Testing for spatial autocorrelation in Stata*

Keisuke Kondo†

Version: March 23, 2025
(moransi: version 1.31)

Abstract

This paper introduces the new Stata command moransi, which allows users to easily compute global and local Moran's I statistics in Stata. The fundamental feature of the moransi command is that the spatial weight matrix is constructed internally within a sequence of the program code. The additional information required in the dataset to implement this command are the latitude and longitude of regions. This paper presents two applied examples of the moransi command to deepen the understanding of global and local spatial autocorrelation.

Keywords: moransi, Moran's I, Global indicators of spatial association, local indicators of spatial association, Spatial lag

1 Introduction

There is an increasing demand for spatial analysis using Stata among researchers, as the Stata version 15 newly released the Sp estimation commands in 2017 (StataCorp, 2023). The increasing availability of geographically disaggregated data and shapefile is making spatial analysis easier.

The objective of this paper is to introduce the novel Stata command, moransi, which facilitates the calculation of global and local Moran's I statistics (Moran, 1950; Anselin, 1995; Kondo, 2018). In the literature on spatial statistical analysis, spatial autocorrelation constitutes a pivotal concept, which can be further categorized into two distinct classes. Primarily, global spatial autocorrelation quantifies the degree to which regions exhibit interdependence. The global Moran's I statistic is a principal statistical approach to test for spatial autocorrelation. Secondarily, the local Moran's I statistic is used to identify local spatial clusters in terms of local similarities to neighboring regions.[1]

*I thank Jesús Lucindo for his helpful comments and suggestions. This is a research outcome undertaken at the Research Institute of Economy, Trade and Industry and the Research Institute for Economics and Business Administration at Kobe University. This study was supported by JSPS KAKENHI Grant Numbers 21K01497. Sample datasets used in this article are available online (URL: https://github.com/keisukekondokk/moransi).

†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]The Getis–Ord $G_i^*(d)$ is also used to detect hot and cold spots as spatial outliers (Getis and Ord, 1992; Ord and Getis, 1995; Anselin, 1995).Kondo (2016) provides the Stata command, getisord, which calculates Getis–Ord $G_i^*(d)$ statistic.

The concept of local indicator of spatial association (LISA), which was initially proposed by Anselin (1995), establishes a coherent link between global and local Moran's I statistics. To illustrate, the global Moran's I statistic can be interpreted as the average of local Moran's I statistics for individual regions. As discussed later in two applied examples, this integration enables a coherent interpretation of local Moran's I statistics through the lens of the Moran scatterplot, such as high-high and low-low spatial clusters.

The `moransi` command facilitates the calculation of global and local Moran's I in Stata. Stata users may frequently encounter challenges when constructing a spatial weight matrix. For example, there are some helpful packages for Moran's I in Stata. Pisati (2001) provides the `spatgsa` command. Additionally, Jeanty (2010) offers the `splagvar` command. However, it should be noted that the use of these packages requires the exogenous inclusion of the spatial weight matrix.[2]

The fundamental feature of the `moransi` command is that the spatial weight matrix is constructed internally within a sequence of the program code. This method was originally employed by Kondo (2016). Although a disadvantage of this method is its computational inefficiency, which stems from the fact that the spatial weight matrix is constructed every time, automating the construction of the spatial weight matrix provides a more intuitive manipulability for users.

In order to execute the `moransi` command, it is necessary to provide the latitude and longitude in the dataset. Even in the absence of an adequate shapefile of the study area, the `moransi` command enables researchers to implement spatial analysis using the latitude and longitude of the regions. In the event that a dataset is missing coordinate information, a recent geocoding technique can be employed to add this information.

This paper presents two interesting and compelling examples of the `moransi` command. First, the `moransi` command can identify unemployment clusters in Japan using municipal unemployment rates. Second, the `moransi` command can capture residential clusters in the Tokyo metropolitan area using Grid Square Statistics from the population census in Japan. The present study utilizes elementary yet efficacious spatial analyses to ascertain that residential clusters are formed along urban railroads when combined with data from the rail network.

The rest of this paper is organized as follows. Section 2 explains details of spatial weight matrix and Moran's I statistic. Section 3 describes the `moransi` command. Section 4 offers two applied examples using the `moransi` command. Finally, Section 5 presents the conclusions.

2 Spatial autocorrelation

Based on Cliff and Ord (1970), Anselin (1995), and Sokal et al. (1998), this section explains details of global and local Moran's I statistics.

[2]Stata version 15 or later offers the `spset` command, which facilitates keeping the consistency of the spatial weight matrix. On Stata version 14 or earlier, matching regional IDs between data and spatial weight matrix was not easy since regions with missing values were dropped in some situations.

2.1 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 1 (i.e., row-standardization). The spatial weight matrix is always row-standardized throughout the paper.

The spatial weight matrix is generally constructed from the contiguity or distance matrix. The interregional contiguity relationship is generally constructed from shapefiles. The distance matrix is easily obtained from longitude and latitude information. If transport network data is available, the spatial weight matrix can also be constructed using road distance or travel time between regions. In circumstances where adequate shapefiles are not available, this paper elects to utilize the distance matrix.[3]

Various types of spatial weight matrices are proposed in the literature. The `moransi` command deals with four types of spatial weight matrices.

The first case of power functional type is shown below:

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

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:

$$w_{ij} = \begin{cases} \dfrac{\exp(-\delta d_{ij})}{\sum_{j=1}^{n} \exp(-\delta d_{ij})}, & \text{if} \quad d_{ij} < d, \quad i \neq j, \quad \delta > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where δ is the distance decay parameter.

The third case considers a uniform weight for regions located within d km as follows:

$$w_{ij} = \begin{cases} \dfrac{I(d_{ij} < d)}{\sum_{j=1}^{n} I(d_{ij} < d)}, & \text{if} \quad d_{ij} < d, \quad i \neq j, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

[3] 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.

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 shorter than the threshold distance d and 0 otherwise.

The fourth case considers the k-nearest neighbor weight as follows:

$$w_{ij} = \begin{cases} \dfrac{I(d_{ij} \leq d_{ij,(k)})}{\sum_{j=1}^{n} I(d_{ij} \leq d_{ij,(k)})}, & \text{if} \quad i \neq j, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $I(d_{ij} \leq d_{ij,(k)})$ is the indicator function that takes the value of 1 if a bilateral distance between i and j, d_{ij}, is shorter than or equal to the distance of the kth nearest neighbor $d_{ij,(k)}$ and 0 otherwise.

2.2 Spatial weight matrix with weight variable

The degree of interregional dependence may vary according to economic relationships between regions even if a geographical distance is identical. It is possible to consider economic distance using an additional weight variable in the spatial weight matrix.

In line with the gravity equation (Anderson, 1979), values in origin and destination regions i and j are incorporated into the spatial weight matrix. For example, Molho (1995) uses employment size in destination region j as a weight variable when constructing the spatial weight matrix. It is noteworthy that the value of the origin region i is offset by the row-standardization.

In accordance with Kondo (2015b), the `moransi` command facilitates the incorporation of a weight variable v into the spatial weight matrix. The power functional form of the spatial weight matrix in Equation (1) is extended as follows:

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

where v_j is a value of weight variable v in region j.

Second, the exponential type of spatial weight matrix in Equation (2) is extended as follows:

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

Third, the binary type of spatial weight matrix in Equation (3) is extended as follows:

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

Fourth, the k-nearest neighbor type of spatial weight matrix in Equation (4) is extended as follows:

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

Note that, in the context of the spatial econometric model, elements of the spatial weight matrix are assumed to be non-stochastic and exogenous (Anselin, 1988, 2006). The exogeneity assumption for the spatial weight matrix might be violated if a weight variable v_j is endogenous.

2.3 Global Moran's I

The formula of Moran's I is given by

$$I = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2}, \tag{5}$$

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:

$$I = \frac{\boldsymbol{z}^\top \boldsymbol{W} \boldsymbol{z}}{\boldsymbol{z}^\top \boldsymbol{z}}. \tag{6}$$

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

Moran's I lies within the range $[-1, 1]$.[4] 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 for spatial autocorrelation can be conducted under the null hypothesis of the spatial randomization, under which the statistic asymptotically follows a standard normal distribution. The test statistic $z(I)$ is computed as follows:

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

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 first term on the right hand side in the variance is given by

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

[4]This is not guaranteed when the spatial weight matrix is not row-standardized.

where m_h is the hth sample moment about the sample mean:

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

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

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

Note that S_0 is equal to n since the the spatial weight matrix is row-standardized. See Cliff and Ord (1970) for further details.

2.4 Local Moran's I

The formula of local Moran's I_i is given by

$$I_i = z_i \sum_{j=1}^{n} w_{ij} z_j, \tag{7}$$

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 $\boldsymbol{W}$.

The term on the right hand side of Equation (7) is the product of the standardized value of region i, z_i, and its spatial lag. Note that if the both are negative, the product takes a positive value, which complicates the direct interpretation of local Moran's I. However, the statistical significance of local Moran's I statistics provides a powerful interpretation through the lens of the Moran scatterplot, which integrates the global and local spatial autocorrelation.

As a class of the LISA, the sum of local Moran's I_i is connected to the global Moran's I, as follows:

$$I = \frac{1}{\sum_{i=1}^{n} z_i^2} \sum_{i=1}^{n} I_i.$$

The hypothesis testing for local spatial autocorrelation can also be conducted under the null hypothesis of the spatial randomization (Anselin, 1995).

The test statistic of local Moran's I_i, $z(I_i)$, is computed as follows:

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

where $\mathrm{E}(I_i)$ is the expected value of I_i of region i and $\mathrm{Var}(I_i)$ is the variance of I_i of region i under the spatial randomization.

The expected value of local Moran's I_i, $\mathrm{E}(I_i)$, under the spatial randomization is as follows:

$$\mathrm{E}(I_i) = -\frac{\sum_{j=1}^{n} w_{ij}}{n-1},$$

where $\sum_{j=1}^{n} w_{ij} = 1$ when the spatial weight matrix is row-standardized.

The variance of local Moran's I_i, $\mathrm{Var}(I_i)$, under the spatial randomization is as follows:

$$\mathrm{Var}(I_i) = \mathrm{E}(I_i^2) - [\mathrm{E}(I_i)]^2.$$

where

$$\mathrm{E}(I_i^2) = \frac{\sum_{i=1}^{n} w_{ij}^2 (n - m_4/m_2^2)}{n-1} + \frac{\sum_{k=1}^{n}\sum_{h=1}^{n} w_{ik} w_{ih} (2m_4/m_2^2 - n)}{(n-1)(n-2)}.$$

See Anselin (1995) and Sokal et al. (1998) for further details.

2.5 Moran scatter plot

Anselin (1995) proposes a Moran scatterplot, which illustrates a spatial autocorrelation in terms of Moran's I. Consider the following regression without constant term:

$$\boldsymbol{W}\boldsymbol{z} = \alpha \boldsymbol{z} + \text{residuals} \tag{8}$$

where $\boldsymbol{W}\boldsymbol{z}$ is called the spatial lag of the variable $\boldsymbol{z}$, and residuals indicate that any statistical assumption on error terms is not considered.

The OLS estimator of the coefficient α is obtained by

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

which is equal to the formula of the Moran's I in Equation (6). Therefore, the Moran scatter plot is a graphical representation of the correlation between $\boldsymbol{W}\boldsymbol{z}$ and $\boldsymbol{z}$. In the case where the variable is not standardized or centered around the mean, it is necessary to include the constant term, thereby centering the variable around its mean.

3 Implementation in Stata

3.1 Syntax

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

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 types of spatial weight matrix must be specified: `bin` (binary), `knn` (k-nearest neighbor), `exp` (exponential), or `pow` (power). The parameter k must be specified for the k-nearest neighbor weights as follows: `swm(knn #)`. 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, with the unit of distance specified by the `dunit(km|mi)` option.

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

`largesize` is designed to enhance the efficiency of calculation for the large-sized spatial weight matrices. The `largesize` option is not used in the default setting.

`approx` uses bilateral distance approximated by the simplified version of the Vincenty formula (Vincenty, 1975). The `approx` option is not used in the default setting.

`wvar(`*varname*`)` specifies a weight variable for the spatial weight matrix. Weight variable 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.

`detail` displays descriptive statistics of distance for lower triangular elements of the distance matrix. The `detail` option is not used in the default setting.

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

`graph` draws a Moran scatterplot. The `graph` option is not used in the default setting.

`replace` is used to overwrite the existing output variables in the dataset. The `replace` 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)`	parameter δ of distance decay or k of knn
`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 triangular distance matrix	`r(W)`	spatial weight matrix
Macros			
`r(cmd)`	`moransi`	`r(varname)`	name of variable
`r(swm)`	type of spatial weight matrix	`r(dunit)`	unit of distance
`r(dist_type)`	exact or approximation	`r(wvar)`	name of weight variable

❑ **Technical note**

If the calculation speed is too slow due to the large spatial weight matrices, it is recommended to use the `largesize` and `approx` options at the same time. The duration of the calculation process is contingent upon the CPU performance.

When the spatial weight matrix is too large for the computer specs (e.g., the memory size is small), the `moransi` command may not calculate Moran's I statistic (Stata program or computer may freeze). For example, about $51,842 \times 51,842$ spatial weight matrix uses about 20 GB of memory space during the calculation process. In addition, the `nomatsave` option is recommended to release the memory space after the calculation.

❏

4 Example

This section illustrates the use of the `moransi` command in Stata. Two applied examples are provided below.

4.1 Municipal Unemployment Rates in Japan

This subsection investigates the spatial distribution of unemployment in Japan. The sample data are taken from Kondo (2015b), who investigates the spatial autocorrelation of municipal unemployment rates and unemployment clusters in Japan. Municipal unemployment rates used in Kondo (2015b) are based on the 1980–2005 Population Censuses in Japan.

Figure 1 illustrates geographical distribution of unemployment rates in 2005.[5] The municipalities are categorized into nine 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.

To assess the spatial autocorrelation, the fundamental procedure of the `moransi` command is outlined below:

(*Continued on next page*)

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

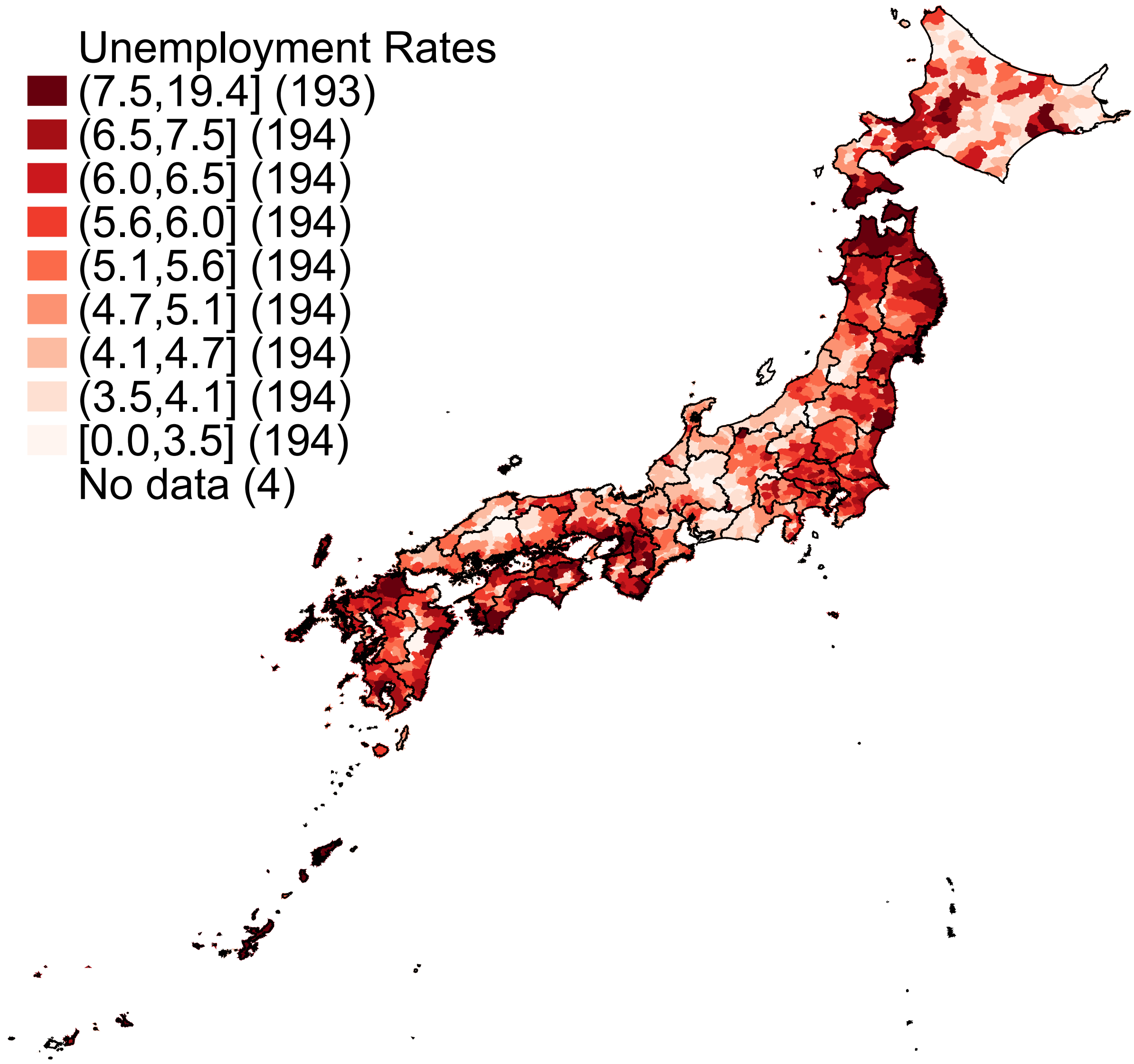

Figure 1: Municipal unemployment rates in 2005

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) .

```
. moransi std_ur2005, lon(lon) lat(lat) swm(pow 2) dist(.) dunit(km) det graph replace

Size of spatial weight matrix: 1745 * 1745
Calculating bilateral distance...

Completed:  10%
Completed:  20%
Completed:  30%
Completed:  40%
Completed:  50%
Completed:  60%
Completed:  70%
Completed:  80%
Completed:  90%
Completed: 100%

Calculating Moran´s I Statistics...

Distance by Vincenty formula (unit: km)
```

	Obs.	Mean	S.D.	Min.	Max.
Distance	1523385	603.951	432.634	0.854	2961.185

```
Distance threshold (unit: km):     .

Summary of Global Moran´s I Statistic                   Number of Obs. =      1745
```

Variable	Moran´s I	E(I)	SE(I)	Z(I)	p-value
std_ur2005	0.49629	-0.00057	0.01019	48.73934	0.00000

```
Null Hypothesis: Spatial Randomization

Summary of Local Moran´s I Statistics                   Number of Obs. =      1745
```

std_ur2005	Obs.	p < 0.10	p < 0.05	p < 0.01
1: High-High	526	205	188	156
2: High-Low	256	5	3	3
3: Low-High	157	20	14	9
4: Low-Low	806	245	206	130

```
Null Hypothesis: Spatial Randomization

splag_std_ur2005_p was generated in the dataset.
lmoran_i_std_ur2005_p was generated in the dataset.
lmoran_e_std_ur2005_p was generated in the dataset.
lmoran_v_std_ur2005_p was generated in the dataset.
lmoran_z_std_ur2005_p was generated in the dataset.
lmoran_p_std_ur2005_p was generated in the dataset.
lmoran_cat_std_ur2005_p was generated in the dataset.

. graph export "fig/FIG_moran_ur2005.svg", replace
(file fig/FIG_moran_ur2005.svg written in SVG format)
```

(*Continued on next page*)

The `moransi` command displays a summary result of the global and local Moran's I. In this case, the Moran's I is 0.496 and statistically significant at the 1 % level. The spatial weight matrix is based on the power functional form with the distance decay parameter $\delta = 2$. Additionally, the `detail` option displays descriptive statistics of distance for lower triangular elements of the distance matrix.[6]

The `moransi` command includes the `graph` option, which automatically draws a Moran scatterplot. However, since the spatial lag of the variable is stored in the dataset, users also can make a Moran scatterplot manually using the outcome variables as follows:[7]

```
. twoway (scatter splag_std_ur2005_p std_ur2005, ms(oh) yaxis(1 2) xaxis(1 2)) ///
>        (lfit splag_std_ur2005_p 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)8, 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)8, 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_moran_ur2005.svg", replace
(file FIG_moran_ur2005.svg written in SVG format)
```

Figure 2 shows a Moran scatter plot, which is made in combination with the standard Stata command `twoway scatter`. The line through the origin depicted in Figure 2 indicates the regression line in Equation (8), which is equal to the Moran's I statistic.

The `moransi` command shows the summary results of local Moran's I statistics based on the Moran scatterplot. There are 526 municipalities in the high-high category, and 205 municipalities are statistically significant at the 10 % level and detected as high-high spatial clusters. In turn, there are 806 municipalities in the low-low category, and 245 municipalities are statistically significant at the 10 % level and classified as low-low spatial clusters.

Figure 3 shows the statistical significance of the local Moran's I of municipal unemployment rates. This map can be integrated with Moran scatterplot in Figure 2, which provides a coherent interpretation between the global and local spatial autocorrelation.

Figure 4 shows spatial clusters of municipal unemployment rates, based on four categories of the Moran scatterplot in Figure 2. As discussed before, 205 municipalities of high-high unemployment clusters and 245 municipalities of low-low unemployment clusters are visualized on the map.

[6]The supplementary material for replication includes a 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.

[7]The `spgen` command also generates the spatially lagged variables (Kondo, 2015a).

Figure 2: Moran Scatterplot of municipal underemployment 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) .

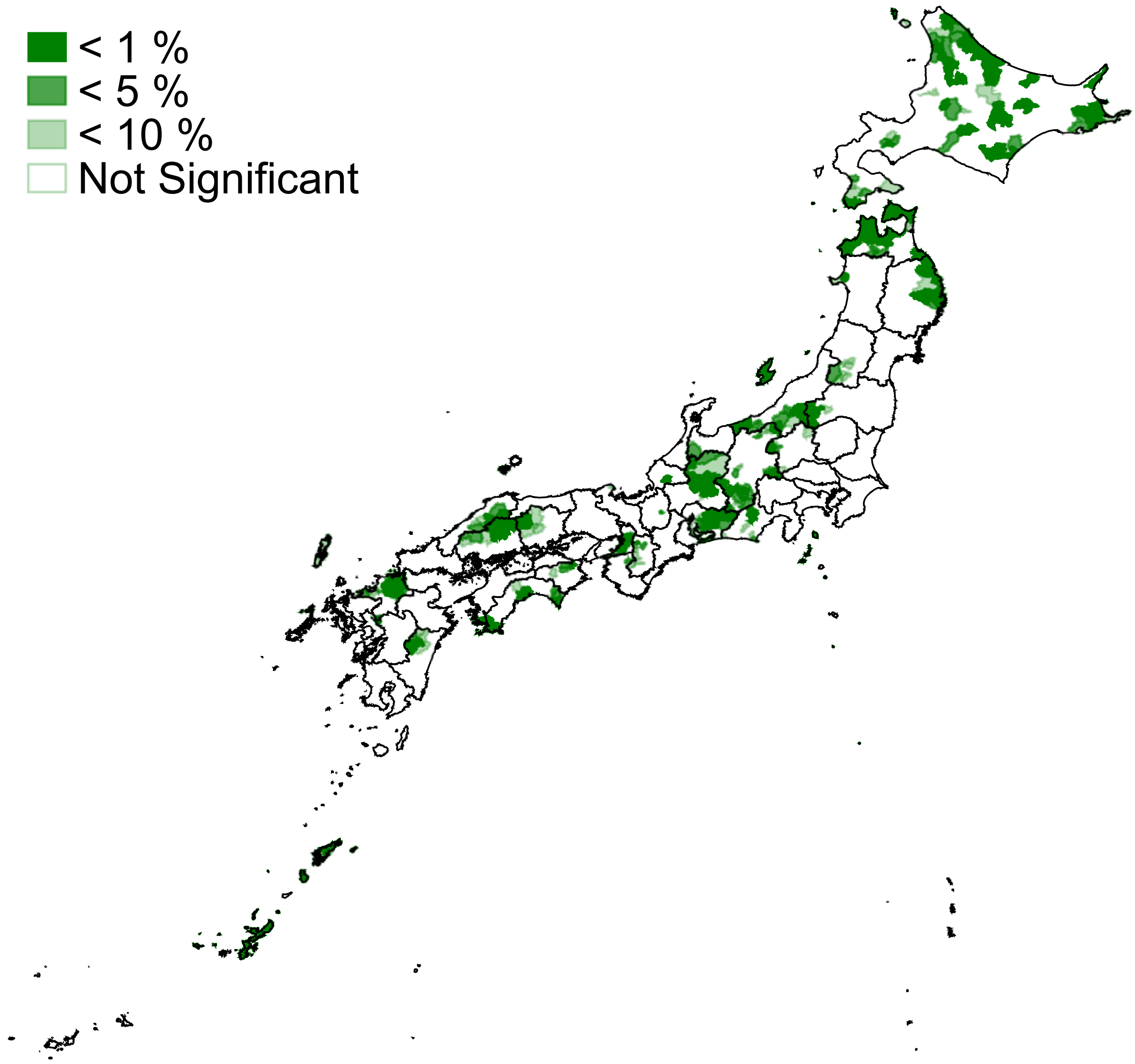

Figure 3: Statistical Significance for Local Moran's I of Municipal Unemployment Rates in 2005

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).

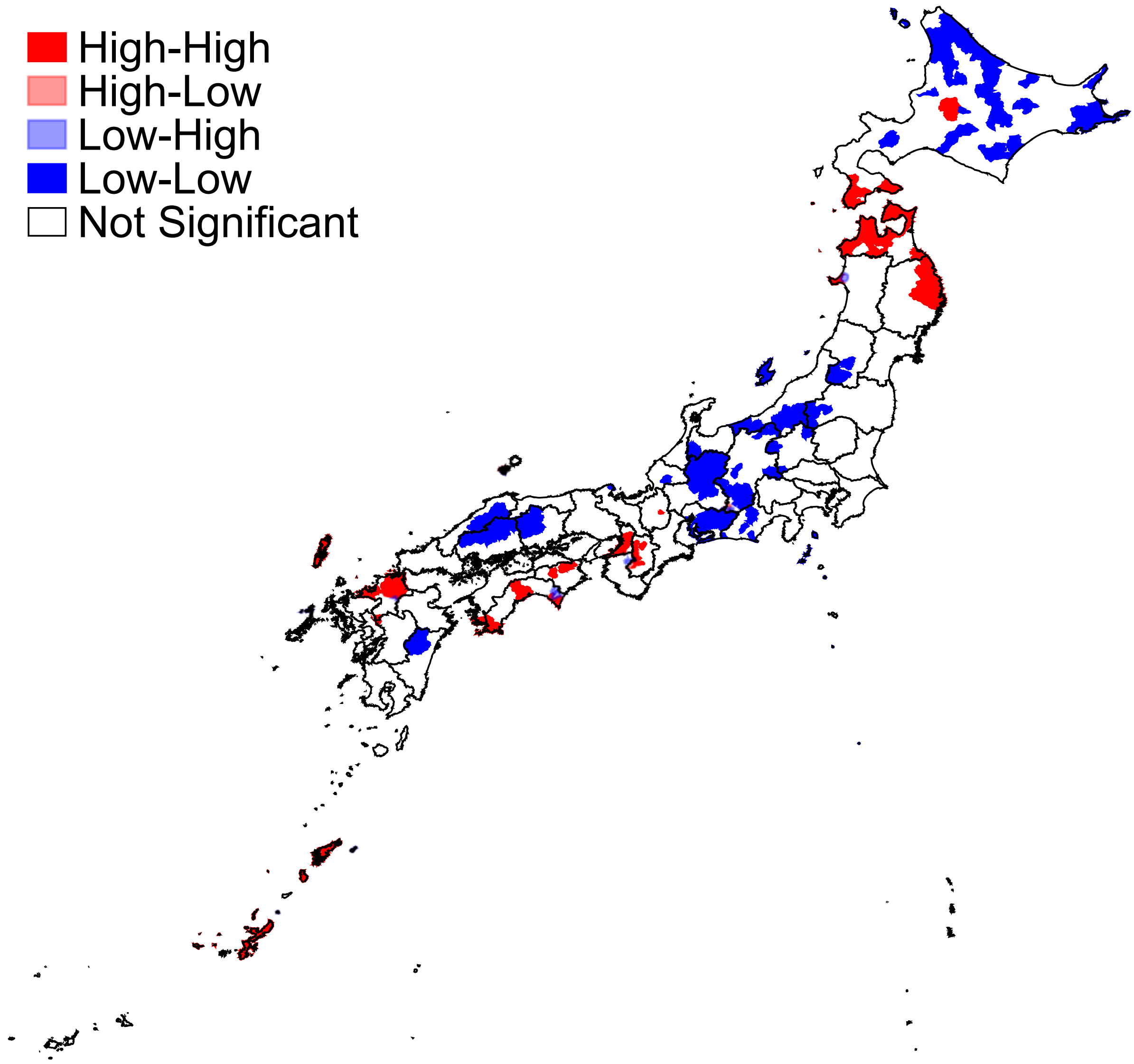

Figure 4: Moran Cluster Map of Municipal Unemployment Rates in 2005

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). Municipalities with statistical significance at the 10 % level are shown.

4.2 Population in Tokyo Metropolitan Area

This subsection investigates a population distribution using the Grid Square Statistics of 2015 Population Census in Japan Ministry of Internal Affairs and Communications (2025a,b). This paper focuses on the residential population in the Tokyo metropolitan area (i.e., Saitama, Chiba, Tokyo, and Kanagawa prefectures). There are several previous studies in the field of geography that discuss population distribution in the corresponding area in terms of the explanatory spatial data analysis (Koizumi, 2010; Monzur, 2017).

As illustrated in Figure 5, the geographical distribution of the residential population in the Tokyo metropolitan area in 2015 is categorized into nine quantile levels. The population is concentrated within 23 wards of Tokyo, and the mesh grid surrounding the central business district near Tokyo Station and the Imperial Palace exhibits areas with few residents. Residential areas extend in a radial pattern toward suburban areas.

The global and local Moran's I statistics were obtained using the `moransi` command, and the spatial weight matrix was based on the power functional form with the distance decay parameter $\delta = 4$. The global Moran's I statistic was found to be 0.84, which is statistically significant at the 1 % level, thereby confirming the presence of spatial similarities in the population distribution. The local Moran's I analysis identified 1,790 residential clusters among 3,106 mesh grids in the high-high category, reaching a 1

Figure 6 presents a Moran scatter plot of residential population, with the regression line corresponding to the Moran's I statistic indicated by the line in the figure. As previously discussed, a clear positive relationship exists between neighboring mesh grids.

Figure 7 shows the statistical significance of the local Moran's I and a radial pattern of residential areas toward suburban areas. The map, which is integrated with the Moran scatterplot in Figure 6, provides an important implication of the spatial analysis.

Figure 8 shows 1,706 residential clusters detected by local Moran's I statistics in the Tokyo metropolitan area, with the rail network depicted according to Ministry of Land, Infrastructure, Transport and Tourism (2025). Notably, high-high residential clusters are found along the rail network, indicating that enhanced transport accessibility increases the desirability of a location for habitation. Additionally, the analysis reveals a clear limit to the distance from the central business districts of the Tokyo metropolitan area.

(*Continued on next page*)

```
. ** Moran´s I
. moransi pop_total_all, lon(lon) lat(lat) swm(pow 4) dist(.) dunit(km) large approx graph re
> place

Size of spatial weight matrix: 13291 * 13291
Calculating bilateral distance...

Completed:  10%
Completed:  20%
Completed:  30%
Completed:  40%
Completed:  50%
Completed:  60%
Completed:  70%
Completed:  80%
Completed:  90%
Completed: 100%

Calculating Moran´s I Statistics...

Distance by simplified version of Vincenty formula (unit: km)

Summary of Global Moran´s I Statistic                         Number of Obs. =     13291
```

Variable	Moran´s I	E(I)	SE(I)	Z(I)	p-value
pop_total_all	0.84325	-0.00008	0.00469	179.63289	0.00000

```
Null Hypothesis: Spatial Randomization

Summary of Local Moran´s I Statistics                         Number of Obs. =     13291
```

pop_total_all	Obs.	p < 0.10	p < 0.05	p < 0.01
1: High-High	3106	2044	1958	1790
2: High-Low	328	0	0	0
3: Low-High	610	1	0	0
4: Low-Low	9247	0	0	0

```
Null Hypothesis: Spatial Randomization

splag_pop_total_all_p was generated in the dataset.
lmoran_i_pop_total_all_p was generated in the dataset.
lmoran_e_pop_total_all_p was generated in the dataset.
lmoran_v_pop_total_all_p was generated in the dataset.
lmoran_z_pop_total_all_p was generated in the dataset.
lmoran_p_pop_total_all_p was generated in the dataset.
lmoran_cat_pop_total_all_p was generated in the dataset.
```

(*Continued on next page*)

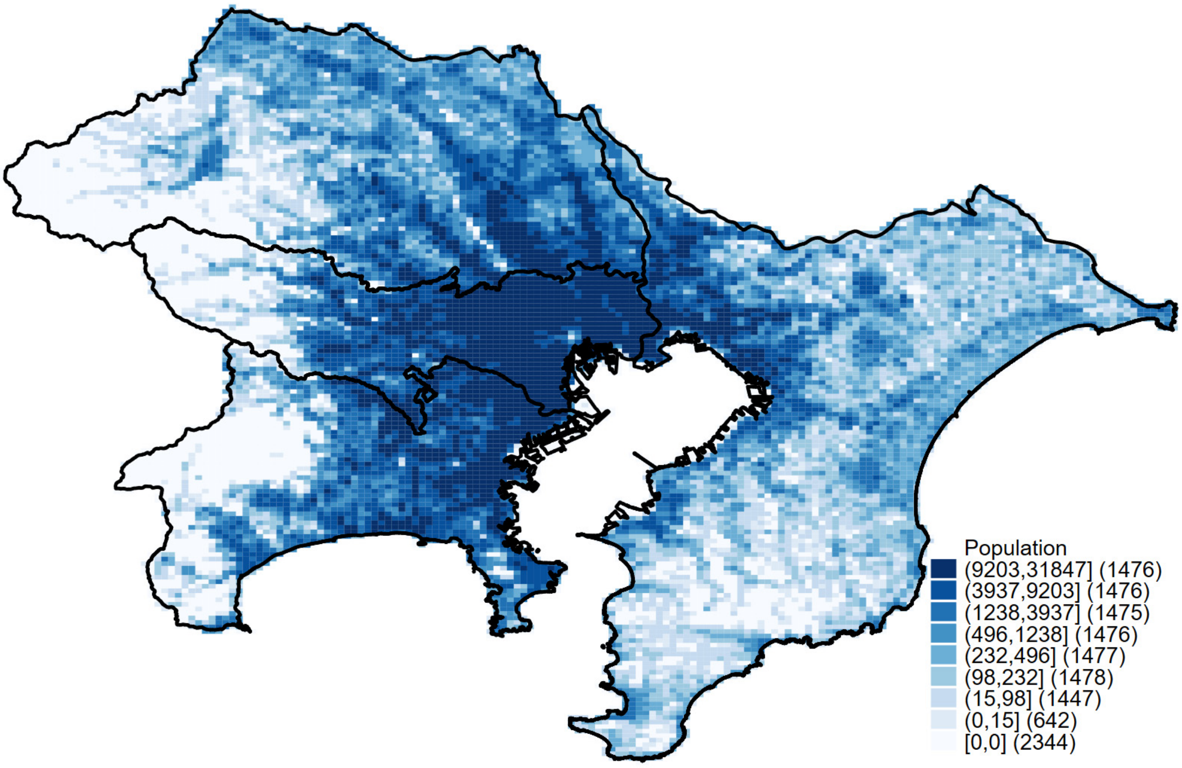

Figure 5: Residential Population Distribution in 2015

Note: Created by the author using the population in 1km-by-1km mesh grid of 2015 Population Census (Statistical Bureau, Ministry of Internal Affairs and Communications of Japan).

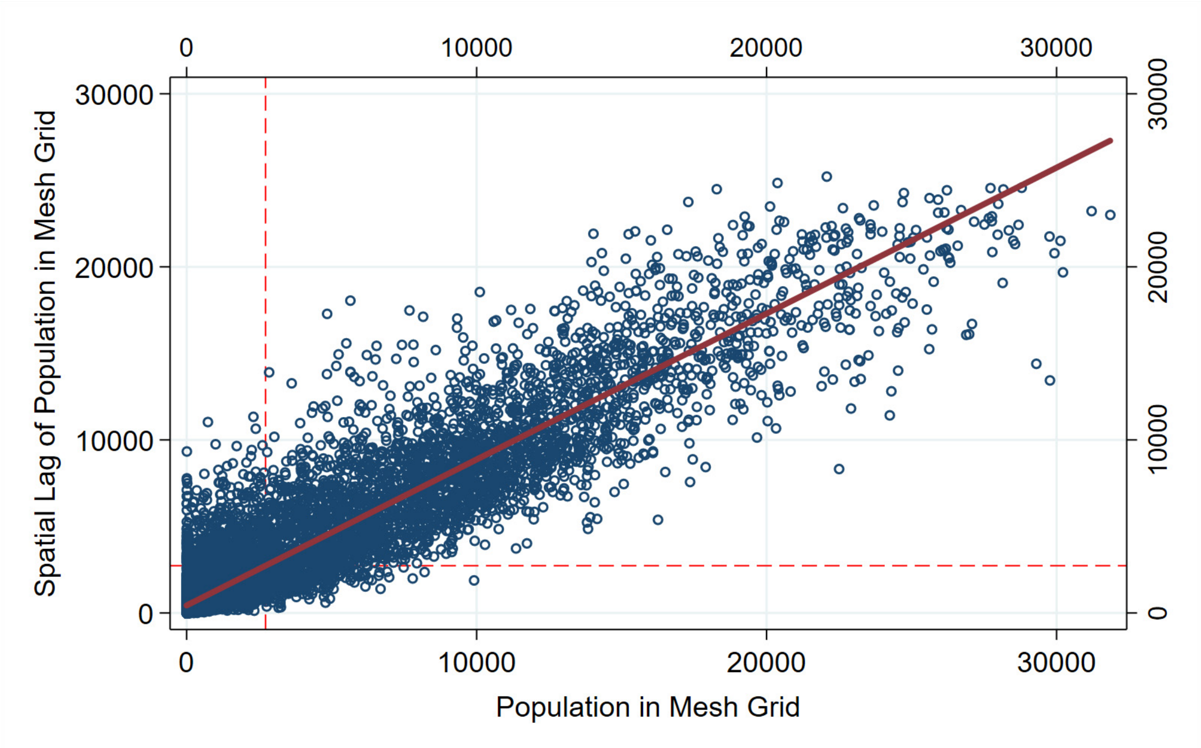

Figure 6: Moran Scatterplot of Population in Tokyo Metropolitan Area

Note: Created by the author using the population in 1km-by-1km mesh grid of 2015 Population Census (Statistical Bureau, Ministry of Internal Affairs and Communications of Japan).

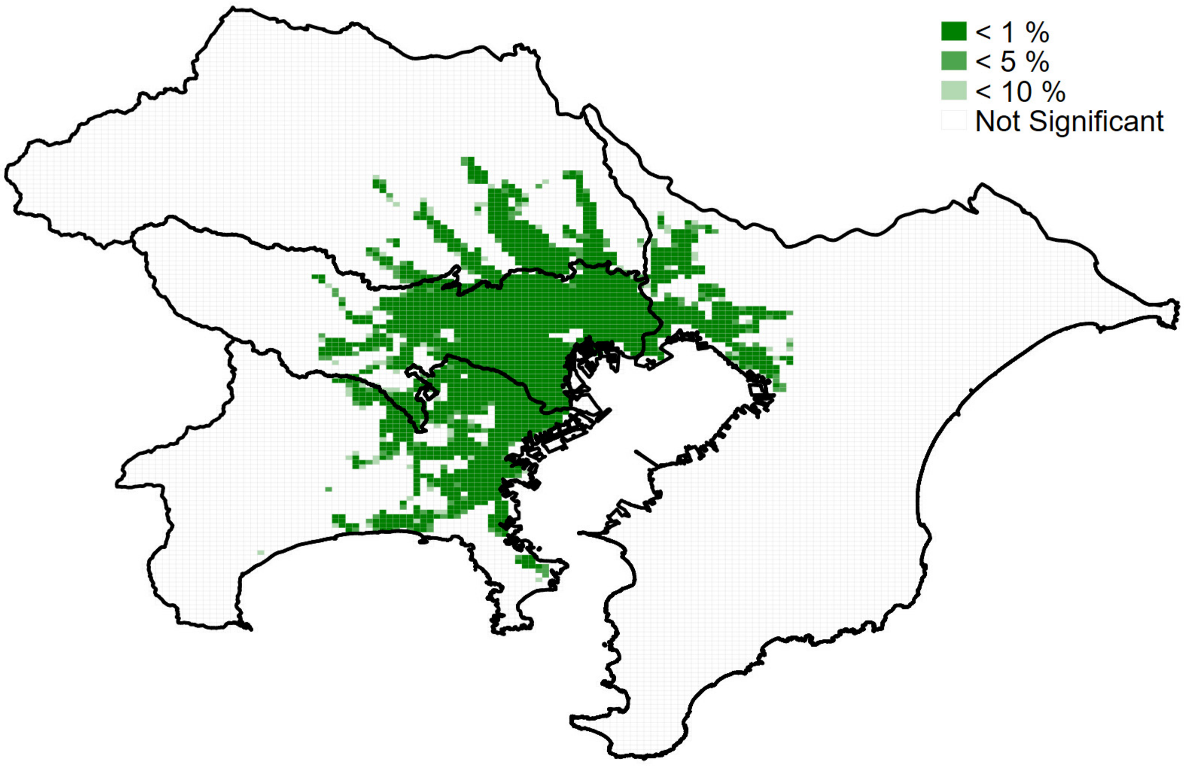

Figure 7: Statistical Significance for Local Moran's I of Population in Tokyo Metropolitan Area
Note: Created by the author using the population in 1km-by-1km mesh grid of 2015 Population Census (Statistical Bureau, Ministry of Internal Affairs and Communications of Japan).

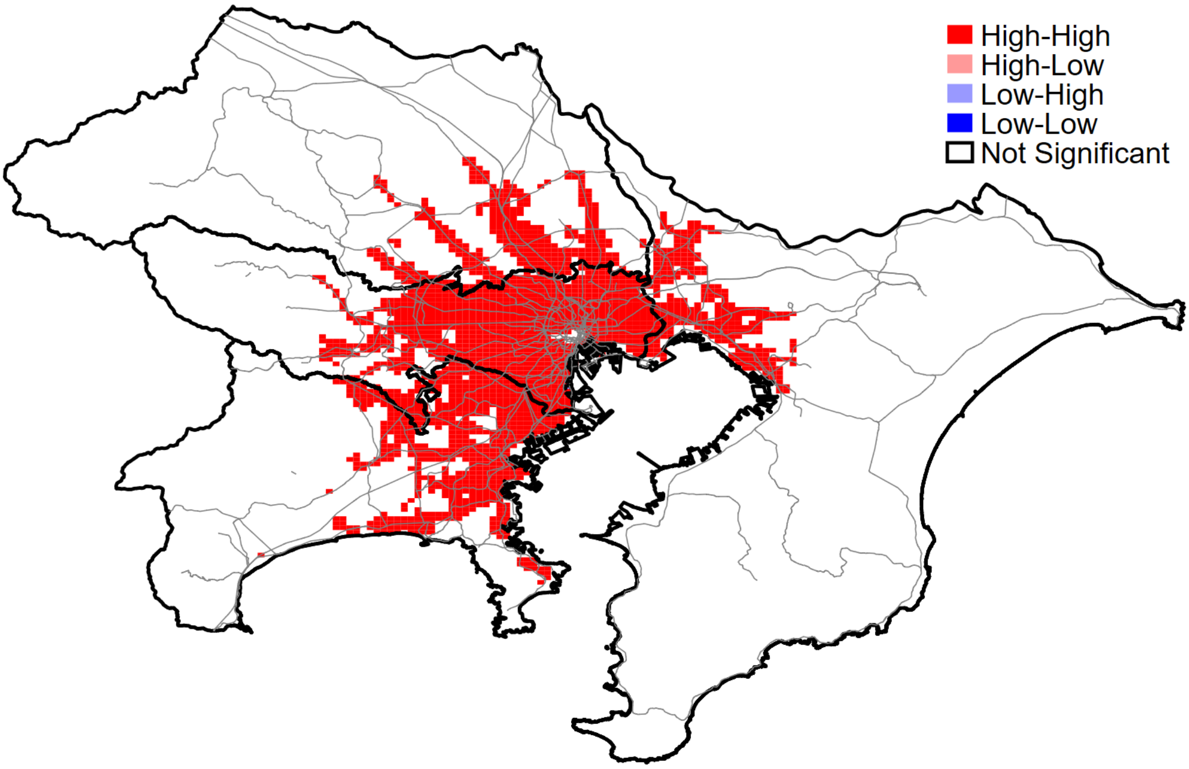

Figure 8: Moran Cluster Map of Population in Tokyo Metropolitan Area

Note: Created by the author using the population in 1km-by-1km mesh grid of 2015 Population Census (Statistical Bureau, Ministry of Internal Affairs and Communications of Japan).

5 Concluding remarks

This paper has introduced the `moransi` command, a novel tool for computing global and local Moran's I in Stata. The `moransi` command offers a more intuitive interface than previous spatial analysis tools, as it does not require the user to construct a spatial weight matrix in advance. While this approach may result in a slight reduction in computational efficiency, it provides a more straightforward and user-friendly interface for Stata users.

To further expand the reader's comprehension of spatial analysis using the `moransi` command, this paper has provided two applied examples of spatial analysis in Japan. First, analysis of the global and local Moran's I revealed spatial clusters of high-high and low-low unemployment rates in Japan. Second, analysis of the global and local Moran's I revealed that residential clusters are formed along with the rail network in Tokyo metropolitan area. The incorporation of additional geographical data facilitates a more profound comprehension of spatial analysis.

It is anticipated that the `moransi` command will contribute to the continued expansion of spatial analysis within the Stata community.

References

Anderson, J. E. 1979. A theoretical foundation for the gravity equation. *American Economic Review* 69(1): 106–116.

Anselin, L. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Press.

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

———. 2006. Spatial econometrics. In *Palgrave Handbook of Econometrics: Econometric Theory*, ed. T. C. Mills and K. Patterson, chap. 26, 901–969. Vol. 1. Basingstoke: Palgrave Macmillan.

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. `https://ideas.repec.org/c/boc/bocode/s456718.html` (Accessed 18 March 2025).

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 the Moran Scatter plot, and calculate Moran's I statistics. Statistical Software Components S457112, Boston College. `http://ideas.repec.org/c/boc/bocode/s457112.html` (Accessed 18 March 2025).

Koizumi, R. 2010. Spatial Patterns of Occupational Structure and Their Changes in Tokyo Metropolitan Area. *Quarterly Journal of Geography* 62(2): 61–70.

Kondo, K. 2015a. SPGEN: Stata module to generate spatially lagged variables. Statistical

Software Components S458105, Boston College.
`http://econpapers.repec.org/software/bocbocode/S458105.htm` (Asscessed 18 March 2025).

———. 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.

———. 2018. MORANSI: Stata module to compute Moran's I. Statistical Software Components S458473, Boston College.
`https://ideas.repec.org/c/boc/bocode/s458473.html` (Asscessed 18 March 2025).

Ministry of Internal Affairs and Communications. 2025a. e-Stat: Statistical and Geographic Information System. `https://www.e-stat.go.jp/gis/statmap-search?type=2`. (Accessed 18 March 2025).

———. 2025b. Grid Square Statistics. `https://www.stat.go.jp/english/data/mesh/index.html`. (Accessed 18 March 2025).

Ministry of Land, Infrastructure, Transport and Tourism. 2025. National Land Information: Railroad. `https://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-N02-2023.html`. (Accessed 18 March 2025).

Molho, I. 1995. Spatial autocorrelation in British unemployment. *Journal of Regional Science* 35(4): 641–658.

Monzur, T. 2017. Spatial Structure of Tokyo Metropolitan Area. *Ritsumeikan Journal of Asia Pacific Studies* 35(1): 44–55.

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.
`https://ideas.repec.org/c/boc/bocode/s456812.html` (Accessed 18 March 2025).

Sokal, R. R., N. L. Oden, and B. A. Thomson. 1998. Local Spatial Autocorrelation in a Biological Model. *Geographical Analysis* 30(4): 331–354.

StataCorp. 2023. *Stata 18 Spatial Autoregressive Models Reference Manual*. College Station, TX: Stata Press.

Vincenty, T. 1975. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review* 23(176): 88–93.