Introduction to ESDA and Spatial Econometric techniques using STATA

Raul Ramos (AQR-IREA, UB)

http://www.raulramos.cat/stata18

June 2018

AQR

Grup d'Anàlisi Quantitativa Regional
Universitat de Barcelona

Universitat de Barcelona

Session 1

Exploratory spatial data analysis (ESDA)

- Thematic mapping
- Multivariate maps
- Spatial point patterns
- Detecting spatial autocorrelation
  - Spatial proximity and the spatial weight matrix
  - Global and local indices of spatial autocorrelation

- **Geographically weighted regression**

A brief overview of software options

From the origins to mainstream:

- SpaceStat – Anselin (1992) + GEODA (2006) + Pysal
- Toolbox for Matlab – Lesage and Pace (2009) – Elhorst
- R spdep and others – Bivand and coauthors (2008)
- STATA Pisati (2001) + user routines
  Drukker et al (2013) sppack -> STATA 15!!!!
- Not GIS software! Try Opensource QGIS (www.qgis.org)

Types of spatial data

- Geostatistical data: continuous spatial data (for example, pollution)

- Lattice/Regional data: discrete spatial data (usually fixed polygons such as counties, provinces or countries)
  Administrative unit – Modifiable Area Units Problem (MAUP)

- Point patterns: location as a random event (crime, accident)

ESDA: thematic mapping

Thematic maps represent the spatial distribution of a phenomenon of interest within a given study area.

Spatial data usually distributed as ESRI shape files. The format uses three files: .shp and .shx files contain the map information while .dbf contains observations on each spatial unit. We need to obtain and translate these files into Stata: *shp2dta* (K. Crow)

Alternative format: Mapinfo interchange format: *mif2dta*

Where can we find shapefiles?

- Stata's user routines: China, US, *worldstat*, …
- http://www.gadm.org/country
- http://ec.europa.eu/eurostat/web/gisco/geodata
- http://www.ine.es/ss/Satellite?L=0&c=Page&cid=1254735116596&p=1254735116596&pagename=ProductosYServicios%2FPYSLayout

In case you need to adapt it, you can use *mergepoly*

ssc install shp2dta

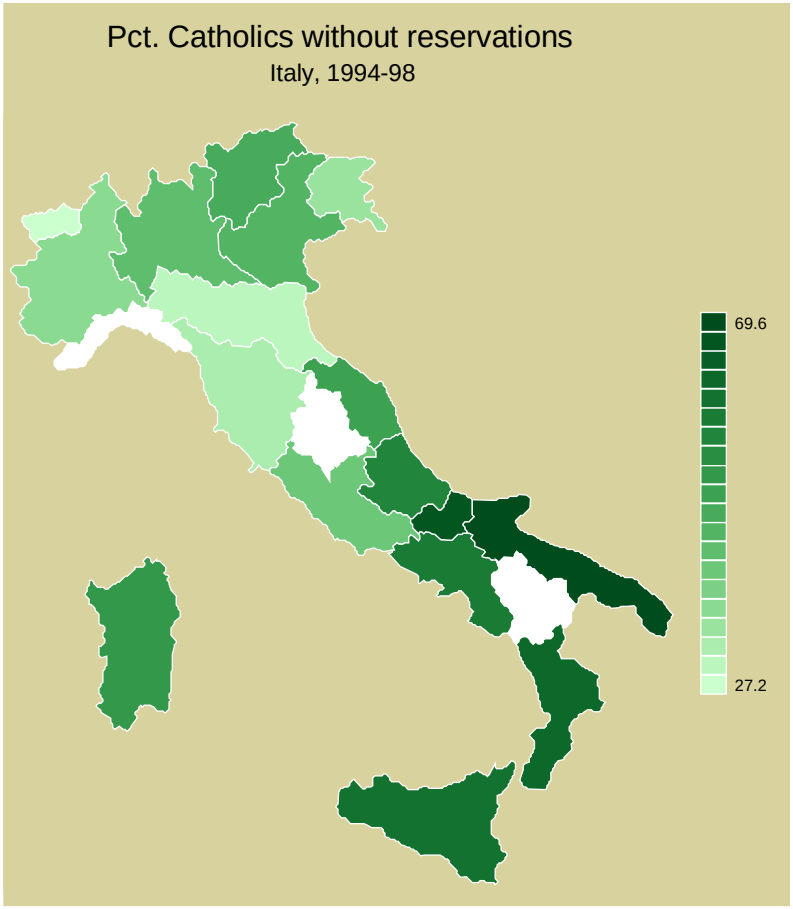
ssc install spmap

- Points (dot map)
- Polygons

Choropleth map

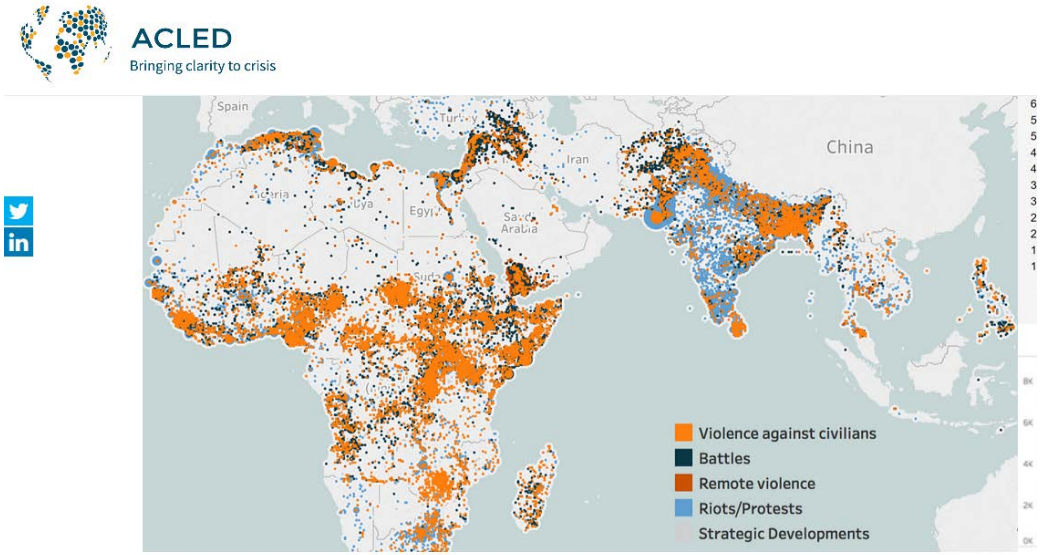
(intensity depends on values)

- Quantiles
- Equal intervals
- Boxplot



ssc install spgrid

ssc install spkde



https://www.acleddata.com/

ESDA: Spatial autocorrelation

Spatial autocorrelation:

“two or more objects that are spatially close tend to be more similar (or more different) to each other with respect a particular characteristic than other that are spatially distant. (spatial clustering)”

How to measure spatial proximity? Spatial weights matrix

Columbus (https://geodacenter.github.io/data-and-lab/columbus)

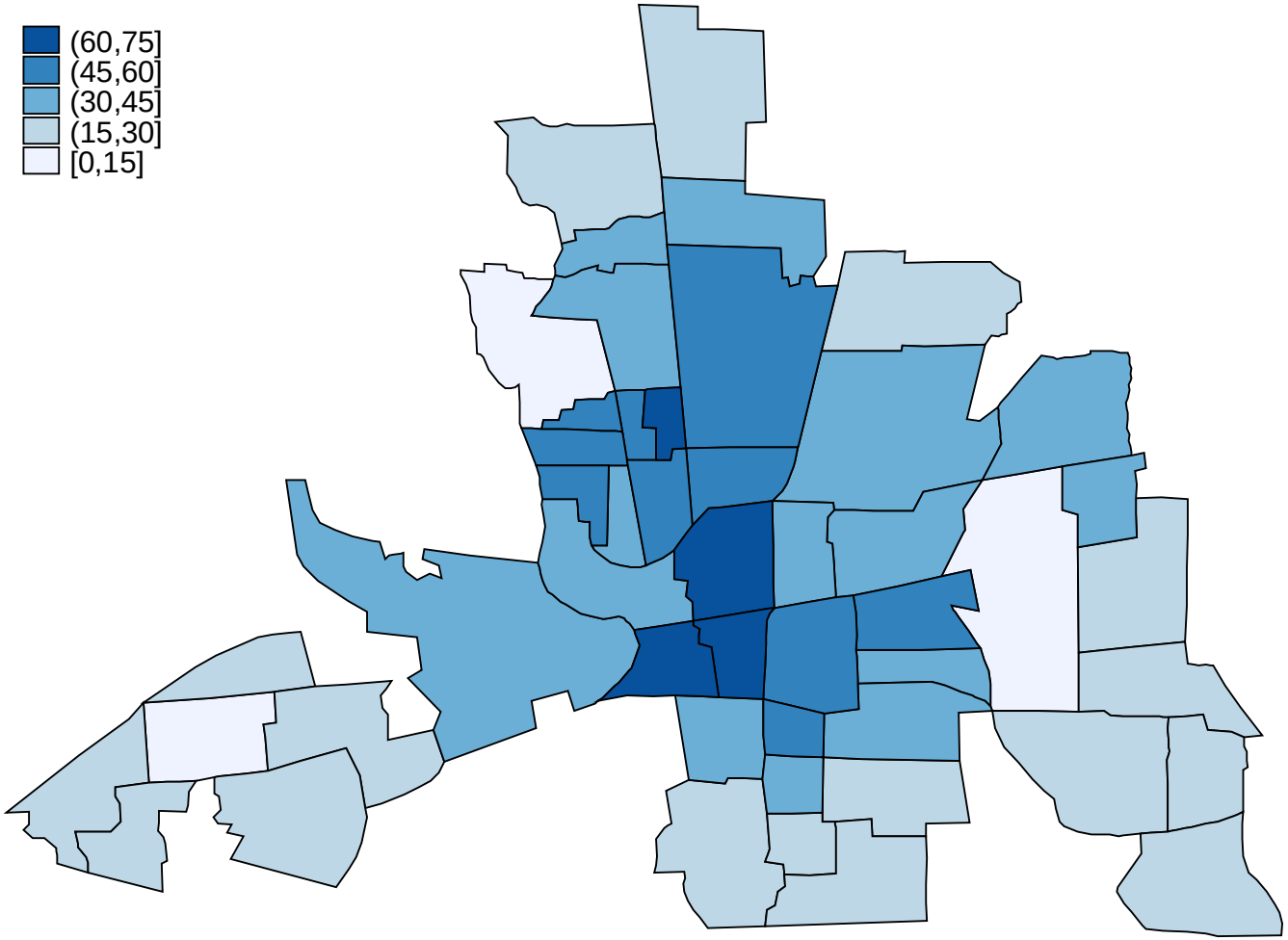
Description: Crime data for 49 neighborhoods in Columbus, OH, 1980

Type = polygon shape file, projected, arbitrary units
Observations = 49
Variables = 20

Source: Anselin, Luc (1988). Spatial Econometrics. Boston, Kluwer Academic, Table 12.1, p. 189.

AREA neighborhood area (computed by ArcView)
PERIMETER neighborhood perimeter (computed by ArcView)
COLUMBUS_ internal polygon ID (generated by ArcView)
COLUMBUS_I internal polygon ID (geneated by ArcView)
POLYID neighborhood ID, used in GeoDa User's Guide and tutorials
NEIG neighborhood ID, used in Spatial Econometrics examples
HOVAL housing value (in $1,000)
INC household income (in $1,000)
CRIME residential burglaries and vehicle thefts per 1000 households
OPEN open space (area)
PLUMB percent housing units without plumbing
DISCBD distance to CBD
X centroid x coordinate (in arbitrary digitizing units)
Y centroid y coordinate (in arbitrary digitizing units)
NSA north-south indicator variable (North = 1)
NSB other north-south indicator variable (North = 1)
EW east-west indicator variable (East = 1)
CP core-periphery indicator variable (Core = 1)
THOUS constant (= 1000)
NEIGNO another neighborhood ID variable (NEIG + 1000)

Property crimes per thousand households



Columbus, Ohio 1980 neighorhood data
Source: Anselin (1988)

Spatial weight matrix (W):

- NxN matrix
- W_{ij}
- $W_{ii}=0$

Different specifications:

- Contiguity matrix: neighbours (1st order and 2nd order)
- Inverse-distance matrix (usually normalized – row)
- Flows (asymmetric) …

net install sppack.pkg

spmat

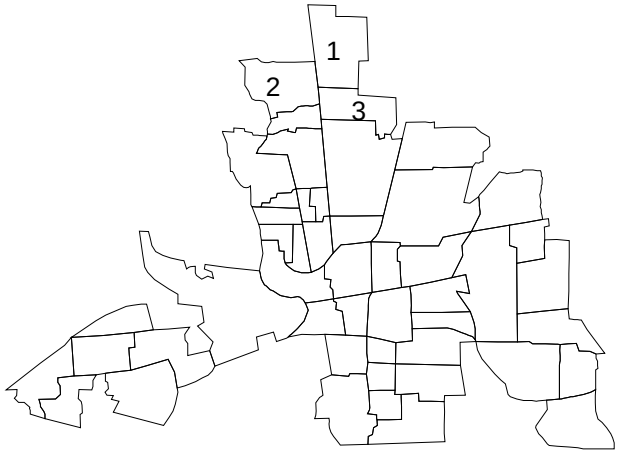
net install sg162.pkg

spatwmat

ssc install spwmatrix

spwmatrix

Normalised contiguity matrix



W[49,49]

	c1	c2	c3	c4	c5	c6	c7
SWMImpo	0	.5	.5	0	0	0	0
SWMImpo	.33333333	0	.33333333	.33333333	0	0	0
SWMImpo	.25	.25	0	.25	.25	0	0
SWMImpo	0	.25	.25	0	.25	0	0
SWMImpo	0	0	.14285714	.14285714	0	.14285714	0
SWMImpo	0	0	0	0	.5	0	0
SWMImpo	0	0	0	0	0	0	0
SWMImpo	0	0	0	.2	.2	0	.2
SWMImpo	0	0	0	0	.16666667	.16666667	0
SWMImpo	0	0	0	0	0	0	0
SWMImpo	0	0	0	0	.25	0	0
SWMImpo	0	0	0	0	0	0	0
SWMImpo	0	0	0	0	0	0	.33333333
SWMImpo	0	0	0	0	0	0	.16666667
SWMImpo	0	0	0	0	.25	0	0
SWMImpo	0	0	0	0	0	0	0
SWMImpo	0	0	0	0	0	0	0
SWMImpo	0	0	0	0	0	0	0

Global indices of spatial autocorrelation

- Moran's I *spatgsa CRIME, w(w1s) moran*

Measures of global spatial autocorrelation

Weights matrix

Name: w1s
Type: Imported (binary)
Row-standardized: Yes

Moran's I

Variables	I	E(I)	sd(I)	z	p-value*
CRIME	0.500	-0.021	0.093	5.589	0.000

*1-tail test

Global indices of spatial autocorrelation

- Getis and Ord *spatgsa CRIME, w(w1) go*

Measures of global spatial autocorrelation

Weights matrix

Name: w1
Type: Imported (binary)
Row-standardized: No

Getis & Ord's G

Variables	G	E(G)	sd(G)	z	p-value*
CRIME	0.128	0.100	0.006	4.638	0.000

*1-tail test

Local indices of spatial autocorrelation

- *spatlsa CRIME, w(w1s) moran id(id) sort*

Measures of local spatial autocorrelation

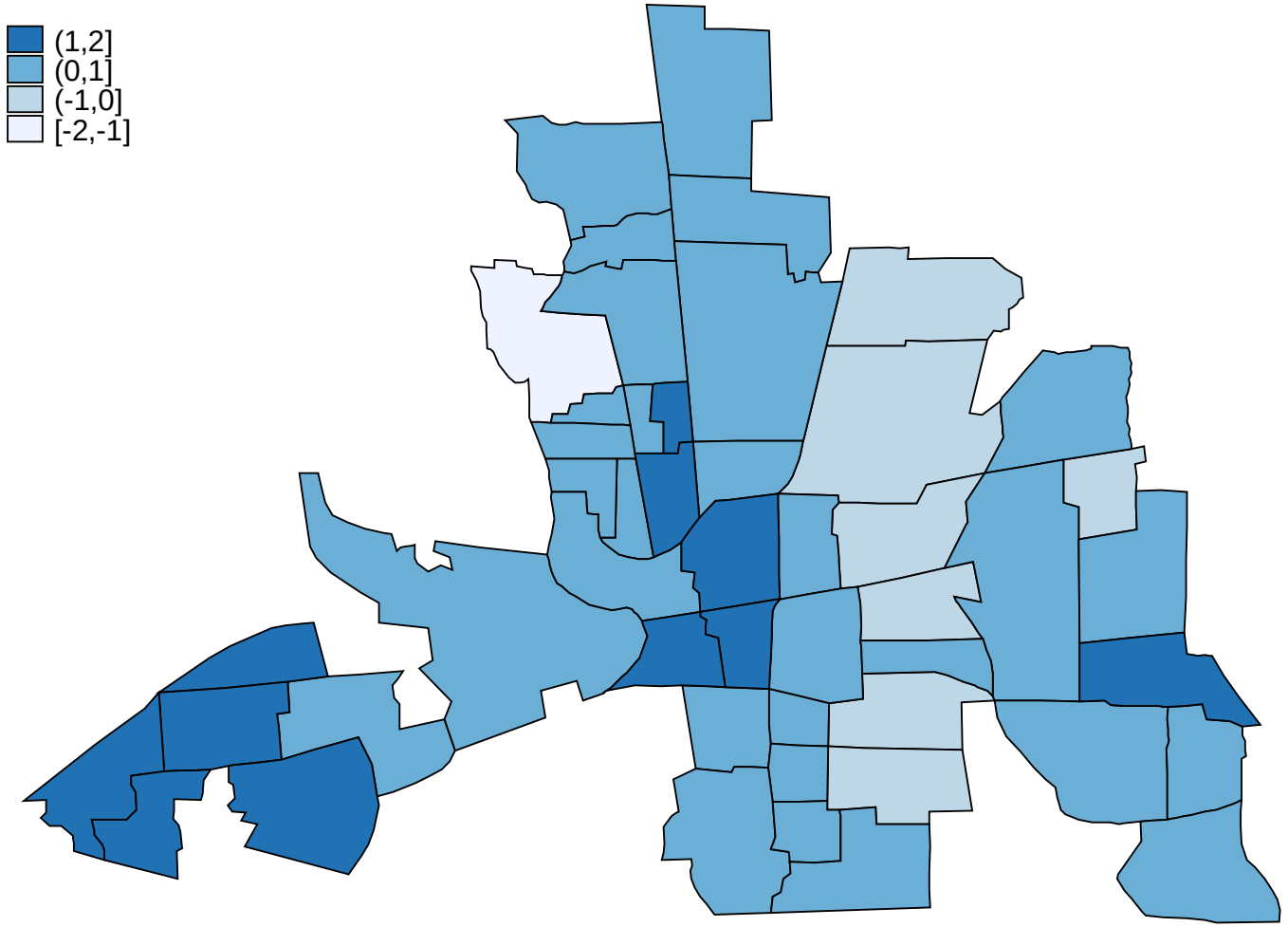
Weights matrix

Name: w1s
Type: Imported (binary)
Row-standardized: Yes

Moran's Ii (CRIME)

id	Ii	E(Ii)	sd(Ii)	z	p-value*
7	-1.861	-0.021	0.478	-3.850	0.000
6	-0.182	-0.021	0.691	-0.233	0.408
27	-0.124	-0.021	0.478	-0.216	0.415
17	-0.108	-0.021	0.558	-0.156	0.438
9	-0.031	-0.021	0.323	-0.030	0.488

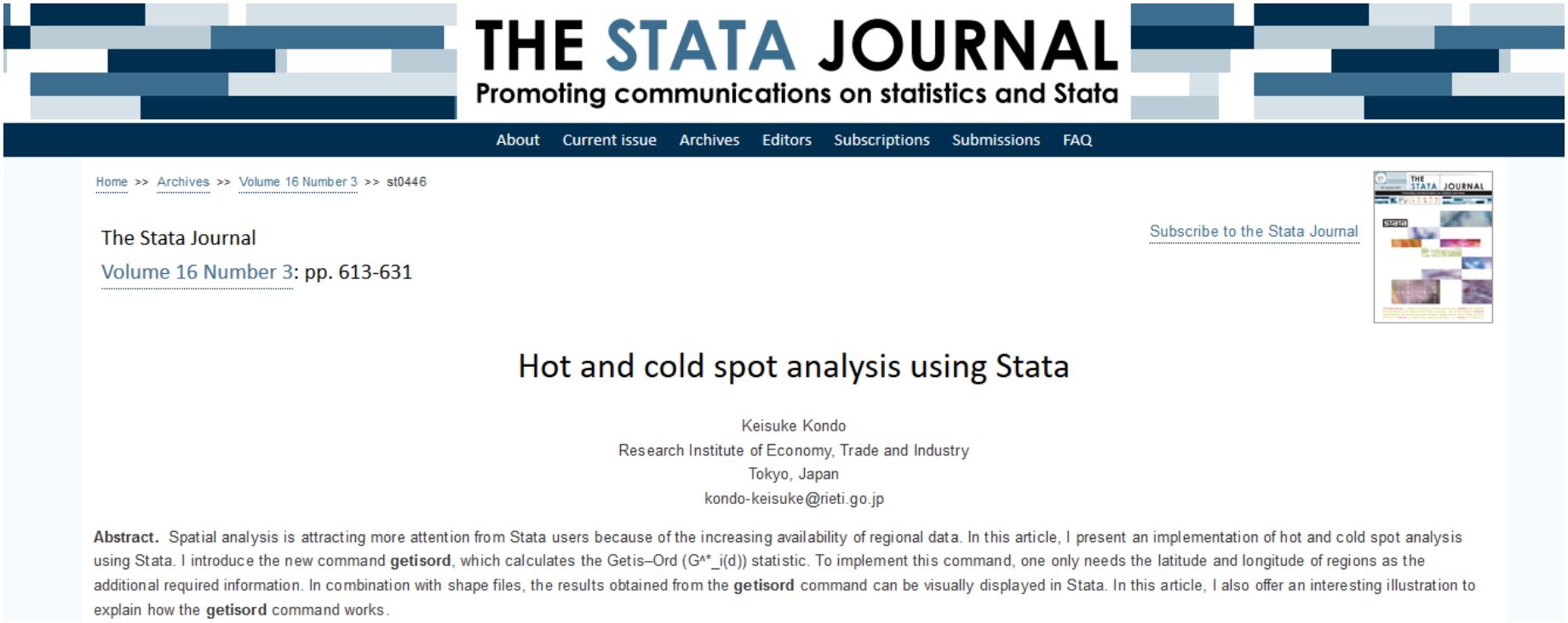
Property crimes per thousand households - LISA Moran



Columbus, Ohio 1980 neighorhood data
Source: Anselin (1988)

Hot and cold spots using *getisord* command

https://www.stata-journal.com/article.html?article=st0446



THE STATA JOURNAL
Promoting communications on statistics and Stata

About Current issue Archives Editors Subscriptions Submissions FAQ

Home >> Archives >> Volume 16 Number 3 >> st0446

The Stata Journal
Volume 16 Number 3: pp. 613-631

Subscribe to the Stata Journal

Hot and cold spot analysis using Stata

Keisuke Kondo
Research Institute of Economy, Trade and Industry
Tokyo, Japan
kondo-keisuke@rieti.go.jp

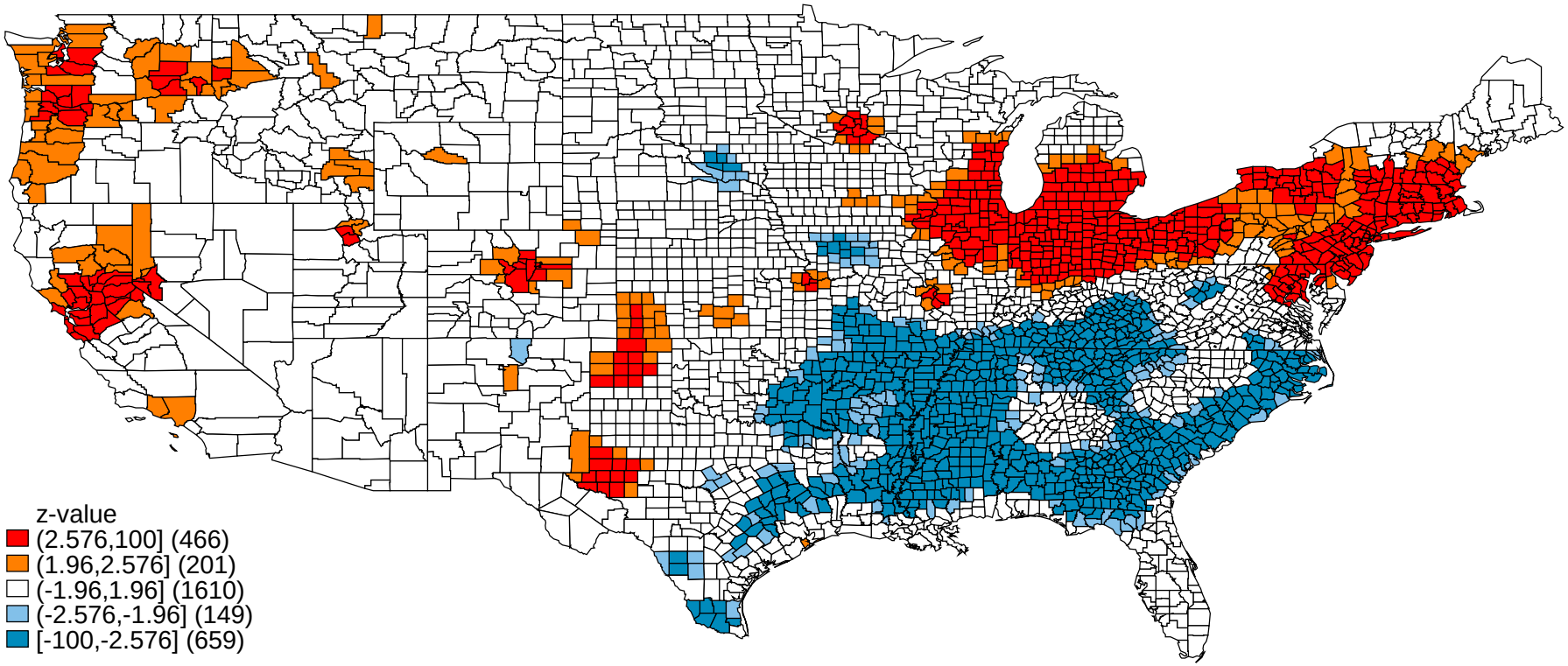
Abstract. Spatial analysis is attracting more attention from Stata users because of the increasing availability of regional data. In this article, I present an implementation of hot and cold spot analysis using Stata. I introduce the new command **getisord**, which calculates the Getis–Ord (G^*_i(d)) statistic. To implement this command, one only needs the latitude and longitude of regions as the additional required information. In combination with shape files, the results obtained from the **getisord** command can be visually displayed in Stata. In this article, I also offer an interesting illustration to explain how the **getisord** command works.

NCOVR (https://geodacenter.github.io/data-and-lab/ncovr/)

Homicides and selected socio-economic characteristics for continental U.S. counties.
Observations = 3,085 Variables = 69 Years = 1960s-90s
Source: S. Messner, L. Anselin, D. Hawkins, G. Deane, S. Tolnay, R. Baller (2000). An Atlas of the Spatial Patterning of County-Level Homicide, 1960-1990. Pittsburgh, PA, National Consortium on Violence Research.

NAME	county name
STATE_NAME	state name
STATE_FIPS	state fips code (character)
CNTY_FIPS	county fips code (character)
FIPS	combined state and county fips code (character)
STFIPS	state fips code (numeric)
COFIPS	county fips code (numeric)
FIPSNO	fips code as numeric variable
SOUTH	dummy variable for Southern counties (South = 1)
HR**	homicide rate per 100,000 (1960, 1970, 1980, 1990)
HC**	homicide count, three year average centered on 1960, 1970, 1980, 1990
PO**	county population, 1960, 1970, 1980, 1990
RD**	resource deprivation 1960, 1970, 1980, 1990
PS**	population structure 1960, 1970, 1980, 1990
UE**	unemployment rate 1960, 1970, 1980, 1990
DV**	divorce rate 1960, 1970, 1980, 1990 (% males over 14 divorced)
MA**	median age 1960, 1970, 1980, 1990
POL**	log of population 1960, 1970, 1980, 1990
DNL**	log of population density 1960, 1970, 1980, 1990
MFIL**	log of median family income 1960, 1970, 1980, 1990
FP**	% families below poverty 1960, 1970, 1980, 1990 (see Codebook for details)
BLK**	% black 1960, 1970, 1980, 1990
GI**	Gini index of family income inequality 1960, 1970, 1980, 1990
FH**	% female headed households 1960, 1970, 1980, 1990

Hot and cold spots using *getisord* command – MFIL59



Geographically weighted regression

Spatial heterogeneity

gwr

gwrgrid

https://www.staff.ncl.ac.uk/m.s.pearce/stbgwr.htm

Stand-alone software (and R packages)

http://gwr.maynoothuniversity.ie/gwr4-software/

https://gwrtools.github.io/

Georgia (https://www.census.gov/geo/maps-data/data/cbf/cbf_counties.html)

1990 Census data for 159 counties of the State of Georgia

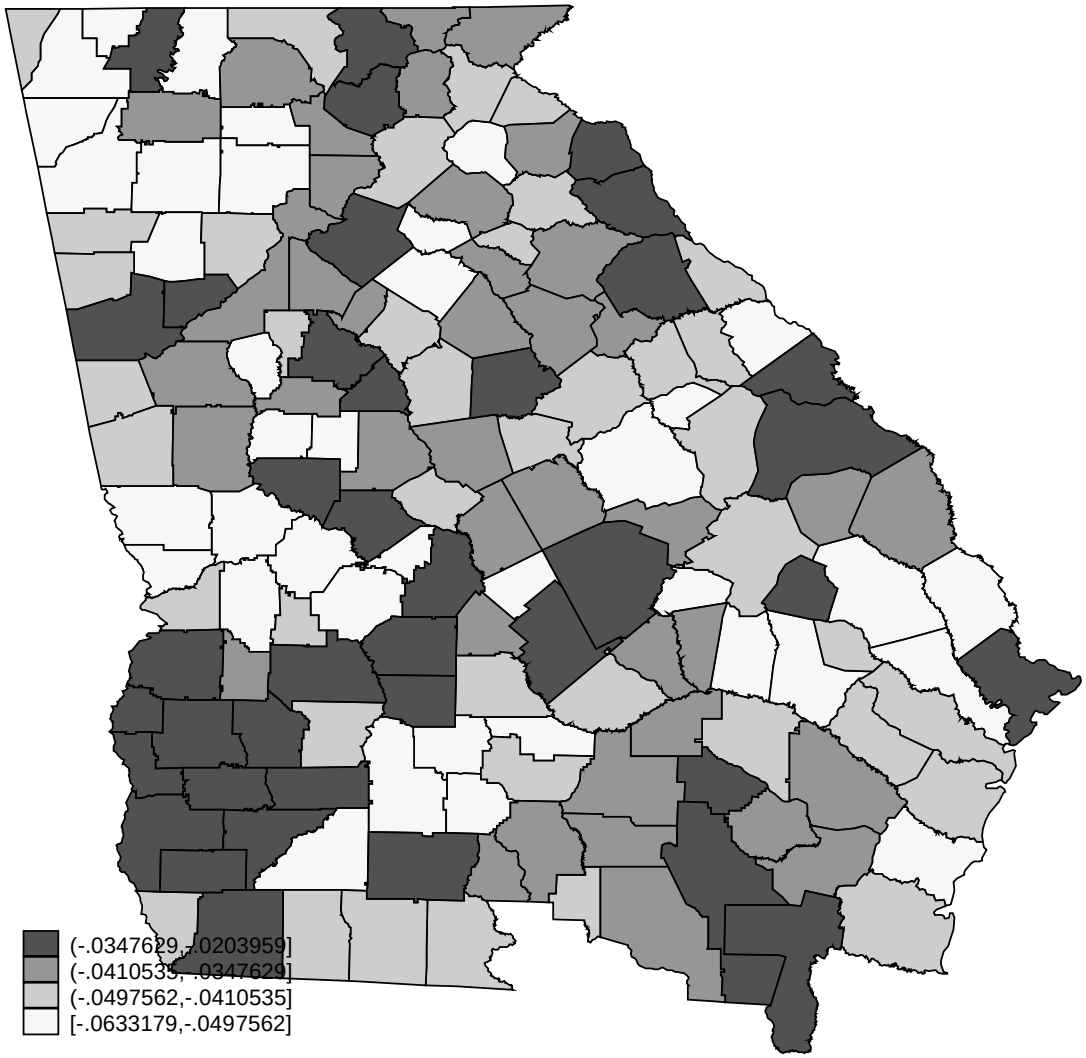
AreaKey	An identification number for each county
Latitude	The latitude of the county centroid
Longitud	The longitude of the county centroid
TotPop90	Population of the county in 1990
PctRural	Percentage of the county population defined as rural
PctBach	Percentage of the county population with a bachelors degree
PctEld	Percentage of the county population aged 65 or over
PctFB	Percentage of the county population born outside the US
PctPov	Percentage of the county population living below the poverty line
PctBlack	Percentage of the county population who are black
ID	a numeric vector of IDs

Geographically weighted regression

```
. regress PctBach TotPop90 PctRural PctEld PctFB PctPov PctBlack

      Source |       SS           df       MS      Number of obs   =       159
-------------+----------------------------------   F(6, 152)       =     46.20
       Model |  3311.91242         6  551.985404   Prob > F        =    0.0000
    Residual |   1816.1638       152  11.9484461   R-squared       =    0.6458
-------------+----------------------------------   Adj R-squared   =    0.6319
       Total |  5128.07623       158  32.4561786   Root MSE        =    3.4567

------------------------------------------------------------------------------
     PctBach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    TotPop90 |   .0000236   4.75e-06     4.96   0.000     .0000142    .0000329
    PctRural |  -.0438531   .0137165    -3.20   0.002    -.0709527   -.0167536
      PctEld |  -.0619128   .1214583    -0.51   0.611    -.3018772    .1780517
       PctFB |   1.256153   .3098017     4.05   0.000       .64408    1.868227
      PctPov |  -.1554096   .0703871    -2.21   0.029     -.294473   -.0163462
    PctBlack |   .0219037   .0252512     0.87   0.387    -.0279849    .0717922
       _cons |   14.77658   1.705775     8.66   0.000     11.40649    18.14667
------------------------------------------------------------------------------
```

```
. sum PctRural, detail

                            PctRural
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -.0630155      -.0633179
 5%    -.0598568      -.0630155
10%    -.0563486      -.0626328       Obs                 159
25%    -.0497562      -.0621394       Sum of Wgt.         159

50%    -.0410535                      Mean           -.0421685
                      Largest         Std. Dev.       .0100101
75%    -.0347629      -.0230648
90%    -.0292019      -.0225735       Variance        .0001002
95%    -.0259903      -.0222334       Skewness        -.127237
99%    -.0222334      -.0203959       Kurtosis        2.332528
```

Geographically weighted regression



Now try what you have learned by answering exercise 1