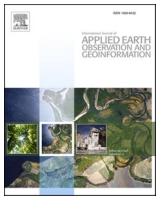
International Journal of Applied Earth Observations and Geoinformation 109 (2022) 102782

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Robust geographical detector

Zehua Zhang, Yongze Song *, Peng Wu

School of Design and the Built Environment, Curtin University, Perth 6102, Australia

ARTICLE INFO

Keywords:
Robust geographical detector
Optimization of spatial zones
Change point detection
Spatial data discretization
Google Earth Engine
Spatial heterogeneity

ABSTRACT

Geographical detector (GD) is a method to measure spatial associations using a power of determinant (PD) value that compares the variance of data within spatial zones and in the whole study area. Recent studies have implemented GD in diverse fields, such as environmental and socio-economic issues. Spatial data discretization is an essential stage for determining zones using explanatory variables. However, the spatial data discretization process has been sensitive to the GD results. To address this issue, this article proposes a Robust Geographical Detector (RGD) to overcome the limitations of the sensitivity in spatial data discretization and estimate robust PD values of explanatory variables using a B-value. The RGD determines spatial zones with numerical interval breaks using an optimization algorithm of variance-based change point detection. In this study, RGD is implemented in a nationwide case study exploring potential factors of nitrogen dioxide (NO_2) density in industrial regions across Australia, where data on both NO_2 and potential factors are sourced from satellite images and remote sensing products using Google Earth Engine. Results show that RGD can effectively explore the maximum PD values of spatial associations between response and explanatory variables due to the optimization algorithm-based spatial zones. In addition, RGD-based PD values are generally higher, more robust, and more stable than GD-based PD values since RGD can guarantee the increment of PD values with the increase of interval numbers, which is a challenge in previous GD models. Finally, RGD could provide a more reliable interpretation of PD as RGD finds optimal intervals-based spatial zones determined by potential factors. This study demonstrates that the developed RGD model can provide robust and reliable solutions to explore spatial associations and identify geographical factors.

1. Introduction

Spatial heterogeneity or spatial disparity is a key characteristic of geographical phenomena in environmental and socio-economic studies (Fotheringham et al., 1998; Weisent et al., 2012; Fang et al., 2017). Spatial heterogeneity refers to a common phenomenon in which a factor's geographic impacts, or distributions of specific features or events, are not consistent across the space (Fotheringham, 2002; Wang et al., 2010). Spatial stratified heterogeneity (SSH) is a form of spatial heterogeneity that can be used to explore spatial associations between geographical variables through the comparison between variance within-strata and variance inter-strata (Wang et al., 2016). Thus, SSH-based models have been widely applied in identifying geographical factors in multiple fields, such as urbanization (Feng et al., 2021), land use and environmental planning (Liu et al., 2019), public health (Li et al., 2021), transport infrastructure (Song et al., 2018a), environmental justice (Dasgupta et al., 2021), ecological protection (Zuo et al., 2021) and climate change (Jiang et al., 2018).

Geographical detector (GD) is a method to measure the SSH by statistical variance (Wang et al., 2016). In GD models, the association between dependent and explanatory variables is quantified using a power of determinant (PD) value, comparing variance within strata and across the whole study area (Wang et al., 2010). GD has been proposed and widely applied in geography for a decade, with solid theories proven. Current GD has been well-developed in both diverse applications and methodology extensions. From the application perspective, GD is a powerful tool for examining spatial differences (Chen et al., 2019), identifying driving factors (He et al., 2019), and supporting spatial advice (Dong et al., 2021). Practicality regarding the spatial analysis advantage of GD has been shown in various studies, from human settlement management to human-environment interaction investigation, at different spatial scales (Raghavan et al., 2013; Qu et al., 2018; Maus et al., 2020; Song et al., 2021). From a methodology extension perspective, optimal parameters regarding break interval and spatial

* Corresponding author.
E-mail address: Yongze.song@curtin.edu.au (Y. Song).

https://doi.org/10.1016/j.jag.2022.102782
Received 5 March 2022; Received in revised form 5 April 2022; Accepted 6 April 2022
Available online 29 April 2022

scale have been investigated to improve GD performance (Cao et al., 2013). A spatial association interactive detector based on GD theory is proposed to quantify spatial associations between spatial cause and effect (Song and Wu, 2021). Moreover, a geographically optimal zones-based heterogeneity model is developed to improve the measure of SSH based on GD (Luo et al., 2022). These advanced GD methods have been applied in infrastructure management and soil moisture modeling.

However, the process of spatial data discretization has been a sensitive stage for exploring spatial associations, computation of PD values, and identification of geographical variables. This means that the changes in the spatial discretization method and the number of spatial zones can usually affect the relative importance of variables. In studies where explanatory variables are continuous numerical data, spatial data discretization is essential before performing GD models (Wang et al., 2016). In natural and social environment studies, continuous numerical data of explanatory variables are common, such as population, economic conditions, wind speed, precipitation, air pollutant indicator, and vegetation coverage. Therefore, developing a practical discretization approach for continuous numeric data is essential for practical implementations of GD models. To address this issue, an optimal parameter-based geographic detector (OPGD) is developed to improve the GD factor detector by providing various discretization strategies based on the statistical distribution of explanatory variables (Song et al., 2020). However, these discretization strategies do not fully address the limitations of deriving reliable strata with robust discretization approaches. In detail, PD values derived from OPGD fluctuate with the increase of interval breaks when selecting optimal discretization parameters (Song et al., 2020; Luo et al., 2021), meaning that the stability and robustness of spatial data discretization are limited. This is because most of the current spatial discretization strategies, including the system developed in OPGD, are performed based on observations of samples instead of in-depth characteristics of data. Therefore, more effective and robust spatial data discretization strategies are required to improve GD modeling.

To address the above issues, this study proposes a Robust Geographical Detector (RGD) to effectively explore more reliable and robust spatial associations between dependent and explanatory variables from a spatial heterogeneity perspective. The RGD determines discretization interval breaks using an optimization algorithm for variance-based change point detection. In this study, RGD is

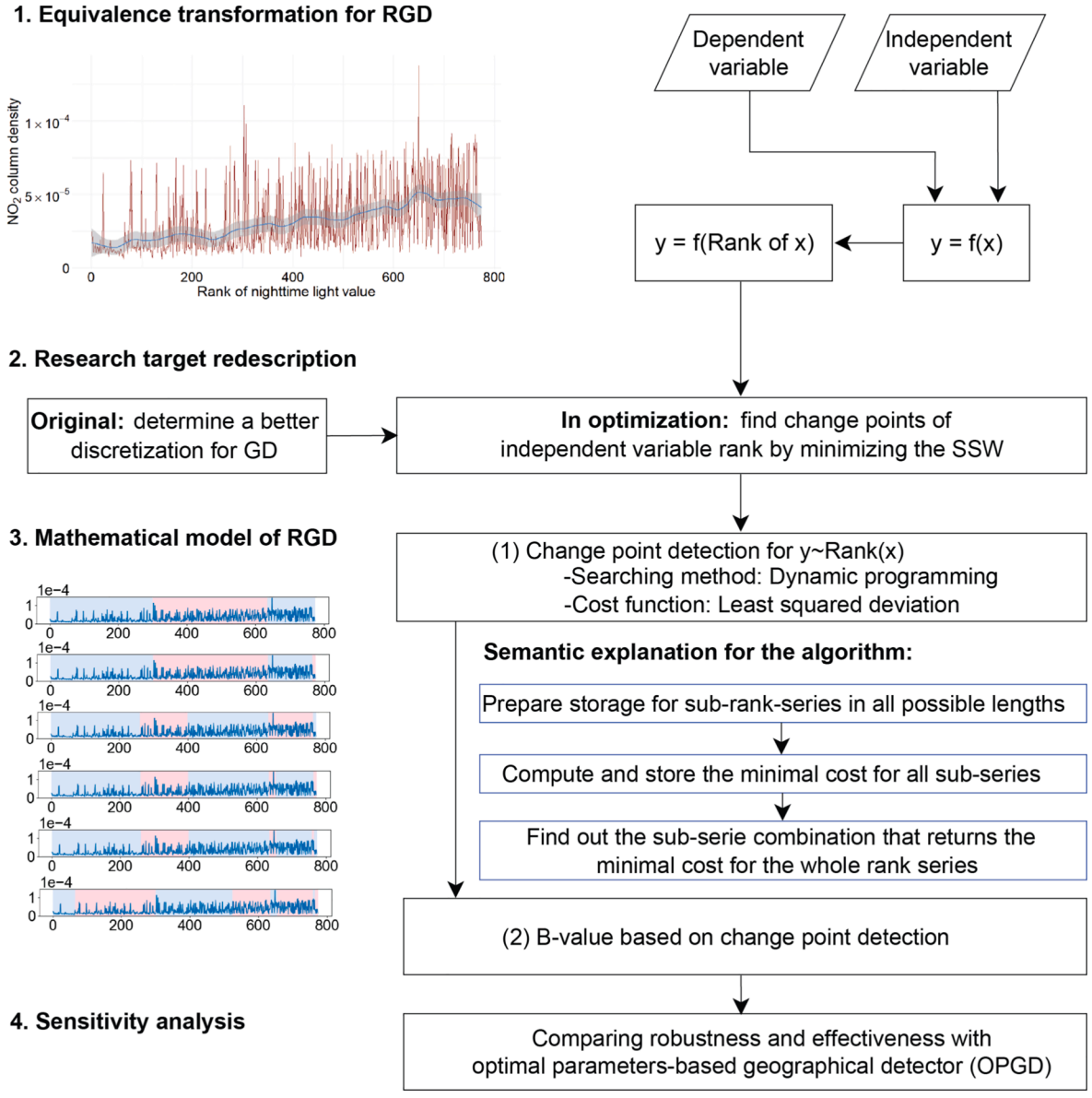


Fig. 1. Process of RGD model for determinant exploration.

implemented in a nationwide case study of exploring potential factors of nitrogen dioxide (NO_2) density in industrial regions across Australia, where data on both NO_2 and potential factors are sourced from satellite images and remote sensing products using Google Earth Engine. Sensitivity analysis is performed to evaluate the effectiveness of RGD for exploring spatial associations.

The remainder of this article is structured as follows. Section 2 describes the detailed steps of the developed RGD model. Section 3 presents the case study of applying RGD in exploring factors affecting air pollutants in the nationwide industrial regions in Australia. Section 4 shows the results of this study, including PD values of potential variables computed using RGD, model evaluation by comparing with GD models, and findings in the case study. Section 5 discusses the advantages of RGD and the contributions of this study, and Section 6 concludes this article.

2. Robust Geographical Detector (RGD)

RGD is an improvement of the geographical detector (GD) with optimal spatial zones determined using optimization of spatial data discretization of explanatory variables. Fig. 1 shows the process of the RGD model for spatial determinant exploration, and the method includes four steps. The first step is the equivalence transformation for RGD using a ranking approach, which guarantees the measure of SSH and creates opportunities for solving an optimization problem. The second step is to redescribe the objective of spatial discretization as an almost-solved optimization problem. This means identifying breaks of spatial discretization is transformed into a change point detection problem, where change points of explanatory variable ranks are specified using a dynamic programming method, and the within the sum of squares (SSW) is minimized using a least squared deviation cost function. The third step is to calculate the PD values of explanatory variables. In RGD, a B-value is used to quantify the PD of variables with the detected robust change points determined spatial zones. Finally, the sensitivity of RGD is evaluated by comparing it with previous GD models. In this study, RGD is implemented in exploring spatial determinants of air pollutants in industrial regions in Australia, which is described in Section 3.

2.1. Equivalence transformation for RGD

The RGD is a variant of GD with a robust optimization discretization strategy for a more effective estimation of spatial stratified heterogeneity. The PD of explanatory variables is computed as a B-value in RGD, as shown in equation (1).

$$B = 1 - \frac{SSW^R}{SST} = 1 - \frac{\sum_{z=1}^{k} N_z \sigma_z^2}{N\sigma^2} \tag{1}$$

where SSW^R is the Sum of Squares Within spatial zones identified using the robust optimization strategy of spatial data discretization for explanatory variables; SST is the Sum of Squares Total of observations in the whole study area; z is an RGD spatial zone; N_z and σ_z^2 are the number and variance of observations within zone z by discretizing an explanatory variable, and N and σ^2 are the number and variance of data in the whole study area. Similar to the Q-value in GD, B-value measures the spatial association between dependent and explanatory variables, and the value ranges from 0 to 1.

The basic idea for equation (1) in the previous OPGD quantile method is how much of the dependent variable's spatial heterogeneity can be explained by dividing the sorted explanatory factor value (Song et al., 2020), identical to dividing the rank of an explanatory factor value. RGD is a function to quantify to what extent spatial heterogeneity of a dependent variable can be explained by ranks of explanatory variables instead of values of explanatory variables in GD. Spatial zone z across the space in equation (1) is determined by discretizing a numerical continuous explanatory variable, which means a division of the study area to show the dependent variable's spatial heterogeneity is based on the segmented sorted and ranked explanatory variable series from the observed minimum to maximum. Suppose dependent and explanatory variables are spatially associated. There should be a consistent mathematical relationship between the sorted dependent variable value (A) and the sorted rank of this explanatory variable (B), meaning that the mapping from A to B is a bijection with simultaneous increase. Namely, explanatory variable discretization for RGD is equivalent to categorizing the sorted rank of explanatory variables. Thus, determining spatial zones using explanatory variables can be converted to deciding spatial zones using the ranks of explanatory variables, which is a robust approach without impacts of outliers and extreme values, to calculate the PD values based on spatial stratified heterogeneity.

Therefore, a rank transformation has two advantages, which could guarantee the measure of the SSH value based on equation (1) and work as an input for the optimization algorithm at the same time. Considering these advantages, RGD accepts the equivalence transformation and investigates how much of the dependent variable's spatial heterogeneity can be explained by the rank of the explanatory variable. This transformation switches the original distribution of explanatory variables into a sequence of natural numbers, starting from value one to the number of total observations. Even with no direct computing advantage, transformed relationships (i.e., the relationship between dependent variable and rank of explanatory variable) can be treated as a simulated signal recorded within a time series, which can be further transferred into an optimization problem.

2.2. Research target redescription

In previous studies, significant efforts have been made to improve GD through various spatial discretization methods. However, it is still a challenge to derive an explicit mathematical approach for reliable and robust modeling. RGD provides a robust solution for addressing this issue. In RGD, the research target can be stated more clearly after the equivalence transformation since explanatory variables are continuous sequences of natural numbers. The relationship between the dependent variable and the rank of the explanatory variable is equivalent to a simulated offline signal series, where the dependent variable acts as a signal pulse and the explanatory variable rank are the time series. Previous GD discretization strategies do not fully explore the relationship between variables when determining the segmentation point and generating spatial zones z. The RGD treats minimizing SSW as an optimal target to segment the transformed signal series. Now, determining a better discretization for GD can be rephrased by a clear optimization problem. Given a simulated signal series, is it possible to find a specified number of segmentation points that could have the least SSW for a dependent variable? The answer to this question is 'yes', and the solution is change point detection (Page, 1955; Truong et al., 2020), introduced in the following section.

2.3. Mathematical model of RGD

The mathematical model of RGD is composed of change point detection for simulated signal series generated from equivalence transformation for variables and B-value derived from change-point segmentation. Change point detection (CPD) is a method to detect time points of a signal series where significantly specified types of changes occur. CPD is composed of the cost function, searching method, and constraints. The cost function defines types of change to detect, and this function is also the optimization target. To minimize the SSW, a least squared deviation cost function is selected for RGD. The searching method is a computing strategy to find required change points, and RGD selects a suitable searching method to overcome past limitations. The previous GD discretization method generates fluctuating spatial stratified heterogeneity value in equation (1) with the increase of interval

number. From a computer science perspective, the lack of learning experience from the relationship between response and explanatory variables and previous segmentation outcomes leads to the unexpected phenomenon. Dynamic programming is based on the view that an optimal solution to the original problem comprises several optimal solutions toward overlapping sub-problems. In CPD, dynamic programming divides a specified number of change points into multiple functions of finding fewer numbers of required points. To determine the K number of break intervals, the dynamic programming-based CPD algorithm would execute (K-1) times and find a change point that meets the optimization goal once. Each K interval determination task sub-process starts from the last executed optimized results. With minimizing SSW as the primary optimization target in each searching sub-task, this bottom-up computing strategy guarantees the increment of spatial stratified heterogeneity value with the increase of specified interval number. Therefore, a dynamic programming searching method is selected to find the required segmentation points. There is no CPD constraint because the number of intervals can be specified by users based on needs. Fundamental ideas of RGD have been translated into the following pseudocodes.

Algorithm: Change point detection for RGD – dynamic programming & least squared deviation.

Input: simulated signal series noted $\{y_x\}_{x=1}^{N}$; cost function $c(y_I) = \sum_{x\in I}||y_x - \overline{y}||_2^2$; specified number of spatial zone K (no less than 2); lists to record costs within each interval given interval number from 1 to K-1, noted as $C_1, C_2, \ldots, C_{k-1}$; an empty list noted L, with a length of K (the top (K-1) elements are to store segmentation points for generating spatial zones, and the last element is the number of observations).

1 **for all** rank series pairs (p, q), $1 \le p < q \le$ N (number of observations) **do**
2 $C_1(p, q) \leftarrow c(\{y_x\}_{x=p}^{q})$
3 **end for**
4 **for all** j from 2 to K-1 **do**
5 **for all** rank series pairs (p, q), $1 \le p < q \le$ N, $q - p \ge j$ **do**
6 $C_j(p, q) \leftarrow \min_{p+j-1\le x<q}(C_{j-1}(p, x) + C_1(x + 1, q))$
7 **end for**
8 **end for**
9 $L[K] \leftarrow$ N; $j \leftarrow K$
10 **while** $j > 1$ **do**
11 $m \leftarrow L(j)$
12 $x^* \leftarrow \text{argmin}_{j-1\le x<m}(C_{j-1}(1, x) + C_1(x + 1, m))$
13 $L[j - 1] \leftarrow x^*$
14 $j \leftarrow j - 1$
15 **end while**
16 **return** list L

Note: Transformed rank observation series and simulated signal series as the algorithm input refer to the same processed data sequence.

The above algorithm tells how RGD intervals are determined using CPD with minimizing SSW as the optimization target. The algorithm from lines 1 to 8 is the preparation for dynamic programming searching, composed of two key steps. The algorithm from lines 1 to 3 is to prepare storage memory for all possible lengths of sub-series. With the least squared deviation as the cost function, Algorithm from lines 4 to 8 is to compute the cost for all sub-series. The rest of the algorithm demonstrates the dynamic programming searching process. This algorithm returns a vector of segmentation points using dynamic programming searching. K-1 change points divide the explanatory variable value range into K groups. It is worth mentioning that Algorithm line 12 is the process of finding the optimal combination of sub-series which minimizes the cost function by searching all possible lengths of sub-series. Where there is an outlier, to minimize the cost function, change point detection will detect the extreme value or outliers and categorize outliers into a new group if it can minimize the cost. Then, explanatory variables are discretized and categorized based on segment value range. The spatial stratified heterogeneity value for RGD is calculated using equation (1) , and we note it as B-value. To distinguish with terms for OPGD, we regard break intervals determined by RGD as robust geographic zones.

3. Case study: Exploring determinants of air pollutants

3.1. Case study background and study area

Industrial regions supporting mining, manufacturing, utility supply, and waste services (hereinafter referred to as key industries) are important to urban expansion and economic development (Ottaviano and Puga, 1998; Delgado et al., 2014). However, economic profits from these industry activities are at the expense of air pollutant emissions, and NO_2 is one of the common pollutants. According to Australian National Pollutant Inventory (NPI) records, key industry activities produced over 98% of NO_2 among all nationwide economic activities in 2020 (Australian Government, 2020). The NO_2 emissions are generated from recorded factories or relevant facilities located in industrial regions. Hence, it is necessary to investigate these industrial regions when monitoring air pollutant emissions. Human activity factors, including nighttime light (Kong et al., 2019), meteorological factors, including wind speed (Davis et al., 2019), and vegetation (Cui et al., 2019), can be influential in airing pollutant emissions in industrial regions. The RGD is utilized to explore spatial stratified heterogeneity in the relationship between satellite measured NO_2 density and selected explanatory variables. B-values are compared with Q-values calculated from optimal parameter geographical detector (OPGD) using a quantile categorization strategy.

The study area contains nationwide Australian industrial regions, mainly designed and built for key industry activities. These industrial regions were identified based on the point of interest (POI) and OpenStreetMap (OSM) land use polygons using kernel density estimation and GIS methods. The practical feasibility of the POIs-based region of interest spatial identification method has been proven in previous research. Selected POIs are spatial points representing locations of facilities or infrastructures supporting key industry activities. These spatial points are collected from multiple sources, including Australian National Pollutant Inventory (NPI) (Australian Government, 2020) and OSM (Geofabrik and OpenStreetMap contributors, 2020). Our POIs-based spatial identification method follows Song's (2018b) methodology framework and makes adjustments for scale parameters by referring to the Australian Statistical Geography Standards (ASGS) (Australian Bureau of Statistics, 2021). Dense POIs regions are identified using kernel density estimations (KDE) using the Epanechnikov kernel. These regions supporting key industry activities are also redefined as areas large enough to be functional areas, which are equivalent to Statistical Area Level 2 (SA2). Therefore, 1000 m, the squared root value of the bottom 99% level of SA2 size, was selected to be the searching radius of the KDE function. The pixel size of our KDE was set to be 194 m, which was equivalent to the median size level of the mesh block (the most acceptable spatial granularity level defined by the Australian Bureau of Statistics). Then, 0.5% of the cumulative distribution function (CDF) is the threshold value for industrial boundary determination. As a result, regions with a POIs density value greater than 1.95, equivalent to 97.5% CDF level, are selected as potential industrial regions. The top 2.5% dense POIs regions, covering an area of 326 square kilometers across Australia, were a part of industrial regions for our study. These dense POIs regions were further processed with the OSM land use polygons to generate results.

According to ABS definition of industry, OSM industrial land use polygons contain factories, warehouses, and workshops mainly for key industries. This spatial data is utilized as supplementary information for dense POIs regions. Raw OSM polygons in 2020 are required to be preprocessed for three reasons. First, raw polygons are coarse in size and contain tiny areas with sizes even less than 5 square meters. Second, entire industrial regions are segmented at a block level, which does not fit our redefinition that industrial regions are areas large enough to preserve functionalities. Third, some industrial polygons are overlapped, which is not consistent with reality. Therefore, a series of GIS preprocessing methods are applied. Firstly, industrial regions wa ith

sizes of less than 5000 square meters are converted to points, merged into POIs for KDE identification, and removed. This spatial magnitude is at the minimal level of a meaningful region that could hold at least one functional facility, according to the ASGS definition. This spatial size is also at the same scale as a single basic facility or infrastructure supporting daily lives in Japan and Canada (Hadjisophocleous & Chen, 2010; Yamaguchi et al., 2012). Then, filtered industrial blocks are added with a 50-meter buffer to contain surrounding roads within industrial regions. Next, buffered blocks are dissolved to form pre-processed regions.

In the last step, identified industrial regions from dense POIs and pre-processed polygons are merged. Merged regions with the size of area less than 0.46 square kilometers (the smallest recorded size of SA2 region representing functional areas according to ASGS definition) are filtered. The final identified industrial regions are areas with a size no less than the minimal level of SA2. As a result, 775 industrial regions sparsely distributed across the nation were identified. Fig. 2 (a) shows Australia's spatial distribution and areas of identified industrial regions. The following analysis is performed within the identified 775 industrial regions in Australia.

3.2. Datasets

In this case study, NO_2 and potential variables that affect NO_2 distributions in the identified industrial regions are collected from satellite images and remote sensing products using Google Earth Engine (GEE) platform (Google Developers and the European Space Agency, 2020). Table 1 shows a brief summary of the data used in this case study. The air pollutant NO_2 density is the dependent variable of this case study. Sentinel-5P products from the European Space Agency (ESA) provide satellite measurement for NO_2 column density. The NO_2 column density data in this case study is collected and processed using GEE.

In addition, a series of potential variable data is collected to explain the spatial pattern of NO_2 in industrial regions. Vegetation is represented by normalized difference vegetation index (NDVI), leaf area index (LAI), and enhanced vegetation index (EVI). High spatial resolution NDVI information is accessed from the Landsat8 collection provided by Google (Google, 2020a). LAI information is derived from GCOM-C/SGLI Level 3 spatially and temporally averaged products from Global Change Observation Mission-Climate, and provided by Google (Google Developers and Global Change Observation Mission, 2020). MODIS Combined 16-Day EVI information is accessed from GEE (Google, 2020b). Wind speed is a meteorological factor that could influence NO_2 density. Wind speed information is generated from TerraClimate datasets accessed from the GEE platform (Google Developers and University of California Merced, 2020). TerraClimate is a high spatial resolution monitoring global climates since 1958, which has been utilized in spatial research. Nighttime light (NTL) is an explanatory variable representing human activity in industrial regions. NTL data is remotely sensed by Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band and provided by Earth Observation Group and Colorado School of Mines (Google Developers and Earth Observation Group, 2020). VIIRS provides monthly updated NTL data and is accessed from the GEE platform.

Table 1
Summary of remote sensing datasets and factors.

Data	Variable	Spatial resolution	Temporal resolution	Unit
Sentinel-5P Nitrogen Dioxide	NO_2 density	1113 m	Daily	mol/m^2
GCOM-C/SGLI L3 product	LAI	4638 m	8-day	m^2/m^2
MODIS Combined EVI	EVI	463 m	16-day	–
Landsat8	NDVI	30 m	18-day	–
TerraClimate climate data	Wind speed	4638 m	Monthly	m/s
VIIRS Day/Night Band	Nighttime light	464 m	Monthly	$nanoWatts/cm^2/sr$

4. Results

4.1. B-values

In RGD, B-values of variables are used to quantify spatial associations between dependent and explanatory variables. Fig. 3 shows the B-values of variables affecting NO_2 density in industrial regions in Australia. B-values of variables are generally increased with the growth of the number of intervals for spatial data discretization and determining spatial zones. However, the increase rates of B-values gradually decrease with the growth of the number of intervals. Thus, the optimal numbers of intervals for discretization are selected when the change rates are lower than 0.05, which has been used in a series of previous studies about spatial discretization (Song et al., 2020; Song and Wu, 2021; Luo et al., 2021, 2022). Results show that the numbers of intervals of the

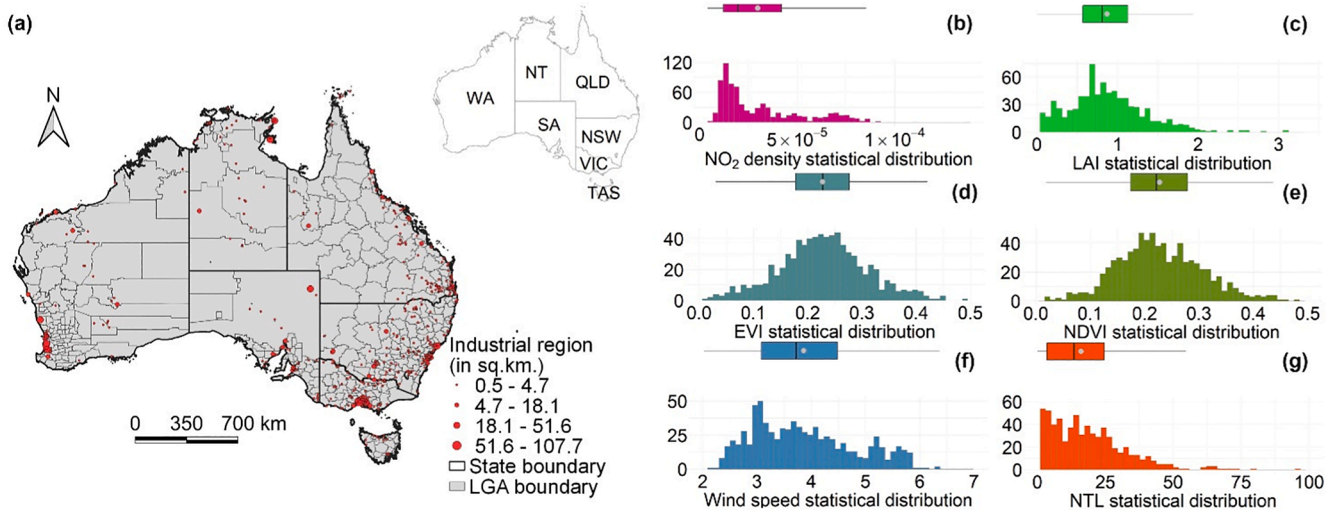


Fig. 2. Study area and data summary. Size and spatial distribution of Australian industrial regions (a), the statistical distribution of dependent variable NO_2 column density in the industrial regions (b), and statistical distributions of explanatory variables leaf area index (LAI) (c), enhanced vegetation index (EVI) (d), normalized difference vegetation index (NDVI) (e), wind speed (f), and nighttime light (NTL) (g).

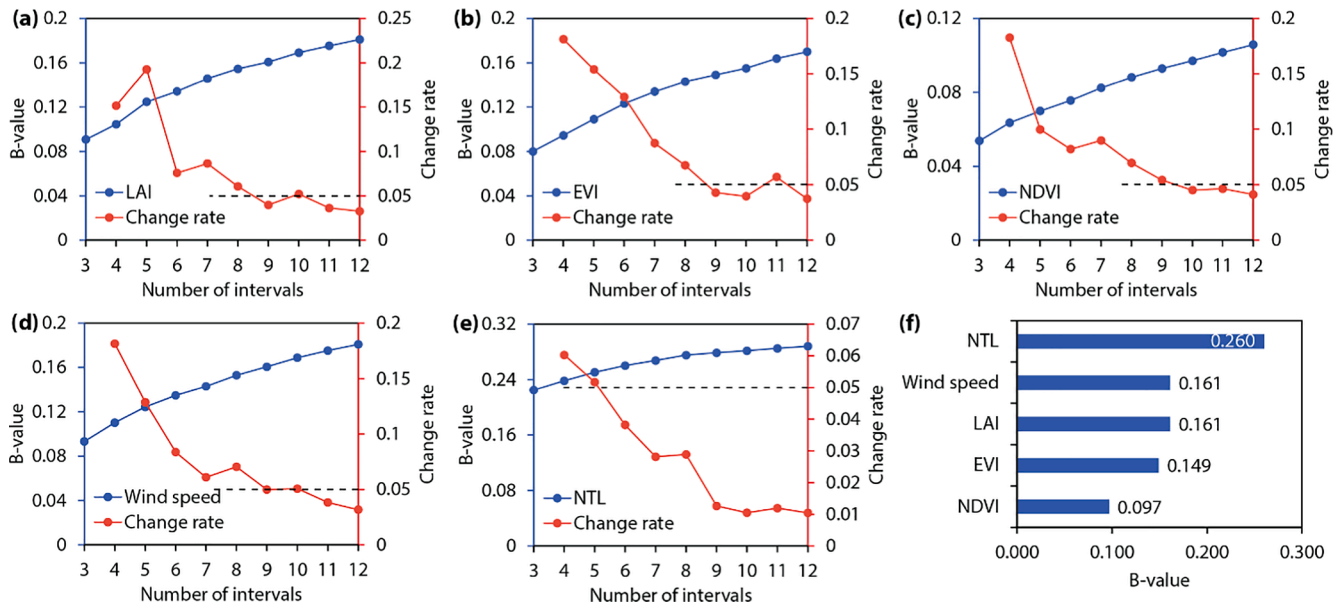


Fig. 3. Process of selecting optimal numbers of intervals for the robust spatial discretization for variables LAI (a), EVI (b), NDVI (c), wind speed (d), and NTL (e) in RGD, and B-values of variables in affecting NO_2 density in industrial regions.

optimal discretization for variables LAI, EVI, NDVI, wind speed, and NTL are 9, 9, 10, 9, and 5, respectively. Fig. 3 (f) shows the B-values of variables with the optimal discretization. NTL has the highest contribution to spatial patterns of NO_2 density in the industrial regions with a contribution of 0.260 (p less than 0.01). The contributions of wind speed, LAI, EVI, and NDVI are 0.161 (p less than 0.01), 0.161 (p less than 0.01), 0.149 (p less than 0.01), and 0.097 (p less than 0.01), respectively. This means that industrial and human activities have higher contributions to the spatial pattern of air pollutants than climate and vegetation variables in industrial regions.

4.2. Sensitivity analysis

The robustness and reliability of RGD for exploring spatial associations and potential variables are evaluated by comparing it with OPGD, an improved, effective, and commonly used GD model. The sensitivity of RGD and OPGD is assessed with different numbers of intervals ranging from 3 to 12, which are used to determine spatial zones. Fig. 4 compares PD values of five explanatory variables explored by RGD and OPGD, which are computed as B-values and Q-values, respectively. Results show that the RGD is more effective and reliable in exploring spatial

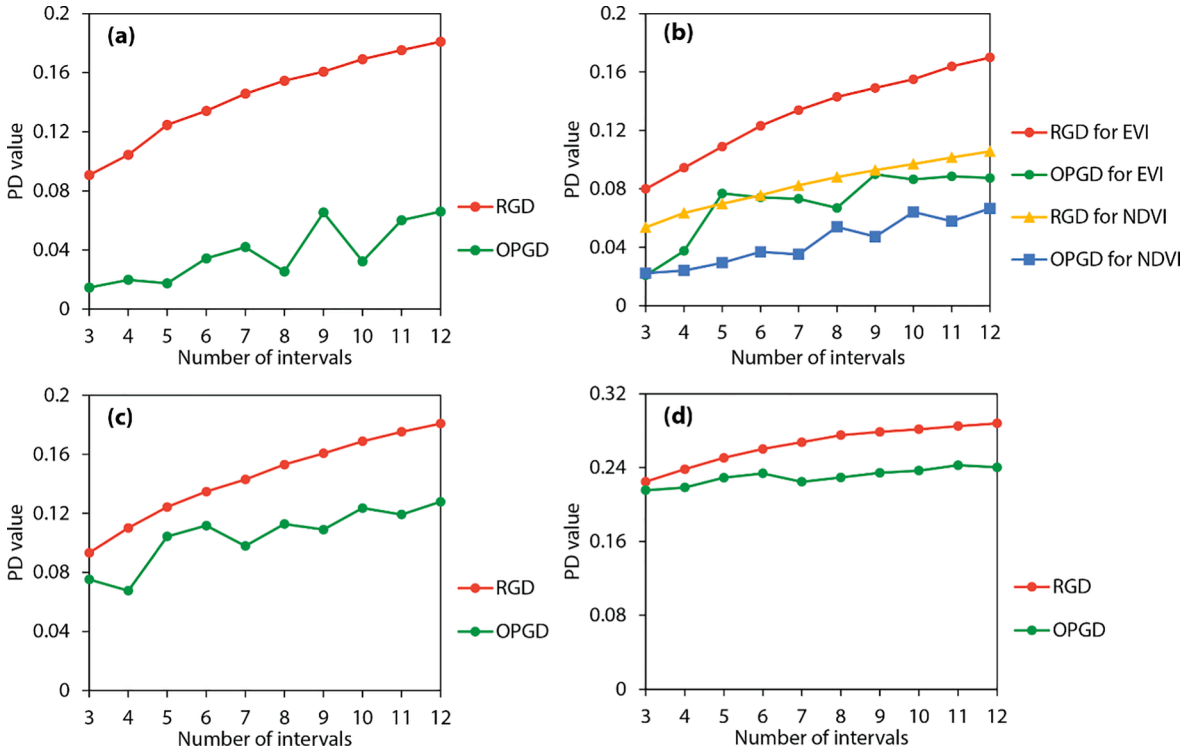


Fig. 4. Comparisons of power of determinant (PD) values, including B-values from RGD and Q-values from OPGD, of variables LAI (a), EVI and NDVI (b), wind speed (c), and NTL (d). Note: All PD values shown in this figure have a statistical significance level with p-values lower than 0.01.

determinants than OPGD models. The advantages of RGD include the following aspects.

First, the comparison between B-values and corresponding Q-values with identical numbers of intervals shows that RGD can determine better spatial zones enabling stronger spatial association between the dependent variable and explanatory variables. Second, RGD guarantees the increment of the B-value with the increase of interval break for explanatory factors, while Q-value from OPGD fluctuates. This phenomenon also confirms the robustness of RGD in assessing spatial associations. B-values based on a higher number of intervals would have more dynamic programming searching processes for finding required breakpoints. The performance of RGD is consistent with the model expectation. In detail, OPGD quantifies the PD value for wind speed from 0.07 with 3 intervals to 0.12 with 12 intervals, and RGD calculates the PD value from 0.09 to 0.18 for 10 intervals. RGD-based B-value of NTL rises from 0.22 to 0.28 with the increase of interval number, which is generally higher than corresponding Q-values from OPGD. It is worth mentioning that the PD value of LAI is at least 200% improved by RGD by comparing B-values and corresponding Q-values. This additional finding is discussed in detail in the next section. For EVI and NDVI, RGD improves no less than 100% of the PD value compared with OPGD results. In summary, NTL is the factor with the highest spatial association with NO_2 density in the Australian industrial region, followed by wind speed and LAI.

4.3. Analysis of RGD-based robust spatial zones

In addition to the robustness of RGD shown in the last section, RGD also provides robust spatial zones in terms of comparing PD ranks of variables between RGD and OPGD. A discretization method is essential for numerical variables before presenting PD values from RGD or OPGD, and how the explanatory factor is discretized would have 'significant impacts' on PD values and result interpretation. We give an example of the 'significant impact' issue in Fig. 5. When assessing OPGD-derived Q-values only, wind speed seems to have the second strongest PD with the dependent variable, while LAI's PD ranked bottom. From Fig. 3 and Fig. 4, OPGD regards LAI as a factor with relatively low association with NO_2 density in industrial regions. However, analysis results from RGD indicate that spatial association between LAI and NO_2 density is far underestimated using previous methods. According to Fig. 3 and Fig. 4, the spatial association between NO_2 density and LAI is as strong as wind speed when using robust geographic zones determined by RGD. PD values of LAI and wind speed are 0.12 with 5 intervals and 0.18 with 12 intervals, respectively. The RGD method does not make a special treatment for LAI but manages to find suitable intervals with minimized SSW for driving factors to be tested.

Further, a map demonstrating robust spatial zones for LAI determined by RGD is generated in Fig. 6. We give an illustrated case with an interval number of five. Five spatial zones are noted from 'A' to 'E', corresponding to the lowest to highest LAI intervals. According to the Australian Remoteness Structure, the 'A' group is distributed in non-urban regions. Group 'B' has the least group elements in Sydney, Melbourne, and Adelaide. 'C' category industrial regions are sparsely distributed across the nation. 'D' and 'E' groups cluster in urban and inner regional areas. As shown in Fig. 6 (f), statistical summaries indicate that urban industrial regions have NO_2 density at multiple levels, and rural industrial regions categorized in the 'A' group maintain a low NO_2 density level.

5. Discussion

5.1. Contributions

This study proposed an RGD model for identifying spatial determinants by optimizing spatial data discretization, deriving robust spatial zones, and exploring robust spatial associations. GD and its improvements, such as OPGD, are widely used approaches for measuring the PD values of explanatory variables in spatial stratified heterogeneity. The spatial data discretization strategy selection can critically affect the measurement of PD values and result interpretation. However, it is still a challenge to discretize continuous numeric variables effectively, robust, and reliable. This study demonstrates that the developed RGD model can provide a robust solution to estimate spatial associations between dependent and explanatory variables. RGD has the following advantages in exploring spatial associations. First, using the robust optimization algorithm, RGD can explore the maximum spatial associations, which are much higher than the PD values explored by OPGD models. Second, RGD guarantees the increment of PD values with the increase of interval numbers as the optimization processes also be extended with the interval increase, which is hardly ensured in previous GD models. Third, RGD is robust for explanatory variables with different statistical distributions. In most previous spatial heterogeneity models, assumptions of statistical distributions of data are required, such as normal distribution in geographically weighted regression, and modeling is affected by outliers. Compared with the OPGD method that uses sorted information only, RGD further utilizes and explores the functionality of rank information. Due to advantages provided by the rank function and change point detection algorithm, RGD can effectively overcome the impacts of outliers and extreme values in explanatory variables, and assumptions of statistical distributions of data are not required. Finally, RGD can provide robust spatial zones for more reliable and practical interpretations of results.

5.2. Future recommendations

This research demonstrates the advantages of RGD in the robust estimation of PD measurement compared with OPGD due to the optimal interval determination using change point detection. The original change point detection method allows researchers to adjust the minimal segmentation length and control size of intervals. In this article, we presented RGD results with a minimal interval length. Future GD-based spatial heterogeneity research could set the minimal segment length parameter based on the case's requirements or specific research targets. The setting of the minimal segment length parameter is related to the

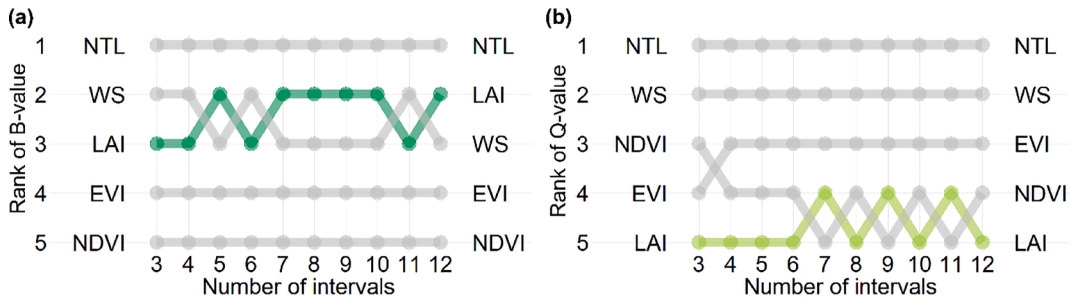


Fig. 5. Ranks of PD values, including RGD-based B-values (a) and OPGD-based Q-values (b).

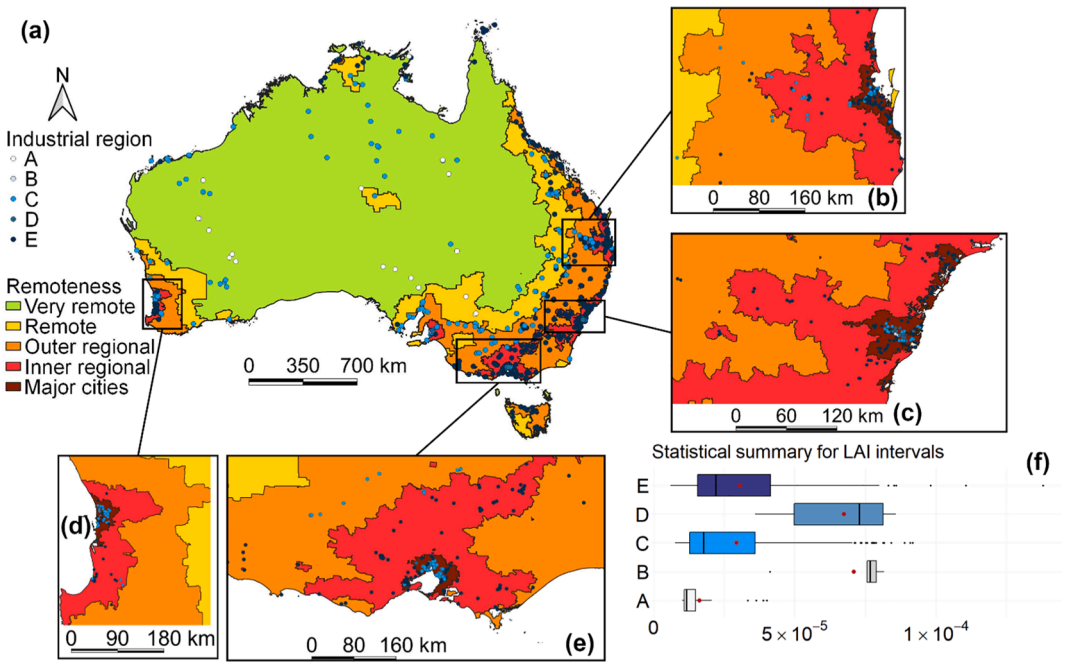


Fig. 6. Robust spatial zones for LAI determined by RGD. (a) Spatial distribution of LAI robust geographical zones with 5 intervals. (b) Brisbane. (c) Sydney. (d) Perth. (e) Melbourne. (f) Statistical summary of NO_2 column density with LAI determined spatial zones.

scale of spatial analysis (Song et al., 2020). Small minimal segment length could enlarge SSW values, and large minimal segment length could provide better spatial visualization and interpretations for large-scale spatial analysis. In addition, the associations between RGD and other GD-based models could be compared in terms of PD values and characteristics of spatial zones. For instance, it is interesting that the quantile OPGD method is a particular case of RGD. When the minimal segment length is equivalent to the number of elements in an equal-sized interval determined by a given number of breakpoints, RGD becomes a quantile OPGD.

6. Conclusion

This study proposes a Robust Geographical detector (RGD) model for exploring robust spatial associations between dependent and explanatory variables. The RGD-based analysis of the case study indicates that RGD can effectively identify the robust PD values of explanatory variables using a rank function and change point detection-based optimization approach for robust spatial data discretization. The analysis and visualization of results and sensitivity analysis for model evaluation demonstrate that RGD can explore the maximum spatial associations and guarantees the stable increase of PD values with the number of intervals. RGD is robust in dealing with variables with different statistical distributions, outliers, and extreme values and provides robust spatial zones for spatial analysis. In summary, RGD delivers a solution for an in-depth understanding of spatial stratified heterogeneity and spatial associations. RGD can be implemented in diverse fields for robust and optimal spatial zones identification, spatial determinant or factor exploration, and assessing spatial disparities.

CRediT authorship contribution statement

Zehua Zhang: Conceptualization, Methodology, Software, Data curation, Formal analysis, Visualization, Writing – original draft. **Yongze Song:** Conceptualization, Methodology, Data curation, Formal analysis, Visualization, Supervision, Writing – original draft, Writing – review & editing. **Peng Wu:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Australian Bureau of Statistics, 2021. Australian statistical geography standard (ASGS) edition 3. https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/latest-release.

Cao, F., Ge, Y., Wang, J.-F., 2013. Optimal discretization for geographical detectors-based risk assessment. GIScience Remote Sens. 50 (1), 78–92. https://doi.org/10.1080/15481603.2013.778562.

Chen, M., Chen, Y., Wang, X., Tan, H., Luo, F., 2019. Spatial difference of transit-based accessibility to hospitals by regions using spatially adjusted ANOVA. Int. J. Environ. Res. Public Health 16 (11), 1923. https://doi.org/10.3390/ijerph16111923.

Cui, Y., Jiang, L., Zhang, W., Bao, H., Geng, B., He, Q., Zhang, L., Streets, D.G., 2019. Evaluation of china's environmental pressures based on satellite NO2 observation and the extended STIRPAT model. Int. J. Environ. Res. Public Health 16 (9), 1487. https://doi.org/10.3390/ijerph16091487.

Dasgupta, S., Wheeler, D., Khaliquzzaman, M., Huq, M., 2021. Siting priorities for congestion-reducing projects in Dhaka: a spatiotemporal analysis of traffic congestion, travel times, air pollution, and exposure vulnerability. Int. J. Sustain. Transp. 1–19 https://doi.org/10.1080/15568318.2021.1969707.

Davis, Z.Y.W., Baray, S., McLinden, C.A., Khanbabakhani, A., Fujs, W., Csukat, C., Debosz, J., McLaren, R., 2019. Estimation of NO_x and SO_2 emissions from Sarnia, Ontario, using a mobile MAX-DOAS (Multi-AXis Differential Optical Absorption Spectroscopy) and a NO_x analyzer. Atmos. Chem. Phys. 19 (22), 13871–13889. https://doi.org/10.5194/acp-19-13871-2019.

Delgado, M., Porter, M.E., Stern, S., 2014. Clusters, convergence, and economic performance. Res. Policy 43 (10), 1785–1799. https://doi.org/10.1016/j.respol.2014.05.007.

Department of the Environment and Energy, Australian Government, 2020. National Pollutant Inventory. Retrieved from http://www.npi.gov.au/npidata/action/load/browse-search/criteria/browse-type/Industry/year/2020.

Dong, F., Zhang, X., Liu, Y., Pan, Y., Zhang, X., Long, R., Sun, Z., 2021. Economic policy choice of governing haze pollution: evidence from global 74 countries. Environ. Sci. Pollut. Res. Int. 28 (8), 9430–9447. https://doi.org/10.1007/s11356-020-11350-6.

Fang, Y., Wang, L., Ren, Z., Yang, Y., Mou, C., Qu, Q., 2017. Spatial heterogeneity of energy-related CO2 emission growth rates around the world and their determinants during 1990–2014. Energies 10 (3), 367. https://doi.org/10.3390/en10030367.

Feng, R., Wang, F., Wang, K., Wang, H., Li, L., 2021. Urban ecological land and natural-anthropogenic environment interactively drive surface urban heat island: An urban agglomeration-level study in China. Environ. Int. 157 (106857), 106857 https://doi.org/10.1016/j.envint.2021.106857.

Fotheringham, A.S., 2002. Geographically weighted regression: The analysis of spatially varying relationships. John Wiley & Sons.

Fotheringham, A.S., Charlton, M.E., Brunsdon, C., 1998. Geographically weighted regression: A natural evolution of the expansion method for spatial data analysis. Environ. Plan. A 30 (11), 1905–1927. https://doi.org/10.1068/a301905.

He, J., Pan, Z., Liu, D., Guo, X., 2019. Exploring the regional differences of ecosystem health and its driving factors in China. Sci. Total Environ. 673, 553–564. https://doi.org/10.1016/j.scitotenv.2019.03.465.

Hadjisophocleous, G., Chen, Z., 2010. A survey of fire loads in elementary schools and high schools. J. Fire. Prot. Eng. 20, 55–71. https://doi.org/10.1177/1042391509360266.

Jiang, X.-T., Wang, Q., Li, R., 2018. Investigating factors affecting carbon emission in China and the USA: A perspective of stratified heterogeneity. J. Cleaner Prod. 199, 85–92. https://doi.org/10.1016/j.jclepro.2018.07.160.

Kong, H., Lin, J., Zhang, R., Liu, M., Weng, H., Ni, R., Chen, L., Wang, J., Yan, Y., Zhang, Q., 2019. High-resolution ($0.05^\circ \times 0.05^\circ$) NO_x emissions in the Yangtze River Delta inferred from OMI. Atmos. Chem. Phys. 19 (20), 12835–12856. https://doi.org/10.5194/acp-19-12835-2019.

Li, H., Jia, P., Fei, T., 2021. Associations between taste preferences and chronic diseases: a population-based exploratory study in China. Public Health Nutr. 24 (8), 2021–2032. https://doi.org/10.1017/S136898002000035X.

Liu, J., Jin, X., Xu, W., Sun, R., Han, B., Yang, X., Gu, Z., Xu, C., Sui, X., Zhou, Y., 2019. Influential factors and classification of cultivated land fragmentation, and implications for future land consolidation: A case study of Jiangsu Province in eastern China. Land Use Policy 88 (104185), 104185. https://doi.org/10.1016/j.landusepol.2019.104185.

Luo, P., Song, Y., Huang, X., Ma, H., Liu, J., Yao, Y., Meng, L., 2022. Identifying determinants of spatio-temporal disparities in soil moisture of the Northern Hemisphere using a geographically optimal zones-based heterogeneity model. ISPRS J. Photogramm. Remote Sens. Off. Public. Int. Soc. Photogramm. Remote Sens. (ISPRS) 185, 111–128. https://doi.org/10.1016/j.isprsjprs.2022.01.009.

Luo, P., Song, Y., Wu, P., 2021. Spatial disparities in trade-offs: economic and environmental impacts of road infrastructure on continental level. GIScience Remote Sens. 58 (5), 756–775. https://doi.org/10.1080/15481603.2021.1947624.

Maus, V., Giljum, S., Gutschlhofer, J., da Silva, D.M., Probst, M., Gass, S.L.B., Luckeneder, S., Lieber, M., McCallum, I., 2020. A global-scale data set of mining areas. Sci. Data 7 (1), 289. https://doi.org/10.1038/s41597-020-00624-w.

Ottaviano, G.I.P., Puga, D., 1998. Agglomeration in the global economy: A survey of the 'new economic geography'. World Econ. 21 (6), 707–731. https://doi.org/10.1111/1467-9701.00160.

Page, E.S., 1955. A test for a change in a parameter occurring at an unknown point. Biometrika 42 (3–4), 523–527. https://doi.org/10.1093/biomet/42.3-4.523.

Qu, Y., Jiang, G., Yang, Y., Zheng, Q., Li, Y., Ma, W., 2018. Multi-scale analysis on spatial morphology differentiation and formation mechanism of rural residential land: A case study in Shandong Province, China. Habitat Int. 71, 135–146. https://doi.org/10.1016/j.habitatint.2017.11.011.

Raghavan, R.K., Brenner, K.M., Harrington Jr, J.A., Higgins, J.J., Harkin, K.R., 2013. Spatial scale effects in environmental risk-factor modelling for diseases. Geospatial Health 7 (2), 169–182. https://doi.org/10.4081/gh.2013.78.

Song, Y., Wright, G., Wu, P., Thatcher, D., McHugh, T., Li, Q., Li, S., Wang, X., 2018a. Segment-based spatial analysis for assessing road infrastructure performance using monitoring observations and remote sensing data. Remote Sens. 10 (11), 1696. https://doi.org/10.3390/rs10111696.

Song, Y., Long, Y., Wu, P., Wang, X., 2018b. Are all cities with similar urban form or not? Redefining cities with ubiquitous points of interest and evaluating them with indicators at city and block levels in China. Int. J. Geogr. Inform. Syst. 32 (12), 2447–2476. https://doi.org/10.1080/13658816.2018.1511793.

Song, Y., Wang, J., Ge, Y., Xu, C., 2020. An optimal parameters-based geographical detector model enhances geographic characteristics of explanatory variables for spatial heterogeneity analysis: cases with different types of spatial data. GIScience Remote Sens. 57 (5), 593–610. https://doi.org/10.1080/15481603.2020.1760434.

Song, Y., Wu, P., 2021. An interactive detector for spatial associations. Geogr. Inform. Syst. 35 (8), 1676–1701. https://doi.org/10.1080/13658816.2021.1882680.

Song, Y., Wu, P., Gilmore, D., Li, Q., 2021. A spatial heterogeneity-based segmentation model for analyzing road deterioration network data in multi-scale infrastructure systems. IEEE Trans. Intell. Transp. Syst. Public. IEEE Intell. Transp. Syst. Council 22 (11), 7073–7083. https://doi.org/10.1109/tits.2020.3001193.

Truong, C., Oudre, L., Vayatis, N., 2020. Selective review of offline change point detection methods. Signal Process. 167 (107299), 107299 https://doi.org/10.1016/j.sigpro.2019.107299.

Wang, J.-F., Li, X.-H., Christakos, G., Liao, Y.-L., Zhang, T., Gu, X., Zheng, X.-Y., 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the heshun region, China. Geogr. Inform. Syst. 24 (1), 107–127. https://doi.org/10.1080/13658810802443457.

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

Weisent, J., Rohrbach, B., Dunn, J.R., Odoi, A., 2012. Socioeconomic determinants of geographic disparities in campylobacteriosis risk: a comparison of global and local modeling approaches. Int. J. Health Geogr. 11 (1), 45. https://doi.org/10.1186/1476-072X-11-45.

Yamaguchi, Y., Suzuki, Y., Yamazaki, M., Shimoda, Y., Murakami, S., Bogaki, K., Matsunawa, K., Kametani, S., Takaguchi, H., Hanzawa, H., Yoshino, H., Asano, Y., Okumiya, M., Murakawa, S., Yoda, H., 2012. Comparison of energy consumption per unit floor area among retail categories based on the database of energy consumption for commercial buildings (decc). J. Environ. Eng. (Trans. AIJ) 77 (681), 889–897. https://doi.org/10.3130/aije.77.889.

Zuo, L., Gao, J., Du, F., 2021. The pairwise interaction of environmental factors for ecosystem services relationships in karst ecological priority protection and key restoration areas. Ecol. Ind. 131 (108125), 108125 https://doi.org/10.1016/j.ecolind.2021.108125.

Datasets

Geofabrik and OpenStreetMap contributors. (2020). Download OpenStreetMap for this region: Australia and Oceania [Data set]. Retrieved from http://download.geofabrik.de/australia-oceania.html.

Google. (2020). Landsat 8 Collection 1 Tier 1 8-Day NDVI Composite [Data set]. Retrieved from https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C01_T1_8DAY_NDVI.

Google. (2020). MODIS Combined 16-Day EVI [Data set]. Retrieved from https://developers.google.com/earth-engine/datasets/catalog/MODIS_MCD43A4_006_EVI.

Google Developers and Earth Observation Group, . VIIRS Stray Light Corrected Nighttime Day/Night Band Composites Version 1. Retrieved from. https://developers.google.com/earth-engine/datasets/catalog/NOAA_VIIRS_DNB_MONTHLY_V1_VCMSLCFG. .

Google Developers and Global Change Observation Mission, . GCOM-C/SGLI L3 Leaf Area Index (V2). Retrieved from. https://developers.google.com/earth-engine/datasets/catalog/JAXA_GCOM-C_L3_LAND_LAI_V2. .

Google Developers and the European Space Agency, . Sentinel-5P. Retrieved from. https://developers.google.com/earth-engine/datasets/catalog/sentinel-5p. .

Google Developers and University of California Merced, . TerraClimate: Monthly Climate and Climatic Water Balance for Global Terrestrial Surfaces. Retrieved from. https://developers.google.com/earth-engine/datasets/catalog/IDAHO_EPSCOR_TERRACLIMATE. .