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

Yongze Song [a,*], Jinfeng Wang [b], Yong Ge [b] and Chengdong Xu [b]

[a] *School of Design and the Built Environment, Curtin University, Perth 6845, Australia; yongze.song@curtin.edu.au (Y.S.);*

[b] *State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Nature Resources Research, Chinese Academy of Sciences, Beijing 100101, China; wangjf@lreis.ac.cn (J.W.); gey@lreis.ac.cn (Y.G.); xucd@lreis.ac.cn (C.X.)*

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

Spatial heterogeneity represents a general characteristic of the inequitable distributions of spatial issues. The spatial stratified heterogeneity analysis investigates the heterogeneity among various strata of explanatory variables by comparing the spatial variance within strata and that between strata. The geographical detector model is a widely used technique for spatial stratified heterogeneity analysis. In the model, the spatial data discretization and spatial scale effects are fundamental issues, but they are generally determined by experience and lack accurate quantitative assessment in previous studies. To address this issue, an optimal parameters-based geographical detector (OPGD) model is developed for more accurate spatial analysis. The optimal parameters are explored as the best combination of spatial data discretization method, break number of spatial strata, and spatial scale parameter. In the study, the OPGD model is applied in three example cases with different types of spatial data, including spatial raster data, spatial point or areal statistical data and spatial line segment data, and an R "GD" package is developed for computation. Results show that the parameter optimization process can further extract geographical characteristics and information contained in spatial explanatory variables in the geographical detector model. The improved model can be flexibly applied in both global and regional spatial analysis for various types of spatial data. Thus, the OPGD model can improve the overall capacity of spatial stratified heterogeneity analysis. The OPGD model and its diverse solutions can contribute to more accurate, flexible and efficient spatial heterogeneity analysis, such as spatial patterns investigation and spatial factor explorations.

Keywords: GIS; spatial analysis; geographical detector; spatial stratified heterogeneity; spatial factors exploration; R package GD

1 Introduction

Spatial heterogeneity is a common property of geographical phenomena. It refers to the uneven distributions of various geospatial attributes within a certain geographical area (Fischer 2010, Wang, Zhang, and Fu 2016). Spatial heterogeneity analysis is widely used in the spatial and spatiotemporal issues in fields of ecology, geology, public health, economy, built environment, etc. The objectives of spatial heterogeneity analysis usually consist of three aspects. The first objective is to explore spatial clusters that are generally defined as spatially high or low value regions (Anselin 1995). Second, spatial heterogeneity analysis can be used to investigate potential factors associated with the uneven spatial distributions (Brunsdon, Fotheringham, and Charlton 1996, Fotheringham, Brunsdon, and Charlton 2003). The third objective includes spatial and spatiotemporal prediction and decision-making based on the spatial heterogeneity (Wang et al. 2014).

In general, spatial heterogeneity can be measured from three perspectives. First, spatial heterogeneity with local clusters is a popular approach that explores the spatially local clustering regions with similarity in geographical attributes. For instance, spatial autocorrelation indicators, such as local indicators of spatial association (LISA) (Anselin 1995) and Getis-Ord Gi (Getis and Ord 1992, Ord and Getis 1995), are used to evaluate if a geographical attribute is spatially clustered. Spatial scan statistics detect spatial clusters by comparing the likelihood ratio within and out of dynamically changed moving windows (Kulldorff 1997). Geographically weighted regression (GWR) and its extended models measure geographically local effects by location-wise coefficients of explanatory variables with distance-decay weights across space (Fotheringham, Brunsdon, and Charlton 2003, Brunsdon, Fotheringham, and Charlton 1996, Huang, Wu, and Barry 2010, Lu et al. 2014, Lu et al. 2017, Ge et al. 2017). The second approach is the spatial stratified heterogeneity analysis, which compares the spatial variance within strata and that between strata (Wang et al. 2010, Wang, Zhang, and Fu 2016). The spatial stratified heterogeneity can be quantified by the geographical detector model (Wang et al. 2010, Luo et al. 2016). The primary advantage of spatial stratified heterogeneity analysis is that no assumptions are required for geographical variables and it reflects the real spatial associations of geographical attributes. Third, spatial scaling structure heterogeneity is a method of characterizing complexity of fractal or scaling structure of geographical attributes (Jiang 2013, 2015). Based on the scaling law that far more small geographical objects exist than large ones, an ht-index is proposed to measure the spatial scaling structure heterogeneity (Jiang and Yin 2014).

The geographical detector model is a promising approach and a primary tool for the spatial stratified heterogeneity analysis. The main idea of geographical detectors is that the study space is divided into sub-regions by variables, and the spatial variance within each sub-region and among different sub-regions are compared to evaluate the determinant power of potential explanatory variables (Wang et al. 2010, Wang, Zhang, and Fu 2016). The general geographical detectors include four parts, where the core part is the factor detector that quantifies the relative importance of different geographical variables. Other three parts are interaction detector, risk detector and ecological detector.

To comprehensively understand applications and model improvements of the geographical detector model, the application trend of the model is reviewed using the Clarivate Analytics' Web of Science database in September 2019. The search is limited to the "English" language and the "topic" search equation is: "geographical detector" OR "geographical detectors" OR "geodetector". As a result, 130 research papers are yield ranging from 2010 to 2019. The overview of global research using the geographical detector model is presented in Figure 1. The annual variation of research using the geographical detector model is compared with the variation of papers citing the publication first proposing the model (Wang et al. 2010), which accumulate to 213 based on the database of the Web of Science. The conceptual structure map generated by the "bibliometrix" R package (Aria and Cuccurullo 2017) presents primary application fields of the geographical detector model. In general, applications of the model are predominant in geographically local determinants or factors exploration, and spatial patterns and heterogeneity investigation. Research topics can be clustered into three categories. The first topic is about disease determinants analysis, air pollution sources studies, and the association between air pollution and disease. The second one includes climate change research and land use driving forces exploration. The last category is water resources and dynamics modelling, such as runoff and precipitation variations. In the current stage, applications of the geographical detector model are primarily clustered in the fields of public health and environment. Therefore, it is necessary to broad the application fields of the model to enhance its capabilities in explaining geographical objects in other fields and integrating with other models. Simultaneously, more studies about improving the model are required for optimal parameters selection and more flexible, applicable and effective studies.

[Figure 1 near here]

In geographical studies, explanatory variables can be continuous and categorical variables, where the continuous variables should be discretized and converted to categorical variables in the geographical detector model. Spatial data discretization is to divide continuous

geographical and geospatial data into several intervals according to physical or statistical characteristics of the data, so that the continuous variable is converted to a categorical variable (Cao, Ge, and Wang 2013). Two common methods for the spatial data discretization are supervised and unsupervised discretization methods. The supervised discretization methods break continuous variables according to certain statistical regulars, such as equal breaks, natural breaks, quantile breaks, geometric breaks and standard deviation breaks. For the unsupervised methods, breaking intervals can be manually defined. The result of spatial discretization for a continuous variable is associated with discretization methods and break numbers (Cao, Ge, and Wang 2013, Ju et al. 2016). Currently, spatial data discretization process is generally performed in terms of professional experience instead of data-driven approaches (Ding et al. 2019, Luo et al. 2019, Duan and Tan 2020). In addition, the spatial scale effect is common in geographical issues and may have critical impacts on the spatial stratified heterogeneity analysis, but it has not been fully investigated and integrated in the model.

To address above issues, this study develops an optimal parameters-based geographical detector (OPGD) model for improving accuracy and effectiveness of spatial analysis. In the OPGD model, the process of spatial data discretization and spatial scales for spatial analysis are optimized and the best parameter combination is determined for the geographical detector model. The OPGD model can provide flexible and comprehensive solutions with a series of visualizations for more effective spatial factor explorations, and spatial patterns and heterogeneity investigation than the geographical detectors model. In the study, the OPGD model is applied in three example cases with different types of spatial data, including spatial raster data, spatial point or areal statistical data and spatial line segment data.

This paper is organized as follows. Section 2 presents a review of the geographical detector model development and applications. Section 3 elaborates the developed OPGD model and its mathematical basis. Section 4 describes different types of spatial data used in three cases for explaining the OPGD model. Section 5, 6 and 7 present the results, discussion and conclusions of the study.

2 Optimal parameters-based geographical detector (OPGD) model

The OPGD model includes five parts: factor detector, parameters optimization, interaction detector, risk detector and ecological detector. The parameters optimization consists of the optimization of spatial discretization and optimization of spatial scale. The schematic

overview of the OPGD is shown in Figure 2, and five parts of the model are explained in following subsections.

[Figure 2 near here]

2.1 Factor detector

As the core part of geographical detector, the factor detector reveals the relative importance of explanatory variables with a Q-statistic. The Q-statistic compares the dispersion variances between observations in the whole study area and strata of variables (Wang et al. 2010, Wang, Zhang, and Fu 2016). The Q value of a potential variable v is computed by:

$$Q_v = 1 - \frac{1}{(N_v-1)\sigma_v^2}\sum_{j=1}^{M}(N_{v,j}-1)\sigma_{v,j}^2 \quad (1)$$

where N_v and σ_v^2 are the number and variance of observations within the whole study area, and $N_{v,j}$ and $\sigma_{v,j}^2$ are the number and variance of observations within the jth ($j = 1, \dots, M$) sub-region of variable v. A large Q value means the relatively high importance of the explanatory variable, due to a small variance within sub-regions and a large variance between sub-regions. In the geographical detector, at least two samples are required in each of strata to compute mean and variance values.

The F-test is utilized to determine whether the variances of observations and stratified observations are significantly different, since the transformed Q value can be tested with the non-central F-distribution:

$$F = \frac{N-M}{M-1}\frac{Q}{1-Q} \sim F(M-1, N-M; \delta) \quad (2)$$

where M is the number of sub-regions, N is the number of observations, and δ is the non-central parameter:

$$\delta = \left[\sum_{j=1}^{M}\bar{Y}_j^2 - \frac{1}{N}\left(\sum_{j=1}^{M}\bar{Y}_j\sqrt{N_j}\right)^2\right]/\sigma^2 \quad (3)$$

where $\bar{Y}_j$ is the mean value of observations within the jth sub-region of variable. Thus, with the given significant level, the null hypothesis $H_0: \sigma_v^2 = \sigma_{v,j}^2$ can be tested by checking $F(M-1, N-M; \delta)$ in the distribution table.

2.2 Parameters optimization

The parameters optimization consists of the optimization of spatial discretization and optimization of spatial scale. In this study, the OPGD model selects a best combination of discretization method and the break number for each geographical continuous variable as the optimal discretization parameters. The Q value computed with the factor detector is used to determine the best parameter combinations. A set of combinations of discretization methods and break numbers are provided for each continuous variable to compute respective Q values.

The optional discretization methods can be a list of supervised and unsupervised discretization methods, and optional break numbers can be an integer sequence in terms of observations and practical requirements. As such, the optional combinations can cover almost all available choices. For a continuous variable, the parameter combination with the highest Q value among all combinations is selected for spatial discretization, since it presents the highest importance of the variable from the perspective of spatial stratified heterogeneity.

The optimization of spatial scale aims at identifying the optimal spatial scale for the spatial stratified heterogeneity analysis. A geographical variable at different spatial scales probably reveal significantly varied geographical characteristics (Roth, Allan, and Erickson 1996, Store and Jokimäki 2003, Chen et al. 2016). The Q values of all explanatory variables with respective optimal spatial discretization parameters at various spatial scales are compared with corresponding spatial scales to investigate their relationships. The assumption of optimal spatial scale selection is that Q values are the highest for most explanatory variables. In the study, the 90% quantile of Q values of all explanatory variables at a spatial scale is computed and used for the comparison of overall Q value trends at different spatial scales. For a set of optional spatial scales, the optimal one is selected when the 90% quantile of Q values of all explanatory variables reach the highest value.

2.3 Interaction detector

The interaction detector determines the interactive impacts of two overlapped spatial variables based on the relative importance of interactions computed with Q values of the factor detector. A spatial interaction is an overlay of two spatial explanatory variables. The interaction detector explores an interaction by the comparison between Q values of the interaction and two single variables. The interactions explain whether the impacts of two spatial variables are weakened, enhanced or independent. The interaction detector explores five interactions, including nonlinear-weaken, uni-variable weaken, bi-variable enhance, independent and nonlinear-enhance (Wang et al. 2010, Wang, Zhang, and Fu 2016) (Table 1). Therefore, the interaction detector result includes both Q values of interactions and types of interaction effects.

[Table 1 near here]

2.4 Risk detector

The risk detector is used to test if spatial patterns represented by mean values are significantly different among sub-regions classified by a categorical or stratified variable. The

difference between mean values of sub-regions η and κ is tested with the *t*-test (Wang et al. 2010, Wang, Zhang, and Fu 2016):

$$t_{\bar{Y}_\eta - \bar{Y}_\kappa} = \left(\bar{Y}_\eta - \bar{Y}_\kappa\right) \Big/ \sqrt{\frac{\sigma^2_{\bar{Y}_\eta}}{N_\eta} + \frac{\sigma^2_{\bar{Y}_\kappa}}{N_\kappa}} \quad (4)$$

where $\bar{Y}_\eta$ and $\bar{Y}_\kappa$ are mean values of observations within sub-regions η and κ, $\sigma^2_{\bar{Y}_\eta}$ and $\sigma^2_{\bar{Y}_\kappa}$ are the variance, and N_η and N_κ are numbers of observations, respectively. The statistic is approximately distributed as Student's *t* with the degree of freedom of:

$$df = \left(\frac{\sigma^2_{\bar{Y}_\eta}}{N_\eta} + \frac{\sigma^2_{\bar{Y}_\kappa}}{N_\kappa}\right) \Big/ \left[\frac{1}{(N_\eta - 1)}\left(\frac{\sigma^2_{\bar{Y}_\eta}}{N_\eta}\right)^2 + \frac{1}{(N_\kappa - 1)}\left(\frac{\sigma^2_{\bar{Y}_\kappa}}{N_\kappa}\right)^2\right] \quad (5)$$

Thus, with a given significant level, the null hypothesis $H_0: \bar{Y}_\eta = \bar{Y}_\kappa$ can be tested with the student-*t* distribution table.

2.5 Ecological detector

The ecological detector is used to test if an explanatory variable has a higher impact than another one. The significance of the different influence of explanatory variables is tested with the *F*-statistic (Wang et al. 2010, Wang, Zhang, and Fu 2016):

$$F = \frac{N_u (N_v - 1) \sum_{j=1}^{M_u} N_{u,j} \sigma^2_{u,j}}{N_v (N_u - 1) \sum_{j=1}^{M_v} N_{v,j} \sigma^2_{v,j}} \quad (6)$$

where N_u and N_v are numbers of observations, M_u and M_v are numbers of sub-regions, and $\sum_{j=1}^{M_u} N_{u,j} \sigma^2_{u,j}$ and $\sum_{j=1}^{M_v} N_{v,j} \sigma^2_{v,j}$ are sums of variance within sub-regions of variables u and v respectively. Thus, with a given significant level, the null hypothesis $H_0: \sum_{j=1}^{M_u} N_{u,j} \sigma^2_{u,j} = \sum_{j=1}^{M_v} N_{v,j} \sigma^2_{v,j}$ is tested with the *F*-distribution table.

In this study, an open-source software package "GD" in R is developed for systematic computation and visualization of the OPGD model. The general calculation process, functions and their relationships in the GD package are introduced in the Supplementary Information 1: Overview of the GD package.

3 Data

The OPGD model can be flexibly applied in the spatial factors exploration and heterogeneity analysis for various types of spatial data. In this study, three example cases with different types of spatial data, including spatial raster data, spatial areal statistical data and spatial line segment data, are investigated using the OPGD model (Table 2). The first case dataset is to investigate impacts of potential variables of human actives and climate on the

vegetation changes, where vegetation coverage conditions are quantified by the normalized difference vegetation index (NDVI), which is a spatial raster variable. The second case is assessing associations between incidence variations of influenza A virus subtype H1N1, a spatial point or areal data, and potential explanatory variables of meteorological conditions and human activities. The third case examines relationships between road damage and variables of vehicles and environment with the spatial line segment data. Descriptions and data sources of the example cases are presented in the following subsections.

[Table 2 near here]

3.1 Spatial raster data of vegetation changes

A major application topic of the spatial stratified heterogeneity analysis is the environment, ecology and forest studies (Ren et al. 2014, Ren et al. 2016). In recent years, an increasing number of researches investigate the comprehensive impacts of human activities and climate conditions on vegetation changes (Du et al. 2017). In this study, vegetation changes are explored using the spatial gridded annual mean NDVI changes from 2010 to 2014 in Inner Mongolia, China, where is one of the major mining regions in China. Respective contributions of human activities and climate conditions on the NDVI changes are explored using the OPGD model. The spatial raster map of NDVI changes and distributions of explanatory variables are shown in Figure 3.

[Figure 3 near here]

The NDVI raster data is derived from the SPOT Vegetation 1-km NDVI Dataset for China since 1998. The climate variables include temperature changes and annual average precipitation from 2010 to 2014, and the climate zone data. The temperature and precipitation data are sourced from the Annual Average Temperature Spatial Interpolation Dataset for China since 1980, and the Annual Precipitation Spatial Interpolation Dataset for China since 1980. The datasets of NDVI, temperature and precipitation are all provided by Data Center for Resources and Environment Science, Chinese Academy of Sciences (RESDC) (http://www.resdc.cn). The climate zone data is from the CliMond Dataset: World Map of The Koppen-Geiger Climate Classification (Kriticos et al. 2012). In the study area, there are five climate zones, including cold desert climate (Bwk), cold semi-arid climate (Bsk), monsoon-influenced humid subtropical climate (Dwa), subtropical highland climate (Dwb) and cold subtropical highland climate (Dwc). Human activity variables consist of coal mining production, gross domestic product (GDP) and population density. County-level annual coal mining production is the average of coal production data from 2011 to 2014, sourced from the

from the Annual Reports of China National Coal Association (www.coalchina.org.cn). Since the data of coal production smaller than 10^7 ton are not available, the variable of coal production classifies the production into five categories, very low, low, medium, high and very high, for reasonable spatial comparison with other explanatory variables. The 1-km gridded GDP data comes from the Gridded Global Datasets for Gross Domestic Product and Human Development Index over 1990-2015 (Kummu, Taka, and Guillaume 2018b, Kummu, Taka, and Guillaume 2018a), and the 1-km gridded population density data is from the Gridded Population of the World, Version 4 (GPWv4) (Center for International Earth Science Information Network - CIESIN - Columbia University 2016). Among the explanatory variables, climate zone data and coal mining production are categorical variables, and others are continuous variables. In addition, six sizes of grids are generated for the NDVI changes data, including 5 km, 10 km, 20 km, 30 km, 40 km and 50 km, to examine which size of grid can better reveal the impacts of potential variables on the changes of NDVI.

3.2 Spatial point or areal data of H1N1 flu incidences

Spatial point or areal data are widely used in spatial analysis. This study explores potential variables of H1N1 flu incidences derived from spatial areal statistical data based on administrative units. The H1N1 flu incidences are collected with provincial statistics in 2013 in China. The explanatory variables include meteorological and environmental variables, and the socio-economic variables. To investigate spatial scale effects, the analysis is performed at 50-km, 100-km and 150-km spatial grids, respectively. Spatial areal data of H1N1 flu incidences and distributions of explanatory variables are mapped in Figure 4.

[Figure 4 near here]

The H1N1 flu incidences data are provincial statistical data sourced from the China Health Statistical Yearbook (National Health Commission of the People's Republic of China 2014). Explanatory variables contain two categories: meteorological and environmental variables, and socio-economic variables. First, the meteorological and environmental data include geographical region and annual average temperature, precipitation and moisture index. The Chinese provinces are categorized into three geographical regions: north-east and north, central and south, and western China. The annual average temperature, precipitation and moisture index data are sourced from the Annual Average Temperature Spatial Interpolation Dataset for China since 1980, the Annual Precipitation Spatial Interpolation Dataset for China since 1980, and the Chinese Meteorological Background - Humidity Index Data. The datasets of temperature, precipitation and humidity index are all provided by Data Center for Resources

and Environment Science, Chinese Academy of Sciences (RESDC) (http://www.resdc.cn). In addition, the socio-economic data consist of the population density, GDP, road density, percentages of sensitive people (children and elders) and urban population among total population, medical cost per capita and the urban-rural consumption ratio. The 1-km gridded GDP data comes from the Gridded Global Datasets for Gross Domestic Product and Human Development Index over 1990-2015 (Kummu, Taka, and Guillaume 2018b, Kummu, Taka, and Guillaume 2018a), and the 1-km gridded population density data is from the Gridded Population of the World, Version 4 (GPWv4) (Center for International Earth Science Information Network - CIESIN - Columbia University 2016). The road density map with the spatial resolution of 1 km is computed with the kernel density function based on the road network distribution. The populations of child and the old are the people younger than 14 years old and older than 65 years old in 2013, respectively (National Bureau of Statistics of China 2015). The sensitive people include both child and the old. The percentage of urban population, medical cost per capita and the urban-rural consumption ratio are all sourced from the China Statistical Yearbook in 2014 (National Bureau of Statistics of China 2015).

3.3 Spatial line segment data of road damage

In addition to the spatial raster data and point or areal data, spatial analysis for line segment data is performed using the OPGD model. In the study, spatial line segment based road damage conditions and potential variables are selected from the road deterioration datasets in the Wheatbelt region in Western Australia, Australia (Song et al. 2018, Song et al. 2019). The road damage conditions are described with the deflection of pavement, which is measured with a Dynatest 8000 series Falling Weight Delectometer (FWD) and calibrated with Calibration Method WA 2060.5 by Main Roads, WA (Main Roads Western Australia 2017a, b). Deflection is a pavement strength indicator that describes the maximum depression on the surface of pavement under a standard load. Explanatory variables include road speed limits, soil types, population within 1 km around the road segments, and annual mean daily volumes of vehicles. Soil type data is sourced from the State of the Environment (SoE) Land Australian Soil Classification Orders dataset in 2016 (Ashton and McKenzie 2001, State of the Environment in Australia 2017). Population within 1 km around the road segments is computed with the population data with 1-km spatial resolution is from Gridded Population of the World fourth version (GPWv4) (Center for International Earth Science Information Network - CIESIN - Columbia University 2016). Traffic volumes are estimated with a segment-based regression kriging (SRK) method. The SRK method is an improved regression kriging method by

integrate the spatial morphological characteristics of road segments and regression kriging model for more accurate spatial prediction of line segment-based observations, such as traffic and road attributes (Song et al. 2019).

4 Results

4.1 Spatial raster data of vegetation changes

In this study, spatial explanatory variables of vegetation changes are investigated using the OPGD model. The OPGD model can simultaneously deal with both categorical and continuous explanatory variables in practical spatial analysis, where categorical variables can be directly used in the geographical detector model, but continuous variables should be discretized with optimal parameters before modelling. Thus, the first step of the OPGD model is the spatial discretization parameters optimization for continuous variables (Figure 5). Results show that the optimal parameter combinations of discretization methods and break numbers are varied for different explanatory variables. The optimal parameter combination for temperature change, precipitation and GDP is the natural break with seven intervals, and that for population density is the quantile break with seven intervals. With the spatial discretization parameters, continuous variables are converted to strata variables, which are equivalent to categorical variables in the geographical detector model. Codes and completed analysis results are provided in the Supplementary Information 2: Computation process of example cases.

[Figure 5 near here]

The next step is to identify contributions of single variables on vegetation changes using the factor detector. Factor detector results include Q values, corresponding significances, and ranks of variables, where the variable (precipitation) with the highest Q value compred with other explanatory variables is highlighted (Figure 6).

[Figure 6 near here]

In the third step, the risk detector provides risk means of spatial zones determined by variables and tests if the risk means of various spatial zones are significantly different (Figure 7). Risk detector results show that data within different intervals of an explanatory variable have significantly varied effects on vegetation changes. For instance, vegetation change in the cold subtropical highland climate (Dwc) region is 0.445, but that in the cold desert climate (Bwk) region is 0.005. To further investigate risk regions of vegetation changes, spatial distributions of risks determined by explanatory variables are mapped on Figure 7b. The variables determined mean vegetation changes are classified into three levels: high values (red), medium values (gray) and low values (blue). Spatial patterns of risk regions explored by

variables tend to be similar that the vegetation change in the eastern region is relatively high and that in the western region is low. However, local patterns explored by various variables are different. For instance, in the northeast region, climate variables, including climate zone, temperature change and precipitation, have more effects on vegetation changes compared with other variables. In the eastern and southern regions, high vegetation changes are closely associated with human activities, such as GDP and population density.

[Figure 7 near here]

The last two parts are interactions between variables explored by the interaction detector (Figure 8a), and the ecological matrix derived by the ecological detector (Figure 8b). In the interaction effect analysis, the interaction with the highest Q value (0.915) is that between precipitation and mining activities. The intermediate computation processes the *t*-test for risk detector, interactions explored by the interaction detector and the *F*-test for ecological detector are presented in the Supplementary Information 2.

[Figure 8 near here]

Finally, when spatial units are grids, a common method for selecting a reasonable grid size is to compare size effects of spatial units using the factor detector. In the study, six sizes of gridded data are contained in the vegetation changes dataset. Figure 9 shows the comparison of the size effects of spatial units. Results show that the Q values of most of the variables are increased from the 5-km to 40-km spatial unit. The 90% quantile of Q values reaches to the highest value when the spatial unit is 40 km and becomes lower after 40-km spatial unit. Thus, we recommend using 40 km as an optional spatial unit for the spatial stratified heterogeneity analysis.

[Figure 9 near here]

In summary, this case study has following findings according to the OPGD-based analysis. First, 40-km spatial grid is an optimal spatial unit for assessing impacts of human activities and climate change on vegetation changes in the study area. In addition, precipitation is the variable with the highest association with the vegetation change. Precipitation and mining activities are enhanced by each other in affecting vegetation change, and their interaction is the major interactive variables in the study area. The variation of vegetation in the north-eastern regions is closely associated with climate variables, and that in the eastern and southern regions is linked with human activity variables, such as GDP and population density.

4.2 Spatial point or areal data of H1N1 flu incidences

The H1N1 flu incidences case is used to demonstrate the OPGD-based analysis for point or areal data, and the comparison of spatial analysis for the whole study area (section 4.2.1) and for geographical sub-regions (section 4.2.2). Full results of the OPGD-based analysis are provided in the Supplementary Information 2.

4.2.1 In the whole study area

In the study, thirteen potential explanatory variables are collected for the analysis of H1N1 flu incidences. The geographical region is a categorical variable, and other environmental and socio-economic conditions presented in Figure 4 are all continuous variables. Results of spatial analysis in the whole study area are presented in Figure 10. The OPGD-based analysis for the whole study area consists of six parts: spatial scale effects analysis, spatial discretization optimization, factor detector, risk detector, interaction detector and ecological detector. The comparison of size effects indicates that relative impacts of meteorological and socio-economic factors are varied with the change of spatial units. In general, Q values of variables temperature, medical cost, and percentage of sensitive population are major contributors to the flu incidence, and they reach the maximum values when the spatial unit is 100 km. The 90% quantiles of Q values show a similar trend. Thus, we recommend choosing 100 km as the optimal spatial unit for spatial analysis. In detail, temperature is the primary contributor to the H1N1 flu incidences. The socio-economic variables medical cost and percentage of sensitive population have higher impacts than meteorological variable precipitation when the spatial unit is smaller than 100 km. The impact of road density is continuously increased with the increase of spatial unit, since the spreading of flu becomes easier with the growth of spatial accessibility of road network that presented by the increase of road density. In the spatial analysis under the 100-km spatial unit, the optimal parameter combination for spatial discretization is selected for each continuous variable. In Figure 10, the temperature variable is used as an example to present the process and result of discretization optimization. The selected optimal combinations of discretization method and break number of all explanatory variables are listed in the Supplementary Information. Results of geographical detectors show that temperature is the major contributor to the flu incidence with the contribution of 49.09%, where southern region is of high temperature driven risks. Effects of the medical cost and percentage of sensible population are enhanced by each other, and their interaction can contribute 79.4% of flu incidence variations.

[Figure 10 near here]

426 *4.2.2 In sub-regions*

427 The OPGD model can be flexibly utilized in terms of objectives of research and
428 characteristics of spatial data. This section presents an example that the study area is divided
429 into three sub-regions based on geographical regions, and the OPGD-based analysis are
430 performed in the sub-regions respectively.

431 Figure 11 shows the spatial analysis for H1N1 flu incidences in sub-regions. Results
432 include four parts of geographical detectors and size effects of spatial unit. Steps of the spatial
433 scale optimization and spatial discretization optimization are similar with processes of the
434 whole study area analysis, and they are presented in the Supplementary Information 2. Spatial
435 units are respectively determined for the spatial analysis of three sub-regions according to the
436 comparison of spatial scale effects. The geographical detector results show that primary
437 explanatory variables and interactive variables are varied among sub-regions. In the northeast,
438 northern and western regions (Figure 11 a and c), socio-economic variables and interactions
439 are major contributors to the flu incidence. In the northeast and northern regions (Figure 11 a),
440 the percentage of urban population contributes most to the flu incidence, and the interaction
441 between percentage of urban population and medical cost per capita has the highest association
442 with flu incidence. In the western region (Figure 11 c), medical cost per capita is the primary
443 single explanatory variable, and the interaction between percentage of urban population and
444 precipitation is the major interactive variable of the flu incidence. However, in central and
445 south regions (Figure 11 b), meteorological variables have higher associations with flu
446 incidence than socio-economic variables. Precipitation, temperature and humidity are top three
447 variables with relatively high associations with flu incidence in central and south regions. The
448 interaction between precipitation and percentage of sensitive population is the primary
449 interactive variable of flu incidence.

450 [Figure 11 near here]

451 ***4.3 Spatial line segment data of road damage***

452 To explore potential variables of road damage, the OPGD model is applied in the analysis
453 for spatial line segment-based road damage data. Figure 12 shows the OPGD-based spatial
454 analysis results, including the spatial discretization optimization and geographical detectors.
455 Computation process and intermediate results are summarized in the Supplementary
456 Information 2. Optimal discretization parameter combinations for the local population and
457 traffic vehicles are quantile breaks with 5 intervals and equal breaks with 7 intervals. Result of
458 factor detector shows that soil type contributes most to the road damage compared with other

variables. Soil type can explain 19.5% of road damage conditions. Results of risk detector indicate that the road segments at the soil type of Podosol have the highest risk of road damage, and those at the soil type of Kandosol have the relatively lowest risk. The interaction detector reveals the impacts of interactions of variables, where the interaction between volumes of vehicles and soil type has the highest contribution (47.12%) that is nonlinearly enhanced by the single variables. Results of ecological detector demonstrate that the impacts of road speed limit are significantly different with other variables.

[Figure 12 near here]

5 Discussion

This study develops an OPGD model for spatial stratified heterogeneity analysis, which is an improvement of the geographical detector model by integrating the parameters optimization. The primary contribution is that the OPGD model can reveal more geographical characteristics and information through the parameter optimization process for spatial discretization and spatial scale. The identification of characteristics of geographical attributes can support more accurate and effective spatial patterns and heterogeneity exploration. In addition, applications of the OPGD model in different types of spatial data, including spatial raster data, spatial point or areal data, and spatial line segment data, demonstrate that more findings can be provided from the perspectives of spatial associations and regional investigations by the analysis based on the in-depth geographical characteristics and information. The innovative findings are critical for practical spatial data analysis and support regional decision making.

In the first case, the OPGD-based spatial analysis provides accurate evidence for regional and interactive impacts of potential variables of vegetation changes. First, for continuous variables, the OPGD model provides an optimization method for determining the best parameter combinations of spatial discretization parameters and spatial scale parameter. In most of previous research, both types of spatial parameters are manually determined in the geographical detector model (Ding et al. 2019, Luo et al. 2019, Duan and Tan 2020). The optimal combinations of discretization method and break number for explanatory variables can reveal more approximately real associations between dependent and independent spatial variables. In the case, 40-km grid is selected as the optimal spatial scale for the vegetation change variables exploration, which is approximate to the spatial units that have been used in vegetation studies at large spatial ranges (Saidaliyeva et al. 2017, Rodríguez-Fernández et al. 2018, Velasquez et al. 2019). The process of spatial scale parameter optimization can indicate

spatial scale effects during the analysis, and the optimal parameter demonstrates a more reasonable spatial unit for spatial analysis. In addition, geographically regional and interactive impacts of potential variables on vegetation changes in the study area are investigated. From the perspective of regional effects of variables, the association between vegetation changes and potential variables is significantly varied in different regions. In north-eastern regions, the vegetation change is closely associated with climate variables, because forest and grassland are major land use types and they are sensitive to temperature and precipitation (Li et al. 2018). In the eastern and southern regions, the vegetation change is linked with human activities, such as GDP and population density. This result is mainly caused by the high dense human activities, low forest coverage and large areas of steppe desert, which is not sensitive to the climate change (Li et al. 2018, Yin et al. 2018). From the perspective of interactive effects of variables, the interaction of precipitation and mining activities is the major interaction variable in the study area, and they are enhanced by each other in affecting vegetation change. It has been widely confirmed that that climate conditions and human activities have combined effects on vegetation changes (Brandt et al. 2017, Wang et al. 2018, Zheng et al. 2019), but the OPGD-based spatial analysis in this study provides a quantitative comparison between effects of single variables and variable interactions from a spatial perspective.

The second case demonstrates that the OPGD model can be flexibly applied in spatial variables exploration in both the whole study area and geographical sub-regions. In the whole study area, temperature is the major contributor to the flu incidence, where southern high-temperature region is of high risks driven by temperature. High temperature and extreme weather usually link with outbreaks of H1N1 flu (Xiao et al. 2013, Chowell et al. 2012, Li, Song, and Wang 2009). Effects of the medical cost and percentage of sensible population are enhanced by each other. The close association between the H1N1 flu with socioeconomic conditions indicates the essential role of public health resources in the variation of flu incidence (Ponnambalam et al. 2012, Kumar et al. 2015, Mulinari et al. 2018). Compared with previous studies, this study provides more details about geographically regional effects of explanatory variables of the flu incidence. In the northeast, northern and western regions, socio-economic variables and interactions are major contributors to the flu incidence, but in central and south regions, meteorological variables have higher associations with flu incidence than socio-economic variables.

The OPGD-based spatial line segment data analysis in the third case indicates the comprehensive impacts of traffic volumes and environment on road damage. The spatial analysis consists of two major findings. First, soil type contributes most to the road damage

compared with traffic conditions and population distributions. In general, different soil types have significantly varied capacity to bear road damage and potential vulnerability to rut formation (Mohtashami et al. 2017). Another finding is that the interaction between traffic volumes and soil type has the highest contribution (47.12%) to road damage, and they are nonlinearly enhanced by each other.

Finally, this study provides an open-source software "GD" package in R for more flexible, efficient and user-friendly computation of the OPGD model. The package can provide sufficient details during computation and generate diverse statistics and visualizations of spatial analysis results. Meanwhile, computation speed can be significantly improved by the package. A simulation data is sampled from the disease mapping case of the Excel-based software (http://www.geodetector.org/), and it contains three explanatory variables for disease incidence. Results show that the time consuming is linearly increased with the number of samples. When the sample size reaches to 1000, 10 000, 100 000, only 0.05 s, 0.14 s and 1.55 s are used for simultaneous computation of four parts of geographical detectors by the GD package, respectively. The package has strong capability in dealing with big quantity spatial data. More sample units will improve the accuracy, but the marginal benefit might be tiny if sample units are large than 50-100 in each stratum.

6 Conclusions

This study demonstrates that the parameter optimization can further extract information contained in geographical explanatory variables for the geographical detector model. The developed OPGD model improves the capacity of the geographical detector model with a parameter optimization method to optimize both spatial discretization parameters (discretization method and break number) and the spatial scale parameter. The OPGD model can provide a comprehensive solution for spatial stratified heterogeneity analysis through more accurate and effective extraction of geographical characteristics of explanatory variables. The developed open-source software package can show a full picture of the spatial stratified heterogeneity analysis at all stages. The OPGD-based analysis for three example cases with different types of spatial data provide comprehensive benchmarks for broadening application scenarios, ways, and fields.

Supplementary Materials

The following are the Supplementary Information to this article. Supplementary Information 1: *Overview of the GD package*. Supplementary Information 2: *Computation process and results of example cases.*

Acknowledgments

This research was partially supported by the National Natural Science Foundation of China (No 41421001): Statistics for Spatiaotemporal Big Data, and the National Natural Science Foundation of China (No 41531179): Spatial Sampling Trinity Theory. The authors wish to thank the R community for the efforts during releasing the package. The R package, its manual and all case datasets are available at https://cran.r-project.org/web/packages/GD/.

Conflicts of Interest

The authors have declared that they have no completing interests.

Reference

Anselin, Luc. 1995. "Local indicators of spatial association—LISA." *Geographical analysis* 27 (2):93-115.

Aria, Massimo, and Corrado Cuccurullo. 2017. "bibliometrix: An R-tool for comprehensive science mapping analysis." *Journal of Informetrics* 11 (4):959-975.

Ashton, LJ, and NJ McKenzie. 2001. Conversion of the atlas of Australian soils to the Australian Soil Classification. CSIRO Land and Water (unpublished).

Brandt, Martin, Kjeld Rasmussen, Josep Peñuelas, Feng Tian, Guy Schurgers, Aleixandre Verger, Ole Mertz, John RB Palmer, and Rasmus Fensholt. 2017. "Human population growth offsets climate-driven increase in woody vegetation in sub-Saharan Africa." *Nature Ecology & Evolution* 1 (4):1-6.

Brunsdon, Chris, A Stewart Fotheringham, and Martin E Charlton. 1996. "Geographically weighted regression: a method for exploring spatial nonstationarity." *Geographical analysis* 28 (4):281-298.

Cao, Feng, Yong Ge, and Jin-Feng Wang. 2013. "Optimal discretization for geographical detectors-based risk assessment." *GIScience & remote sensing* 50 (1):78-92.

Center for International Earth Science Information Network - CIESIN - Columbia University. 2016. Gridded Population of the World, Version 4 (GPWv4): Population Density Adjusted to Match 2015 Revision UN WPP Country Totals. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC).

Chen, Qi, Ronald E McRoberts, Changwei Wang, and Philip J Radtke. 2016. "Forest aboveground biomass mapping and estimation across multiple spatial scales using model-based inference." *Remote Sensing of Environment* 184:350-360.

Chowell, Gerardo, Sherry Towers, Cécile Viboud, Rodrigo Fuentes, Viviana Sotomayor, Lone Simonsen, Mark A Miller, Mauricio Lima, Claudia Villarroel, and Monica Chiu. 2012. "The influence of climatic conditions on the transmission dynamics of the 2009 A/H1N1 influenza pandemic in Chile." *BMC infectious diseases* 12 (1):298.

Ding, Yueting, Ming Zhang, Xiangyan Qian, Chengren Li, Sai Chen, and Wenwen Wang. 2019. "Using the geographical detector technique to explore the impact of

596 socioeconomic factors on PM2. 5 concentrations in China." *Journal of cleaner*
597 *production* 211:1480-1490.
598 Du, Ziqiang, Xiaoyu Zhang, Xiaoming Xu, Hong Zhang, Zhitao Wu, and Jing Pang. 2017.
599 "Quantifying influences of physiographic factors on temperate dryland vegetation,
600 Northwest China." *Scientific reports* 7:40092.
601 Duan, Qianwen, and Minghong Tan. 2020. "Using a geographical detector to identify the key
602 factors that influence urban forest spatial differences within China." *Urban Forestry*
603 *& Urban Greening*:126623.
604 Fischer, Manfred M. 2010. "Handbook of Applied Spatial Analysis." *Journal of Geographical*
605 *Systems* 10 (2):109-139.
606 Fotheringham, A Stewart, Chris Brunsdon, and Martin Charlton. 2003. *Geographically*
607 *weighted regression: the analysis of spatially varying relationships*: John Wiley &
608 Sons.
609 Ge, Yong, Yongze Song, Jinfeng Wang, Wei Liu, Zhoupeng Ren, Junhuan Peng, and Binbin
610 Lu. 2017. "Geographically weighted regression- based determinants of malaria
611 incidences in northern China." *Transactions in GIS* 21 (5):934-953.
612 Getis, Arthur, and J. K. Ord. 1992. "The Analysis of Spatial Association by Use of Distance
613 Statistics." *Geographical Analysis* 24 (3):189–206.
614 Huang, Bo, Bo Wu, and Michael Barry. 2010. "Geographically and temporally weighted
615 regression for modeling spatio-temporal variation in house prices." *International*
616 *Journal of Geographical Information Systems* 24 (3):383-401.
617 Jiang, Bin. 2013. "Head/Tail Breaks: A New Classification Scheme for Data with a Heavy-
618 Tailed Distribution." *Professional Geographer* 65 (3):482-494.
619 Jiang, Bin. 2015. "Geospatial analysis requires a different way of thinking: the problem of
620 spatial heterogeneity." *Geojournal* 80 (1):1-13.
621 Jiang, Bin, and Junjun Yin. 2014. "Ht-Index for Quantifying the Fractal or Scaling Structure
622 of Geographic Features." *Annals of the Association of American Geographers* 104
623 (3):530-540.
624 Ju, Hongrun, Zengxiang Zhang, Lijun Zuo, Jinfeng Wang, Shengrui Zhang, Xiao Wang, and
625 Xiaoli Zhao. 2016. "Driving forces and their interactions of built-up land expansion
626 based on the geographical detector–a case study of Beijing, China." *International*
627 *Journal of Geographical Information Science* 30 (11):2188-2207.
628 Kriticos, Darren J, Bruce L Webber, Agathe Leriche, Noboru Ota, Ian Macadam, Janice
629 Bathols, and John K Scott. 2012. "CliMond: global high- resolution historical and
630 future scenario climate surfaces for bioclimatic modelling." *Methods in Ecology and*
631 *Evolution* 3 (1):53-64.
632 Kulldorff, Martin. 1997. "A spatial scan statistic." *Communications in Statistics-Theory and*
633 *methods* 26 (6):1481-1496.
634 Kumar, Pratyush, Anurag Sachan, Atul Kakar, and Atul Gogia. 2015. "Socioeconomic impact
635 of the recent outbreak of H1N1." *Current Medicine Research and Practice* 5 (4):163-
636 167.
637 Kummu, M., M. Taka, and J. H. A. Guillaume. 2018a. Data from: Gridded global datasets for
638 Gross Domestic Product and Human Development Index over 1990-2015. Dryad Data
639 Repository.
640 Kummu, Matti, Maija Taka, and Joseph HA Guillaume. 2018b. "Gridded global datasets for
641 Gross Domestic Product and Human Development Index over 1990–2015." *Scientific*
642 *data* 5:180004.
643 Li, Chunlan, Jun Wang, Richa Hu, Shan Yin, Yuhai Bao, and Desalegn Yayeh Ayal. 2018.
644 "Relationship between vegetation change and extreme climate indices on the Inner
645 Mongolia Plateau, China, from 1982 to 2013." *Ecological Indicators* 89:101-109.

Li, Wei, YanLing Song, and ChangKe Wang. 2009. "Comparability analysis between the climate characteristics of early summer in China and the meteorological conditions during the periods that the A (H1N1) flu spread in America and broke out in Mexico." *Science & Technology Review* (11):19-22.
Lu, Binbin, Chris Brunsdon, Martin Charlton, and Paul Harris. 2017. "Geographically weighted regression with parameter-specific distance metrics." *International Journal of Geographical Information Systems* 31 (5):982-998.
Lu, Binbin, Martin Charlton, Paul Harris, and A. Stewart Fotheringham. 2014. "Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data." *International Journal of Geographical Information Science* 28 (4):660-681.
Luo, Lili, Kun Mei, Liyin Qu, Chi Zhang, Han Chen, Siyu Wang, Di Di, Hong Huang, Zhenfeng Wang, and Fang Xia. 2019. "Assessment of the Geographical Detector Method for investigating heavy metal source apportionment in an urban watershed of Eastern China." *Science of the Total Environment* 653:714-722.
Luo, Wei, Jaroslaw Jasiewicz, Tomasz Stepinski, Jinfeng Wang, Chengdong Xu, and Xuezhi Cang. 2016. "Spatial association between dissection density and environmental factors over the entire conterminous United States." *Geophysical Research Letters* 43 (2):n/a-n/a.
Main Roads Western Australia. 2017a. Calibration Of Falling Weight Deflectometers, Calibration Method WA 2060.5.
Main Roads Western Australia. 2017b. "Falling Weight Deflectometer." https://www.mainroads.wa.gov.au/BuildingRoads/StandardsTechnical/MaterialsEngineering/Pages/Falling_Weight_Deflectometer.aspx (Accessed on Jan 2018).
Mohtashami, Sima, Lars Eliasson, Gunnar Jansson, and Johan Sonesson. 2017. "Influence of soil type, cartographic depth-to-water, road reinforcement and traffic intensity on rut formation in logging operations: a survey study in Sweden." *Silva Fenn* 51 (5):14.
Mulinari, Shai, Maria Wemrell, Björn Rönnerstrand, SV Subramanian, and Juan Merlo. 2018. "Categorical and anti-categorical approaches to US racial/ethnic groupings: Revisiting the National 2009 H1N1 Flu Survey (NHFS)." *Critical Public Health* 28 (2):177-189.
National Bureau of Statistics of China. 2015. *China Statistical Yearbook 2014*: China Statistics Press.
National Health Commission of the People's Republic of China. 2014. *China Health Statistical Yearbook*: China Union Medical University Press.
Ord, J. K., and Arthur Getis. 1995. "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application." *Geographical Analysis* 27 (4):286-306.
Ponnambalam, L, Lee Samavedham, HR Lee, and CS Ho. 2012. "Understanding the socioeconomic heterogeneity in healthcare in US counties: the effect of population density, education and poverty on H1N1 pandemic mortality." *Epidemiology & Infection* 140 (5):803-813.
Ren, Yin, Lu-Ying Deng, Shu-Di Zuo, Xiao-Dong Song, Yi-Lan Liao, Cheng-Dong Xu, Qi Chen, Li-Zhong Hua, and Zheng-Wei Li. 2016. "Quantifying the influences of various ecological factors on land surface temperature of urban forests." *Environmental pollution* 216:519-529.
Ren, Yin, Luying Deng, Shudi Zuo, Yunjian Luo, Guofan Shao, Xiaohua Wei, Lizhong Hua, and Yusheng Yang. 2014. "Geographical modeling of spatial interaction between human activity and forest connectivity in an urban landscape of southeast China." *Landscape ecology* 29 (10):1741-1758.
Rodríguez-Fernández, Nemesio J, Arnaud Mialon, Stéphane Mermoz, Alexandre Bouvet, Philippe Richaume, Ahmad Al Bitar, Amen Al-Yaari, Martin Brandt, Thomas

Kaminski, and Thuy Le Toan. 2018. "SMOS L-band vegetation optical depth is highly sensitive to aboveground biomass." IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium.
Roth, Nancy E, J David Allan, and Donna L Erickson. 1996. "Landscape influences on stream biotic integrity assessed at multiple spatial scales." *Landscape ecology* 11 (3):141-156.
Saidaliyeva, Zarina, Ian Davenport, Mohamad Nobakht, Kevin White, and Maria Shahgedanova. 2017. "The use of remotely-sensed snow, soil moisture and vegetation indices to develop resilience to climate change in Kazakhstan." EGU General Assembly Conference Abstracts.
Song, Yongze, Xiangyu Wang, Graeme Wright, Dominique Thatcher, Peng Wu, and Pascal Felix. 2019. "Traffic Volume Prediction With Segment-Based Regression Kriging and its Implementation in Assessing the Impact of Heavy Vehicles." *IEEE Transactions on Intelligent Transportation Systems* 20 (1):232-243.
Song, Yongze, Graeme Wright, Peng Wu, Dominique Thatcher, Tom McHugh, Qindong Li, Shuk Li, and Xiangyu Wang. 2018. "Segment-Based Spatial Analysis for Assessing Road Infrastructure Performance Using Monitoring Observations and Remote Sensing Data." *Remote Sensing* 10 (11):1696.
State of the Environment in Australia. 2017. "2016 SoE Land Australian Soil Classification orders."
Store, Ron, and Jukka Jokimäki. 2003. "A GIS-based multi-scale approach to habitat suitability modeling." *Ecological modelling* 169 (1):1-15.
Velasquez, Patricio, Jed O Kaplan, Martina Messmer, Patrick Ludwig, and Christoph C Raible. 2019. "High resolution simulations of climate and vegetation in Europe at the Last Glacial Maximum." Geophysical Research Abstracts.
Wang, Jin-Feng, Tong-Lin Zhang, and Bo-Jie Fu. 2016. "A measure of spatial stratified heterogeneity." *Ecological Indicators* 67:250-256.
Wang, Jin- Feng, Xin- Hu Li, George Christakos, Yi- Lan Liao, Tin Zhang, Xue Gu, and Xiao- Ying Zheng. 2010. "Geographical Detectors- Based Health Risk Assessment and its Application in the Neural Tube Defects Study of the Heshun Region, China." *International Journal of Geographical Information Science* 24 (1):107-127.
Wang, Jinfeng, Yong Ge, Lianfa Li, Bin Meng, Jilei Wu, Yanchen Bo, Shihong Du, Yilan Liao, Maogui Hu, and Chengdong Xu. 2014. "Spatiotemporal data analysis in geography." *Acte Geographica Sinica* 69 (9):1326-1345.
Wang, Lanhui, Feng Tian, Yuhang Wang, Zhendong Wu, Guy Schurgers, and Rasmus Fensholt. 2018. "Acceleration of global vegetation greenup from combined effects of climate change and human land management." *Global change biology* 24 (11):5484-5499.
Xiao, Hong, HuaiYu Tian, XiaoLing Lin, LiDong Gao, XiangYu Dai, XiXing Zhang, BiYun Chen, Jian Zhao, and JingZhe Xu. 2013. "Influence of extreme weather and meteorological anomalies on outbreaks of influenza A (H1N1)." *Chinese Science Bulletin* 58 (7):741-749.
Yin, He, Dirk Pflugmacher, Ang Li, Zhengguo Li, and Patrick Hostert. 2018. "Land use and land cover change in Inner Mongolia-understanding the effects of China's re-vegetation programs." *Remote Sensing of Environment* 204:918-930.
Zheng, Kai, Jian-Zhou Wei, Jiu-Ying Pei, Hua Cheng, Xu-Long Zhang, Fu-Qiang Huang, Feng-Min Li, and Jian-Sheng Ye. 2019. "Impacts of climate change and human activities on grassland vegetation variation in the Chinese Loess Plateau." *Science of the Total Environment* 660:236-244.

Figure 1. Overview of the global research using the geographical detector model regarding the conceptual structure map and annual distributions of publications.

Figure 2. Schematic overview of the optimal parameters based geographical detector (OPGD) model

Figure 3. Spatial distributions of vegetation changes and explanatory variables. (a) Study area, (b) NDVI change, (c) Climate zone, (d) Temperature change, (e) Precipitation, (f) Mining production, (g) GDP, and (h) Population density.

Figure 4. Spatial distributions of H1N1 flu incidences and explanatory variables

Figure 5. OPGD-based explanatory variables exploration of vegetation changes: Processes (a) and results (b) of parameter optimization for spatial data discretization.

Figure 6. OPGD-based explanatory variables exploration of vegetation changes: Contributions of single variables on vegetation changes investigated by the factor detector.

Figure 7. OPGD-based explanatory variables exploration of vegetation changes: Vegetation changes in variable determined spatial zones computed by the risk detector, including risk mean values (a), spatial distributions of high, medium and low levels of mean vegetation changes (b), and the risk detector result (c).

Figure 8. OPGD-based explanatory variables exploration of vegetation changes: Interaction detector (a) and ecological detector (b) results.

Figure 9. Comparison of size effects of spatial units for Q values and the 90% quantile of explanatory variables.

Figure 10. Results of OPGD-based analysis for H1N1 flu incidences. (a) Spatial scale effects and spatial unit selection; (b) spatial discretization optimization for the variable temperature; (c) factor detector; (d) risk mean values of variables geographical region and temperature; (e) risk matrixes of variables geographical region and temperature; (f) interaction detector; and (g) ecological detector.

Figure 11. Spatial analysis for H1N1 flu incidences in the sub-regions: (a) northeast and north; (b) central and south and (c) western China.

773 Figure 12. Spatial analysis for road damage conditions. (a) Processes and results of spatial
774 discretization optimization for continuous variables; (b) factor detector; (c) risk detector; (d)
775 interaction detector; and (e) ecological detector.

776 Table 1. Interactions between two explanatory variables and their interactive impacts

777 Table 2. A summary of cases with different types of spatial data
778

Table 1. Interactions between two explanatory variables and their interactive impacts

Geographical interaction relationship	Interaction
$Q_{u\cap v} < \min(Q_u, Q_v)$ [1]	Nonlinear-weaken: Impacts of single variables are nonlinearly weakened by the interaction of two variables.
$\min(Q_u, Q_v) \le Q_{u\cap v} \le \max(Q_u, Q_v)$	Uni-variable weaken: Impacts of single variables are uni-variable weakened by the interaction.
$\max(Q_u, Q_v) < Q_{u\cap v} < (Q_u + Q_v)$	Bi-variable enhance: Impact of single variables are bi-variable enhanced by the interaction.
$Q_{u\cap v} = (Q_u + Q_v)$	Independent: Impacts of variables are independent.
$Q_{u\cap v} > (Q_u + Q_v)$	Nonlinear-enhance: Impacts of variables are nonlinearly enhanced.

[1] Q_u is the Q value of variable u, Q_v is the Q value of variable v, and $Q_{u\cap v}$ is the Q value of the interaction between variables u and v.

Table 2. A summary of cases with different types of spatial data

Case	Description	Study area	Type of sample	Explanatory variables and variable types
Vegetation change variables exploration	Impacts of human activities and climate on the vegetation changes	Whole area	Vari ables derived from raster data	Categorical variables: Climate zone and mining production Continuous variables: Temperature change, precipitation, GDP and population density
H1N1 flu incidence variables exploration	Associations of meteorological conditions and human activities with H1N1 flu incidences	Whole area and sub-regions	Spatial point or areal data	Categorical variable: Geographical region Continuous variables: Temperature, precipitation, humidity index, population density, GDP, road density, percentage of sensitive people, percentage of urban population and medical cost per capita
Road damage variables exploration	Impacts of vehicles and environment on road damage	Whole area	Spatial line segment	Categorical variables: Traffic speed and soil type Continuous variables: Population within 1 km of road segments and daily traffic volumes

Figure 1

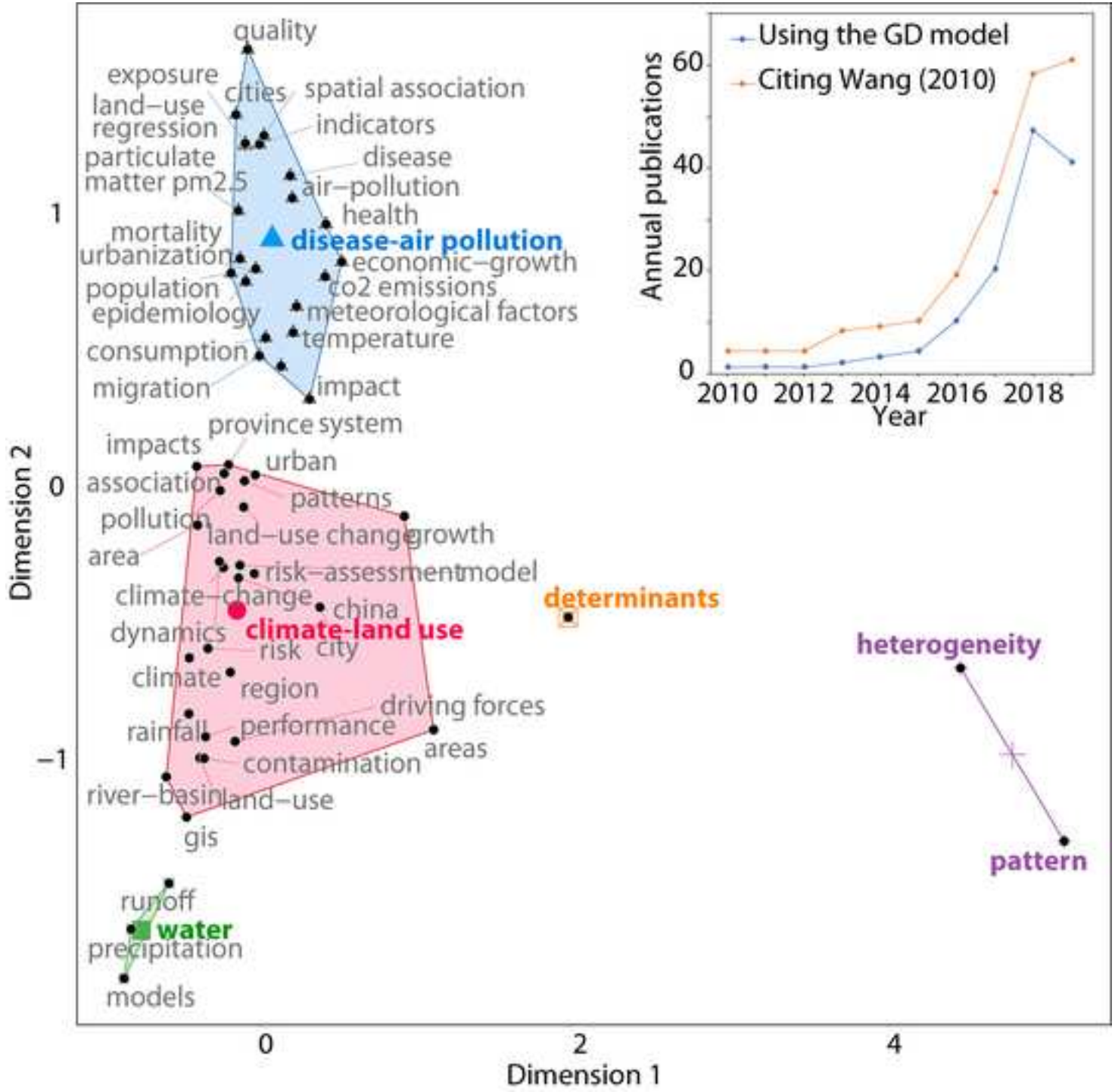


Figure 2

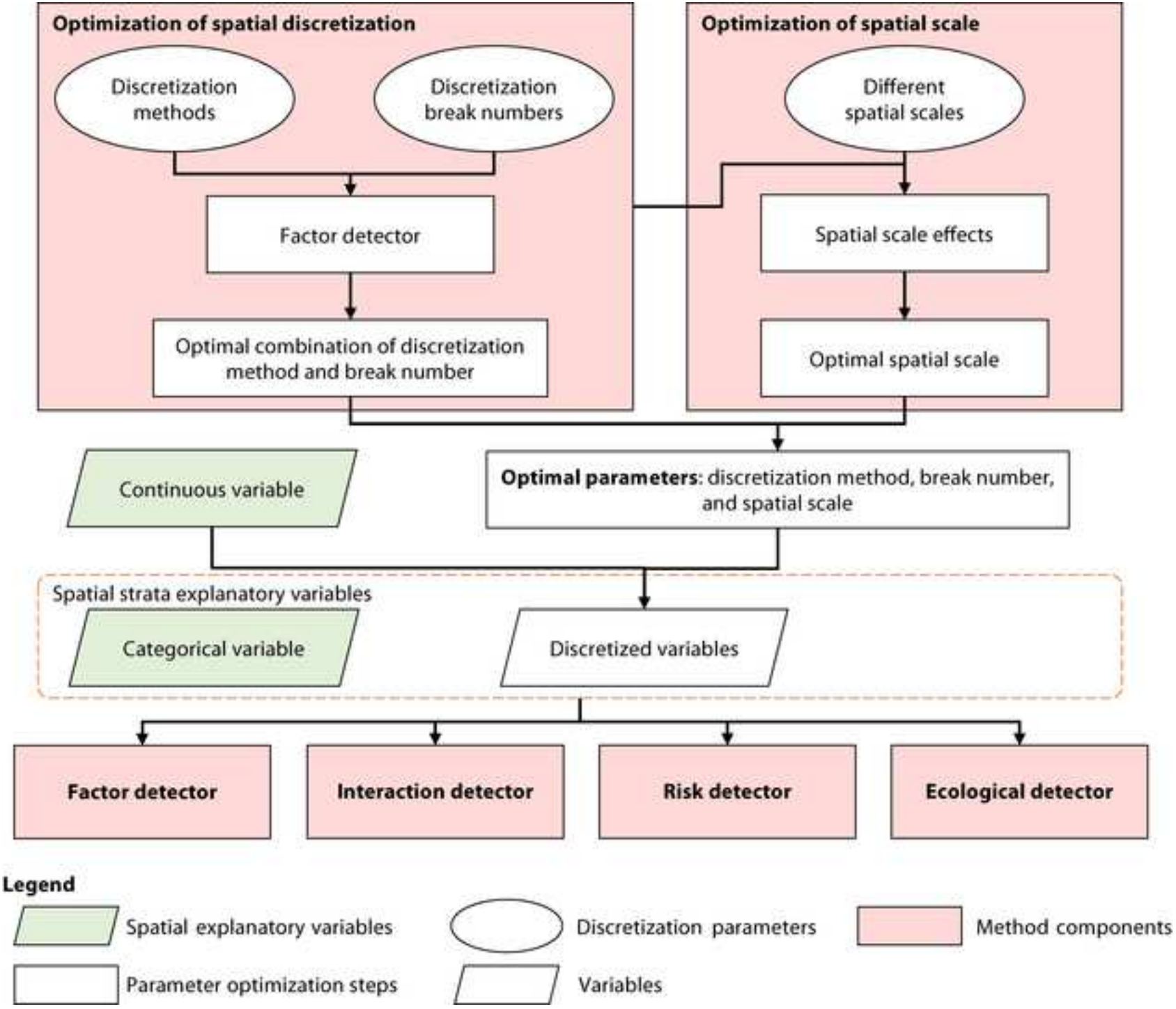


Figure 3

Click here to access/download;Figure;Figure 3 data 1.tif

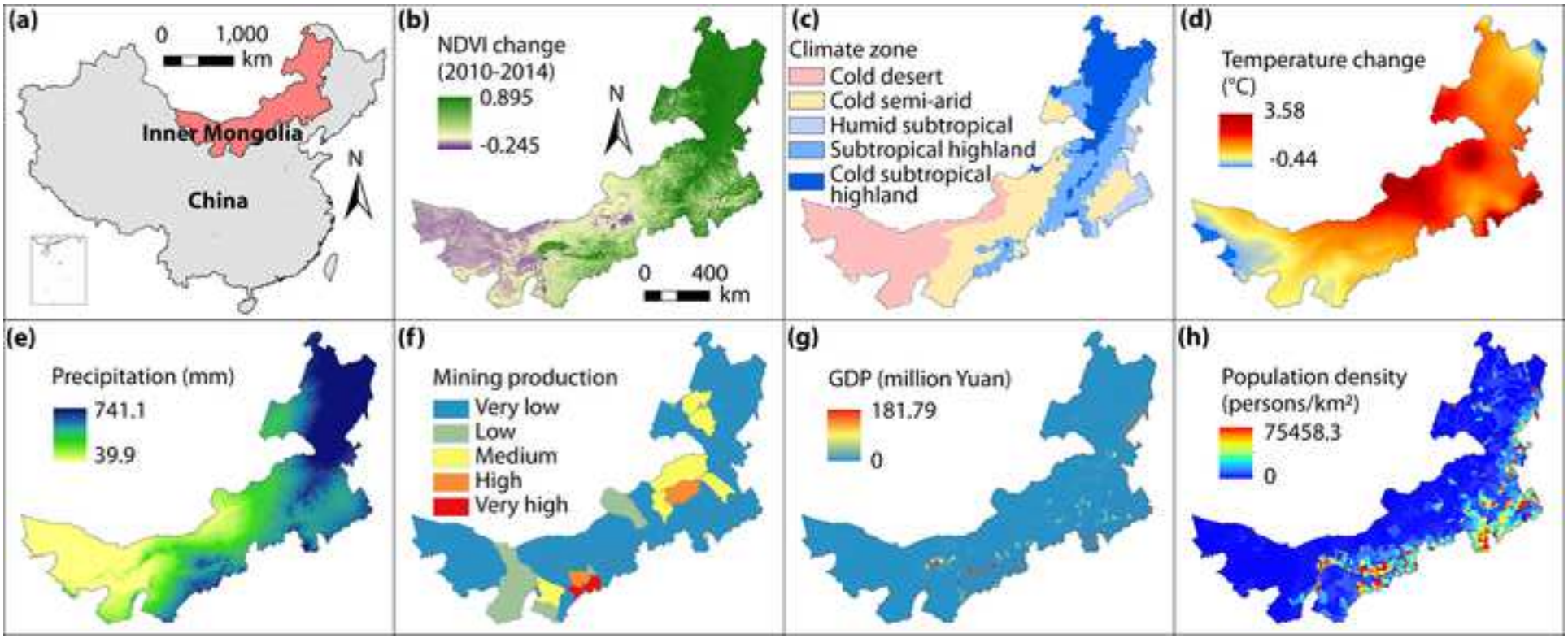


Figure 5

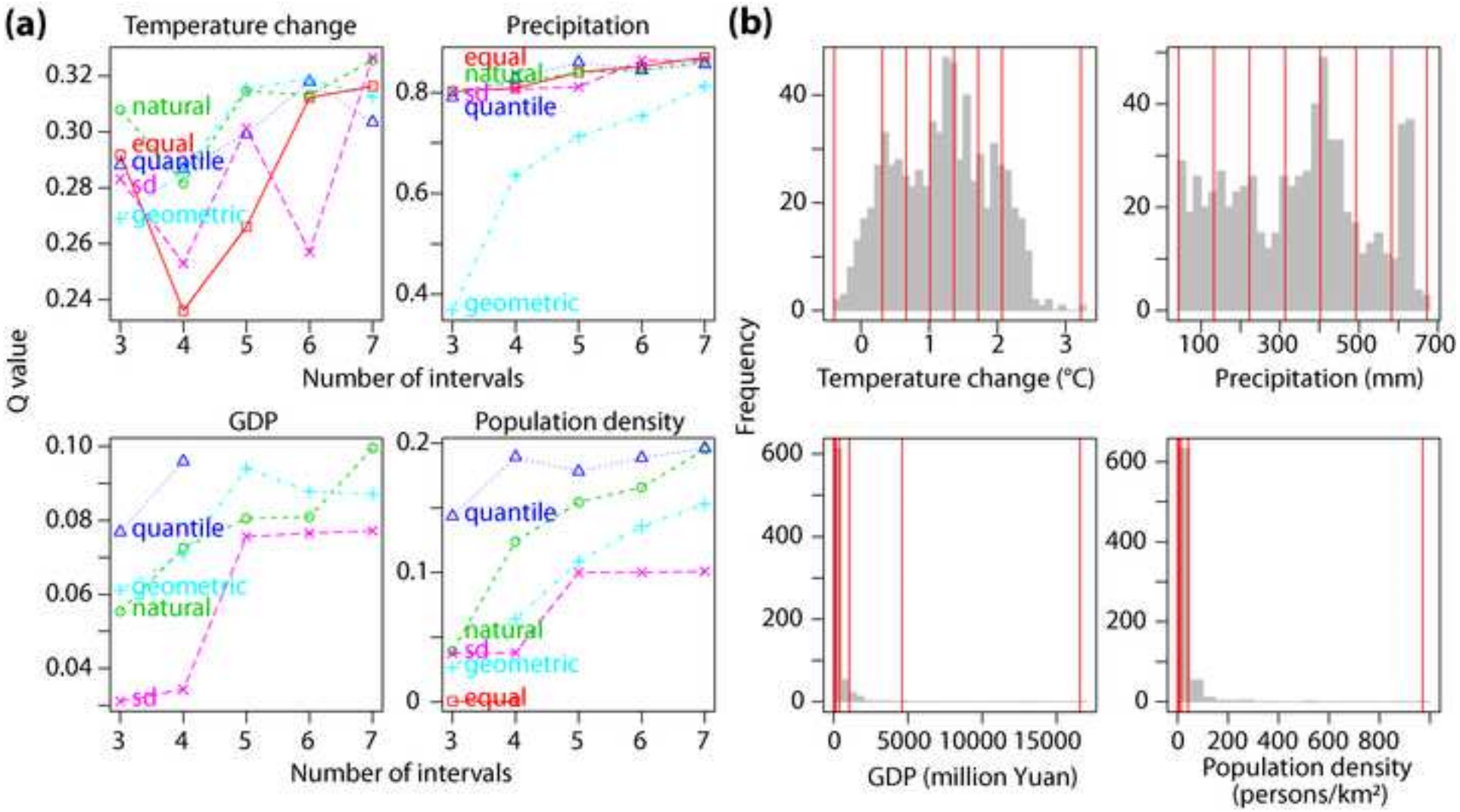


Figure 6

Click here to access/download;Figure;Figure 6 part 2.tif

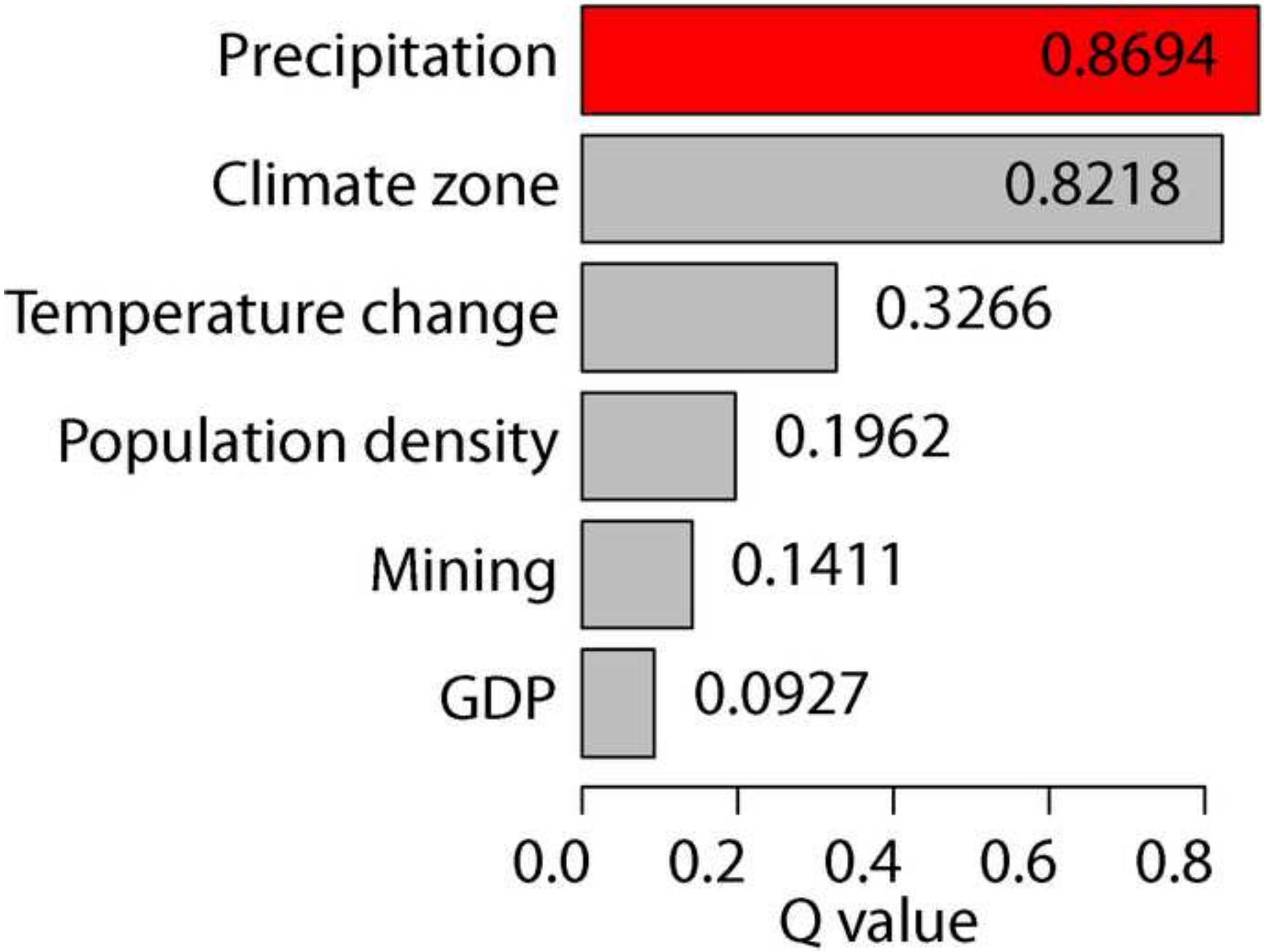


Figure 7

Click here to access/download;Figure;Figure 7 part 3.tif

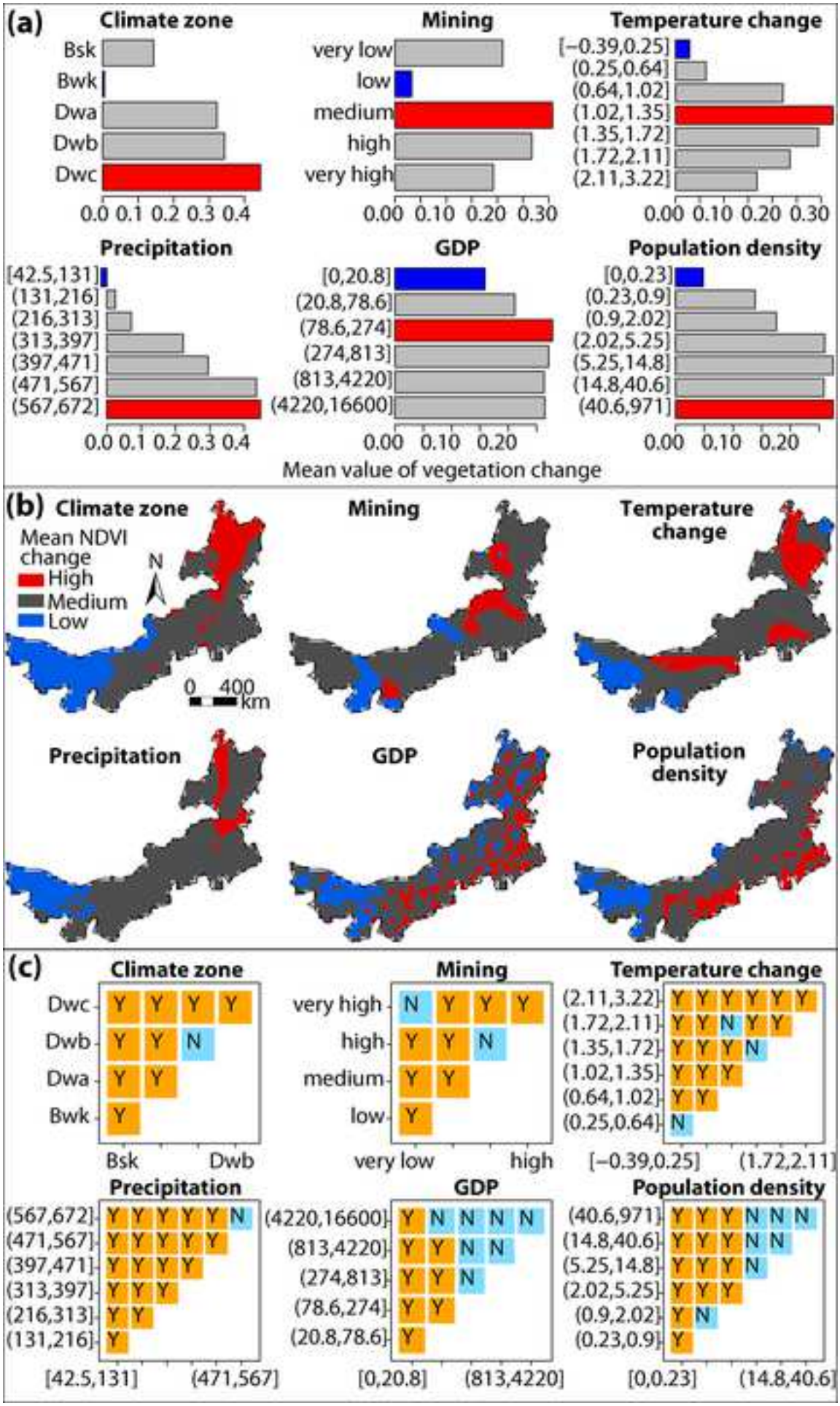


Figure 8

Click here to access/download;Figure;Figure 8 part 4.tif

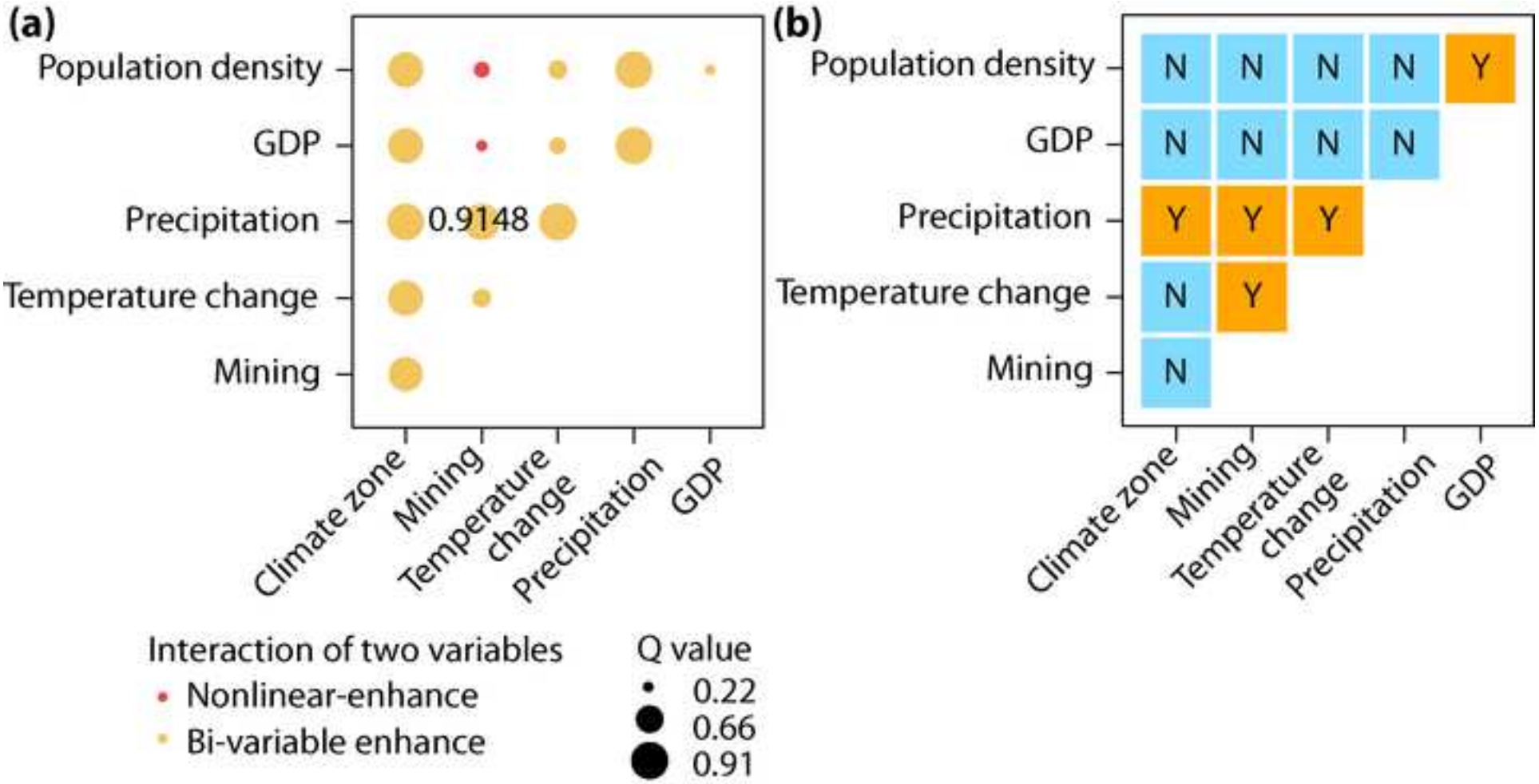


Figure 10

Click here to access/download;Figure;Figure 10 case 2.tif

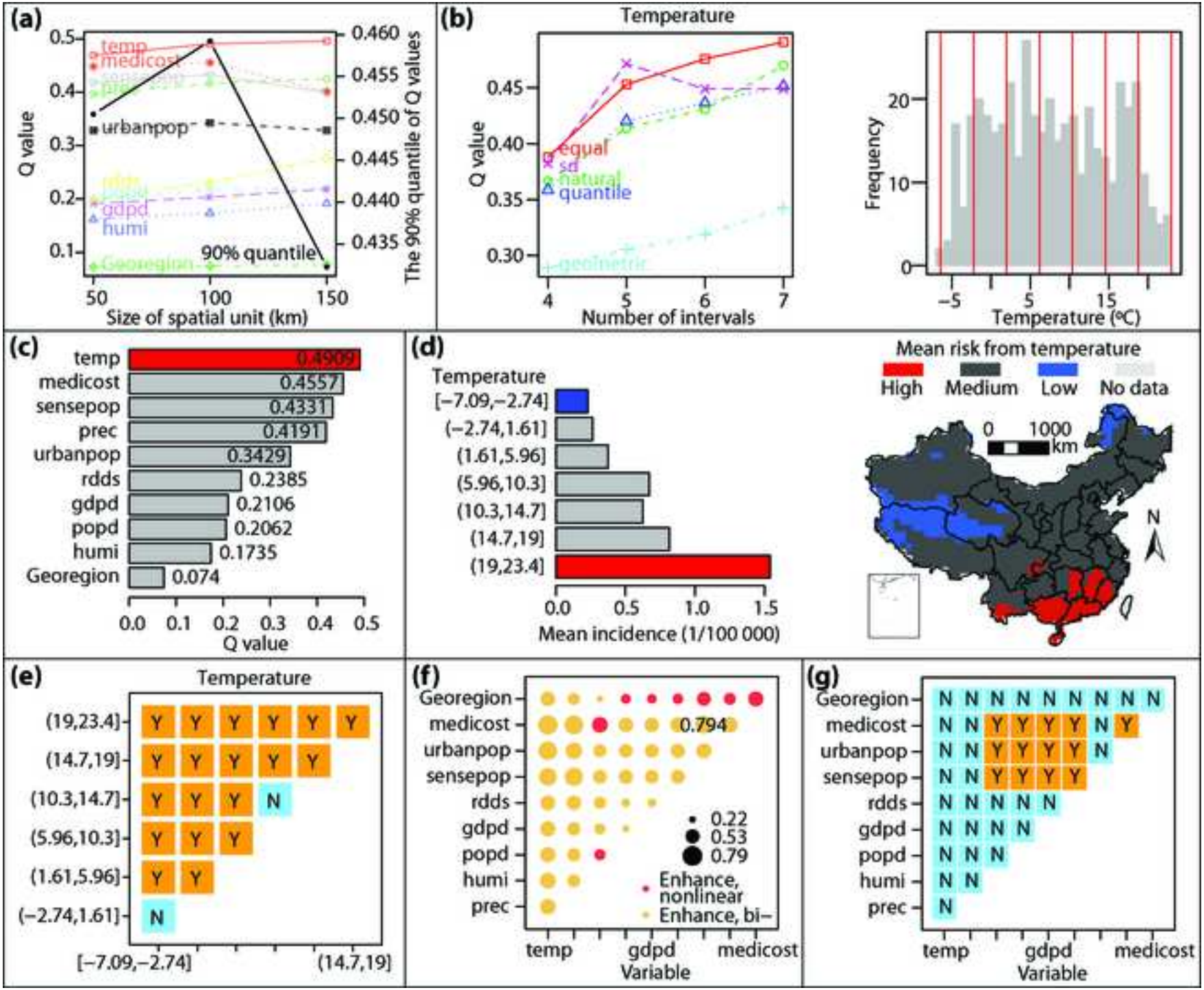


Figure 11

Click here to access/download;Figure;Figure 11 case 2 subregions.tif

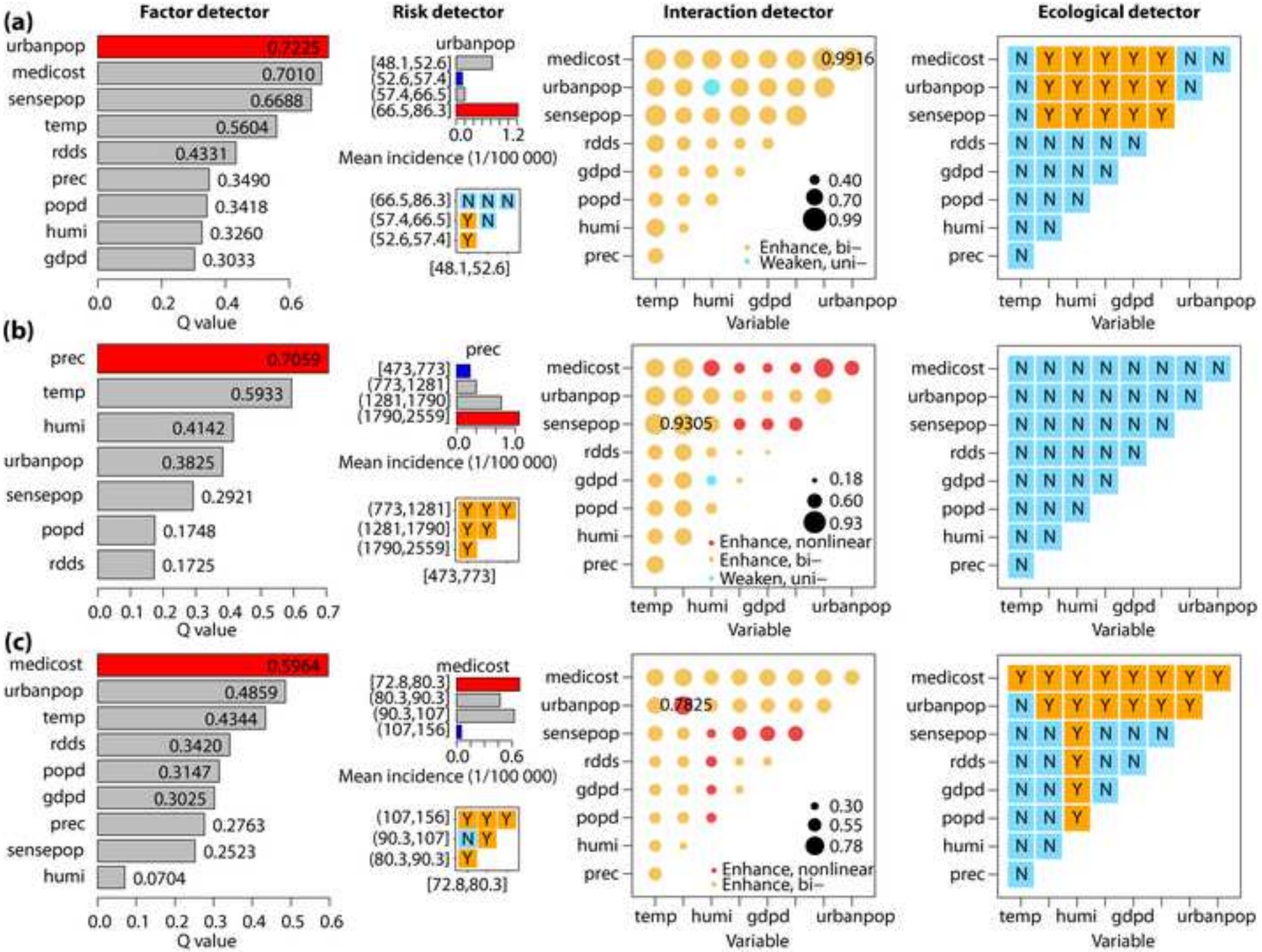
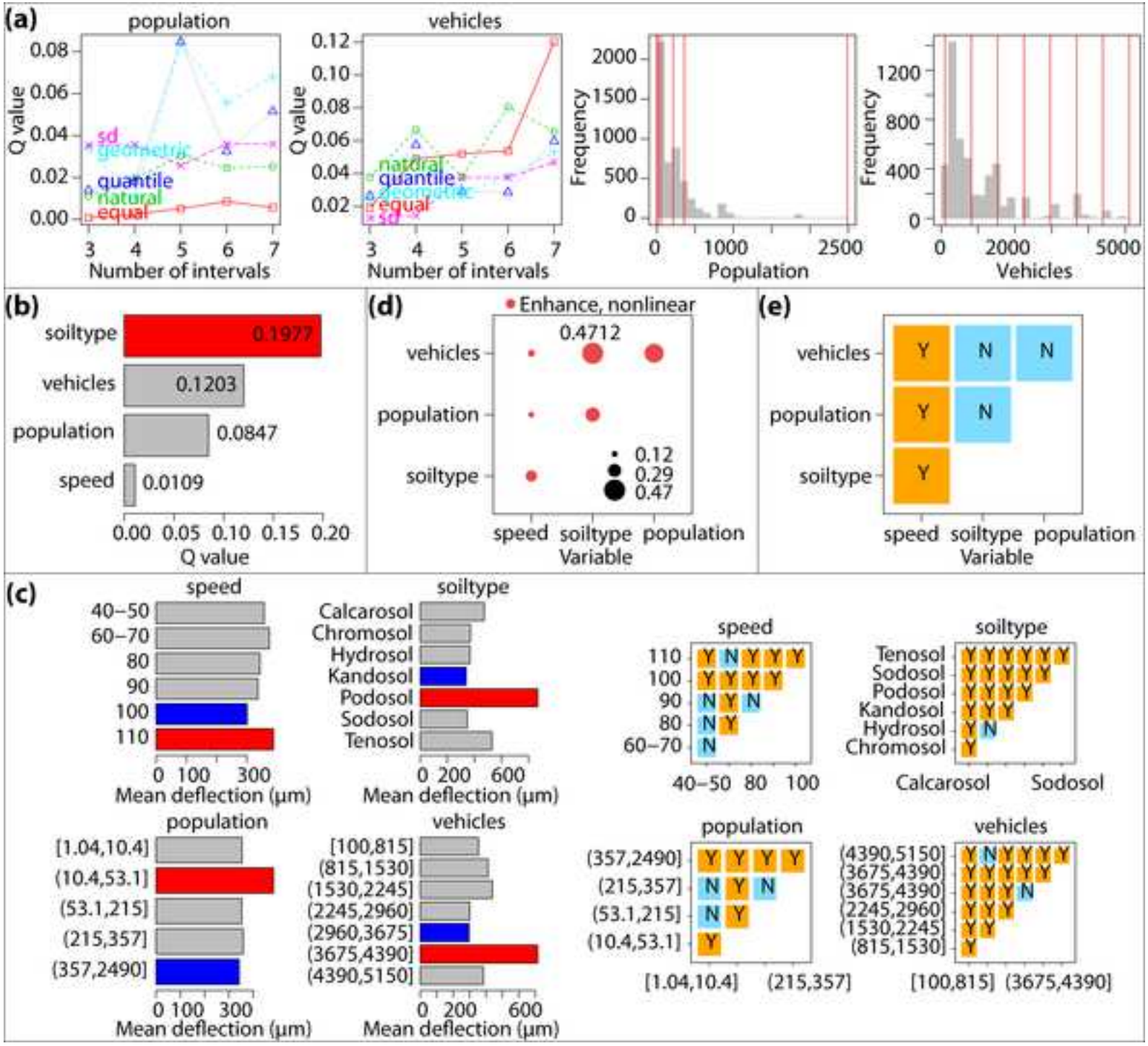


Figure 12

Click here to access/download;Figure;Figure 12.tif



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.

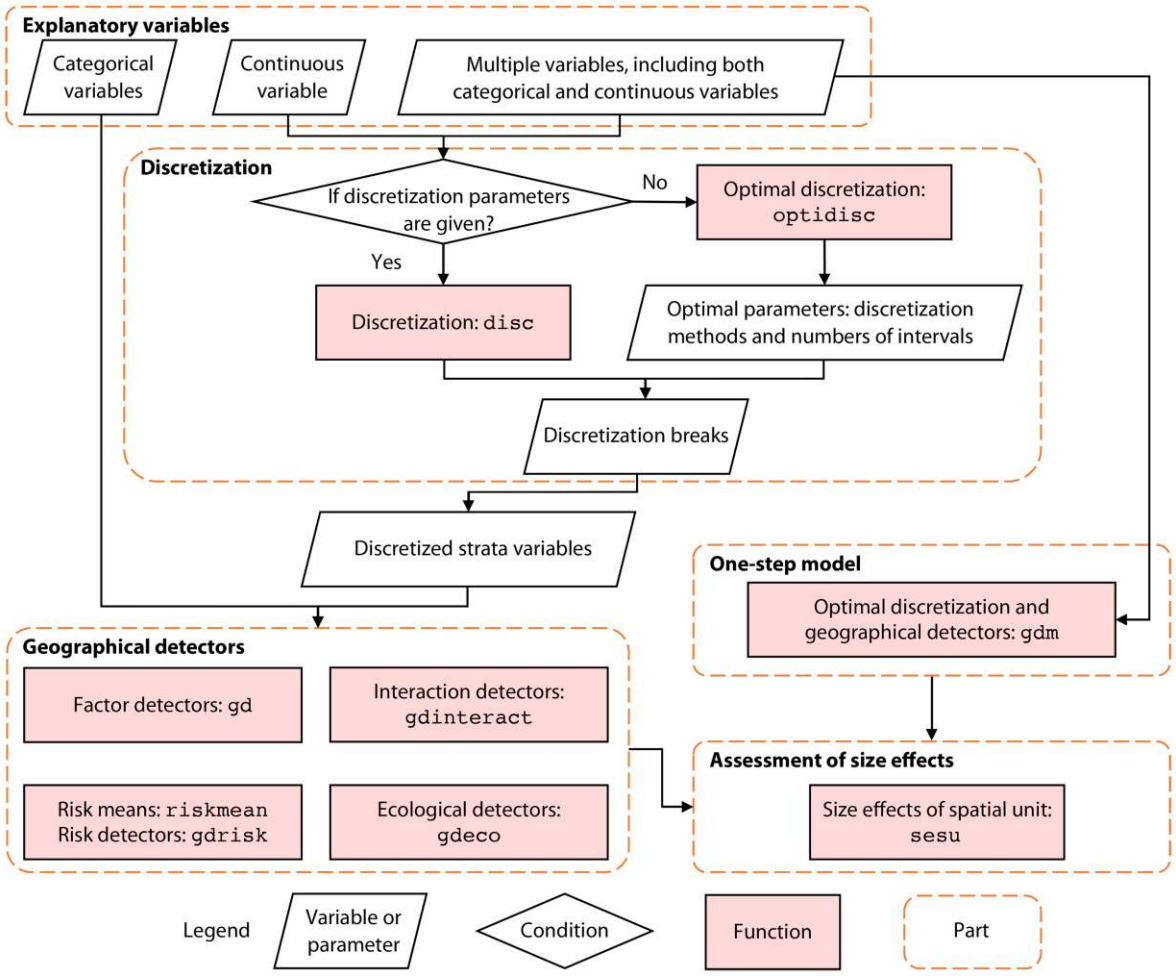


Figure S1. General calculation process and the relationships of functions in GD package

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 Q 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 Q values of two variables, the Q 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.

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 2: Computation process and results of example cases

Note:

R package "GD: Geographical Detectors" version 1.7.

The codes can be run by R (>= 3.4.0).

Orders of codes are consistent with the orders in the article.

4.1 Spatial raster data of vegetation changes

```
library("GD")
data(ndvi_40)
head(ndvi_40)
```

```
##   NDVIchange Climatezone Mining Tempchange Precipitation   GDP Popdensity
## 1    0.11599         Bwk    low    0.25598        236.54 12.55    1.44957
## 2    0.01783         Bwk    low    0.27341        213.55  2.69    0.80124
## 3    0.13817         Bsk    low    0.30247        448.88 20.06   11.49432
## 4    0.00439         Bwk    low    0.38302        212.76  0.00    0.04620
## 5    0.00316         Bwk    low    0.35729        205.01  0.00    0.07482
## 6    0.00838         Bwk    low    0.33750        200.55  0.00    0.54941
```

Optimal discretization for multiple spatial variables (Figure 5)

```
## set optional discretization methods and numbers of intervals
discmethod <- c("equal","natural","quantile","geometric","sd")
discitv <- 3:7
## optimal discretization
ndvi.test <- ndvi_40
odc1 <- optidisc(NDVIchange ~ ., data = ndvi.test[, -(2:3)], discmethod, discitv)
odc1
## plot optimal discretization processes and results
plot(odc1)
## convert continuous variables to strata variables based on discretization breaks
ndvi.test[, 4:7] <- do.call(cbind, lapply(1:4, function(x)
  data.frame(cut(ndvi.test[, x+3], unique(odc1[[x]]$itv), include.lowest = TRUE))))
```

Factor detector (Figure 6):

```
## factor detector
mvgd <- gd(NDVIchange ~ ., data = ndvi.test)
mvgd
plot(mvgd)
```

Risk detector (Figure 7):

```
## risk detector: risk means
mvrm <- riskmean(NDVIchange ~ ., data = ndvi.test)
mvrm
plot(mvrm)
## risk detector: risk matrix
mvgr <- gdrisk(NDVIchange ~ ., data = ndvi.test)
mvgr
plot(mvgr)
```

Interaction detector (Figure 8a):

```
## interaction detector
mvgi <- gdinteract(NDVIchange ~ ., data = ndvi.test)
mvgi
plot(mvgi)
```

Ecological detector (Figure 8b):

```
## ecological detector
mvge <- gdeco(NDVIchange ~ ., data = ndvi.test)
mvge
plot(mvge)
```

Results of the optimal discretization and geographical detectors by the one-step function "gdm" for exploring factors of vegetation changes.

```
data("ndvi_40")
## set optional parameters of optimal discretization
discmethod <- c("equal","natural","quantile","geometric","sd")
discitv <- 3:7
## "gdm" function (~ 10.7 s)
ndvigdm <- gdm(NDVIchange ~ .,
               continuous_variable = c("Tempchange", "Precipitation", "GDP", "Popdensity"),
               data = ndvi_40,
               discmethod = discmethod, discitv = discitv)
ndvigdm
```

```
## Explanatory variables include 4 continuous variables.
##
## optimal discretization result of Tempchange
## method             :  sd
## number of intervals:  7
## intervals:
##  -0.39277 0.3106004 0.6614902 1.01238 1.36327 1.71416 2.065049 3.22051
## numbers of data within intervals:
##  93 102 82 151 99 100 86
##
## optimal discretization result of Precipitation
## method             :  equal
## number of intervals:  7
## intervals:
##  42.51 132.5014 222.4929 312.4843 402.4757 492.4671 582.4586 672.45
## numbers of data within intervals:
##  108 107 81 137 136 56 88
##
```

```
## optimal discretization result of GDP
## method             :  natural
## number of intervals:  7
## intervals:
##  0 9.9235 33.406 100.6513 289.6999 887.9468 4598.732 16589.09
## numbers of data within intervals:
##  325 84 79 90 84 46 5
##
## optimal discretization result of Popdensity
## method             :  quantile
## number of intervals:  7
## intervals:
##  0 0.2301671 0.9002686 2.021381 5.253373 14.80794 40.57514 970.8682
## numbers of data within intervals:
##  102 102 102 101 102 102 102
##
## Geographical detectors results:
##
## Factor detector:
##        variable         qv          sig
## 1   Climatezone 0.82183348 7.176240e-10
## 2        Mining 0.14111542 6.608981e-10
## 3    Tempchange 0.32655370 3.094152e-10
## 4 Precipitation 0.86935048 2.518923e-10
## 5           GDP 0.09514837 2.347511e-09
## 6    Popdensity 0.19618770 2.146283e-10
##
## Risk detector:
## Climatezone
##   itv    meanrisk
## 1 Bsk 0.143572961
## 2 Bwk 0.004536505
## 3 Dwa 0.321735000
## 4 Dwb 0.343155655
## 5 Dwc 0.444868361
##
## Mining
##         itv   meanrisk
## 1  very low 0.21008297
## 2       low 0.03294513
## 3    medium 0.30733460
## 4      high 0.26695286
## 5 very high 0.19176875
##
## Tempchange
##              itv   meanrisk
## 1 [-0.393,0.311] 0.03237419
## 2  (0.311,0.661] 0.07312216
## 3   (0.661,1.01] 0.22091646
## 4    (1.01,1.36] 0.32457258
## 5    (1.36,1.71] 0.29258313
## 6    (1.71,2.07] 0.23839880
## 7    (2.07,3.22] 0.17535547
##
```

```
## Precipitation
##          itv    meanrisk
## 1 [42.5,133] -0.01608009
## 2  (133,222]  0.02460271
## 3  (222,312]  0.07515988
## 4  (312,402]  0.22508642
## 5  (402,492]  0.31663221
## 6  (492,582]  0.45128393
## 7  (582,672]  0.44689250
##
## GDP
##                  itv  meanrisk
## 1           [0,9.92] 0.1460310
## 2        (9.92,33.4] 0.2295733
## 3         (33.4,101] 0.2217597
## 4          (101,290] 0.2818238
## 5          (290,888] 0.2605960
## 6      (888,4.6e+03] 0.2748348
## 7 (4.6e+03,1.66e+04] 0.2642400
##
## Popdensity
##           itv   meanrisk
## 1    [0,0.23] 0.04825588
## 2  (0.23,0.9] 0.13872392
## 3  (0.9,2.02] 0.17562520
## 4 (2.02,5.25] 0.25928089
## 5 (5.25,14.8] 0.27412971
## 6 (14.8,40.6] 0.25845363
## 7  (40.6,971] 0.27435755
##
## Climatezone
##   interval  Bsk  Bwk  Dwa  Dwb  Dwc
## 1      Bsk <NA> <NA> <NA> <NA> <NA>
## 2      Bwk    Y <NA> <NA> <NA> <NA>
## 3      Dwa    Y    Y <NA> <NA> <NA>
## 4      Dwb    Y    Y    N <NA> <NA>
## 5      Dwc    Y    Y    Y    Y <NA>
##
## Mining
##    interval very low  low medium high very high
## 1  very low     <NA> <NA>   <NA> <NA>      <NA>
## 2       low        Y <NA>   <NA> <NA>      <NA>
## 3    medium        Y    Y   <NA> <NA>      <NA>
## 4      high        Y    Y      N <NA>      <NA>
## 5 very high        N    Y      Y    Y      <NA>
##
## Tempchange
##         interval [-0.393,0.311] (0.311,0.661] (0.661,1.01] (1.01,1.36]
## 1 [-0.393,0.311]           <NA>          <NA>         <NA>        <NA>
## 2  (0.311,0.661]              Y          <NA>         <NA>        <NA>
## 3   (0.661,1.01]              Y             Y         <NA>        <NA>
## 4    (1.01,1.36]              Y             Y            Y        <NA>
## 5    (1.36,1.71]              Y             Y            Y           N
## 6    (1.71,2.07]              Y             Y            N           Y
```

```
## 7    (2.07,3.22]             Y            Y           N           Y
##   (1.36,1.71] (1.71,2.07] (2.07,3.22]
## 1        <NA>        <NA>        <NA>
## 2        <NA>        <NA>        <NA>
## 3        <NA>        <NA>        <NA>
## 4        <NA>        <NA>        <NA>
## 5        <NA>        <NA>        <NA>
## 6           Y        <NA>        <NA>
## 7           Y           Y        <NA>
##
## Precipitation
##     interval [42.5,133] (133,222] (222,312] (312,402] (402,492] (492,582]
## 1 [42.5,133]       <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 2  (133,222]          Y      <NA>      <NA>      <NA>      <NA>      <NA>
## 3  (222,312]          Y         Y      <NA>      <NA>      <NA>      <NA>
## 4  (312,402]          Y         Y         Y      <NA>      <NA>      <NA>
## 5  (402,492]          Y         Y         Y         Y      <NA>      <NA>
## 6  (492,582]          Y         Y         Y         Y         Y      <NA>
## 7  (582,672]          Y         Y         Y         Y         Y         N
##   (582,672]
## 1      <NA>
## 2      <NA>
## 3      <NA>
## 4      <NA>
## 5      <NA>
## 6      <NA>
## 7      <NA>
##
## GDP
##             interval [0,9.92] (9.92,33.4] (33.4,101] (101,290] (290,888]
## 1           [0,9.92]     <NA>        <NA>       <NA>      <NA>      <NA>
## 2        (9.92,33.4]        Y        <NA>       <NA>      <NA>      <NA>
## 3         (33.4,101]        Y           N       <NA>      <NA>      <NA>
## 4          (101,290]        Y           Y          Y      <NA>      <NA>
## 5          (290,888]        Y           N          N         N      <NA>
## 6      (888,4.6e+03]        Y           N          Y         N         N
## 7 (4.6e+03,1.66e+04]        Y           N          N         N         N
##   (888,4.6e+03] (4.6e+03,1.66e+04]
## 1          <NA>               <NA>
## 2          <NA>               <NA>
## 3          <NA>               <NA>
## 4          <NA>               <NA>
## 5          <NA>               <NA>
## 6          <NA>               <NA>
## 7             N               <NA>
##
## Popdensity
##      interval [0,0.23] (0.23,0.9] (0.9,2.02] (2.02,5.25] (5.25,14.8]
## 1    [0,0.23]     <NA>       <NA>       <NA>        <NA>        <NA>
## 2  (0.23,0.9]        Y       <NA>       <NA>        <NA>        <NA>
## 3  (0.9,2.02]        Y          N       <NA>        <NA>        <NA>
## 4 (2.02,5.25]        Y          Y          Y        <NA>        <NA>
## 5 (5.25,14.8]        Y          Y          Y           N        <NA>
## 6 (14.8,40.6]        Y          Y          Y           N           N
```

```
## 7  (40.6,971]        Y          Y          Y           N           N
##   (14.8,40.6] (40.6,971]
## 1        <NA>       <NA>
## 2        <NA>       <NA>
## 3        <NA>       <NA>
## 4        <NA>       <NA>
## 5        <NA>       <NA>
## 6        <NA>       <NA>
## 7           N       <NA>
##
## Interaction detector:
##        variable Climatezone Mining Tempchange Precipitation    GDP Popdensity
## 1   Climatezone          NA     NA         NA            NA     NA          NA
## 2        Mining      0.8345     NA         NA            NA     NA          NA
## 3    Tempchange      0.8538 0.4223         NA            NA     NA          NA
## 4 Precipitation      0.9016 0.8861     0.9158            NA     NA          NA
## 5           GDP      0.8571 0.2438     0.3886        0.8959     NA          NA
## 6    Popdensity      0.8599 0.3588     0.4352        0.9035 0.2166          NA
##
## Ecological detector:
##        variable Climatezone Mining Tempchange Precipitation  GDP Popdensity
## 1   Climatezone        <NA>   <NA>       <NA>          <NA> <NA>       <NA>
## 2        Mining           N   <NA>       <NA>          <NA> <NA>       <NA>
## 3    Tempchange           N      Y       <NA>          <NA> <NA>       <NA>
## 4 Precipitation           Y      Y          Y          <NA> <NA>       <NA>
## 5           GDP           N      N          N             N <NA>       <NA>
## 6    Popdensity           N      N          N             N    N       <NA>
```

```
plot(ndvigdm)
```

Codes and data of the computation processes of the t-test for risk detector, interactions explored by the interaction detector and the F-test for ecological detector.

```
## t-test of risk detector for the variable Climatezone
ndvigdm$Risk.detector$Climatezone
```

```
##    itv1 itv2          t        df           sig risk
## 1   Bsk  Bwk  17.730358 315.88563  2.637662e-49    Y
## 2   Bsk  Dwa -11.809163  52.13343  2.360201e-16    Y
## 3   Bsk  Dwb -19.658136 327.45054  2.276521e-57    Y
## 4   Bsk  Dwc -37.998231 304.88974 1.143504e-117    Y
## 5   Bwk  Dwa -22.770854  38.27911  7.381701e-24    Y
## 6   Bwk  Dwb -40.614567 215.91365 4.556501e-103    Y
## 7   Bwk  Dwc -81.337846 296.50585 8.422103e-205    Y
## 8   Dwa  Dwb  -1.395461  55.29265  1.684563e-01    N
## 9   Dwa  Dwc  -8.808167  38.78614  8.589747e-11    Y
## 10  Dwb  Dwc -12.080308 215.01449  5.524023e-26    Y
```

```
## t-test of risk detector for the variable Tempchange
ndvigdm$Risk.detector$Tempchange
```

```
##              itv1          itv2           t       df          sig risk
## 1  [-0.393,0.311] (0.311,0.661]  -2.1784882 192.2786 3.058620e-02    Y
## 2  [-0.393,0.311]  (0.661,1.01]  -8.1978707 141.2153 1.345932e-13    Y
## 3  [-0.393,0.311]   (1.01,1.36] -16.0313363 234.0217 1.570643e-39    Y
## 4  [-0.393,0.311]   (1.36,1.71] -12.8583388 182.2434 2.393170e-27    Y
```

```
## 5  [-0.393,0.311]   (1.71,2.07] -10.9120573 189.5401 7.974504e-22    Y
## 6  [-0.393,0.311]   (2.07,3.22]  -7.4129666 170.1484 5.545589e-12    Y
## 7   (0.311,0.661]  (0.661,1.01]  -6.2085679 153.7650 4.757633e-09    Y
## 8   (0.311,0.661]   (1.01,1.36] -13.0734367 236.4896 9.325721e-30    Y
## 9   (0.311,0.661]   (1.36,1.71] -10.3773420 194.7078 2.309357e-20    Y
## 10  (0.311,0.661]   (1.71,2.07]  -8.3244706 199.8569 1.320714e-14    Y
## 11  (0.311,0.661]   (2.07,3.22]  -5.0504897 182.3930 1.063394e-06    Y
## 12   (0.661,1.01]   (1.01,1.36]  -4.4238549 156.7164 1.806223e-05    Y
## 13   (0.661,1.01]   (1.36,1.71]  -2.8636285 165.3820 4.731052e-03    Y
## 14   (0.661,1.01]   (1.71,2.07]  -0.7301574 154.8633 4.663970e-01    N
## 15   (0.661,1.01]   (2.07,3.22]   1.8775727 153.0666 6.234221e-02    N
## 16    (1.01,1.36]   (1.36,1.71]   1.5434067 214.0154 1.242095e-01    N
## 17    (1.01,1.36]   (1.71,2.07]   4.4408817 231.4842 1.387933e-05    Y
## 18    (1.01,1.36]   (2.07,3.22]   7.5356809 203.4925 1.565776e-12    Y
## 19    (1.36,1.71]   (1.71,2.07]   2.5434236 194.1126 1.175653e-02    Y
## 20    (1.36,1.71]   (2.07,3.22]   5.4107475 182.9856 1.947797e-07    Y
## 21    (1.71,2.07]   (2.07,3.22]   3.0896365 181.6485 2.319198e-03    Y
```

```
## interactions explored by the interaction detector
ndvigdm$Interaction.detector$Interaction
```

```
##            var1          var2        qv1        qv2      qv12
## 1    Climatezone        Mining 0.82183348 0.14111542 0.8344738
## 2    Climatezone    Tempchange 0.82183348 0.32655370 0.8537915
## 3    Climatezone Precipitation 0.82183348 0.86935048 0.9015820
## 4    Climatezone           GDP 0.82183348 0.09514837 0.8571228
## 5    Climatezone    Popdensity 0.82183348 0.19618770 0.8599232
## 6         Mining    Tempchange 0.14111542 0.32655370 0.4223200
## 7         Mining Precipitation 0.14111542 0.86935048 0.8861185
## 8         Mining           GDP 0.14111542 0.09514837 0.2438202
## 9         Mining    Popdensity 0.14111542 0.19618770 0.3587628
## 10    Tempchange Precipitation 0.32655370 0.86935048 0.9158048
## 11    Tempchange           GDP 0.32655370 0.09514837 0.3885994
## 12    Tempchange    Popdensity 0.32655370 0.19618770 0.4351518
## 13 Precipitation           GDP 0.86935048 0.09514837 0.8958678
## 14 Precipitation    Popdensity 0.86935048 0.19618770 0.9035324
## 15           GDP    Popdensity 0.09514837 0.19618770 0.2165923
##           interaction
## 1        Enhance, bi-
## 2        Enhance, bi-
## 3        Enhance, bi-
## 4        Enhance, bi-
## 5        Enhance, bi-
## 6        Enhance, bi-
## 7        Enhance, bi-
## 8  Enhance, nonlinear
## 9  Enhance, nonlinear
## 10       Enhance, bi-
## 11       Enhance, bi-
## 12       Enhance, bi-
## 13       Enhance, bi-
## 14       Enhance, bi-
## 15       Enhance, bi-
```

```
## F-test of ecological detector
ndvigdm$Ecological.detector$Ecological
```

```
##             var1          var2         f          sig eco
## 1    Climatezone        Mining 0.2080971 1.000000e+00   N
## 2    Climatezone    Tempchange 0.2637874 1.000000e+00   N
## 3    Climatezone Precipitation 1.3616512 1.987113e-05   Y
## 4    Climatezone           GDP 0.1968338 1.000000e+00   N
## 5    Climatezone    Popdensity 0.2209808 1.000000e+00   N
## 6         Mining    Tempchange 1.2676168 7.915003e-04   Y
## 7         Mining Precipitation 6.5433453 0.000000e+00   Y
## 8         Mining           GDP 0.9458750 7.709827e-01   N
## 9         Mining    Popdensity 1.0619118 2.115358e-01   N
## 10    Tempchange Precipitation 5.1619268 0.000000e+00   Y
## 11    Tempchange           GDP 0.7461836 9.999515e-01   N
## 12    Tempchange    Popdensity 0.8377230 9.908628e-01   N
## 13 Precipitation           GDP 0.1445553 1.000000e+00   N
## 14 Precipitation    Popdensity 0.1622888 1.000000e+00   N
## 15           GDP    Popdensity 1.1226767 6.143133e-02   N
```

Comparison of size effects of spatial units (Figure 9).

```
ndvilist <- list(ndvi_5, ndvi_10, ndvi_20, ndvi_30, ndvi_40, ndvi_50)
su <- c(5,10,20,30,40,50) ## sizes of spatial units
## set optional parameters of optimal discretization
discmethod <- c("equal","natural","quantile","geometric","sd")
discitv <- 3:7
## "gdm" function (~ 108 s)
gdlist <- list()
for (i in 1:6){
    gdlist[[i]] <- gdm(NDVIchange ~ .,
      continuous_variable = c("Tempchange", "Precipitation", "GDP", "Popdensity"),
      data = ndvilist[[i]], discmethod = discmethod, discitv = discitv)
}
## size effects of spatial units
sesu(gdlist, su)
```

4.2 Spatial point or areal data of H1N1 flu incidences

4.2.1 In the whole study area

```
data(h1n1_100)
head(h1n1_100)
```

```
##   H1N1   temp     prec    humi   popd   gdpd  rdds sensepop urbanpop medicost
## 1 2.02 22.257 2169.194 28.0085 171.24 0.4074 23.33  26.8913    52.74  94.1805
## 2 2.02 22.730 2131.414 44.7943 213.10 0.7876 26.55  26.8913    52.74  94.1805
## 3 2.02 23.288 2438.123 45.8407 288.81 1.5207 38.17  26.8913    52.74  94.1805
## 4 1.45 22.914 2251.993 24.3231 719.93 1.9710 53.61  23.9246    67.76  64.8081
## 5 1.45 22.566 2355.137 33.6370 791.68 3.4102 57.40  23.9246    67.76  64.8081
## 6 1.09 21.169 1658.817 34.9150  90.36 0.2519 17.24  28.0352    40.48  77.7068
##   Georegion
## 1         S
## 2         S
## 3         S
```

```
## 4         S
## 5         S
## 6         W
```

Results of OPGD-based analysis for H1N1 flu incidences (Figure 10).

```
h1n1list <- list(h1n1_50, h1n1_100, h1n1_150)
su <- c(50, 100, 150)
## set optional parameters of optimal discretization
discmethod <- c("equal","natural","quantile","geometric","sd")
discitv <- 4:7
continuous_variable <- colnames(h1n1_50)[-c(1,11)]
## "gdm" function (~ 67 s)
gdlist <- list()
for (i in 1:3){
  gdlist[[i]] <- gdm(H1N1 ~ .,
                continuous_variable = continuous_variable,
                data = h1n1list[[i]],
                discmethod = discmethod, discitv = discitv)
}
## size effects of spatial units
sesu(gdlist, su)
```

```
## recalculation with 100-km spatial unit
discmethod <- c("equal","natural","quantile","geometric","sd")
discitv <- 4:7
continuous_variable <- colnames(h1n1_100)[-c(1,11)]
h1n1.gdm.100 <- gdm(H1N1 ~ .,
                continuous_variable = continuous_variable,
                data = h1n1_100,
                discmethod = discmethod, discitv = discitv)
h1n1.gdm.100
```

```
## Explanatory variables include 9 continuous variables.
##
## optimal discretization result of temp
## method             :  equal
## number of intervals:  7
## intervals:
##  -7.092 -2.740286 1.611429 5.963143 10.31486 14.66657 19.01829 23.37
## numbers of data within intervals:
##  78 175 178 167 148 172 69
##
## optimal discretization result of prec
## method             :  natural
## number of intervals:  7
## intervals:
##  13.13 218.9193 407.3756 611.4574 828.5553 1176.74 1750.94 2569.669
## numbers of data within intervals:
##  171 153 196 204 108 120 35
##
## optimal discretization result of humi
## method             :  quantile
## number of intervals:  6
## intervals:
```

```
##  -109.0914 -54.95307 -29.4622 -8.1786 12.91037 41.38707 177.9288
## numbers of data within intervals:
##  165 164 165 164 164 165
##
## optimal discretization result of popd
## method             :  natural
## number of intervals:  6
## intervals:
##  0 15.4227 81.1539 225.1411 469.9451 989.8658 2786.56
## numbers of data within intervals:
##  471 140 184 100 71 21
##
## optimal discretization result of gdpd
## method             :  natural
## number of intervals:  7
## intervals:
##  0 0.059117 0.288647 0.666049 1.548748 3.410617 7.495601 22.1713
## numbers of data within intervals:
##  460 134 138 123 76 42 14
##
## optimal discretization result of rdds
## method             :  natural
## number of intervals:  7
## intervals:
##  0 3.9791 11.2851 21.8487 37.4985 62.3373 131.1288 316.5
## numbers of data within intervals:
##  396 150 145 153 73 55 15
##
## optimal discretization result of sensepop
## method             :  sd
## number of intervals:  6
## intervals:
##  18.5058 22.74193 24.14647 25.551 26.95554 28.36007 29.76461 31.3683
## numbers of data within intervals:
##  212 33 56 75 318 250 43
##
## optimal discretization result of urbanpop
## method             :  sd
## number of intervals:  4
## intervals:
##  23.71 35.0363 46.59943 58.16257 86.3
## numbers of data within intervals:
##  133 369 292 193
##
## optimal discretization result of medicost
## method             :  quantile
## number of intervals:  7
## intervals:
##  60.1877 64.8081 72.8208 79.1247 86.1034 95.1002 143.9295 158.2044
## numbers of data within intervals:
##  149 137 141 229 115 82 134
##
## Geographical detectors results:
##
```

```
## Factor detector:
##     variable         qv          sig
## 1       temp 0.49088033 1.838595e-10
## 2       prec 0.40566048 3.619388e-10
## 3       humi 0.17352575 2.380844e-10
## 4       popd 0.22000829 4.038251e-10
## 5       gdpd 0.21462378 7.726012e-10
## 6       rdds 0.23398371 7.034787e-10
## 7   sensepop 0.43308776 3.418482e-10
## 8   urbanpop 0.34291104 6.720092e-10
## 9   medicost 0.45571331 3.509079e-10
## 10 Georegion 0.07399135 6.481682e-11
##
## Risk detector:
## temp
##             itv  meanrisk
## 1 [-7.09,-2.74] 0.2285897
## 2  (-2.74,1.61] 0.2637714
## 3   (1.61,5.96] 0.3753933
## 4   (5.96,10.3] 0.6703593
## 5   (10.3,14.7] 0.6264189
## 6     (14.7,19] 0.8155233
## 7     (19,23.4] 1.5410145
##
## prec
##                   itv  meanrisk
## 1          [13.1,219] 0.6188889
## 2           (219,407] 0.4271242
## 3           (407,611] 0.4032143
## 4           (611,829] 0.3638725
## 5      (829,1.18e+03] 0.6881481
## 6 (1.18e+03,1.75e+03] 1.0233333
## 7 (1.75e+03,2.57e+03] 1.7288571
##
## humi
##             itv  meanrisk
## 1    [-109,-55] 0.4675152
## 2   (-55,-29.5] 0.5257927
## 3 (-29.5,-8.18] 0.5613333
## 4  (-8.18,12.9] 0.3378659
## 5   (12.9,41.4] 0.6747561
## 6    (41.4,178] 0.9697576
##
## popd
##              itv  meanrisk
## 1       [0,15.4] 0.3786624
## 2    (15.4,81.2] 0.6187857
## 3     (81.2,225] 0.8997826
## 4      (225,470] 0.8733000
## 5      (470,990] 0.6371831
## 6 (990,2.79e+03] 0.9028571
##
## gdpd
##              itv  meanrisk
```

```
## 1     [0,0.0591] 0.3791739
## 2 (0.0591,0.289] 0.6033582
## 3  (0.289,0.666] 0.8801449
## 4   (0.666,1.55] 0.9006504
## 5    (1.55,3.41] 0.7182895
## 6     (3.41,7.5] 0.6107143
## 7     (7.5,22.2] 1.0228571
##
## rdds
##           itv  meanrisk
## 1    [0,3.98] 0.3558586
## 2 (3.98,11.3] 0.5077333
## 3 (11.3,21.8] 0.8104828
## 4 (21.8,37.5] 0.9307843
## 5 (37.5,62.3] 0.7598630
## 6  (62.3,131] 0.6310909
## 7   (131,316] 0.9920000
##
## sensepop
##           itv  meanrisk
## 1 [18.5,22.7] 0.5559434
## 2 (22.7,24.1] 1.3181818
## 3 (24.1,25.6] 0.8432143
## 4   (25.6,27] 0.7149333
## 5   (27,28.4] 0.5564465
## 6 (28.4,29.8] 0.2841600
## 7 (29.8,31.4] 1.6716279
##
## urbanpop
##           itv  meanrisk
## 1   [23.7,35] 0.0300000
## 2   (35,46.6] 0.7440108
## 3 (46.6,58.2] 0.4399315
## 4 (58.2,86.3] 0.9071503
##
## medicost
##           itv   meanrisk
## 1 [60.2,64.8] 0.48510067
## 2 (64.8,72.8] 0.64489051
## 3 (72.8,79.1] 1.14609929
## 4 (79.1,86.1] 0.60742358
## 5 (86.1,95.1] 0.82208696
## 6  (95.1,144] 0.25109756
## 7   (144,158] 0.04186567
##
## Georegion
##   itv  meanrisk
## 1   N 0.3687603
## 2   S 0.8471154
## 3   W 0.5708451
##
## temp
##        interval [-7.09,-2.74] (-2.74,1.61] (1.61,5.96] (5.96,10.3] (10.3,14.7]
## 1 [-7.09,-2.74]          <NA>         <NA>        <NA>        <NA>        <NA>
```

```
## 2  (-2.74,1.61]            N         <NA>        <NA>        <NA>        <NA>
## 3   (1.61,5.96]            Y            Y        <NA>        <NA>        <NA>
## 4   (5.96,10.3]            Y            Y           Y        <NA>        <NA>
## 5   (10.3,14.7]            Y            Y           Y           N        <NA>
## 6     (14.7,19]            Y            Y           Y           Y           Y
## 7     (19,23.4]            Y            Y           Y           Y           Y
##   (14.7,19] (19,23.4]
## 1      <NA>      <NA>
## 2      <NA>      <NA>
## 3      <NA>      <NA>
## 4      <NA>      <NA>
## 5      <NA>      <NA>
## 6      <NA>      <NA>
## 7         Y      <NA>
##
## prec
##              interval [13.1,219] (219,407] (407,611] (611,829] (829,1.18e+03]
## 1          [13.1,219]       <NA>      <NA>      <NA>      <NA>           <NA>
## 2           (219,407]          Y      <NA>      <NA>      <NA>           <NA>
## 3           (407,611]          Y         N      <NA>      <NA>           <NA>
## 4           (611,829]          Y         N         N      <NA>           <NA>
## 5      (829,1.18e+03]          N         Y         Y         Y           <NA>
## 6 (1.18e+03,1.75e+03]          Y         Y         Y         Y              Y
## 7 (1.75e+03,2.57e+03]          Y         Y         Y         Y              Y
##   (1.18e+03,1.75e+03] (1.75e+03,2.57e+03]
## 1                <NA>                <NA>
## 2                <NA>                <NA>
## 3                <NA>                <NA>
## 4                <NA>                <NA>
## 5                <NA>                <NA>
## 6                <NA>                <NA>
## 7                   Y                <NA>
##
## humi
##        interval [-109,-55] (-55,-29.5] (-29.5,-8.18] (-8.18,12.9] (12.9,41.4]
## 1    [-109,-55]       <NA>        <NA>          <NA>         <NA>        <NA>
## 2   (-55,-29.5]          N        <NA>          <NA>         <NA>        <NA>
## 3 (-29.5,-8.18]          Y           N          <NA>         <NA>        <NA>
## 4  (-8.18,12.9]          Y           Y             Y         <NA>        <NA>
## 5   (12.9,41.4]          Y           Y             Y            Y        <NA>
## 6    (41.4,178]          Y           Y             Y            Y           Y
##   (41.4,178]
## 1       <NA>
## 2       <NA>
## 3       <NA>
## 4       <NA>
## 5       <NA>
## 6       <NA>
##
## popd
##         interval [0,15.4] (15.4,81.2] (81.2,225] (225,470] (470,990]
## 1       [0,15.4]     <NA>        <NA>       <NA>      <NA>      <NA>
## 2    (15.4,81.2]        Y        <NA>       <NA>      <NA>      <NA>
## 3     (81.2,225]        Y           Y       <NA>      <NA>      <NA>
```

```
## 4      (225,470]        Y           Y          N      <NA>      <NA>
## 5      (470,990]        Y           N          Y         Y      <NA>
## 6 (990,2.79e+03]        Y           Y          N         N         Y
##   (990,2.79e+03]
## 1           <NA>
## 2           <NA>
## 3           <NA>
## 4           <NA>
## 5           <NA>
## 6           <NA>
##
## gdpd
##        interval [0,0.0591] (0.0591,0.289] (0.289,0.666] (0.666,1.55]
## 1     [0,0.0591]       <NA>           <NA>          <NA>         <NA>
## 2 (0.0591,0.289]          Y           <NA>          <NA>         <NA>
## 3  (0.289,0.666]          Y              Y          <NA>         <NA>
## 4   (0.666,1.55]          Y              Y             N         <NA>
## 5    (1.55,3.41]          Y              N             Y            Y
## 6     (3.41,7.5]          Y              N             Y            Y
## 7     (7.5,22.2]          Y              Y             N            N
##   (1.55,3.41] (3.41,7.5] (7.5,22.2]
## 1        <NA>       <NA>       <NA>
## 2        <NA>       <NA>       <NA>
## 3        <NA>       <NA>       <NA>
## 4        <NA>       <NA>       <NA>
## 5        <NA>       <NA>       <NA>
## 6           N       <NA>       <NA>
## 7           Y          Y       <NA>
##
## rdds
##      interval [0,3.98] (3.98,11.3] (11.3,21.8] (21.8,37.5] (37.5,62.3]
## 1    [0,3.98]     <NA>        <NA>        <NA>        <NA>        <NA>
## 2 (3.98,11.3]        Y        <NA>        <NA>        <NA>        <NA>
## 3 (11.3,21.8]        Y           Y        <NA>        <NA>        <NA>
## 4 (21.8,37.5]        Y           Y           N        <NA>        <NA>
## 5 (37.5,62.3]        Y           Y           N           Y        <NA>
## 6  (62.3,131]        Y           Y           Y           Y           N
## 7   (131,316]        Y           Y           N           N           N
##   (62.3,131] (131,316]
## 1       <NA>      <NA>
## 2       <NA>      <NA>
## 3       <NA>      <NA>
## 4       <NA>      <NA>
## 5       <NA>      <NA>
## 6       <NA>      <NA>
## 7          Y      <NA>
##
## sensepop
##      interval [18.5,22.7] (22.7,24.1] (24.1,25.6] (25.6,27] (27,28.4]
## 1 [18.5,22.7]        <NA>        <NA>        <NA>      <NA>      <NA>
## 2 (22.7,24.1]           Y        <NA>        <NA>      <NA>      <NA>
## 3 (24.1,25.6]           Y           Y        <NA>      <NA>      <NA>
## 4   (25.6,27]           Y           Y           N      <NA>      <NA>
## 5   (27,28.4]           N           Y           Y         Y      <NA>
```

```
## 6 (28.4,29.8]           Y           Y           Y         Y         Y
## 7 (29.8,31.4]           Y           Y           Y         Y         Y
##   (28.4,29.8] (29.8,31.4]
## 1        <NA>        <NA>
## 2        <NA>        <NA>
## 3        <NA>        <NA>
## 4        <NA>        <NA>
## 5        <NA>        <NA>
## 6        <NA>        <NA>
## 7           Y        <NA>
##
## urbanpop
##      interval [23.7,35] (35,46.6] (46.6,58.2] (58.2,86.3]
## 1   [23.7,35]      <NA>      <NA>        <NA>        <NA>
## 2   (35,46.6]         Y      <NA>        <NA>        <NA>
## 3 (46.6,58.2]         Y         Y        <NA>        <NA>
## 4 (58.2,86.3]         Y         Y           Y        <NA>
##
## medicost
##      interval [60.2,64.8] (64.8,72.8] (72.8,79.1] (79.1,86.1] (86.1,95.1]
## 1 [60.2,64.8]        <NA>        <NA>        <NA>        <NA>        <NA>
## 2 (64.8,72.8]           Y        <NA>        <NA>        <NA>        <NA>
## 3 (72.8,79.1]           Y           Y        <NA>        <NA>        <NA>
## 4 (79.1,86.1]           Y           N           Y        <NA>        <NA>
## 5 (86.1,95.1]           Y           Y           Y           Y        <NA>
## 6  (95.1,144]           Y           Y           Y           Y           Y
## 7   (144,158]           Y           Y           Y           Y           Y
##   (95.1,144] (144,158]
## 1       <NA>      <NA>
## 2       <NA>      <NA>
## 3       <NA>      <NA>
## 4       <NA>      <NA>
## 5       <NA>      <NA>
## 6       <NA>      <NA>
## 7          Y      <NA>
##
## Georegion
##   interval    N    S    W
## 1        N <NA> <NA> <NA>
## 2        S    Y <NA> <NA>
## 3        W    Y    Y <NA>
##
## Interaction detector:
##     variable   temp   prec   humi   popd   gdpd   rdds sensepop urbanpop
## 1       temp     NA     NA     NA     NA     NA     NA       NA       NA
## 2       prec 0.5837     NA     NA     NA     NA     NA       NA       NA
## 3       humi 0.5754 0.4510     NA     NA     NA     NA       NA       NA
## 4       popd 0.5483 0.4984 0.3903     NA     NA     NA       NA       NA
## 5       gdpd 0.5483 0.5043 0.3906 0.2309     NA     NA       NA       NA
## 6       rdds 0.5482 0.5343 0.4138 0.2427 0.2444     NA       NA       NA
## 7   sensepop 0.7005 0.7133 0.6048 0.5992 0.6137 0.6125       NA       NA
## 8   urbanpop 0.6902 0.6333 0.4970 0.5177 0.5204 0.5369   0.7739       NA
## 9   medicost 0.7732 0.7568 0.6910 0.5951 0.6038 0.6254   0.7940   0.6127
## 10 Georegion 0.5266 0.4431 0.2183 0.3298 0.3261 0.3459   0.5874   0.4472
```

```
##    medicost Georegion
## 1        NA         NA
## 2        NA         NA
## 3        NA         NA
## 4        NA         NA
## 5        NA         NA
## 6        NA         NA
## 7        NA         NA
## 8        NA         NA
## 9        NA         NA
## 10   0.6297         NA
##
## Ecological detector:
##     variable temp prec humi popd gdpd rdds sensepop urbanpop medicost Georegion
## 1       temp <NA> <NA> <NA> <NA> <NA> <NA>     <NA>     <NA>     <NA>      <NA>
## 2       prec    N <NA> <NA> <NA> <NA> <NA>     <NA>     <NA>     <NA>      <NA>
## 3       humi    N    N <NA> <NA> <NA> <NA>     <NA>     <NA>     <NA>      <NA>
## 4       popd    N    N    N <NA> <NA> <NA>     <NA>     <NA>     <NA>      <NA>
## 5       gdpd    N    N    N    N <NA> <NA>     <NA>     <NA>     <NA>      <NA>
## 6       rdds    N    N    N    N    N <NA>     <NA>     <NA>     <NA>      <NA>
## 7   sensepop    N    N    Y    Y    Y    Y     <NA>     <NA>     <NA>      <NA>
## 8   urbanpop    N    N    Y    Y    Y    Y        N     <NA>     <NA>      <NA>
## 9   medicost    N    N    Y    Y    Y    Y        N        Y     <NA>      <NA>
## 10 Georegion    N    N    N    N    N    N        N        N        N      <NA>
```

```r
plot(h1n1.gdm.100)
```

Spatial analysis for data in three sub-regions (Figure 11).

```r
## data in different regions
h1n1list <- list(h1n1_50, h1n1_100, h1n1_150)
h1n1.N <- lapply(h1n1list, function(x) x[which(x$Georegion == "N"), 1:10])
h1n1.S <- lapply(h1n1list, function(x) x[which(x$Georegion == "S"), 1:10])
h1n1.W <- lapply(h1n1list, function(x) x[which(x$Georegion == "W"), 1:10])
## select spatial unit
su <- c(50, 100, 150)
## set optional parameters of optimal discretization
discmethod <- c("equal","natural","quantile","geometric","sd")
discitv <- c(3:4)
continuous_variable <- colnames(h1n1_50)[-c(1,11)]
## "gdm" for region "N"
gdlist <- list()
for (i in 1:3){
  gdlist[[i]] <- gdm(H1N1 ~ .,
                continuous_variable = continuous_variable,
                data = h1n1.N[[i]],
                discmethod = discmethod, discitv = discitv)
}
sesu(gdlist, su) # => 150 km
## "gdm" for region "S"
gdlist <- list()
for (i in 1:3){
  gdlist[[i]] <- gdm(H1N1 ~ .,
                continuous_variable = continuous_variable,
                data = h1n1.S[[i]],
```

```
                discmethod = discmethod, discitv = discitv)
}
sesu(gdlist, su) # => 150 km
## "gdm" for region "W"
gdlist <- list()
for (i in 1:3){
  gdlist[[i]] <- gdm(H1N1 ~ .,
                continuous_variable = continuous_variable,
                data = h1n1.W[[i]],
                discmethod = discmethod, discitv = discitv)
}
sesu(gdlist, su) # => 150 km
## recalculation for datasets in different regions
## region "N"
h1n1.N.150 <- gdm(H1N1 ~ .,
                continuous_variable = continuous_variable,
                data = h1n1.N[[3]],
                discmethod = discmethod, discitv = discitv)
h1n1.N.150
plot(h1n1.N.150)
## region "S"
h1n1.S.150 <- gdm(H1N1 ~ .,
                continuous_variable = continuous_variable,
                data = h1n1.S[[3]],
                discmethod = discmethod, discitv = discitv)
h1n1.S.150
plot(h1n1.S.150)
## region "W"
h1n1.W.150 <- gdm(H1N1 ~ .,
                continuous_variable = continuous_variable,
                data = h1n1.W[[3]],
                discmethod = discmethod, discitv = discitv)
h1n1.W.150
plot(h1n1.W.150)
```

4.3 Spatial line segment data of road damage

```
data(road_GD)
head(road_GD)
```

```
##    damage speed soiltype population vehicles
## 1 325.772   110  Tenosol     227.42      420
## 2 325.772   110 Hydrosol     227.42      420
## 3 325.772   110 Hydrosol     227.42      420
## 4 325.772   110 Hydrosol     227.42      420
## 5 325.772   110 Hydrosol     227.42      420
## 6 325.772   110 Hydrosol     227.42      420
```

Results of OPGD-based analysis for the road damage conditions (Figure 12).

```
## set optional discretization methods and numbers of intervals
discmethod <- c("equal","natural","quantile","geometric","sd")
discitv <- c(3:7)
```

```
continuous_variable <- colnames(road_GD)[c(4,5)]
## geographical detectors with optimal parameters
gdm_road <- gdm(damage ~ .,
                continuous_variable = continuous_variable,
                data = road_GD,
                discmethod = discmethod, discitv = discitv)
gdm_road
```

```
## Explanatory variables include 2 continuous variables.
##
## optimal discretization result of population
## method             :  quantile
## number of intervals:  5
## intervals:
##  1.04 10.4 53.13 215.19 357.15 2489.72
## numbers of data within intervals:
##  1001 1008 1038 982 971
##
## optimal discretization result of vehicles
## method             :  equal
## number of intervals:  7
## intervals:
##  100 815 1530 2245 2960 3675 4390 5105
## numbers of data within intervals:
##  2986 1020 468 115 224 117 70
##
## Geographical detectors results:
##
## Factor detector:
##     variable        qv          sig
## 1      speed 0.0108854 2.334147e-05
## 2   soiltype 0.1977133 3.204566e-10
## 3 population 0.0846682 6.368206e-11
## 4   vehicles 0.1202810 3.536698e-10
##
## Risk detector:
## speed
##     itv meanrisk
## 1 40-50 356.1679
## 2 60-70 374.0222
## 3    80 342.7952
## 4    90 336.2123
## 5   100 300.0822
## 6   110 387.7802
##
## soiltype
##          itv meanrisk
## 1 Calcarosol 473.1282
## 2  Chromosol 370.2571
## 3   Hydrosol 366.9375
## 4   Kandosol 340.3882
## 5    Podosol 869.3997
## 6    Sodosol 347.1542
## 7    Tenosol 530.9144
```

```
##
## population
##              itv meanrisk
## 1    [1.04,10.4] 354.3701
## 2    (10.4,53.1] 484.5460
## 3     (53.1,215] 353.8543
## 4      (215,357] 358.7460
## 5 (357,2.49e+03] 342.4888
##
## vehicles
##                   itv meanrisk
## 1           [100,815] 354.0679
## 2      (815,1.53e+03] 412.7972
## 3 (1.53e+03,2.24e+03] 438.7350
## 4 (2.24e+03,2.96e+03] 300.5194
## 5 (2.96e+03,3.68e+03] 296.8623
## 6 (3.68e+03,4.39e+03] 715.4692
## 7 (4.39e+03,5.10e+03] 384.2573
##
## speed
##   interval 40-50 60-70   80   90  100  110
## 1    40-50  <NA>  <NA> <NA> <NA> <NA> <NA>
## 2    60-70     N  <NA> <NA> <NA> <NA> <NA>
## 3       80     N     Y <NA> <NA> <NA> <NA>
## 4       90     N     Y    N <NA> <NA> <NA>
## 5      100     Y     Y    Y    Y <NA> <NA>
## 6      110     Y     N    Y    Y    Y <NA>
##
## soiltype
##     interval Calcarosol Chromosol Hydrosol Kandosol Podosol Sodosol Tenosol
## 1 Calcarosol       <NA>      <NA>     <NA>     <NA>    <NA>    <NA>    <NA>
## 2  Chromosol          Y      <NA>     <NA>     <NA>    <NA>    <NA>    <NA>
## 3   Hydrosol          Y         N     <NA>     <NA>    <NA>    <NA>    <NA>
## 4   Kandosol          Y         Y        Y     <NA>    <NA>    <NA>    <NA>
## 5    Podosol          Y         Y        Y        Y    <NA>    <NA>    <NA>
## 6    Sodosol          Y         Y        Y        Y       Y    <NA>    <NA>
## 7    Tenosol          Y         Y        Y        Y       Y       Y    <NA>
##
## population
##         interval [1.04,10.4] (10.4,53.1] (53.1,215] (215,357] (357,2.49e+03]
## 1    [1.04,10.4]        <NA>        <NA>       <NA>      <NA>           <NA>
## 2    (10.4,53.1]           Y        <NA>       <NA>      <NA>           <NA>
## 3     (53.1,215]           N           Y       <NA>      <NA>           <NA>
## 4      (215,357]           N           Y          N      <NA>           <NA>
## 5 (357,2.49e+03]           Y           Y          Y         Y           <NA>
##
## vehicles
##              interval [100,815] (815,1.53e+03] (1.53e+03,2.24e+03]
## 1           [100,815]      <NA>           <NA>                <NA>
## 2      (815,1.53e+03]         Y           <NA>                <NA>
## 3 (1.53e+03,2.24e+03]         Y              Y                <NA>
## 4 (2.24e+03,2.96e+03]         Y              Y                   Y
## 5 (2.96e+03,3.68e+03]         Y              Y                   Y
## 6 (3.68e+03,4.39e+03]         Y              Y                   Y
```

```
## 7 (4.39e+03,5.10e+03]         Y              N                   Y
##   (2.24e+03,2.96e+03] (2.96e+03,3.68e+03] (3.68e+03,4.39e+03]
## 1                <NA>                <NA>                <NA>
## 2                <NA>                <NA>                <NA>
## 3                <NA>                <NA>                <NA>
## 4                <NA>                <NA>                <NA>
## 5                   N                <NA>                <NA>
## 6                   Y                   Y                <NA>
## 7                   Y                   Y                   Y
##   (4.39e+03,5.10e+03]
## 1                <NA>
## 2                <NA>
## 3                <NA>
## 4                <NA>
## 5                <NA>
## 6                <NA>
## 7                <NA>
##
## Interaction detector:
##     variable  speed soiltype population vehicles
## 1      speed     NA       NA         NA       NA
## 2   soiltype 0.2428       NA         NA       NA
## 3 population 0.1173   0.3158         NA       NA
## 4   vehicles 0.1438   0.4712     0.4237       NA
##
## Ecological detector:
##     variable speed soiltype population vehicles
## 1      speed  <NA>     <NA>       <NA>     <NA>
## 2   soiltype     Y     <NA>       <NA>     <NA>
## 3 population     Y        N       <NA>     <NA>
## 4   vehicles     Y        N          N     <NA>
```

```
plot(gdm_road)
```