Testing for global spatial autocorrelation in Stata*

Keisuke Kondo†

March 31, 2018
(`moransi`: version 1.00)

Abstract

This paper introduces the new Stata command `moransi`, which computes Moran's I statistic to test for global spatial autocorrelation in Stata. The additional information required to implement this command are the latitude and longitude of regions. A practical example is also provided in this paper.

Keywords: `moransi`, Moran's I, global spatial autocorrelation

1 Introduction

The newly developed Stata command, `moransi`, enables researchers to easily calculate Moran's I statistic to test for global spatial autocorrelation in Stata (Moran, 1950). In the literature on spatial statistical analysis, spatial autocorrelation is an important concept, which is further divided into two classes. First, global spatial autocorrelation measures the extent to which regions are interdependent. The Moran's I is a main statistical approach to test for global spatial autocorrelation. In turn, local spatial autocorrelation captures spots showing high spatial autocorrelation locally. The Getis–Ord $G_i^*(d)$ and local Moran's I_i are used to detect hot and cold spots as spatial outliers (Getis and Ord, 1992; Ord and Getis, 1995; Anselin, 1995).[1]

Some researchers have already developed helpful packages for Moran's I in Stata. For example, Pisati (2001) provides the `spatgsa` command. In addition, Jeanty (2010) also offers the `splagvar` command. However, some researchers might have difficulties when using these commands since the spatial weight matrix is exogenously included.

The `moransi` command is simpler than others developed ever before since researchers do not need to consider constructing the spatial weight matrix in advance. Matching regional IDs between

*This is a research outcome undertaken at the Research Institute of Economy, Trade and Industry. Sample datasets used in this article are available online (URL: `https://sites.google.com/site/keisukekondokk/`).

†Research Institute of Economy, Trade and Industry. 1-3-1 Kasumigaseki, Chiyoda-ku, Tokyo, 100–8901, Japan. (e-mail: `kondo-keisuke@rieti.go.jp`).

[1]Kondo (2016) provides the Stata command, `getisord`, which calculates Getis–Ord $G_i^*(d)$ statistic.

data and spatial weight matrix is not easy.[2] The `moransi` command solves this issue by facilitating a computing procedure of spatial weight matrix.

The key feature of the `moransi` command is that the spatial weight matrix is endogenously constructed in a sequence of the program code and not exogenously included into Stata as a matrix type.[3] The additional information required to implement this command are the latitude and longitude of regions. Even if a dataset has no coordinate information (i.e., latitude and longitude), a recent geocoding technique facilitates adding this information to the dataset.

The rest of this article is organized as follows. Section 2 explains the basic idea of a spatial lagged variable. Section 3 describes the `moransi` command. Section 4 offers an example using the `moransi` command. Finally, Section 5 presents the conclusions.

2 Moran's I

Based on Cliff and Ord (1970) and Anselin (1995), this section explains details of Moran's I.

2.1 Formula

The formula of Moran's I is given by

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2}, \tag{1}$$

where n is the number of regions, z_i is the value of region i of variable $\boldsymbol{z}$, which is standardized or centered to the mean, and w_{ij} is the ijth element of the row-standardized spatial weight matrix $\boldsymbol{W}$. This formula can be expressed using the matrix form as follows:

$$I = \frac{\boldsymbol{z}^\top \boldsymbol{W} \boldsymbol{z}}{\boldsymbol{z}^\top \boldsymbol{z}}. \tag{2}$$

Note again that $\boldsymbol{W}$ is a row-standardized spatial weight matrix.

Moran's I lies within the range $[-1, 1]$. When values in the variable $\boldsymbol{z}$ are randomly distributed in space, the statistic asymptotically tends to zero. When a positive (negative) value of Moran's I is observed, this indicates that positive (negative) spatial autocorrelation exists across the regions; that is, the regions neighboring a region with high (low) value also show high (low) value.

The hypothesis testing can be conducted under the null hypothesis of the spatial randomization, under which the statistic asymptotically follows a standard normal distribution. The test statistic

[2] Stata 15 offers the `spset` command, which facilitates keeping the consistency.

[3] This method is originally employed by Kondo (2016). There is a disadvantage of calculation inefficiency because the spatial weight matrix is constructed every time. However, automating the construction of the spatial weight matrix provides a more intuitive manipulation for users.

$z(I)$ is computed as follows:[4]

$$z(I) = \frac{I - \mathrm{E}(I)}{\sqrt{\mathrm{Var}(I)}}$$

where $\mathrm{E}(I)$ is the expected value of I and $\mathrm{Var}(I)$ is the variance of I under the spatial randomization, and these terms are calculated as follows:

$$\mathrm{E}(I) = -\frac{1}{n-1} \quad \text{and} \quad \mathrm{Var}(I) = \mathrm{E}(I^2) - [\mathrm{E}(I)]^2.$$

The second term on the right hand side in the variance is given by

$$\mathrm{E}(I^2) = \frac{n\left[(n^2 - 3n + 3)S_1 - nS_2 + 3S_0^2\right] - m_4/m_2^2[(n^2 - n)S_1 - 2nS_2 + 6S_0^2]}{(n-1)(n-2)(n-3)S_0^2}, \tag{3}$$

where m_k is the kth sample moment about the sample mean:

$$\frac{m_4}{m_2^2} = \frac{1/n\sum_{i=1}^{n} z_i^4}{(1/n\sum_{i=1}^{n} z_i^2)^2}, \tag{4}$$

and the terms S_0, S_1, and S_2 denote, respectively,

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}, \quad S_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij} + w_{ji})^2, \quad \text{and} \quad S_2 = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{ij} + \sum_{j=1}^{n} w_{ji}\right)^2. \tag{5}$$

Note that S_0 is equal to n since the the spatial weight matrix is row-standardized.

2.2 Spatial weight matrix

The matrix that expresses spatial structure is called the spatial weight matrix, which plays an important role in spatial analysis. The spatial weight matrix $\boldsymbol{W}$ takes the following formula:

$$\boldsymbol{W} = \begin{pmatrix} 0 & w_{1,2} & w_{1,3} & \cdots & w_{1,n} \\ w_{2,1} & 0 & w_{2,3} & \cdots & w_{2,n} \\ w_{3,1} & w_{3,2} & 0 & \cdots & w_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{n,1} & w_{n,2} & w_{n,3} & \cdots & 0 \end{pmatrix},$$

where diagonal elements take the value of 0, and the sum of each row takes the value of (row-standardization).

Various types of spatial weight matrices are proposed in the literature. The `moransi` com-

[4]The ESRI ArcGIS online manual also explains the mathematical formula: "How Spatial Autocorrelation (Global Moran's I) works" (URL: `https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-spatial-autocorrelation-moran-s-i-spatial-st.htm`).

mand deals with three types of spatial weight matrices.[5] The spatial weight matrix is always row-standardized throughout the paper.

The first case of power functional type is shown below:

$$w_{ij} = \begin{cases} \dfrac{d_{ij}^{-\delta}}{\sum_{j=1}^{n} d_{ij}^{-\delta}}, & \text{if} \quad d_{ij} < d, \quad i \neq j, \quad \delta > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where δ is a distance decay parameter and d is a threshold distance.

The second case of the exponential type of spatial weight matrix is shown as follows:

$$w_{ij} = \begin{cases} \dfrac{\exp(-\delta d_{ij})}{\sum_{j=1}^{n} \exp(-\delta d_{ij})}, & \text{if} \quad d_{ij} < d, \quad i \neq j, \quad \delta > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where δ is the distance decay parameter. The distance decay pattern differs between the two types of spatial weight matrix.

The third case considers a uniform weight as follows:

$$w_{ij} = \begin{cases} \dfrac{I(d_{ij} < d)}{\sum_{j=1}^{n} I(d_{ij} < d)}, & \text{if} \quad d_{ij} < d, \quad i \neq j, \\ 0, & \text{otherwise}. \end{cases} \tag{8}$$

where $I(d_{ij} < d)$ is the indicator function that takes the value of 1 if a bilateral distance between i and j (d_{ij}) is less than the threshold distance d and 0 otherwise.

2.3 Moran scatter plot

Anselin (1995) proposes a Moran scatter plot, which illustrates a spatial autocorrelation for Moran's I. Consider the following regression without constant term:

$$\boldsymbol{W}\boldsymbol{z} = \alpha \boldsymbol{z} + \text{residuals} \tag{9}$$

where residuals indicate that any statistical assumption on error terms is not considered. The OLS estimator of the coefficient α is obtained by

$$\hat{\alpha} = \frac{\boldsymbol{z}^{\top}\boldsymbol{W}\boldsymbol{z}}{\boldsymbol{z}^{\top}\boldsymbol{z}}, \tag{10}$$

which is equal to the formula of the Moran's I in Equation 2. In other words, the Moran scatter

[5] A commonly used spatial weight matrix is constructed by a contiguity matrix, whose element w_{ij} takes a value of 1 if two regions i and j share the same border and 0 otherwise. Note that the `moransi` command is limited to a distance-based spatial weight matrix.

plot illustrates the correlation between $\boldsymbol{W}\boldsymbol{z}$ and $\boldsymbol{z}$

3 Implementation in Stata

3.1 Syntax

`moransi` *varname* [*if*] [*in*] , `lat(`*varname*`) lon(`*varname*`) swm(`*swmtype*`) dist(`#`) dunit(km|mi)`
[`nomatsave dms approx detail`]

3.2 Options

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

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

`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 functional types of spatial weight matrix as follows: `swm(exp #)` and `swm(pow #)`.

`dist(#)` specifies the threshold distance # for the spatial weight matrix. The unit of distance is specified by the `dunit(km|mi)` option. Regions located within the threshold distance # including the own region take a value of 1 in the binary spatial weight matrix or a positive value in the non-binary spatial weight matrix, and 0 otherwise.

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

`nomatsave` does not save the bilateral distance matrix `r(D)` on the memory. The `nomatsave` option is not used in the default setting.

`dms` converts the degrees, minutes and seconds (DMS) format to a decimal. The `dms` option is not used in the default setting.

`approx` uses bilateral distance approximated by the simplified version of the Vincenty formula. The `approx` option is not used in the default setting.

`detail` displays descriptive statistics of distance. The `detail` option is not used in the default setting.

3.3 Output

3.3.1 Stored results

The `moransi` command stores the following results in r-class.

Scalars

`r(I)`	Moran's I statistic	`r(EI)`	expected value of I
`r(seI)`	standard error of I	`r(zI)`	z-value of I
`r(pI)`	p-value of I	`r(N)`	number of observations
`r(td)`	threshold distance	`r(dd)`	distance decay parameter
`r(dist_mean)`	mean of distance	`r(dist_sd)`	standard deviation of distance
`r(dist_min)`	minimum value of distance	`r(dist_max)`	maximum value of distance

Matrices

`r(D)`	lower triangle distance matrix

Macros

`r(cmd)`	`moransi`	`r(varname)`	name of variable
`r(swm)`	type of spatial weight matrix	`r(dist_type)`	exact or approximation

Technical note

When the spatial weight matrix is too large for the computer specs, the `moransi` command may not calculate Moran's I statistic (The computer may freeze). The `nomatsave` option is recommended to save the memory space in the case of a large-sized spatial weight matrix.

4 Example

This section illustrates the use of the `moransi` command in Stata. In this paper, the sample data are taken from Kondo (2015b), who investigates the spatial autocorrelation of municipal unemployment rates in Japan.

Figure 1 illustrates geographical distribution in unemployments rates using the dataset of Kondo (2015b).[6] The municipalities are categorized into seven quantile levels. It can be seen that municipalities with high unemployment rates have neighbors with similar characteristic, suggesting a positive spatial autocorrelation in municipal unemployment rates.

The basic manipulation is conducted as follows:

```
. use "DTA_ur_1980_2005_all.dta", clear

. moransi ur2005, lon(lon) lat(lat) swm(pow 2) dist(.) dunit(km)
Size of spatial weight matrix:     1745
Calculating bilateral distance...
Calculating spatial weight matrix...

Distance by Vincenty formula (unit: km)

Moran´s I Statistic                                 Number of Obs =     1745
```

Variable	Moran´s I	E(I)	SE(I)	Z(I)	p-value
ur2005	0.49629	-0.00057	0.01019	48.73934	0.00000

```
Null Hypothesis: Spatial Randomization
```

[6] Stata 14 or lower version can depict maps, like Figure 1, using the `shp2dta` that command converts shape files to a DTA file (Crow, 2015) and the `spmap` command that illustrates data on map (Pisati, 2008) in . Stata 15 provides corresponding official commands `spshape2dta` and `grmap`.

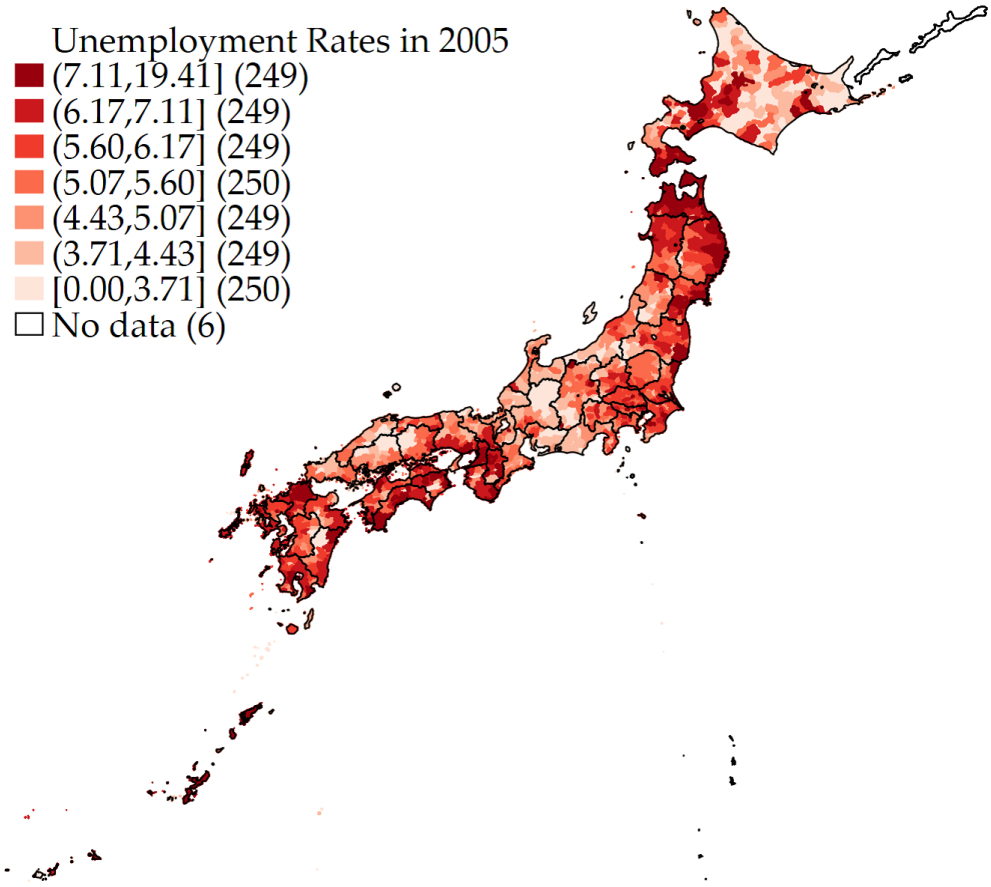

Figure 1: Municipal unemployment rates in 2005

Note: Created by the author using the dataset of Kondo (2015b). Original data source of municipal unemployment rates is Population Census (Statistical Bureau, Ministry of Internal Affairs and Communications of Japan) .

The `moransi` command displays a summary result of the Moran's I. In this case, the Moran's I is 0.496 and statistically significant at the 1% level. The Web Supplement offers the comparison program between the `spatgsa` command developed by Pisati (2001), the `splagvar` command developed by Jeanty (2010), and the `moransi` command. These three commands show the same calculation results.

Moran scatter plot can be depicted using the `scatter` command and `spgen` command developed by Kondo (2015a). The following command is one example, which generates Figure 2:

(*Continued on next page*)

```
. use "DTA_ur_1980_2005_all.dta", clear

. egen std_ur2005 = std(ur2005)

. spgen std_ur2005, lon(lon) lat(lat) swm(pow 2) dunit(km) dist(.)
Size of spatial weight matrix:     1745
Calculating bilateral distance...
Calculating spatial weight matrix...

Distance by Vincenty formula (unit: km)
splag1_std_ur2005_p is generated in the dataset.

. rename splag1_std_ur2005_p w_std_ur2005

. twoway (scatter w_std_ur2005 std_ur2005, ms(oh) yaxis(1 2) xaxis(1 2)) /*
>         */ (lfit w_std_ur2005 std_ur2005, lw(medthick) estopts(nocons)), /*
>         */ ytitle("W.Standardized Unemployment Rates", tstyle(size(large)) axis(1)) /*
>         */ xtitle("Standardized Unemployment Rates", tstyle(size(large)) height(6) axis(1)) /*
>         */ ytitle("", axis(2)) /*
>         */ xtitle("", axis(2)) /*
>         */ ylabel(-2(2)6, ang(h) labsize(large) format(%2.0f) nogrid axis(1)) /*
>         */ xlabel(-4(2)12, labsize(large) format(%2.0f) nogrid axis(1)) /*
>         */ ylabel(-2(2)6, ang(h) labsize(large) format(%2.0f) nogrid axis(2)) /*
>         */ xlabel(-4(2)12, labsize(large) format(%2.0f) nogrid axis(2)) /*
>         */ ysize(3) xsize(4) /*
>         */ yline(0, lwidth(thin) lcolor(gray) lpattern(dash)) /*
>         */ xline(0, lwidth(thin) lcolor(gray) lpattern(dash)) /*
>         */ legend(off) /*
>         */ graphregion(color(white) fcolor(white))

. graph export "FIG_map_ur2005.png", as(png) width(1600) height(1200) replace
(file FIG_map_ur2005.png written in PNG format)
```

5 Concluding remarks

This paper has introduced the new command `moransi`, which easily computes Moran's I in Stata to test for global spatial autocorrelation. An advantage of the `moransi` command is that although the computational efficiency is partly lost, it provides an easy and intuitive manipulation for researchers.

References

Anselin, L. 1995. Local indicators of spatial association—LISA. *Geographical Analysis* 27(2): 93–115.

Cliff, A. D., and J. K. Ord. 1970. Spatial autocorrelation: a review of existing and new measures with applications. *Economic Geography* 46: 269–292.

Crow, K. 2015. SHP2DTA: Stata module to converts shape boundary files to Stata datasets. Statistical Software Components S456718, Boston College.
(URL: `https://ideas.repec.org/c/boc/bocode/s456718.html`).

Getis, A., and J. K. Ord. 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24(3): 189–206.

Jeanty, P. W. 2010. SPLAGVAR: Stata module to generate spatially lagged variables, construct

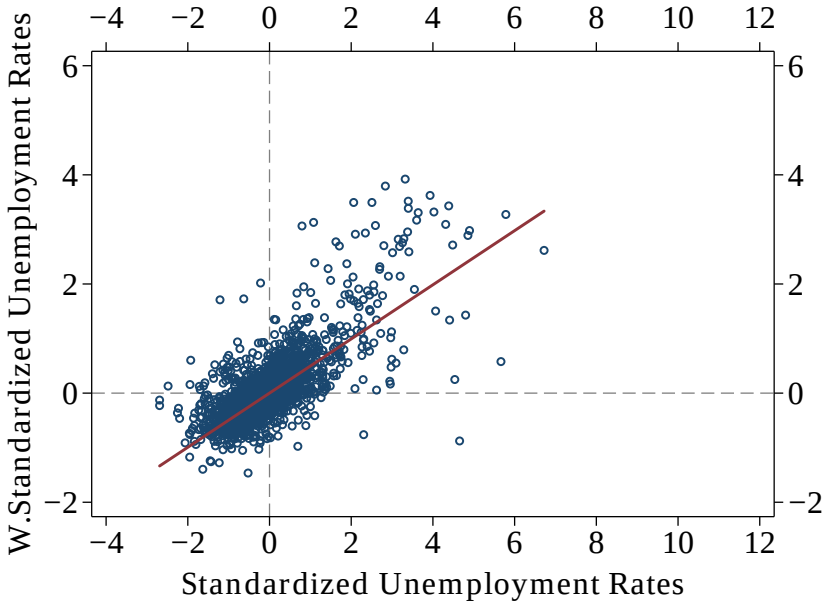

Figure 2: Moran Scatterplot of municipal underemployment rates

Note: Created by the author using the dataset of Kondo (2015b). Original data source of municipal unemployment rates is Population Census (Statistical Bureau, Ministry of Internal Affairs and Communications of Japan) .

the Moran Scatter plot, and calculate Moran's I statistics. Statistical Software Components S457112, Boston College.

(URL: `http://ideas.repec.org/c/boc/bocode/s457112.html`).

Kondo, K. 2015a. SPGEN: Stata module to generate spatially lagged variables. Statistical Software Components S458105, Boston College.

(URL: `http://econpapers.repec.org/software/bocbocode/S458105.htm`).

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

———. 2016. Hot and cold spot analysis using Stata. *Stata Journal* 16(3): 613–631.

Moran, P. A. P. 1950. Notes on continuous stochastic phenomena. *Biometrika* 37(1/2): 17–23.

Ord, J. K., and A. Getis. 1995. Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis* 27(4): 286–306.

Pisati, M. 2001. Tools for spatial data analysis. *Stata Technical Bulletin* 60: 21–37.

———. 2008. SPMAP: Stata module to visualize spatial data. Statistical Software Components S456812, Boston College.

(URL: `https://ideas.repec.org/c/boc/bocode/s456812.html`).