Introduction to Spatial Stratified Heterogeneity Models in R

Yongze Song

April 2021

To cite this document in publications, please use:

Song, Y and Wu, P (2021). "An interactive detector for spatial associations". International Journal of Geographical Information Science. doi:10.1080/13658816.2021.1882680.

1. Introduction

Models of spatial stratified heterogeneity can be used to identify associations of geographical attributes through the comparison between regional variances and the global variance of data (Wang et al. 2010, Wang et al. 2016). The geographical detector (GD) is a widely used model to estimate power of determinants (PD) regarding the theory of spatial stratified heterogeneity (Wang et al. 2010).

In the past a few years, GD model has been improved from multiple methodological perspective. For instance, to accurately estimate PD of geographical variables, an optimal parameters-based geographical detector (OPGD) model was developed with the optimization of methods and numbers of spatial discretization, and spatial scale effects (Cao et al 2013, Song et al 2020a). In addition, a spatial association detector (SPADE) model to more accurately explain spatial autocorrelation of variables in GD model (Cang et al 2018). To reduce the finely divided zones by the overlap of multiple variables and improve the power of interactive determinant (PID) between a response variable and explanatory variables, an interactive detector for spatial associations (IDSA) model was developed with the spatial autocorrelation of each explanatory variable and the optimization of spatial units based on spatial fuzzy overlay (Song et al 2021).

2. R Packages

We have developed "GD" and "IDSA" R packages to implement above models in practice. In this document, primary steps of using "GD" and "IDSA" R packages are introduced. More details about using "GD" package can be found in paper Song et al (2020a) and the Vignettes of "GD" R package (https://cran.r-

Table 1: A brief summary of spatial stratified heterogeneity models and R packages in this document.

Model	R package and primary functions
Geographical detector (GD) (Wang et al 2010; 2016)	"GD" package (Song et al 2020a): functions 'gd', 'riskmean', 'gdrisk', 'gdinteract', and 'gdeco'.
Optimal parameters-based geographical detector (OPGD) (Song et al 2020a)	"GD" package (Song et al 2020a): functions 'gdm', 'optidisc', and 'sesu'.
Spatial association detector (SPADE) (Cang et al 2018)	"IDSA" package (Song et al 2021): function 'spade'.
Interactive detector for spatial associations (IDSA) (Song et al 2021)	"IDSA" package (Song et al 2021): function 'idsa'.

project.org/web/packages/GD/vignettes/GD.html). More details about using "IDSA" package can be found in paper Song et al (2021).

```
## install and library pacakges
install.packages("GD")
install.packages("IDSA")
library("GD")
library("IDSA")
```

3. Dataset

In this document, we use the simulation data in paper Song et al (2021) to show applications of "GD" and "IDSA" R packages.

```
library("IDSA")
data(sim)
head(sim)
## visualize simulation data
library(ggplot2)
library(RColorBrewer)
library(wesanderson)
# plot y
ggplot(sim, aes(x = lo, y = la, fill = y)) +
  geom_tile() +
  scale_fill_gradientn(colours = wes_palette("Zissou1", 100,
                       type = "continuous")) +
  scale_x_discrete(expand = c(0, 0)) +
  scale_y_discrete(expand = c(0, 0)) +
  coord_equal()
# plot xa
ggplot(sim, aes(x = lo, y = la, fill = xa)) +
  geom_tile() +
  scale_fill_gradientn(colours = brewer.pal(n = 8, name = "YlGn")) +
  scale_x_discrete(expand = c(0, 0)) +
  scale_y_discrete(expand = c(0, 0)) +
  coord_equal()
```

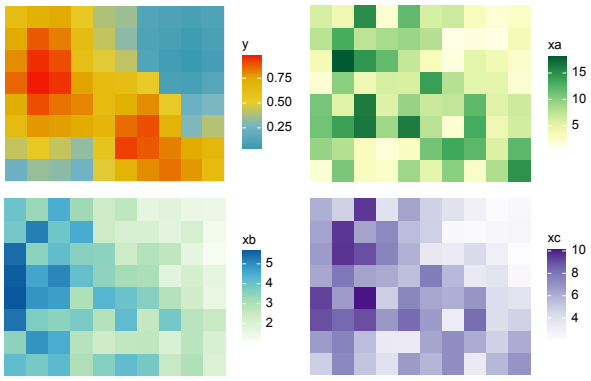

Figure 1: Spatial distributions of response and independent variables in simulation data, which has been partially shown in Song et al (2021)

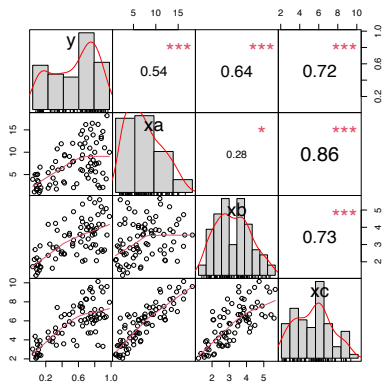

Figure 2: Pearson correlations of variables visualized with "PerformanceAnalytics" R package

4. Applications

4.1 Geographical detector (GD) model

The GD model includes four parts: factor detector, risk detector, interaction detector, and ecological detector. The calculation of GD model using "GD" package is shown below.

```
## spatial discretization for variables
library(IDSA)
sim.disc <- do.call(cbind, lapply(1:3, function(x){
  data.frame(discretize(sim[,x+3], 3, method = "quantile"))
}))
names(sim.disc) <- paste("h", 1:3, sep = "")
sim <- cbind(sim, sim.disc)

## GD model
library(GD)
## factor detector
g1 <- gd(y ~ h1 + h2 + h3, data = sim)
g1
plot(g1)
## risk detector: risk mean and detector
rm1 <- riskmean(y ~ h1 + h2 + h3, data = sim)
rm1
plot(rm1)
gr1 <- gdrisk(y ~ h1 + h2 + h3, data = sim)
gr1
plot(gr1)
## interaction detector
gi1 <- gdinteract(y ~ h1 + h2 + h3, data = sim)
gi1
plot(gi1)
## ecological detector
gd1 <- gdeco(y ~ h1 + h2 + h3, data = sim)
gd1
plot(gd1)
```

More cases of using "GD" R package for GD modeling can be found at Song et al (2020a), and other typical cases include identifying geographical determinants (Song et al 2018) and segmenting data by maximizing

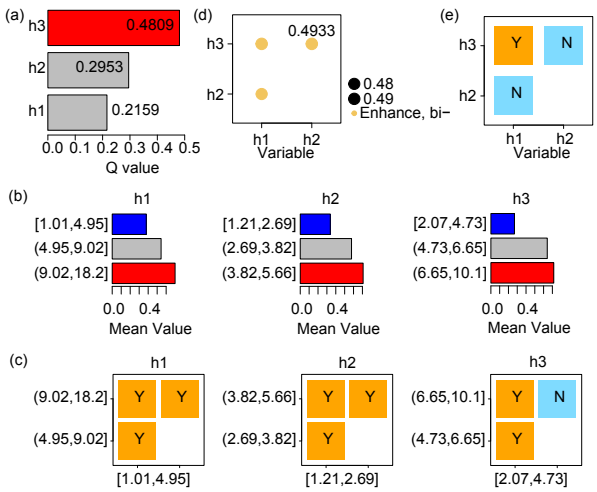

Figure 3: Geographical detector modelling for simulation data: factor detector (a), risk detector (b and c), interaction detector (d), and ecological detector (e)

spatial heterogeneity (Song et al 2020b).

4.2 Optimal parameters-based geographical detector (OPGD) model

In the OPGD model, parameters to be optimized include the number and method of spatial data discretization and scale effects. The "GD" package provides a one-step function for performing optimal discretization and geographical detectors at the same time.

```
## optional methods: equal, natural, quantile, geometric, sd and manual
discmethod <- c("equal","quantile")
discitv <- c(3:6)
## "gdm" function
gdm1 <- gdm(y ~ xa + xb + xc,
            continuous_variable = c("xa", "xb", "xc"),
            data = sim,
            discmethod = discmethod, discitv = discitv)
gdm1
plot(gdm1)
```

4.2.1 Optimization of spatial data discretization

In "GD" package, `optidisc` function is use to determine the optimal combination of the number and method for spatial discretization. In the paper Song et al (2021), we have recommended two optional strategies for the optimization of spatial data discretization in terms of the number of observations.

```
## set optional discretization methods and numbers of intervals
discmethod <- c("equal","quantile")
discitv <- c(3:6)
## optimal discretization
odc1 <- optidisc(y ~ xa + xb + xc, data = sim,
                 discmethod, discitv)
odc1
plot(odc1)
```

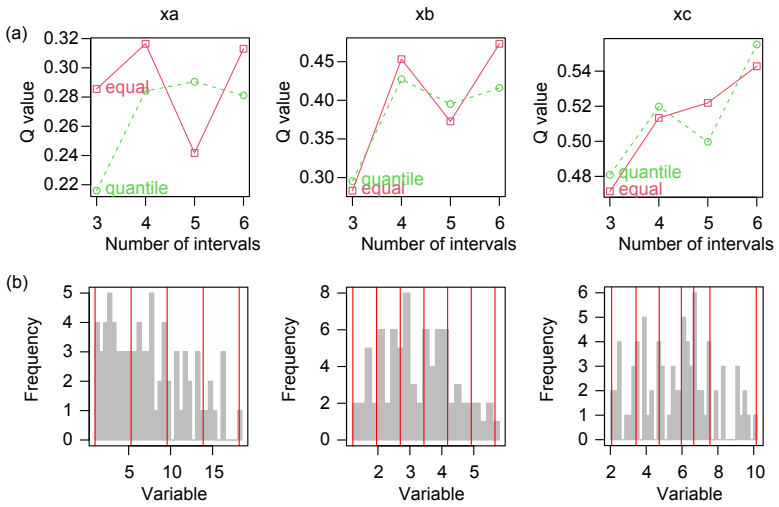

Figure 4: Processes of the optimization of spatial data discretization (a) and results (b)

4.2.2 Optimization of spatial scale effects

In "GD" package, `sesu` function is used to examine scale effects. Cases of scale effects can be found in paper Song et al (2020a) and the Vignettes of "GD" R package (https://cran.r-project.org/web/packages/GD/vignettes/GD.html).

4.3 Spatial association detector (SPADE) model

In SPADE model, spatial autocorrelation (dependence) of variables is integrated in GD model. In "IDSA" package, `spade` function is used to calculate PD, or the power of spatial and multilevel discretization determinant (PSMD), of SPADE model. The codes below are used to calculate SPADE-based PD in Figure 2 in paper Song et al (2021).

```
library("IDSA")
data(sim)
## SPADE-based PD
q.spade <- spade(formula = y ~ xa + xb, location = c("lo", "la"),
                 data = sim, ndisc = c(4, 4), methoddisc = "quantile")
q.spade
```

4.4 Interactive detector for spatial associations (IDSA) model

In "IDSA" package, `idsa` function is used to calculate PID of IDSA model.

```
idsa.ab <- idsa(y ~ xa + xb, location = c("la", "lo"), data = sim, c(4, 4),
                methoddisc = "quantile", methodoverlay = "fuzzyAND")
idsa.ab$qs.interaction
```

Following codes are used to perform the significant difference test for the zone pairs determined by the interaction of variables in the IDSA model.

```
r3 <- gdrisk(y ~ overlay, data = idsa.ab$data)
plot(r3) # Figure 4(c) in Song et al (2021)
```

The effectiveness of IDSA can be compared with that of GD-ID (interaction detector) and SPADE-ID models.

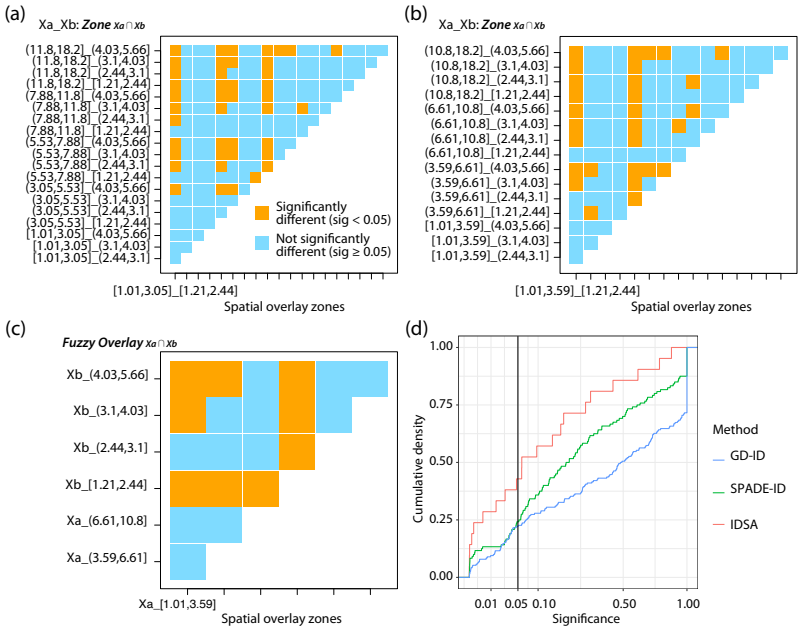

Figure 5: Difference test for the zone pairs determined by the interaction of variables in the GD-ID (a), SPADE-ID (b), and IDSA (c) models, and corresponding cumulative significance distributions (d). (Figure 4 in Song et al (2021))

```
# interaction variable derived from GD-ID
sim$xa2 <- discretize(sim$xa, 5, method = "quantile")
sim$xb2 <- discretize(sim$xb, 4, method = "quantile")
sim$xid.gd <- do.call(paste, c(sim[, c("xa2", "xb2")], sep = "_"))
# significant difference test
level1 <- expand.grid(levels(sim$xb2), levels(sim$xa2))
level1 <- do.call(paste, c(level1[,c(2:1)], sep = "_"))
sim$xid.gd <- factor(sim$xid.gd, levels = level1)
r1 <- gdrisk(y ~ xid.gd, data = sim)
plot(r1) # Figure 4(a) in Song et al (2021)

# interaction variable derived from SPADE-ID
xid.spade <- idsa(y ~ xa + xb, location = c("la", "lo"), data = sim, c(4, 4),
                  methoddisc = "quantile", methodoverlay = "intersection")
xid.spade$qs.interaction
# significant difference test
r2 <- gdrisk(y ~ overlay, data = xid.spade$data)
plot(r2)  # Figure 4(b) in Song et al (2021)

# plot cumulative significance distributions: Figure 4(d) in Song et al (2021)
sigratio <- rbind(data.frame(method = rep("idsa", nrow(r3$overlay)), sig = r3$overlay$sig),
                  data.frame(method = rep("spadeid", nrow(r2$overlay)), sig = r2$overlay$sig),
                  data.frame(method = rep("gdid", nrow(r1$xid.gd)), sig = r1$xid.gd$sig))
sigratio$method <- factor(sigratio$method, levels = c("idsa", "spadeid", "gdid"))
library(ggplot2)
ggplot(sigratio, aes(sig, color = method)) + stat_ecdf(geom = "step") +
  scale_x_sqrt(breaks = c(0, 0.01, 0.05, 0.1, 0.5, 1)) +
  geom_vline(xintercept = 0.05) +
```

```
  theme_bw()
```

It should be noted that model assumptions should be tested and satisfied when using SPADE and IDSA models, while no assumptions are required in GD and OPGD models. Following data preprocessing and tests are required in IDSA models. First, spatial autocorrelation should be tested using Moran's I or other indicators. Variables without significant spatial autocorrelation may cause biased estimations. In addition, outliers should be identified and removed, and normal distribution of data are also required, similar with other spatial models based on spatial autocorrelation. Finally, variable selection is also needed before modeling. The commonly used methods for variable selection includes correlation analysis, multicollinearity analysis, and step-wise linear regression.

Reference

Cang X and Luo W (2018). Spatial association detector (SPADE). International Journal of Geographical Information Science, 32 (10), 2055–2075. doi: 10.1080/13658816.2018.1476693

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

Song Y and Wu P (2021). "An interactive detector for spatial associations". International Journal of Geographical Information Science. doi: 10.1080/13658816.2021.1882680

Song Y, Wang J, Ge Y and Xu C (2020a). "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 Sensing, 57(5), pp. 593-610. doi: 10.1080/15481603.2020.1760434.

Song Y, Wright G, Wu P, Thatcher D, McHugh T, Li Q, Li SJ and Wang X (2018). "Segment-Based Spatial Analysis for Assessing Road Infrastructure Performance Using Monitoring Observations and Remote Sensing Data". Remote Sensing, 10(11), pp. 1696. doi: 10.3390/rs10111696.

Song Y, Wu P, Gilmore D and Li Q (2020b). "A Spatial Heterogeneity-Based Segmentation Model for Analyzing Road Deterioration Network Data in Multi-Scale Infrastructure Systems." IEEE Transactions on Intelligent Transportation Systems. doi: 10.1109/TITS.2020.3001193.

Wang J, Li X, Christakos G, Liao Y, Zhang T, Gu X and Zheng X (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), pp. 107-127. doi: 10.1080/13658810802443457.

Wang J, Zhang T and Fu B (2016). "A measure of spatial stratified heterogeneity." Ecological Indicators, 67, pp. 250-256. doi: 10.1016/j.ecolind.2016.02.052.