# 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.

**Keywords:** st0446, getisord, Getis–Ord $G_i^*(d)$, local spatial autocorrelation, shape file

## 1 Introduction

Spatial analysis is becoming more popular because of the increasing availability of geographically disaggregated data and map files (for example, shape files); hence, there is a growing demand among researchers worldwide for spatial analysis using Stata. However, Stata packages that specialize in spatial analysis currently have some computational difficulties, especially in how a spatial weight matrix is included and processed in Stata. In this article, I aim to fill this gap by introducing the new command `getisord` for hot and cold spot analysis.

Our socioeconomic activities are concentrated in certain locations in the real world, and the spatial pattern is not randomly distributed. Thus, one of the purposes of spatial analysis is to describe how our socioeconomic activities are distributed in space. To detect clusters (that is, groups of spatially contiguous areas) in a geographic space, Getis and Ord (1992) developed the Getis–Ord $G_i^*(d)$ statistic.

Spatial autocorrelation is an important concept in this literature and comprises two strands. Global spatial autocorrelation, such as Moran's $I$, looks at the overall spatial interdependence between regions and tests the degree to which a region and its neighboring regions are, on average, mutually correlated. On the other hand, local spatial autocorrelation, such as Getis–Ord $G_i^*(d)$, is motivated by the idea that the spatial association may be locally heterogeneous, even if a global spatial autocorrelation is not observed.

The Getis–Ord $G_i^*(d)$ statistic tests whether a region and its neighboring regions form a spatial cluster for each region. In other words, hot and cold spots are detected

as spatial outliers.[1] Furthermore, the method developed by Getis and Ord (1992) is extended to the generalized case by Ord and Getis (1995) to flexibly consider the degrees of spatial connections.[2]

The `getisord` command introduced in this article allows us to calculate Getis–Ord $G_i^*(d)$ with binary and nonbinary spatial weight matrices. To implement the `getisord` command, you need geographic information on the latitude and longitude; that is, researchers can easily conduct hot and cold spot analysis as long as the datasets contain basic regional information, such as the zip code, city code, and city name. Current geocoding techniques facilitate the addition of the latitude and longitude into the dataset, even if a suitable shape file of the corresponding area is not available.[3]

In the existing literature, Pisati (2001) provides the `spatlsa` command, which includes a calculation function for the Getis–Ord $G_i^*(d)$ statistic.[4] The `spatlsa` command computes the geographic distance between locations under the projected coordinate system (expressed in kilometers or miles). In turn, the `getisord` command calculates the great-circle distance between locations by using the Vincenty (1975) formula based on the geographic (or spherical) coordinate system (expressed in longitude and latitude), which is directly compatible with current geocoding techniques. This coding method facilitates the calculation of geographic distance in Stata, independently of the map projections used in Geographic Information System software.

The `getisord` command becomes a more powerful tool if a shape file is available. Visualization helps us foster a better understanding of the spatial analysis. Fortunately, Stata already provides the `shp2dta` command, which converts a shape file to a Stata `.dta` file (Crow 2006). In addition, results obtained from the `getisord` command can be visualized in Stata in combination with the `spmap` command, which displays regional data in a map (Pisati 2007).

In this article, I provide an interesting illustration of the `getisord` command. Using the U.S. county data on median family income from 1959 to 1989, the `getisord` command clarifies which counties in the United States have formed high-income clusters from spatial and dynamic perspectives. A similar analysis is conducted in Kondo (2015b), which detects Japanese unemployment clusters that permanently show high unemployment rates regardless of temporal fluctuations from 1980 to 2005.

---

1. Intuitively, the hot (cold) spot is detected when a region and its neighboring regions have similar values and higher (lower) values than the average. The local Moran's $I_i$ statistic, which is proposed by Anselin (1995) as a new class of local indicator of spatial association, can also be used for spatial clustering detection like the Getis–Ord $G_i^*(d)$ statistic. However, the local indicator of spatial association is designed to decompose a global indicator of spatial association into the contribution of each observation.

2. Getis and Ord (1992) also proposed the Getis–Ord $G_i(d)$ statistic, which does not include the value of region $i$. In this article, I focus only on Getis–Ord $G_i^*(d)$, which is frequently used in empirical analyses.

3. For example, the Google Maps geocoding application programming interface is publicly available.

4. See Pisati (2012) for a comprehensive review of the Stata package on spatial analysis including `spatlsa`.

In addition, I discuss the possibility of developing Stata packages for spatial statistics and spatial econometrics. Researchers often face difficulty in dealing with spatial weight matrices in Stata, which makes spatial analysis difficult within the Stata framework. An outstanding command for constructing a spatial weight matrix is the `spmat` command offered by Drukker et al. (2013a), which can construct a spatial weight matrix from a contiguity matrix by using a shape file and can construct an inverse-distance matrix by using coordinate variables (for example, longitude and latitude). Then, the spatial weight matrix is exogenously included in Stata as a matrix type when researchers do spatial econometric analysis with the `spreg` command (Drukker, Prucha, and Raciborski 2013b). The key feature of the `getisord` command is that the spatial weight matrix is endogenously constructed in a sequence of steps in the program code.

A spatial weight matrix based on a contiguity matrix might be dominant in the literature, but constructing one using a shape file may prove difficult because a suitable shape file may not be available in some situations. However, geographic information on the latitude and longitude is easily available via current geocoding techniques. In that sense, Stata packages are expected to be extended to allow construction of a spatial weight matrix based on a distance matrix. Therefore, this article contributes to the construction method of a spatial weight matrix in Stata, which facilitates further development of Stata packages for spatial analysis.[5]

The rest of this article is organized as follows. Section 2 reviews the Getis–Ord $G_i^*(d)$ statistic, and section 3 explains how bilateral distance is measured in Stata. Section 4 describes the `getisord` command, and section 5 offers an example. Section 6 concludes.

## 2 Detecting hot and cold spots

Getis and Ord (1992) developed a method for hot and cold spot analysis in a geographic space. The Getis–Ord $G_i^*(d)$ statistic (Getis and Ord 1992) for variable $x_i$ of region $i$ is calculated as

$$G_i^*(d) = \frac{\sum\limits_{j=1}^{N} w_{ij}(d)x_j}{\sum\limits_{j=1}^{N} x_j} \tag{1}$$

where $N$ is the number of regions and $w_{ij}(d)$ denotes the $ij$th element of the spatial weight matrix as

$$w_{ij}(d) = \begin{cases} 1, & \text{if} \quad d_{ij} < d, \quad \text{for all } i,j \\ 0, & \text{otherwise} \end{cases}$$

where $d$ is the threshold distance for the spatial weight matrix. Note that the diagonal elements take the value of 1 because $d_{ii} = 0$. Hereafter, I use the notation $w_{ij}(d)$ to denote a general form of spatial weight matrix, including a nonbinary spatial weight matrix.

---

5. Using a method wherein a spatial weight matrix is endogenously constructed in the program code, Kondo (2015a) developed the `spgen` command, which computes spatially lagged variables in Stata.

The essence of Getis–Ord $G_i^*(d)$ is as follows. The numerator in (1) gives the local sum of variable $x$ within a circle of $d$ radius from the base point (for example, centroid) of region $i$, and the denominator in (1) gives the total sum of variable $x$ for all the regions. The Getis–Ord $G_i^*(d)$ statistic evaluates the ratio of the local sum to the total sum for each region. Therefore, hot and cold spots are detected as spatial outliers.

Ord and Getis (1995) extended this statistic to the case of the nonbinary spatial weight matrix. The generalized formula of Getis–Ord $G_i^*(d)$ is defined for both the binary and nonbinary spatial weight matrices. The standardized Getis–Ord $G_i^*(d)$, which is equivalent to the $z$-value of Getis–Ord $G_i^*(d)$, is given by

$$\text{Standardized } G_i^*(d) = \frac{G_i^*(d) - E\{G_i^*(d)\}}{\sqrt{\text{Var}\{G_i^*(d)\}}}$$

Under the complete spatial randomness, the expectation and the variance of the Getis–Ord $G_i^*(d)$ are, respectively, derived as

$$E\{G_i^*(d)\} = \frac{\sum\limits_{j=1}^{N} w_{ij}(d)}{N}$$

$$\text{Var}\{G_i^*(d)\} = \frac{N \sum\limits_{j=1}^{N} w_{ij}^2(d) - \left\{\sum\limits_{j=1}^{N} w_{ij}(d)\right\}^2}{N^2(N-1)} \left(\frac{s}{\bar{x}}\right)^2$$

$$\tag{2}$$

where $\bar{x}$ is the sample mean and $s^2$ is the sample variance.[6]

The distribution of the standardized $G_i^*(d)$ approaches a standard normal distribution as $N$ approaches infinity. When the standardized $G_i^*(d)$ takes a positive (negative) value and falls within the critical region, region $i$ is identified as a hot (cold) spot. The critical values for hot and cold spots are approximately $\pm 1.96$ and $\pm 2.58$ at the 5% and 1% significance levels, respectively. In a similar manner, the $p$-value is obtained from the cumulative distribution function of the standard normal distribution.

For the standardized Getis–Ord $G_i^*(d)$, several types of nonbinary spatial weight matrices can be considered. For example, the case of the negative power function is

$$w_{ij}(d) = \begin{cases} (a + d_{ij})^{-\delta}, & \text{if } d_{ij} < d, \quad \text{for all } i, j, \quad \delta > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $\delta$ is a distance decay parameter and $a$ is the constant value added to avoid $w_{ii} = \infty$ because of $d_{ii} = 0$. Ord and Getis (1995) consider the case of the inverse-distance matrix ($\delta = 1$, $d = \infty$).

---

6. The standardized Getis–Ord $G_i^*(d)$ cannot be defined if the numerator of the variance in (2) takes the value of 0. In this case, try a different threshold distance $d$ in the binary spatial weight matrix or a different distance decay parameter $\delta$ in the nonbinary spatial weight matrix.

Another case is the exponential type of spatial weight matrix:

$$w_{ij}(d) = \begin{cases} \exp(-\delta d_{ij}), & \text{if} \quad d_{ij} < d, \quad \text{for all } i, j, \quad \delta > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $\delta$ is the distance decay parameter. Compared with the power function type of spatial weight matrix, an advantage of the exponential type is that it is unnecessary to decide constant $a$ beforehand because the own weight is $w_{ii} = 1$ for $d_{ii} = 0$.

# 3 Measuring distance

I use the Vincenty formula to measure bilateral distance between regions (Vincenty 1975). To speed up computational time in the case where the number of regions is too large, the `getisord` command offers a simplified version of the Vincenty formula.

## 3.1 Vincenty formula

The Vincenty formula is commonly used to measure the geographic distance between two points on Earth. Unlike the spherical law of cosines and the haversine formula, Vincenty (1975) proposed a method measuring geodesic distance under the condition where the shape of the Earth is ellipsoidal.

Given latitudes and longitudes of two locations, the `getisord` command measures the bilateral distance using the Vincenty formula, which considers that the shape of Earth is not a perfect sphere. See Vincenty (1975) for further details of the calculation procedure.

## 3.2 Simplified version of Vincenty formula

The Vincenty formula contains an iteration procedure to improve accuracy. However, the computation of bilateral distance takes considerable time when the number of regions is large.[7] To shorten computational time, the `getisord` command offers the `approx` option, which measures bilateral distance using a simplified version of the Vincenty formula.

The great-circle distance is basically calculated by

$$d_{ij} = r \times \theta$$

where $r$ is the radius of Earth ($\approx$ 6,378.137 kilometers or 3,963.189 miles) and $\theta$ is the central angle of the arc between two locations. Given the latitudes and longitudes of two locations, $\theta$ is calculated by the spherical law of cosines or the haversine formula.

---

7. Because the distance matrix is symmetric, iterations are needed for each of $N(N-1)/2$ elements.

Let $\phi_i$ and $\lambda_i$ denote the latitude and longitude of region $i$ in radians. Following Vincenty (1975), $\theta$ is calculated from the inverse of $\tan\theta = \sin\theta/\cos\theta$ as follows:

$$\theta = \arctan\left[\frac{\sqrt{\{\cos(\phi_j)\sin(\Delta\lambda)\}^2 + \{\cos(\phi_i)\sin(\phi_j) - \sin(\phi_i)\cos(\phi_j)\cos(\Delta\lambda)\}^2}}{\sin(\phi_i)\sin(\phi_j) + \cos(\phi_i)\cos(\phi_j)\cos(\Delta\lambda)}\right]$$

where $\Delta\lambda = \lambda_j - \lambda_i$.

This approximation works well even without iteration of the exact process in the Vincenty formula. We will examine the comparison between the exact and the approximated procedures of the Vincenty formula in an applied example in section 5.

## 4  Implementation in Stata

### 4.1  Description

The `getisord` command calculates the Getis–Ord $G_i^*(d)$ statistic of *varname*.

### 4.2  Syntax

`getisord` *varname* $\left[\,if\,\right]$ $\left[\,in\,\right]$, `lat(`*varname*`)` `lon(`*varname*`)` `swm(`*swmtype*`)`
  `dist(`#`)` `dunit(km`|`mi)` $\left[\texttt{dms} \underline{\texttt{approx}} \underline{\texttt{const}}\texttt{ant(\#)} \texttt{detail} \underline{\texttt{genallbin}}\right]$

### 4.3  Options

`lat(`*varname*`)` specifies the variable of latitude in the dataset. The decimal format is expected in the default setting. A positive value denotes the north latitude, whereas a negative value denotes the south latitude. `lat()` is required.

`lon(`*varname*`)` specifies the variable of longitude in the dataset. The decimal format is expected in the default setting. A positive value denotes the east longitude, whereas a negative value denotes the west longitude. `lon()` is required.

`swm(`*swmtype*`)` specifies a type of spatial weight matrix. One of the following three types of spatial weight matrix must be specified: `bin` (binary), `exp` (exponential), or `pow` (power). The distance decay parameter # must be specified for the exponential and power function types of spatial weight matrix as follows: `swm(exp #)` and `swm(pow #)`. `swm()` is required.

`dist(`#`)` specifies the threshold distance # for the spatial weight matrix. The unit of distance is specified by the `dunit()` option. Regions located within the threshold distance # take a value of 1 in the binary spatial weight matrix or a positive value in the nonbinary spatial weight matrix, and take 0 otherwise. `dist()` is required.

dunit(km | mi) specifies the unit of distance. Either km (kilometers) or mi (miles) must be specified. dunit() is required.

dms converts the degrees, minutes, and seconds format to a decimal format.

approx uses the bilateral distance approximated by the simplified version of the Vincenty formula.

constant(#), when swm(pow #) is used, specifies a constant term # in a unit specified by the dunit() option, which is added to the bilateral distance, to avoid the denominator of the spatial weight matrix taking a value of 0. The constant(#) option must be specified when swm(pow #) is used.

detail displays summary statistics of the bilateral distance.

genallbin generates three additional outcome variables (the unstandardized Getis–Ord $G_i^*(d)$, its expected value, and the standard deviation) only when swm(bin) is specified.

## 4.4 Output

### Outcome variables

In the default setting, the getisord command generates two outcome variables in the dataset (go_z_*varname*_*swmtype* and go_p_*varname*_*swmtype*). When the binary spatial weight matrix, swm(bin), is specified, the genallbin option becomes valid, and getisord generates an additional three outcome variables (go_*varname*_*swmtype*, go_e_*varname*_*swmtype*, and go_sd_*varname*_*swmtype*).

go_z_*varname*_*swmtype* is the standardized Getis–Ord $G_i^*(d)$ statistic of *varname*, which is equivalent to the *z*-value of Getis–Ord $G_i^*(d)$. The *varname* is automatically inserted, and the suffix b, e, or p is also inserted in accordance with *swmtype*: b for swm(bin), e for swm(exp #), and p for swm(pow #).

go_p_*varname*_*swmtype* is the *p*-value of the standardized Getis–Ord $G_i^*(d)$ statistic of *varname*.

go_*varname*_*swmtype* is the Getis–Ord $G_i^*(d)$ statistic of *varname* in (1). This is generated using the genallbin option only when swm(bin) is specified.

go_e_*varname*_*swmtype* is the expected value of the Getis–Ord $G_i^*(d)$ statistic of *varname* in (2). This is generated using the genallbin option only when swm(bin) is specified.

go_sd_*varname*_*swmtype* is the standard deviation of the Getis–Ord $G_i^*(d)$ statistic of *varname* in (2). This is generated using the genallbin option only when swm(bin) is specified.

❏ **Technical note**

The `getisord` command works in combination with shape files of the corresponding area. The standardized Getis–Ord $G_i^*(d)$ obtained by the `getisord` command can be visually displayed in a map. Fortunately, Stata already has useful commands that convert shape files to `.dta` files (the `shp2dta` command) and depict colorful maps (the `spmap` command).[8] In the next section, we will look at an interesting illustration of the `getisord` command in combination with the `shp2dta` and `spmap` commands.

❏

**Stored results**

`getisord` stores the following in `r()`:

Scalars
| | | | |
|---|---|---|---|
| `r(N)` | number of observations | `r(dist_sd)` | standard deviation of distance |
| `r(td)` | threshold distance | `r(dist_min)` | minimum value of distance |
| `r(dd)` | distance decay parameter | `r(dist_max)` | maximum value of distance |
| `r(cons)` | constant for `swm(pow #)` | `r(HS)` | number of hot spots ($p < 5\%$) |
| `r(dist_mean)` | mean of distance | `r(CS)` | number of cold spots ($p < 5\%$) |

Macros
| | | | |
|---|---|---|---|
| `r(cmd)` | `getisord` | `r(dunit)` | unit of distance |
| `r(varname)` | name of variable | `r(dist_type)` | exact or approximation |
| `r(swm)` | type of spatial weight matrix | | |

Matrices
| | |
|---|---|
| `r(D)` | lower triangle distance matrix |

# 5 Example

## 5.1 Basic manipulation

I illustrate use of the `getisord` command with the National Consortium on Violence Research (NCOVR) dataset, which is publicly available from the GeoDa Center for Geospatial Analysis and Computation at Arizona State University.[9] The NCOVR dataset contains U.S. shape files at the county level and regional information on homicide, population, labor, and households from 1959 to 1991. The NCOVR dataset coherently has 3,085 counties between 1959 and 1991 accounting for changed county boundaries during this period. In this example, I use the logarithm of median family income in 1959, 1969, 1979, and 1989 to examine dynamic aspects of high-income spots as groups of spatially contiguous counties.

The NCOVR dataset contains three files: `NAT.dbf`, `NAT.shp`, and `NAT.shx`. To begin, the shape file must be converted to the Stata `.dta` format by using the `shp2dta` command:

```
. shp2dta using "NAT", data(nat-d) coor(nat-c) genid(id) genc(cntrd)
  (output omitted)
```

---

8. See Crow (2006) and Pisati (2007) for more details about `shp2dta` and `spmap`, respectively.
9. See https://geodacenter.asu.edu/.

This command creates two files, `nat-d.dta` and `nat-c.dta`, in the current directory. The `genc()` option creates variables of the latitude and longitude in the dataset (in the above case, `y_cntrd` and `x_cntrd`, respectively).

This dataset is now ready for the hot and cold spot analysis because the geographic information on the latitude and longitude is already included.[10] In the following example, I use `getisord` to consider the binary spatial weight matrix with a threshold distance of $d = 50$ kilometers.

```
. use nat-d
. getisord MFIL59, lat(y_cntrd) lon(x_cntrd) swm(bin) dist(50) dunit(km) approx
> detail
Distance by simplified version of Vincenty formula (unit: km)
```

|  | Obs. | Mean | S.D. | Min. | Max |
|---|---|---|---|---|---|
| Distance | 4757070 | 1360.707 | 799.540 | 0.854 | 4566.705 |

```
Getis-Ord G*i(d) Statistics
```
                                                    Number of Obs =      3085

| Variable | z<=-2.58 | -2.58<z<=-1.96 | -1.96<z<1.96 | 1.96<=z<2.58 | 2.58<=z |
|---|---|---|---|---|---|
| MFIL59 | 378 | 177 | 2151 | 171 | 208 |

```
go_z_MFIL59_b and go_p_MFIL59_b are generated in the dataset.
. return list
scalars:
                    r(N) =  3085
                   r(td) =  50
                   r(dd) =  .
                 r(cons) =  .
            r(dist_mean) =  1360.706941777694
              r(dist_sd) =  799.5398649929353
             r(dist_min) =  .8540492034745548
             r(dist_max) =  4566.705435790723
                   r(HS) =  379
                   r(CS) =  555
macros:
                  r(cmd) : "getisord"
              r(varname) : "MFIL59"
                  r(swm) : "binary"
                r(dunit) : "km"
            r(dist_type) : "approximation"
matrices:
                    r(D) :  3085 x 3085
```

Because of the large number of counties, I used the `approx` option to shorten the computational time.

---

10. Confirm whether the latitude and longitude are set exactly from the shape file. Some shape files have no geographic information on the latitude and longitude.

The `getisord` command displays summary statistics of the distance matrix in the top table because I used the `detail` option. The number of observations denotes the number of elements in the lower triangle distance matrix $[= N(N-1)/2]$. In this case, the mean distance between two counties is approximately 1,361 kilometers. The minimum and maximum distances are 0.854 kilometers and 4,566.705 kilometers, respectively.[11]

In the lower table, `getisord` displays a summary of the hypothesis testing results of the complete spatial randomness at the 5% and 1% levels. The numbers of hot spot counties on the median family income at the 5% and 1% levels are 171 and 208, respectively. On the other hand, the numbers of cold spot counties on the median family income at the 5% and 1% levels are 177 and 378, respectively. These results are stored in `r()`. In this example, the `getisord` command generates two variables in the dataset: $z$-values of Getis–Ord $G_i^*(d)$ (`go_z_MFIL59_b`) and $p$-values of Getis–Ord $G_i^*(d)$ (`go_p_MFIL59_b`).

▷ **Example**

The `getisord` command can specify two types of nonbinary spatial weight matrices as follows:

```
. getisord MFIL59, lat(y_cntrd) lon(x_cntrd) swm(exp 0.03) dist(50) dunit(km)
> approx detail
```
  (*output omitted*)
```
. getisord MFIL59, lat(y_cntrd) lon(x_cntrd) swm(pow 1) dist(50) dunit(km) approx
> constant(1) detail
```
  (*output omitted*)

The `swm()` option for the nonbinary spatial weight matrix requires the distance decay parameter. Here, distance decay parameters are specified as $\delta = 0.03$ in (4) and as $\delta = 1$ in (3).

◁

## 5.2 Mapping results

Visualization is a useful method to promote a better understanding of empirical results. The `getisord` command works in combination with the `spmap` command, which displays a map in Stata. An example is given below:

---

11. In this shape file, the minimum distance is observed between Henry, VA $(36.68377, -79.87409)$ and Martinsville, VA $(36.68407, -79.86453)$. The maximum distance is observed between San Mateo, CA $(37.42419, -122.3202)$ and Washington, ME $(45.04659, -67.63785)$. The numbers in parentheses denote the latitude and longitude of the centroid in a decimal format. The negative longitude value denotes the west longitude.

```
. use nat-d, clear
. getisord MFIL59, lat(y_cntrd) lon(x_cntrd) swm(bin) dist(50) dunit(km) approx
> detail
```
 (*output omitted*)
```
. spmap go_z_MFIL59_b using "nat-c", id(id)
>    clm(custom) clb(-100 -2.576 -1.960 1.960 2.576 100)
>    fcolor(ebblue eltblue white orange red) legtitle("{it: z}-value")
>    legstyle(1) legcount legend(size(*1.8))
```

After implementing the `getisord` command, the `spmap` command visualizes $z$-values of Getis–Ord $G_i^*(d)$ (`go_z_MFIL59_b`) in the map. Figure 1 is created by this command.[12]

## 5.3 Empirical application

Figures 1–4 present results of a dynamic hot spot analysis on the median family income at the U.S. county level from 1959 to 1989. The binary spatial weight matrix with a threshold distance of 50 kilometers is used.

The hot spot counties are scattered across the United States. From 1959 to 1989, the biggest spots are concentrated in the Northeastern region: Boston, MA; Rochester and New York, NY; Philadelphia and Pittsburgh, PA; Washington, DC; Cincinnati, Columbus, and Cleveland, OH; Detroit, MI; Indianapolis, IN; Chicago, IL; and Milwaukee, WI. However, the spatial pattern dynamically changes between 1959 and 1989. In particular, the hot spot areas in OH, MI, IN, IL, and WI become centered on big cities.

In the Western region, Seattle, WA; Portland, OR; San Francisco, Sacramento, San Jose, and Los Angeles, CA; Salt Lake City, UT; and Denver, CO are classified as hot spots from 1959 to 1989.

Few hot spots are identified in the Southern region in 1959. However, hot spot counties gradually emerge over time. Atlanta, GA is an outstanding place that has expanded the hot spot area outward from 1969 to 1989. In a similar manner, Nashville, TN, and Houston and Dallas, TX appear as hot spots over time.

In this example, I have examined how the spatial pattern of high-income counties has changed dynamically in the United States. The `getisord` command provides a flexible extension for the spatial weight matrix to satisfy a variety of researchers' demands.

---

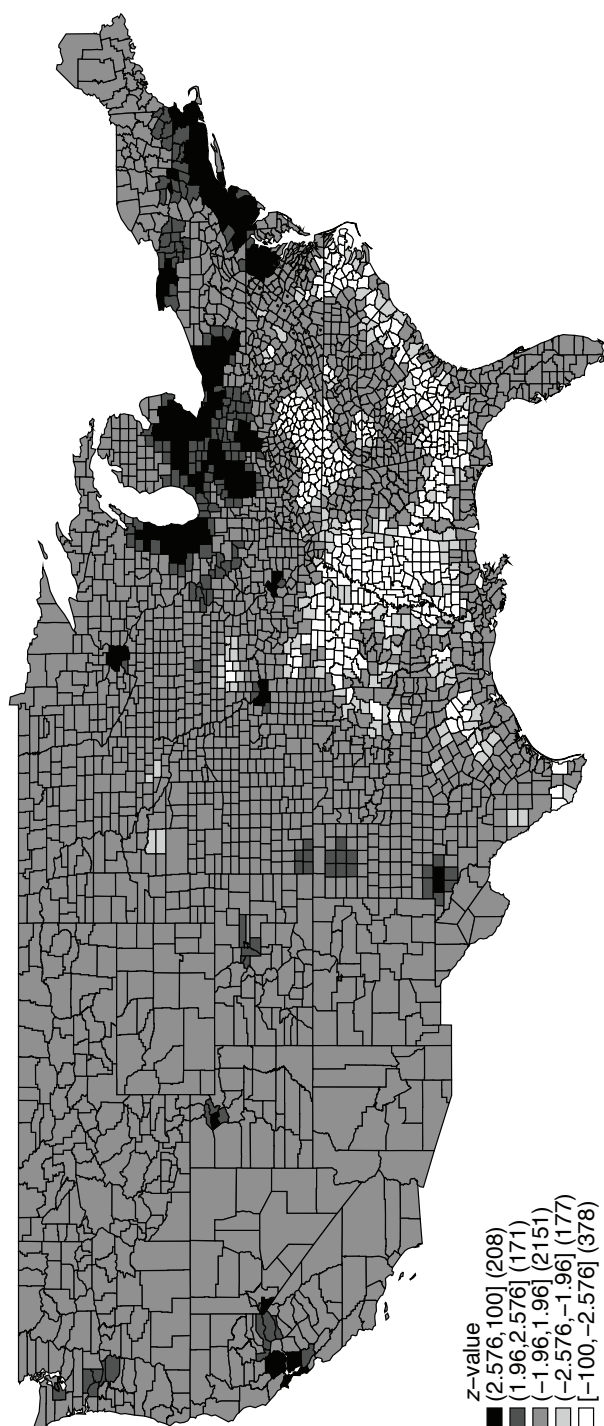12. Although figure 1 is shown in black and white, this example produces a color map.

Figure 1. Mapping Getis–Ord $G_i^*(d)$ of median family income in 1959, $d = 50$ kilometers

Note: The $z$-values of Getis–Ord $G_i^*(d)$ are calculated by the `getisord` command. They are illustrated by the `spmap` command. The original NCOVR dataset is taken from the GeoDa Center for Geospatial Analysis and Computation (https://geodacenter.asu.edu/).
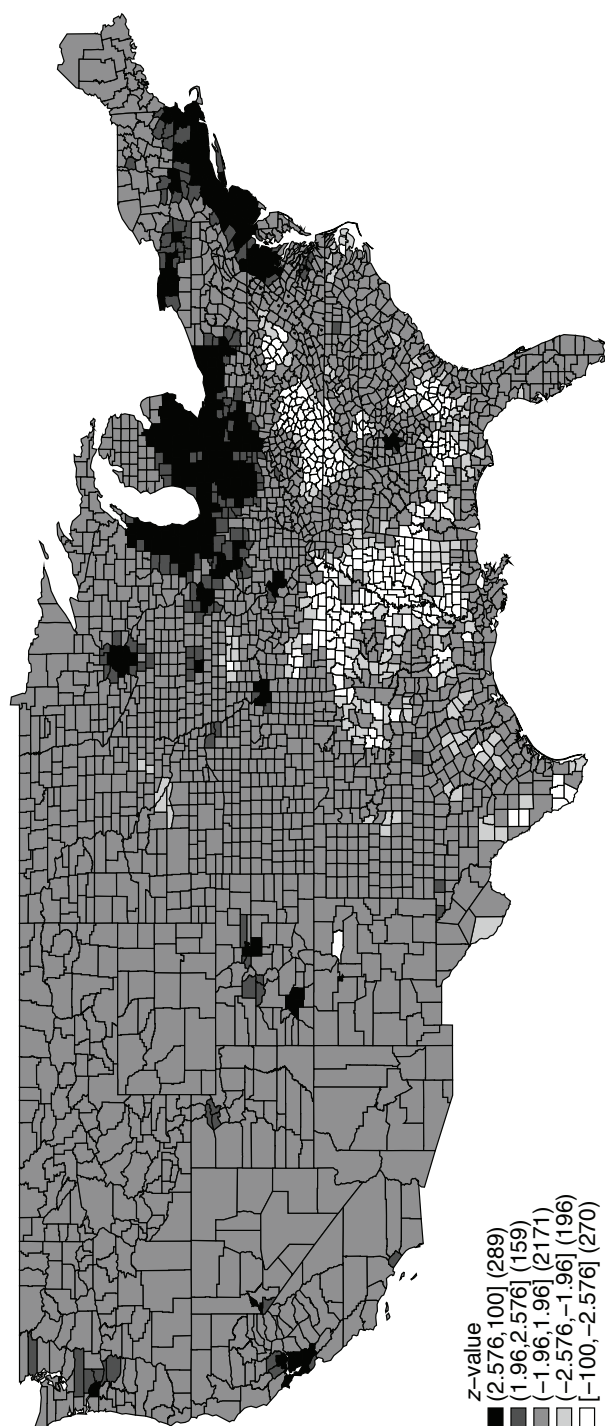
Figure 2. Mapping Getis–Ord $G_i^*(d)$ of median family income in 1969, $d = 50$ kilometers

Note: The $z$-values of Getis–Ord $G_i^*(d)$ are calculated by the getisord command. They are illustrated by the spmap command. The original NCOVR dataset is taken from the GeoDa Center for Geospatial Analysis and Computation (https://geodacenter.asu.edu/).
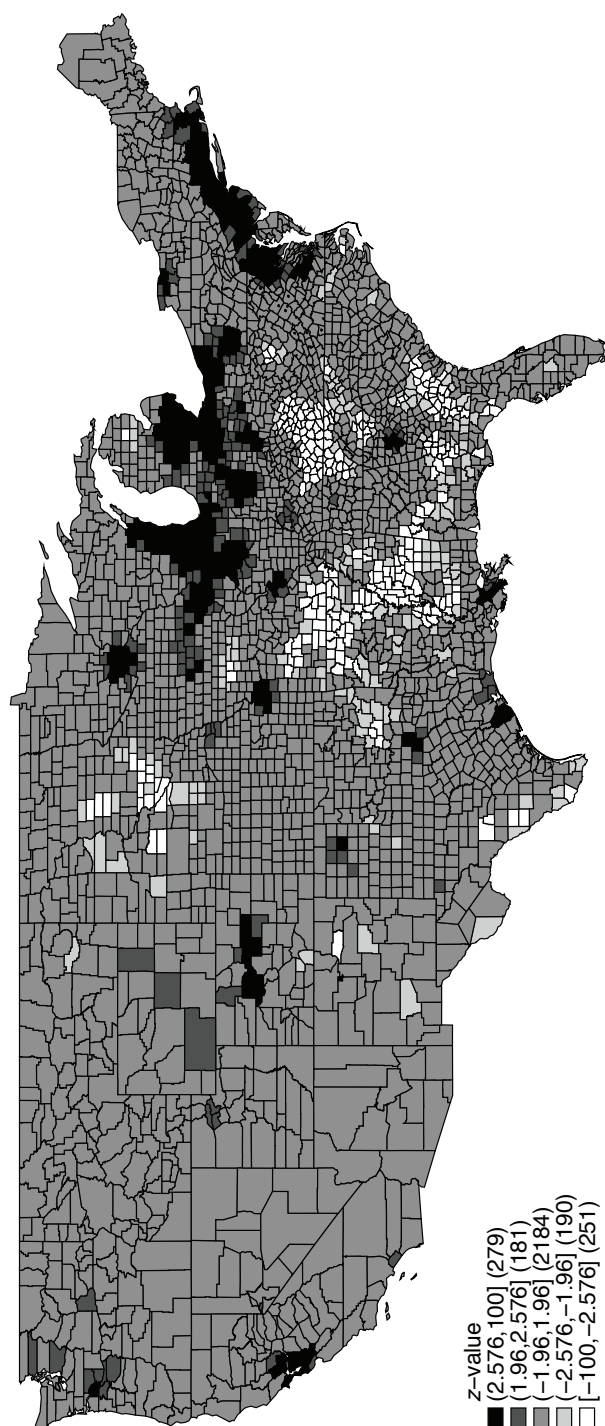
Figure 3. Mapping Getis–Ord $G_i^*(d)$ of median family income in 1979, $d = 50$ kilometers

Note: The $z$-values of Getis–Ord $G_i^*(d)$ are calculated by the `getisord` command. They are illustrated by the `spmap` command. The original NCOVR dataset is taken from the GeoDa Center for Geospatial Analysis and Computation (https://geodacenter.asu.edu/).

Figure 4. Mapping Getis–Ord $G_i^*(d)$ of median family income in 1989, $d = 50$ kilometers

Note: The $z$-values of Getis–Ord $G_i^*(d)$ are calculated by the getisord command. They are illustrated by the spmap command. The original NCOVR dataset is taken from the GeoDa Center for Geospatial Analysis and Computation (https://geodacenter.asu.edu/).
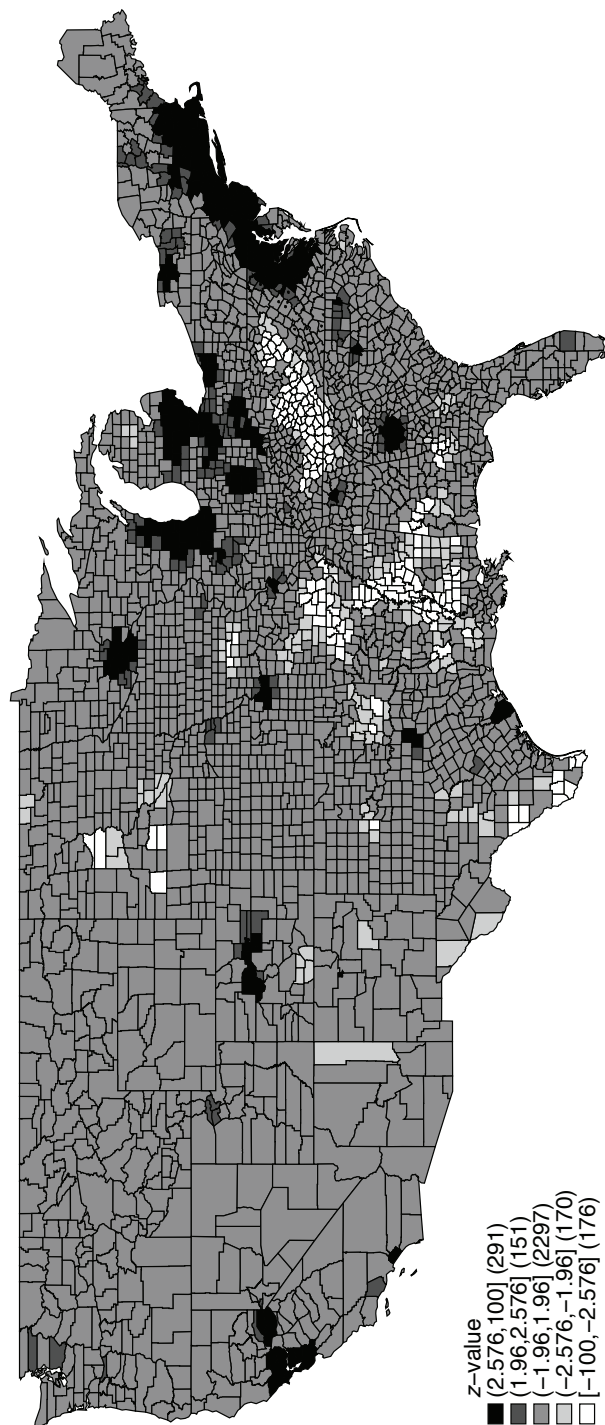
## 5.4   Comparison of distance

The `getisord` command offers the `approx` option, which uses bilateral distance approximated by the simplified version of the Vincenty formula. The comparison between the two types of bilateral distances obtained from the exact and the approximated processes of the Vincenty formula is shown below:

```
. use nat-d, clear
. getisord MFIL59, lat(y_cntrd) lon(x_cntrd) swm(bin) dist(50) dunit(km) detail
Distance by Vincenty formula (unit: km)
```

|          | Obs.    | Mean     | S.D.    | Min.  | Max      |
|----------|---------|----------|---------|-------|----------|
| Distance | 4757070 | 1360.816 | 800.466 | 0.855 | 4572.731 |

```
Getis-Ord G*i(d) Statistics
```
Number of Obs =     3085

| Variable | z<=-2.58 | -2.58<z<=-1.96 | -1.96<z<1.96 | 1.96<=z<2.58 | 2.58<=z |
|----------|----------|----------------|--------------|--------------|---------|
| MFIL59   | 378      | 177            | 2150         | 172          | 208     |

```
go_z_MFIL59_b and go_p_MFIL59_b are generated in the dataset.
. drop go*
. getisord MFIL59, lat(y_cntrd) lon(x_cntrd) swm(bin) dist(50) dunit(km) approx
> detail
Distance by simplified version of Vincenty formula (unit: km)
```

|          | Obs.    | Mean     | S.D.    | Min.  | Max      |
|----------|---------|----------|---------|-------|----------|
| Distance | 4757070 | 1360.707 | 799.540 | 0.854 | 4566.705 |

```
Getis-Ord G*i(d) Statistics
```
Number of Obs =     3085

| Variable | z<=-2.58 | -2.58<z<=-1.96 | -1.96<z<1.96 | 1.96<=z<2.58 | 2.58<=z |
|----------|----------|----------------|--------------|--------------|---------|
| MFIL59   | 378      | 177            | 2151         | 171          | 208     |

```
go_z_MFIL59_b and go_p_MFIL59_b are generated in the dataset.
```

As you can see, the `approx` option hardly affects the results of Getis–Ord $G_i^*(d)$. The only difference is the number of hot spots at the 5% significance level, which is 172 for the exact Vincenty formula, whereas it is 171 for the simplified version of the Vincenty formula.

Figure 5 shows the histogram of the differences between two distance measures. Only a few percentages of distance exhibit differences of 5 kilometers or more. The average of the absolute differences between two distance measures is small (1.55 kilometers), and the correlation coefficient between two distance measures is 1.00. The maximum difference between two distance measures is 7.46 kilometers.
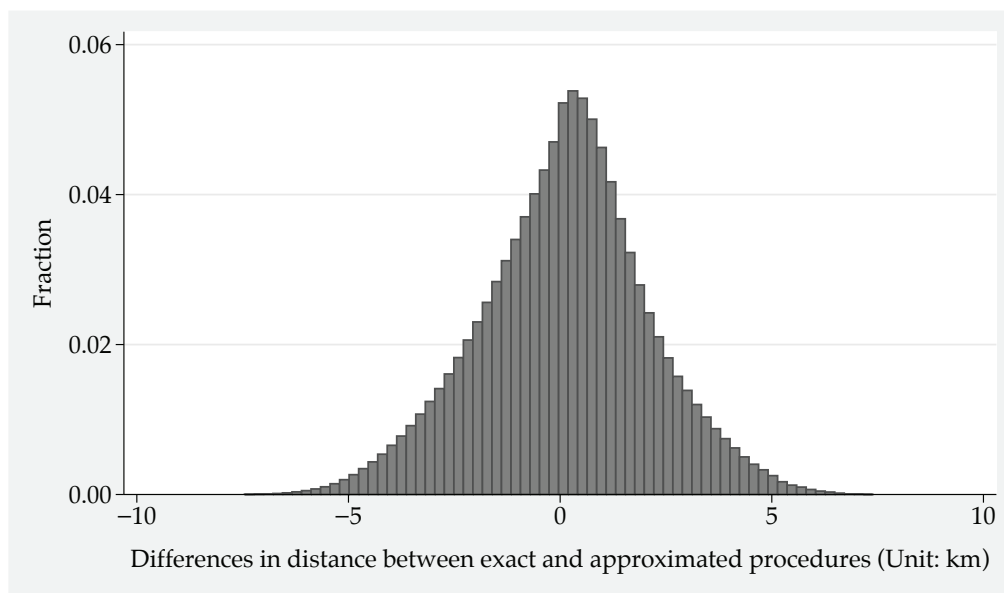
Figure 5. Histogram of differences between two types of distance

Note: The data used here correspond to elements of the lower triangle distance matrix.

To sum up, the `approx` option works well. If the number of regions is too large, the `approx` option enables researchers who want to try various spatial weight matrices to save computational time.

## 6 Concluding remarks

I have introduced the new command `getisord`, which enables Stata users to easily perform hot and cold spot analysis. Given the geographic information on the latitude and longitude, the `getisord` command calculates the Getis–Ord $G_i^*(d)$ statistic with both binary and nonbinary spatial weight matrices. The `spatlsa` command offered by Pisati (2001) also calculates the Getis–Ord $G_i^*(d)$ statistic in Stata. However, it cannot construct a spatial weight matrix based on the geographic distance, a problem that is solved by the `getisord` command. If a shape file is available, the results obtained from `getisord` can be visually illustrated in a map in combination with the `shp2dta` and `spmap` commands in Stata.

Spatial analysis is attracting more attention from Stata users because of increasing availability of regional data. However, difficulties remain in conducting spatial analysis in Stata. An advantage of the `getisord` command is that it does not necessarily require a shape file of the corresponding area because a suitable shape file may not be available. Instead, geographic information on the latitude and longitude is the only requirement;

it is easily added into a dataset directly by using current geocoding techniques. The key feature of `getisord` is that the spatial weight matrix is endogenously constructed in a sequence of steps in the program code and not exogenously included in Stata as a matrix type, which will allow Stata users to more intuitively conduct spatial analysis. I hope that the `getisord` command helps researchers who are interested in hot and cold spot analysis and encourages further extension of packages for spatial analysis in Stata.

# 7    Acknowledgments

# 8    References

Anselin, L. 1995. Local indicators of spatial association—LISA. *Geographical Analysis* 27: 93–115.

Crow, K. 2006. shp2dta: Stata module to convert shape boundary files to Stata datasets. Statistical Software Components S456718, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s456718.html.

Drukker, D. M., H. Peng, I. R. Prucha, and R. Raciborski. 2013a. Creating and managing spatial-weighting matrices with the spmat command. *Stata Journal* 13: 242–286.

Drukker, D. M., I. R. Prucha, and R. Raciborski. 2013b. Maximum likelihood and generalized spatial two-stage least-squares estimators for a spatial-autoregressive model with spatial-autoregressive disturbances. *Stata Journal* 13: 221–241.

Getis, A., and J. K. Ord. 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24: 189–206.

Kondo, K. 2015a. spgen: Stata module to generate spatially lagged variables. Statistical Software Components S458105, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s458105.html.

———. 2015b. Spatial persistence of Japanese unemployment rates. *Japan and the World Economy* 36: 113–122.

Ord, J. K., and A. Getis. 1995. Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis* 27: 286–306.

Pisati, M. 2001. sg162: Tools for spatial data analysis. *Stata Technical Bulletin* 60: 21–37. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 277–298. College Station, TX: Stata Press.

———. 2007. spmap: Stata module to visualize spatial data. Statistical Software Components S456812, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s456812.html.

———. 2012. Exploratory spatial data analysis using Stata. 2012 German Stata Users' Group meeting proceedings. https://ideas.repec.org/p/boc/dsug12/07.html.

Vincenty, T. 1975. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review* 23: 88–93.

**About the author**

Keisuke Kondo is a fellow of the Research Institute of Economy, Trade and Industry. His research interests are spatial economics and spatial econometrics.