northern China. (a) Stratified map of river distance obtained by SBDM. (b) Stratified map of river distance determined by Liang and Yang (2016).

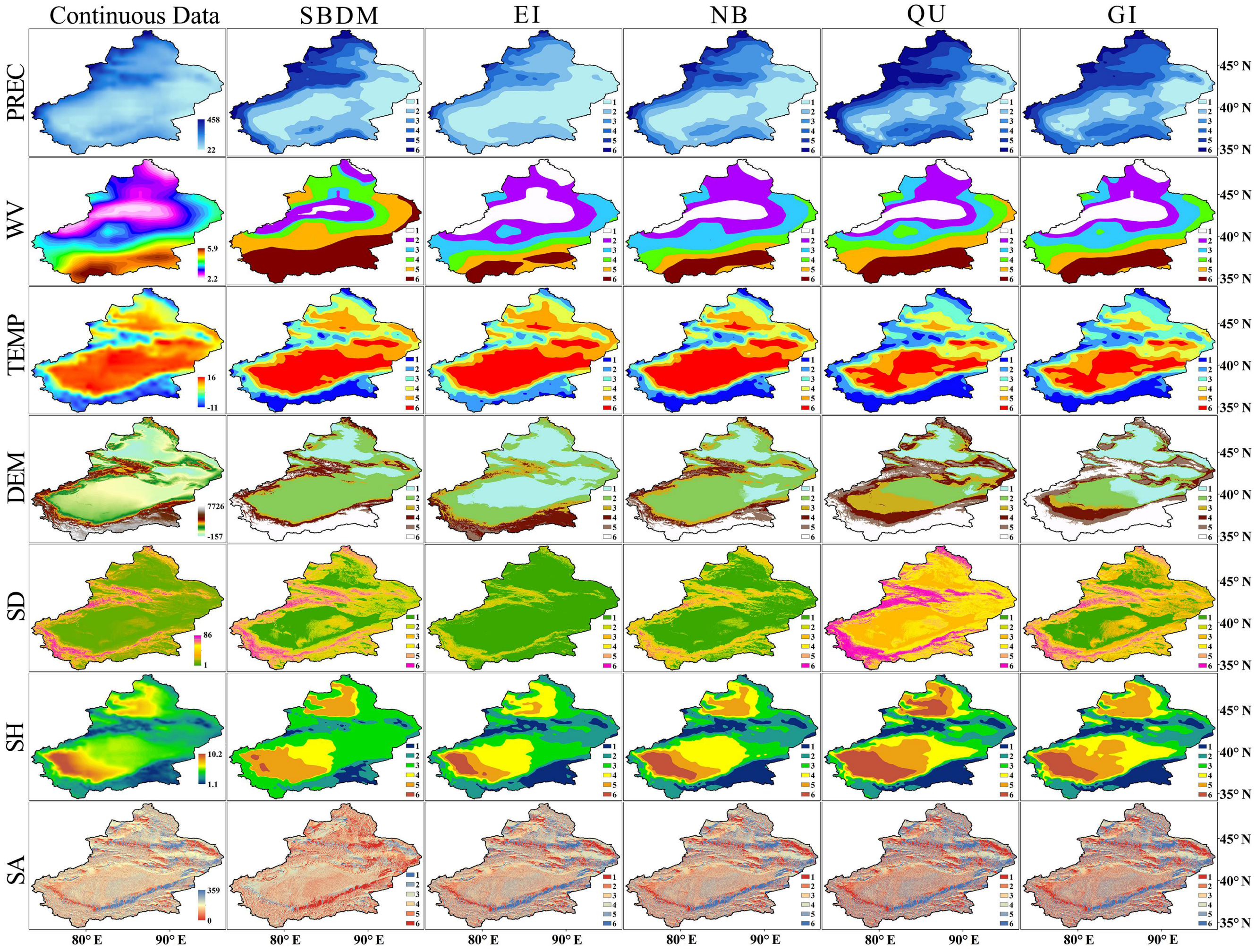


Figure 7. Stratified map of seven factors in Xinjiang, north-western China, with different discretization methods. The first column shows the raw continuous data distribution. Note the distinct differences of the stratified map from SBDM and other discretization methods.

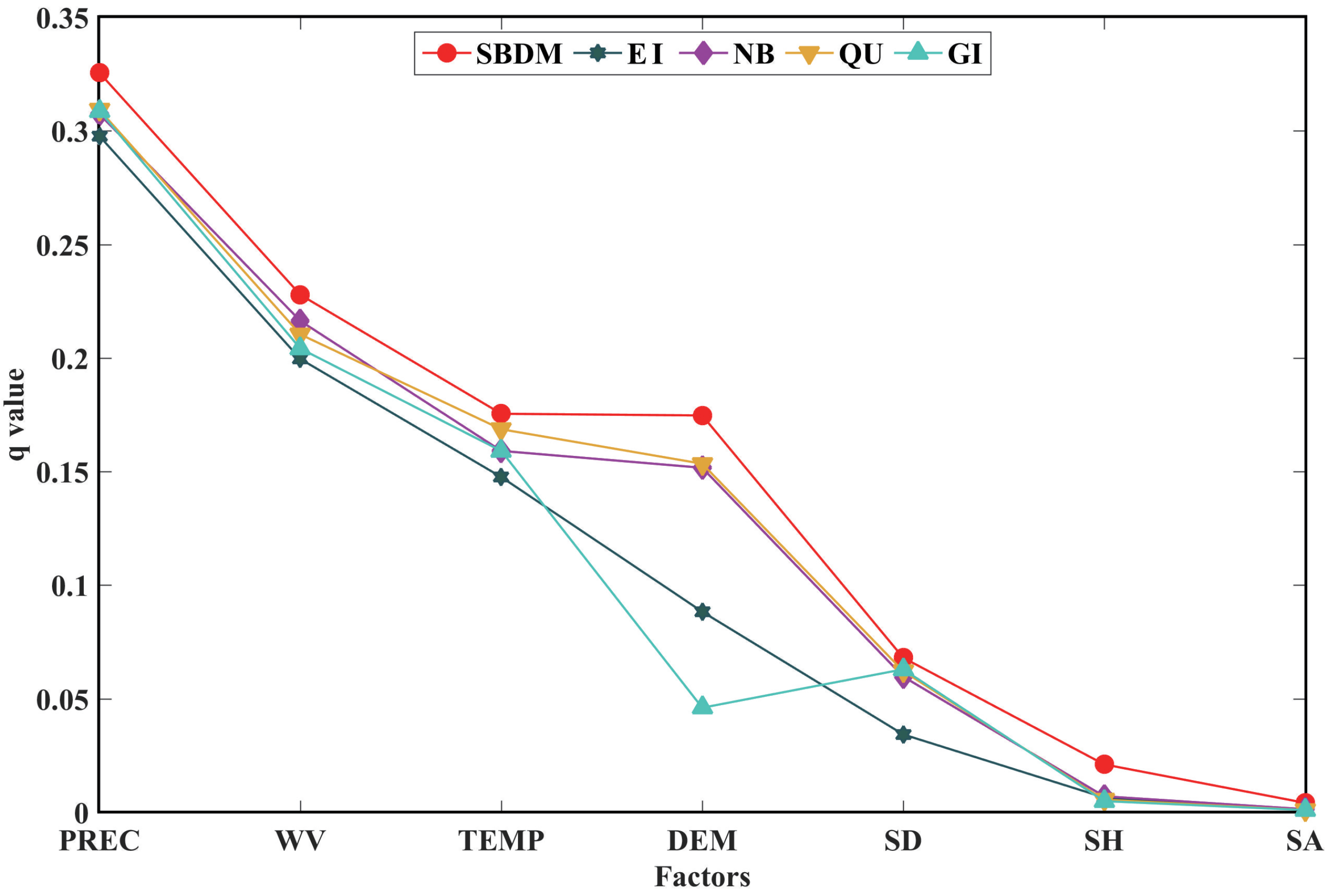


Figure 8. Comparison of the q values for seven factors to the contributions of NDVI spatial pattern in Xinjiang, north-western China, with different discretization methods (see $Factor\_Detector.xlsx$ in the Supporting Information for detailed q values and cut points). Note that the red dots are above other symbols, suggesting the larger q values from SBDM.

Table 1. Stratified results from SBDM and Prior Knowledge for the river impact distance to SCR in the Maowusu (Mu Us) Sandy Land, northern China.

SBDM	Strata	[0, 1000)	[1000, 1414)	[1414, 2236)	[2236, 3000)	[3000, 4242)	[4242, 10049)	[10049, 27730]
	$\overline{SCR}$	0.126	0.225	0.391	0.314	0.235	0.154	0.076
	q	0.154						
Prior Knowledge	Strata	[0, 1000)	[1000, 3000)	[3000, 7000)	[7000, 13000)	[13000,17000)	[17000, 21000)	[21000, 27730]
	$\overline{SCR}$	0.126	0.325	0.188	0.129	0.085	0.034	0.002
	q	0.136						

Table 2. A comparative analysis of the results for the effects of Specific Humidity (SH) factor on NDVI spatial pattern in Xinjiang, northwestern China, with and without downscaling processing of the SBDM software.

Scale	Path	Cut points							USTC	DSTC($\Delta\sigma$)			q value
										1	2	3	
6	—	1*	4*	5*	9*	10*	16*	17*	1 m	—	—	—	0.01574*
4	6-4	2&2*	5&5*	8&8*	14&14*	15&15*	24&24*	25&25*	7.6 m	23 s	35 s	1.5 m	0.01739&0.01739*
2	4-2	5&5*	11&11*	16&16*	28&28*	30&30*	49&49*	51&51*	5.3 h	25 s	2.6 m	2.8 m	0.01870&0.01870*
1	2-1	11&11*	21&21*	31&31*	57&57*	61&61*	97&97*	102&102*	7.6 d	34 s	6.1 m	24 m	0.02111&0.02111*

[a] The data with * markers are calculated without downscaling processing (only from direct upscaling calculation).

[b] The data without * markers are the results from downscaling based on the last scale, e.g., for SH factor, the results (cut points, time costing and q-value) of scale 4 is based on the results of scale 6 when downscaling is applied (downscaling path: 6-4) and scale 2 is based on scale 4 (downscaling path: 4-2).

[c] The USTC and DSTC represent upscaling and downscaling, respectively, time costing for different sliding numbers (($\Delta\sigma$).

[d] The letters “s“, “m“, “h“ and “d“ represent seconds, minutes, hours and days, respectively.