**An optimal parameters-based geographical detector model enhances geographic characteristics of explanatory variables for spatial heterogeneity analysis: Cases with different types of spatial data**

**Supplementary Information 1: Overview of the GD package**

The GD package contains a set of functions of the optimal parameters-based geographical detector (OPGD) model. Results of the GD package based analysis include all intermediate computation processes, spatial stratified analysis results, and the result visualization. The general computation process and relationships of functions for spatial stratified heterogeneity analysis are shown in Figure S1. The functions within GD are briefly described in Table S1, and the usage of functions together with arguments, output function and visualization function are listed in Table S2. The functions include four parts: discretization and optimal discretization, geographical detectors, one-step model and assessment of size effects of spatial units.



**Figure S1.** General calculation process and the relationships of functions in GD package

**Table S1.** Summary of functions in the GD package

|  |  |
| --- | --- |
| **Function** | **Description** |
| *disc* | Function for discretizing continuous data and obtaining the different outputs, including discretization intervals, numbers of values within intervals, and visualization of discretization. |
| *optidisc* | Optimal discretization for continuous variables and visualization. |
| *gd* | Function for calculating power determinant using factor detector of geographical detectors and visualization. |
| *riskmean* | Function for calculating risk means within intervals and visualization. |
| *gdrisk* | Function for risk detector calculation, risk matrix and visualization. |
| *gdinteract* | Function for interaction detector calculation and visualization. |
| *gdeco* | Function for ecological detector calculation, ecological matrix and visualization. |
| *gdm* | A one-step function for optimal discretization and geographical detectors for multiple variables and visualization. |
| *sesu* | Function for comparison of size effects of spatial units in spatial heterogeneity analysis. |

**Table S2.** Usage of functions in the GD package

|  |  |  |  |
| --- | --- | --- | --- |
| **Function** | **Usage** | **Arguments** | **Output and visualization functions** |
| *disc* | *disc(var, n, method = "quantile", ManualItv)* | ***var***: A numeric vector of continuous variable  ***n***: The number of intervals  ***method***: A character of discretization method. Both supervised and unsupervised discretization methods are available in the function. The supervised discretization methods include equal breaks, natural breaks, quantile breaks, geometric breaks and standard deviation breaks, and the unsupervised method supports manually defined breaking intervals.  ***ManualItv***: A numeric vector of manual intervals | ***print***() and ***plot***() |
| *optidisc* | *optidisc(formula, data, discmethod = discmethod, discitv = discitv)* | ***formula***: A formula of response and explanatory variables  ***data***: A data.frame includes response and explanatory variables ***discmethod***: A character vector of discretization methods  ***discitv***: A numeric vector of numbers of intervals | ***print***() and ***plot***() |
| *gd* | *gd(formula, data = NULL)* | ***formula***: A formula of response and explanatory variables  ***data***: A data.frame includes response and explanatory variables | ***print***() and ***plot***() |
| *riskmean* | *riskmean(formula, data = NULL)* |
| *gdrisk* | *gdrisk(formula, data = NULL)* |
| *gdinteract* | *gdinteract(formula, data = NULL)* |
| *gdeco* | *gdeco(formula, data = NULL)* |
| *gdm* | *gdm(formula, continuous\_variable = NULL, data = NULL, discmethod, discitv)* | ***formula***: A formula of response and explanatory variables  ***continuous\_variable***: A vector of continuous variable names  ***data***: A data.frame includes response and explanatory variables  ***discmethod***: A character vector of discretization methods  ***discitv***: A numeric vector of numbers of intervals | ***print***() and ***plot***() |
| *sesu* | *sesu(gdlist, su)* | ***gdlist***: A list of gdm result or gd result  ***su***: A vector of sizes of spatial units |  |

If the explanatory variables contain continuous variables, the continuous variables should be discretized. The GD package provides two options of discretization: discretization with the user defined parameters, i.e. the combination of a discretization method and a break number, and optimal discretization that the best parameter combination is selected from a series of combinations.

1. For the discretization with the user defined parameters, the ***disc*** function provides five supervised discretization methods, including equal breaks, natural breaks, quantile breaks, geometric breaks and standard deviation breaks, and the unsupervised methods that the breaking intervals can be manually defined. The ***disc*** function also visualizes the discretization results.
2. For the optimal discretization process, users can provide a series of combinations of the discretization methods and the numbers of intervals, then utilize the ***optidisc*** function to select the best parameter combinations for discretizing variables. In addition, the process of selecting the best parameter combinations and the discretization results can be visualized with the ***optidisc*** function.

Once the continuous variables are discretized, the next step is to perform the four parts of geographical detectors: factor detector, interactive detector, risk detector and ecological detector. Functions in the four parts of geographical detectors are explained as follows.

1. The ***gd*** function is used to calculate values of variables, together with the significance level.
2. For the risk detector, the ***riskmean*** function generates the mean risk values of sub-regions, and the ***gdrisk*** function assesses the significant difference of risks among sub-regions with the results of t-test value, degree of freedom, significance and the risk factor of two sub-regions. If the difference between two sub-regions is significant within a threshold of significance level (e.g. 0.05), the risk factor of two sub-regions is marked with “Y”, otherwise, it is marked with “N”. The function also forms a matrix of the risk factors and visualizes the matrix.
3. The ***gdinteract*** function is applied on computing the interactive impact of two variables. The results include the respective values of two variables, the value of the interaction, and the type of interaction, such as nonlinear enhance.
4. For the ecological detector, the ***gdeco*** function evaluates if the impacts of two explanatory variables are significantly different with the results of F-test value, significance and the ecological factor of two variables. The function also generates a matrix of ecological factors and visualizes the matrix.

In addition, for ease of use the package, a one-step function ***gdm*** is provided for straight forward performing both optimal discretization and geographical detectors to derive all analysis results and visualizations. Results of the one-step function contains all intermediate computation processes and the OPGD-based analysis.