Introduction to spatial econometric analysis: Creating spatially lagged variables in Stata[*]

Keisuke Kondo[†]

Version of this manual: June 17, 2021
(spgen: version 1.40)

Abstract

This article introduces the new Stata command spgen, which computes spatially lagged variables in Stata. The only additionally information required to implement this command are the latitude and longitude of regions. The spgen command facilitates spatial econometric analysis in Stata. In this article, I offer an interesting illustration for spatial econometric analysis using the spgen command.

Keywords: spgen, spatially lagged variable, spatial weight matrix, spatial econometrics

1 Introduction

Spatial econometric analysis has gained attention from researchers and policy-makers, and demand for its use is continuously growing among Stata users. The Sp commands are newly provided on Stata 15 or later and facilitate handling of spatial data and estimation of spatial econometric models.

The key idea of spatial econometrics is expressed as the spatial lag, which originally derives from the concept of time lag in a time series analysis. Unlike time series data, it is often assumed that observations are independent of each other in the cross-section data. The motivation of spatial econometrics starts from the idea that regions are not independent, but interdependent. Therefore, spatial econometrics aims to measure impacts arising from a spatially dependent structure using spatially lagged variables.

The computation of spatially lagged variables requires a spatial weight matrix, which mathematically describes the spatially dependent structures in the matrix. However, researchers might have

[*] I thank Christopher Barrie and Sara Nowak for discovering typos in the manual and the help file. I also thank Shouro Dasgupta for his bug reporting. Moreover, I thank Helge Arends, Stefan Döringm, Yeheskel Hasenfeld, Ariel Karlinsky, Bill McCarthy, Miriam Romero, Elena Nikulina, and Kate Pankowska for their feedback and comments on the program usage. 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://keisukekondokk.github.io/).

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

difficulties constructing the spatial weight matrix before computing spatially lagged variables. Although the spatial weight matrix is often constructed from the shapefile in spatial statistical packages, the shapefile of the corresponding study area is not always available. The newly developed Stata command, `spgen`, solves this issue by facilitating a computing procedure of spatial weight matrix.[1]

The `spgen` command provides the extended function to calculate spatial lag of variables without any complicated procedure. Although Stata 15 or later provides the `spgenerate` command, which also calculates the spatial lag of variable, the spatial weight matrix must be constructed in andvance using the `spmatrix` command.[2] In addition, Jeanty (2010) offers the `splagvar` command that calculates spatially lagged variables, which also requires the spatial weight matrix in advance.

The `spgen` command computes spatially lagged variables using the geographical information of latitude and longitude in the dataset.[3] Although this is also achieved by the Sp commands on Stata 15 or later, the key feature of the `spgen` 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.[4] Furthermore, the `spgen` command offers flexible extensions for constructing spatial weight matrix based on a distance matrix.

The `spgen` command also contributes to the literature on economic geography. For example, population potential proposed by Stewart (1947) and the market potential proposed by Harris (1954) can be easily calculated by the `spgen` command. Thus, it is expected that the `spgen` command advances the empirical literature on economic geography.

The rest of this article is organized as follows. Section 2 explains the basic idea of a spatial lagged variable. Section 3 describes the `spgen` command. Section 4 offers an illustration of spatial econometric analysis with the `spgen` command, and Section 5 presents the conclusions.

2 Spatially lagged variable

2.1 Basic idea of spatial lag

The spatial lag is defined as analogous to the time lag in a time series analysis (LeSage and Pace, 2009). In time series literature, it is common to consider time dependence between times t and $t-1$ by including a lagged variable. Spatial econometrics incorporates spatial lag into cross-sectional analysis to consider spatial dependence between own region and neighboring regions.

Suppose that there are n regions. The two dimensional spatial information is mathematically expressed by the matrix. The matrix that expresses spatial structures is called the spatial weight matrix, which plays an important role in spatial econometric analysis. The spatial weight matrix $\boldsymbol{W}$

[1] The Sp commands on Stata 15 or later also implements spatial analysis without shapefile.

[2] See the `spmat` command for Stata 14 or ealiear (Drukker et al., 2013).

[3] Even if an original dataset has no coordinate information (i.e., latitude and longitude), a recent geocoding technique facilitates adding this information to the dataset.

[4] This method is originally employed by Kondo (2016).

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 w_{ij} is the weight between regions i and j that defines the degree of interdependence, diagonal elements take the value of 0, and the sum of each row takes the value of 1 (i.e., row-standardization).

Let $\boldsymbol{x}$ denote the vector of a variable. Then, this spatially lagged variable can be mathematically expressed by $\boldsymbol{Wx}$ as follows:

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix}, \quad \boldsymbol{Wx} = \begin{pmatrix} \sum_{j=1}^{n} w_{1j}x_j \\ \sum_{j=1}^{n} w_{2j}x_j \\ \sum_{j=1}^{n} w_{3j}x_j \\ \vdots \\ \sum_{j=1}^{n} w_{nj}x_j \end{pmatrix}.$$

Notice that each element of spatially lagged variable $\boldsymbol{Wx}$ expresses the weighted average of the neighboring regions of region i. For this reason, the diagonal elements must be zero to exclude the own regional values in the spatially lagged variable.

Similar to the time lag, the spatial lag can also define higher orders. For example, the second order spatial lag of variable $\boldsymbol{x}$ can be defined as

$$\boldsymbol{W}^2\boldsymbol{x} = \boldsymbol{W} \times (\boldsymbol{Wx}).$$

By iterative procedure, we can easily derive the pth order spatial lag of variable $\boldsymbol{x}$ as $\boldsymbol{W}^p\boldsymbol{x}$.

2.2 Spatial weight matrix

The spatial weight matrix plays an important role in spatial analysis. An important point is whether or not the spatial weight matrix is row-standardized. The row-standardization indicates that the sum of each row is equal to 1. The spatial weight matrix is generally row-standardized in the context of spatial econometrics.

Various types of spatial weight matrices are proposed in the literature. The spgen command deals with four types of spatial weight matrices.[5] The first case of spatial weight matrix is based on the

[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 spgen command is limited to a distance-based spatial weight matrix.

power functional form as follows:

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

where δ is a distance decay parameter and d is a threshold distance. Note that missing value is generated if region i shares the same location point as region j because of the inverse of zero.

Second, the 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 distance decay pattern differs between the two types of spatial weight matrix.

Until now, it has been considered that weights decay with increasing distance. As the third case, it is also possible to consider a uniform weight as follows:

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

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.

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

$$w_{ij} = \begin{cases} \dfrac{I(d_{ij} < d_{ij,(k)})}{\sum_{j=1}^{n} I(d_{ij} < d_{ij,(k)})}, & \text{if} \quad i \neq j, \quad k = 1, 2, \ldots, n-1, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

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

2.3 Spatial weight matrix with weight variable

One might want to combine economic distance as a weight variable with geographical distance in the spatial weight matrix. The degree of interregional dependence will vary according to economic relationships between regions even if a geographical distance is identical.

The `spgen` command allows to incorporate a weight variable v into 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. Note that the value of origin region i is offset by the row-standardization.

The power functional form of spatial weight matrix (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 variable v in region j.

Second, the exponential type of spatial weight matrix (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 (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 (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 the exogeneity assumption for the spatial weight matrix might be violated. In the spatial econometric model, elements of the spatial weight matrix are assumed to be non-stochastic and exogenous (e.g., Anselin, 1988, 2006).

3 Implementation in Stata

3.1 Syntax

```
spgen varlist [if] [in] , lat(varname) lon(varname) swm(swmtype) dist(#) dunit(km|mi)
[ order(#) wvar(varname) rowif(varname) suffix(string) nostd nomatsave dms approx detail
largesize 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 four 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 as a natural number ($k = 1, 2, 3, \dots$) as follows: `swm(knn #)`. The distance decay parameter δ must be specified for the exponential and power functional forms 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.

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

`order(#)` computes #th order spatial lag of *varname*. Only an integer is allowed. The default setting is the 1st order.

`wvar(`*varname*`)` specifies a weight variable for the spatial weight matrix. Weight variable is not used in the default setting.

`rowif(`*varname*`)` an indicator variable that takes the value 1 for observations for which a user calculates spatially lagged variables and 0 otherwise. The `rowif(`*varname*`)` option is not used in the default setting.

`suffix(`*string*`)` appends a suffix to names of output variables. It is helpful when `spgen` is used in `foreach` or `forvalues`. The `suffix(`*string*`)` option is not used in the default setting.

`nostd` uses the spatial weight matrix that is not row-standardized. The `nostd` option is not used in the default setting.

`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.

`largesize` is used for large sized data. When this option is specified, `nomatsave`, `approx`, and `order(1)` options are automatically applied. The `detail` option displays only minimum and maximum distances. The `largesize` 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 Outcome variables

The `spgen` command creates spatially lagged variables for each of *varlist* in the dataset.

`splag`#_*varname*_*swmtype*[_*wvar*] [*suffix*] is a #th order spatially lagged variable of each *varname* of *varlist*. The value # in `order(#)` option is inserted after `splag`. The *varname* and `swmtype` are automatically inserted. Either `b`, `k`, `e`, or `p` is inserted in accordance with *swmtype*: `b` for `swm(bin)`, `k` for `swm(knn #)`, `e` for `swm(exp #)`, and `p` for `swm(pow #)`. The _*wvar* is optionally inserted as the name of weight variable specified in `wvar()` option. The *suffix* is optionally inserted as a string specified in `rowif()` option if used.

3.3.2 Stored results

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

Scalars			
`r(N)`	number of observations	`r(K)`	number of variables
`r(td)`	threshold distance	`r(dd)`	parameter of distance decay or knn
`r(od)`	order of spatial lag	`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	`r(W)`	spatial weight matrix
Macros			
`r(cmd)`	`spgen`	`r(varlist)`	names of specified variables
`r(swm)`	type of spatial weight matrix	`r(swm_std)`	row-standardization of spatial weight matrix
`r(dunit)`	unit of distance	`r(dist_type)`	exact or approximation
`r(weight)`	variable name specified in wvar()	`r(wtype)`	type of weight: `odweight` or `dweight`
`r(rowif)`	variable name specified in rowif()		

❏ Technical note

When the spatial weight matrix is too large for the computer specs (e.g., the memory size is small), the computer may freeze. For example, about $51,842 \times 51,842$ spatial weight matrix uses 20 GB of memory space during the calculation process. A useful way for large-sized data is to use the `largesize` option. This option avoids matrix manipulation during the calculation process of spatial lagged variables to save memory space. This is faster when the data is large. When this option is specified, `nomatsave`, `approx`, and `order(1)` options are automatically applied.

❏

4 Examples

4.1 Example 1: Basic manipulation

First of all, I illustrate the use of the `spgen` command with the Columbus dataset used by Anselin (1988). In the literature of spatial econometrics, many studies use this dataset as a benchmark analysis for spatial econometrics. Although the Columbus dataset contains the two variables on the locational coordinate information (`X` and `Y`), these are expressed in arbitrary digitizing units. In this study, I modify the original Columbus dataset by adding the geographical information on latitude and longitude (`x_cntrd` and `y_cntrd`).[6]

In this example, I demonstrate how to calculate a spatially lagged variable for crime rate (`CRIME`). The following command computes a spatially lagged variable using a power functional form of spatial weight matrix:

```
. use "columbus.dta", clear

. spgen CRIME, lat(y_cntrd) lon(x_cntrd) swm(pow 8) dist(.) dunit(km)
Size of spatial weight matrix: 49 * 49
Calculating bilateral distance...

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

splag1_CRIME_p was generated in the dataset.
```

After the implementation of the above command, `splag1_CRIME_p` is generated in the dataset. In the above example, the threshold distance d in the spatial weight matrix is not necessarily specified in the `dist()` option when `swm(pow #)` or `swm(exp #)` is used. The threshold distance d in the `dist()` option is ignored when `swm(kmm #)` is specified. Therefore, a useful way is to put the dot (`.`) in the `dist()` option

4.1.1 Multiple variables

The `spgen` command works for multiple variables (`spgen` ver. 1.40 or later). The sample code is given below:

(*Continued on next page*)

[6] The Columbus dataset is publicly available from GeoDa (`https://geodacenter.github.io/`).

```
. use "columbus.dta", clear

. spgen CRIME INC HOVAL, lat(y_cntrd) lon(x_cntrd) swm(pow 8) dist(.) dunit(km)
Size of spatial weight matrix: 49 * 49
Calculating bilateral distance...

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

splag1_CRIME_p was generated in the dataset.
splag1_INC_p was generated in the dataset.
splag1_HOVAL_p was generated in the dataset.
```

4.1.2 `replace` option

The `spgen` command returns error message in the default if the outcome variables already exist in the dataset. To overwrite the existing variables in the dataset, the `replace` option is used as follows:

```
. spgen CRIME INC HOVAL, lat(y) lon(x) swm(pow 8) dist(.) dunit(km) replace
  (output omitted)
```

4.1.3 `rowif()` option

The `rowif()` option allows to calculate spatial lag of the variable for a targeted geographical unit. The rowif() option is convenient if a user is interested in the spatial lag of some specific regions in the dateset. The example is as follows:

(*Continued on next page*)

```
. use "columbus.dta", clear

. gen flag_rowif = NEIG < 15

. spgen CRIME, lat(y_cntrd) lon(x_cntrd) swm(pow 8) dist(.) dunit(km) rowif(flag_rowif)
ROWIF option returns spatial lags for observations with flag_rowif = 1
Size of spatial weight matrix: 14 * 49
Calculating spatial lagged variable...

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

splag1_CRIME_p was generated in the dataset.
```

In the second line, the dummy variable is generated for targeted regions. In the third line, this dummy variable is specified in the `rowif()` option. The `spgen` command calculates the spatial lag of `CRIME` only for 14 regions.

4.1.4 `suffix()` option

The `suffix()` option is helpful to store additional information in the name of outcome variable. The example is given as follows:

```
. forvalues i = 1(1)10 {
.     spgen CRIME INC HOVAL, lat(y) lon(x) swm(pow `i´) dist(.) dunit(km) suffix(_dd`i´)
.}
  (output omitted)
```

4.2 Example 2: Compute local sum

The `spgen` command calculates the local sum of neighboring regions within a circle of radius d km for region i. For ease of explanation, consider a dataset that contains three variables: latitude (`y`), longitude (`x`), and one variable (`var1`). Using `swm(bin)` and `nostd` options, the local sum of the variable (`var1`) is calculated as follows:

```
. spgen var1, lat(y) lon(x) swm(bin) dist(5) dunit(km) nostd
  (output omitted)
```

This command calculates the local sum of neighboring regions located within a circle of radius 5 km except region i.

4.3 Example 3: Compute market potential

The market potential of Harris (1954), which is often used in economic geography literature, can be easily calculated by the spgen command. The market potential MP_i is calculated as the inverse-distance-weighted sum of income (or, gross regional products): $\mathrm{MP}_i = \sum_{j=1}^{n} Y_j d_{ij}^{-1}$, for all i, j, where Y_i is income of region i.

For ease of explanation, consider a dataset that contains three variables: latitude (y), longitude (x), and one variable (value_added). Using swm(pow 1) and nostd options, the following command computes the logarithm of market potential lnmp in Stata:

```
. spgen value_added, lat(y) lon(x) swm(pow 1) dist(.) dunit(km) nostd
  (output omitted)
. gen lnmp = log(value_added + spgen1_value_added_p)
```

In the second line, the diagonal element is added because the spgen command use zero for the diagonal elements ($w_{ii} = 0$). Note that own region's income or GDP may be also weighted by internal distance to consider differences in area. For example, see Head and Mayer (2010) for further discussion.

4.4 Example 4: Calculate spatially lagged variables in panel data

The spgen command calculate spatially lagged variables in long-style panel data using the forvalues and foreach loop . Consider a dataset that contains the following variables: latitude (y), longitude (x), some variables (var1, var2, var3, var4, and var5), id (id) and year (year). The time span ranges from 2010 to 2015. The spatially lagged variable of var1–var5 (wvar1–wvar5) can be calculated as follows:

```
. xtset id year
  (output omitted)

. local VARLIST var1 var2 var3 var4 var5

. foreach VAR in `VARLIST´ {
  2.         gen w`VAR´ = .
  3. }
  (output omitted)

. forvalues i = 2010(1)2015 {
  2.         spgen `VARLIST´ if year == `i´, lat(y_cntrd) lon(x_cntrd) swm(pow 1) dist(.) dunit(km)
  3.         foreach VAR in `VARLIST´ {
  4.                 replace w`VAR´ = splag1_`VAR´_p if year == `i´ & w`VAR´ == .
  5.                 drop splag1_`VAR´_p
  6.         }
  7. }
  (output omitted)
```

4.5 Example 5: spgenerate of Stata ver. 15 or later

The spgen command can be used in collaboration with Sp commands on Stata 15 or later.

4.5.1 `spmatrix` command

The **spgen** command stores the spatial weight matrix in `r(W)` after the implementation. This spatial weight matrix `r(W)` is stored in Mata. Then, the **spgen** command can import the spatial weight matrix generated by the **spgen** command. The example is available below:

```
. use "columbus.dta", clear

. spset NEIG, coord(x_cntrd y_cntrd) coordsys(latlong)
  Sp dataset columbus.dta
                data:  cross sectional
     spatial-unit id:  _ID (equal to NEIG)
         coordinates:  _CY, _CX (latitude-and-longitude, kilometers)
    linked shapefile:  none

. spgen _ID, lat(y_cntrd) lon(x_cntrd) swm(pow 8) dist(.) dunit(km)
Size of spatial weight matrix: 49 * 49
Calculating bilateral distance...
┌────────────────┐
│Completed:  10%│
│Completed:  20%│
│Completed:  30%│
│Completed:  40%│
│Completed:  50%│
│Completed:  60%│
│Completed:  70%│
│Completed:  80%│
│Completed:  90%│
│Completed: 100%│
└────────────────┘
splag1__ID_p was generated in the dataset.

. mata: W = st_matrix("r(W)")

. mata: id = st_data(., "_ID")

. spmatrix spfrommata W = W id, replace normalize(row)

. spmatrix dir

────────────────────────────────────────────────────────────────────────
   Weighting matrix name           N x N      Type         Normalization
────────────────────────────────────────────────────────────────────────
                       W        49 x 49     custom           row
────────────────────────────────────────────────────────────────────────

. spmatrix summarize W

Weighting matrix  W
───────────────────────────────────────
           Type │               custom
  Normalization │                  row
      Dimension │              49 x 49
Elements        │
   minimum      │                    0
   minimum > 0  │             8.50e-12
   mean         │             .0204082
   max          │             .9981432
───────────────────────────────────────

  (output omitted)
```

4.5.2 spgenerate **command**

Once the spatial weight matrix is imported by the `spmatrix` command, the spatial lag of the variable can be calculated by the `spgenerate` command, similar to the `spgen` command. The example is available below:

```
. use "columbus.dta", clear

. spset NEIG, coord(x_cntrd y_cntrd) coordsys(latlong)
  Sp dataset columbus.dta
                data:  cross sectional
     spatial-unit id:  _ID (equal to NEIG)
         coordinates:  _CY, _CX (latitude-and-longitude, kilometers)
    linked shapefile:  none

. spgen _ID, lat(y_cntrd) lon(x_cntrd) swm(pow 8) dist(.) dunit(km)
  (output omitted)

. mata: W = st_matrix("r(W)")

. mata: id = st_data(., "_ID")

. spmatrix spfrommata W = W id, replace normalize(row)
  (output omitted)

. **
. spgenerate w1CRIME = W*CRIME
. spgenerate w2CRIME = W*w1CRIME
```

4.5.3 spregress **command**

Once the spatial weight matrix is imported by the `spmatrix` command, spatial econometric models are estimated by the `spregress` command. The example is available below:

(*Continued on next page*)

```
. use "columbus.dta", clear

. spset NEIG, coord(x_cntrd y_cntrd) coordsys(latlong)
  Sp dataset columbus.dta
                data:  cross sectional
     spatial-unit id:  _ID (equal to NEIG)
         coordinates:  _CY, _CX (latitude-and-longitude, kilometers)
    linked shapefile:  none

. spgen _ID, lat(y_cntrd) lon(x_cntrd) swm(pow 8) dist(.) dunit(km)
  (output omitted)

. mata: W = st_matrix("r(W)")

. mata: id = st_data(., "_ID")

. spmatrix spfrommata W = W id, replace normalize(row)

. spmatrix dir
  (output omitted)
. spmatrix summarize W
  (output omitted)

. **
. spregress CRIME INC HOVAL, ml dvarlag(W)
  (49 observations)
  (49 observations (places) used)
  (weighting matrix defines 49 places)

Performing grid search ... finished

Optimizing concentrated log likelihood:

Iteration 0:   log likelihood = -180.51657
Iteration 1:   log likelihood = -180.48637
Iteration 2:   log likelihood = -180.48637

Optimizing unconcentrated log likelihood:

Iteration 0:   log likelihood = -180.48637
Iteration 1:   log likelihood = -180.48637  (backed up)

Spatial autoregressive model                    Number of obs     =         49
Maximum likelihood estimates                    Wald chi2(3)      =     105.95
                                                Prob > chi2       =     0.0000
Log likelihood = -180.48637                     Pseudo R2         =     0.5842

------------------------------------------------------------------------------
       CRIME |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
CRIME        |
         INC |  -.9852786   .3047046    -3.23   0.001    -1.582489   -.3880685
       HOVAL |  -.2668222   .0829736    -3.22   0.001    -.4294475   -.1041968
       _cons |   43.27086   7.071231     6.12   0.000     29.41151    57.13022
-------------+----------------------------------------------------------------
W            |
       CRIME |   .4247278   .0998508     4.25   0.000     .2290238    .6204317
-------------+----------------------------------------------------------------
var(e.CRIME)|   84.49397   17.53173                      56.26122    126.8944
------------------------------------------------------------------------------
Wald test of spatial terms:          chi2(1) = 18.09       Prob > chi2 = 0.0000


  (output omitted)
```

5 Empirical applications for spatial econometric analysis

5.1 Empirical application 1: Anselin's Columbus dataset

I provide an empirical application of Moran's scatter plot and estimation of spatial econometric model using Anselin's Columbus dataset Anselin (1988).

Anselin (1995) proposes a Moran scatter plot, which illustrates a spatial autocorrelation for Moran's I. The formula of the Moran's I is

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

where $\boldsymbol{z}$ is a vector of standardized variable and $\boldsymbol{W}$ is a row-standardized spatial weight matrix.

The idea of the Moran scatter plot is as follows. Consider a regression: $\boldsymbol{W}\boldsymbol{z} = \alpha \boldsymbol{z} + \text{residuals}$, where residuals indicate that any statistical assumption on error terms is not considered. Deriving the OLS estimator, it is clear that the estimate $\hat{\alpha}$ is equal to the formula of the Moran's I. In other words, the Moran scatter plot illustrates the relationship between $\boldsymbol{W}\boldsymbol{z}$ and $\boldsymbol{z}$

Figure 1 illustrates a spatial variation in crime rates (`CRIME`) in the Columbus dataset. Panel (a) visualizes geographical distribution of crime rates, showing that nearby regions tend to have similar crime rates in geographic space.[7] After implementing the **`spgen`** command, the **`twoway scatter`** command can visualize the spatial autocorrelation, as shown in Panel (b). As mentioned earlier, the slope through the origin in a Moran scatter plot is equal to the Moran's I (in this case, $I = 0.608$). The sample code appears below:[8]

(*Continued on next page*)

[7] Figure 1 is created by the **`shp2dta`** that command converts shapefiles to a DTA file (Crow, 2015) and the **`spmap`** command that illustrates data on map (Pisati, 2008). Stata 15 or later includes official commands **`spshape2dta`** that converts a shapefile to a DTA file and **`grmap`** that illustrates data on map.

[8] Kondo (2018) provides **`moransi`** command.

```
. use "columbus.dta", clear

. egen std_CRIME = std(CRIME)

. spgen std_CRIME, lat(y_cntrd) lon(x_cntrd) swm(pow 8) dist(.) dunit(km)
Size of spatial weight matrix: 49 * 49
Calculating bilateral distance...
┌─────────────────┐
│                 │
│Completed:  10%  │
│Completed:  20%  │
│Completed:  30%  │
│Completed:  40%  │
│Completed:  50%  │
│Completed:  60%  │
│Completed:  70%  │
│Completed:  80%  │
│Completed:  90%  │
│Completed: 100%  │
│                 │
└─────────────────┘
splag1_std_CRIME_p was generated in the dataset.

. local VARS splag1_std_CRIME_p std_CRIME

. twoway (scatter `VARS´, ms(Oh) msize(large)) ///
>         (lfit `VARS´, lw(thick) est(nocon)) ///
>         , ///
>         ytitle("{it:Wz}", tstyle(size(large))) ///
>         xtitle("{it:z}", tstyle(size(large)) height(7)) ///
>         ylabel(-2(1)2, format(%2.1f) grid ang(h) labsize(large)) ///
>         xlabel(-2(1)2, format(%2.1f) grid labsize(large)) ///
>         yline(0) ///
>         xline(0) ///
>         aspect(1) ///
>         legend(off) ///
>         graphregion(color(white) fcolor(white))

. graph export "fig/FIG_msp_d8.eps", replace
(file fig/FIG_msp_d8.eps written in EPS format)
```

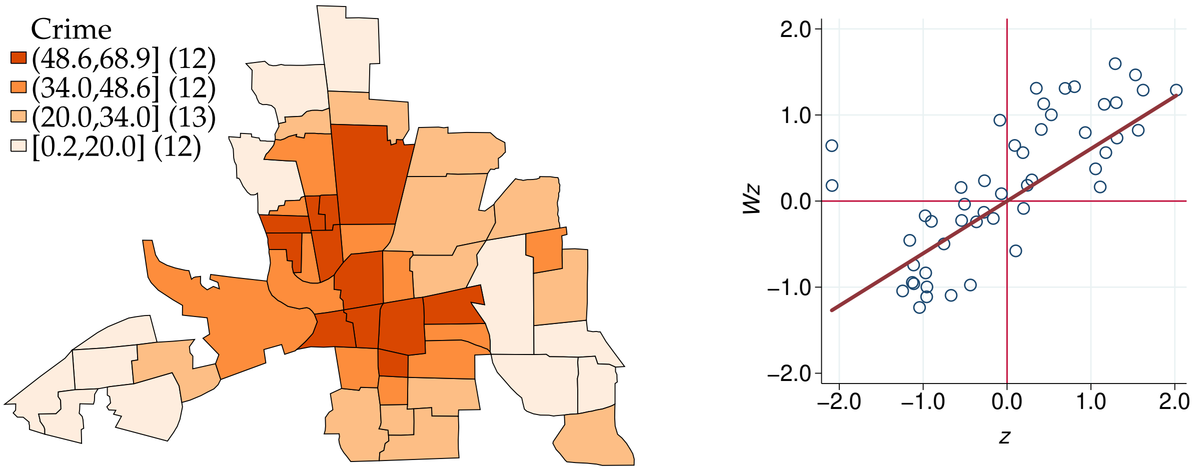
The `spgen` command enables researchers to easily perform spatial econometric analysis in Stata as a simple introduction. Using the Columbus dataset, I demonstrate how the spatial econometric model is estimated with the `spgen` command.

Let $\boldsymbol{y}$, $\boldsymbol{X}$, and $\boldsymbol{u}$ denote an $n \times 1$ vector of dependent variable, an $n \times k$ matrix of explanatory variable, and an $n \times 1$ vector of error terms, respectively. Thus, the spatial lag model, or spatial autoregressive model is given as follows:

$$\boldsymbol{y} = \rho \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{IID}(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{5}$$

where σ^2 is the variance of error terms, $\mathbf{0}$ is the $n \times 1$ vector of the value 0, and $\mathbf{I}$ is the $n \times n$ identity matrix.

An estimation issue for Model (5) is that the spatial lag of dependent variable $\boldsymbol{W}\boldsymbol{y}$ is endogenous, and the OLS estimators are inconsistent. To overcome the endogeneity bias, the spatial econometric model is estimated by maximum likelihood (ML), the method of instrumental variables (IV), or the



(a) Crime Rate

(b) Moran Scatterplot ($I = 0.608$)

Figure 1: Spatial Variation in Crime Rate

Note: Created by the author using the Columbus dataset of Anselin (1988). The distance decay parameter of the spatial weight matrix is $\delta = 8$ in Panel (b).

generalized method of moments (GMM).[9] In this study, I introduce how the `spgen` helps estimation of the spatial econometric model by IV/GMM.[10]

The model estimated by Anselin (1988), which is often used as a benchmark estimation in this literature, takes the following specification:

$$\text{CRIME}_i = \rho\text{WCRIME}_i + \beta_1 + \beta_2\text{INC}_i + \beta_3\text{HOUSE}_i + u_i, \tag{6}$$

where CRIME_i denotes residential burglaries and vehicle thefts per 1,000 households, WCRIME_i denotes the spatial lag of CRIME_i, INC_i denotes household income (in \$1,000) and HOUSE_i denotes housing value (in \$1,000).

Furthermore, the spatially lagged explanatory variables are also considered as follows:

$$\text{CRIME}_i = \beta_1 + \beta_2\text{INC}_i + \beta_3\text{HOUSE}_i + \beta_4\text{WINC}_i + \beta_5\text{WHOUSE}_i + u_i, \tag{7}$$

where WINC_i denotes the spatial lag of INC_i, and WHOUSE_i denotes the spatial lag of HOUSE_i. Model (7) is simply estimated by OLS.

Table 1 presents the estimation results of spatial autoregressive model by OLS, IV/GMM, ML estimations. I follow Anselin and Bera (1998), who compare estimation results between various spec-

[9] See Anselin (2006) and LeSage and Pace (2009) for further discussion.

[10] The IV/GMM estimators used here are consistent, but not efficient. To obtain more efficient estimators, a spatially dependent structure needs to be considered in the variance-covariance matrix. Furthermore, one may be interested in the spatial lag of error terms. These challenges are beyond the scope of the `spgen` command. The Sp commands on Stata 15 or later provide useful commands to estimate spatal econometri models.

ifications and estimation methods. Our results are similar to theirs. However, the differences in estimation results arise from the specification of the spatial weight matrix. Anselin and Bera (1998) use the contiguity matrix, whereas this study uses the distance matrix. We can clearly see that the coefficient estimate of Income in Column (1) differs considerably from the coefficient estimates obtained by the spatial econometric model. Furthermore, the OLS estimates in Column (2) seem to be biased due to the endogeneity of spatially lagged variable $\boldsymbol{Wy}$. Columns (3)–(4) show similar estimation results each other. In the IV/GMM estimation, instrumental variables for the spatially lagged variable $\boldsymbol{Wy}$ include the first, second, and third orders of spatial lags of explanatory variables ($\boldsymbol{WX}$, $\boldsymbol{W}^2\boldsymbol{X}$, $\boldsymbol{W}^3\boldsymbol{X}$), as suggested by Kelejian and Robinson (1993) and Kelejian and Prucha (1998, 1999).[11] Column (5) shows ML estimation results, which are similar to the IV/GMM estimation results. The ML estimation is implemented by the `spregress` with the spatial weight matrix generated from the `spgen` command. Column (6) shows the estimation results of Model (7). The spatial lag of income is statistically significant at the 5% level.

Stata 15 or later provides the Sp commands, which are useful to estimate spatial econometric models. An advantage of the `spgen` is to avoid the matrix manipulation during the estimation when the size of spatial weight matrix is too large. In such a situation, the Sp commands may freeze. Once the spatially lagged variables are obtained by the `spgen` command, researchers can easily estimate spatial econometric model by the standard Stata commands for IV/GMM estimation, such as `ivregress` or `ivreg2`.

5.2 Empirical application 2: Japanese municipal data on crime

The spatial analysis of crime data is extended using Japanese municipal data.[12] Ohtake and Kohara (2010) estimate the impact of unemployment rate on crime rate in Japan. Using the prefecture-level panel data to control for prefectural heterogeneities, they find that there is a significant positive impact of unemployment rate on the crime rate.

A simple extension of the empirical analysis here is to examine spatial spillovers on crime rate across regions. The neighboring unemployment rates might affect the own regional crime rate. The spatial structure on crime is a crucial aspect for anti-crime policy-making, and it is important to clarify whether crime has spread to the surrounding regions. Therefore, the spatial econometric analysis provides an important insight on this issue.[13]

In this empirical illustration, two types of spatial econometric models are estimated. The first specification includes the spatial lag of dependent variable as follows:

$$\mathrm{CRIME}_i = \rho \mathrm{WCRIME}_i + \beta_1 + \beta_2 \mathrm{UNEMP}_i + \beta_3 \mathrm{INCOME}_i + u_i, \tag{8}$$

[11] See Anselin (2006) for a review of theoretical background on spatial econometrics.

[12] See Appendix A for more details on the Japanese municipal data.

[13] Ohtake and Kohara (2010) emphasize the importance of including other explanatory variables to avoid omitted variable bias. However, it is not considered here for simplification.

Table 1: Estimation Results Using Crime Data in Columbus

	Dependent Variable: Crime Rate (per 1,000 Households)					
	OLS	OLS	IV	GMM	ML	OLS
Explanatory Variables	(1)	(2)	(3)	(4)	(5)	(6)
W CrimeRate (ρ)		0.594***	0.440***	0.410***	0.432***	
		(0.105)	(0.106)	(0.087)	(0.102)	
Income	−1.597***	−0.741	−0.964**	−1.628***	−0.975**	−1.236**
	(0.461)	(0.478)	(0.437)	(0.338)	(0.436)	(0.498)
Housing Value	−0.274*	−0.264	−0.267*	−0.049	−0.267*	−0.289*
	(0.163)	(0.160)	(0.153)	(0.075)	(0.153)	(0.163)
W Income						−0.952**
						(0.421)
W Housing Value						−0.027
						(0.112)
Constant	68.619***	33.141***	42.386***	45.624***	42.828***	77.530***
	(4.233)	(7.375)	(6.585)	(5.800)	(6.458)	(5.027)
Number of Observations	49	49	49	49	49	49
Adjusted R^2	0.533	0.685				0.572
Weak IV			10.914	10.914		
Overidentification (p-value)			0.434	0.319		

Note: Heteroskedasticity-consistent standard errors are in parentheses. * denotes statistical significance at the 10% level, ** at the 5% level, and *** at the 1% level. The prefix W indicates spatial lag of the corresponding variable. The `spgen` command with `swm(pow 8)` and `dist(.)` options is used. The instruments for spatial lag of dependent variable are $\boldsymbol{WX}$, $\boldsymbol{W}^2\boldsymbol{X}$, and $\boldsymbol{W}^3\boldsymbol{X}$. Weak IV is the robust Kleinbergen–Paap rk Wald F statistic for test of weak instruments. Overidentification shows p-value of Sagan–Hansen J test. Iterative GMM is used with robust weighting matrix in Column (4). ML estimates are obtained by the `spregress` command in Column (5).

where UNEMP_i is the unemployment rate of municipality i, INCOME_i is the logarithm of average income per capita in municipality i, u_i is an error term, and ρ captures spatial spillover effects on crime rates across municipalities, suggesting that crime rates in neighboring regions also affect the own region's crime rate.

The second specification includes the spatially lagged explanatory variables as follows:

$$\text{CRIME}_i = \beta_1 + \beta_2\text{UNEMP}_i + \beta_3\text{INCOME}_i + \beta_4\text{WUNEMP}_i + \beta_5\text{WINCOME}_i + u_i, \tag{9}$$

where WUNEMP_i is the spatial lag of UNEMP_i, WINCOME_i is the spatial lag of INCOME_i, and β_4 and β_5 capture spatial spillover effects on crime rates among municipalities. However, the spatial spillover process differs from the first specification. The unemployment rates and income per capital in neighboring regions directly affect the own region's crime rate. Anselin (2003) distinguishes both spillover effects: the first is global spatial spillover and the second is local spatial spillover. Unlike the spatial spillover in Model (9), Model (8) captures the two-step spillover process of unemployment rates in neighboring regions affecting the neighboring crime rates, then the neighboring crime rates affecting the own region's crime rate.

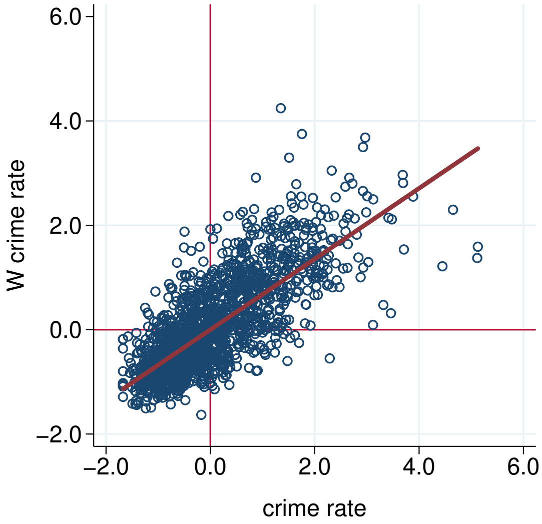


Figure 2: Moran Scatter Plot of Crime Rate in Japan ($I = 0.677$.)

Note: Created by the author. The distance decay parameter of the spatial weight matrix is $\delta = 4$.

Figure 2 illustrates Moran scatter plot of crime rates. The Moran's I is 0.677, and there is a statistically significant, positive spatial autocorrelation.

Table 2 presents the estimation results of the spatial econometric Models (8) and (9). Column (1) shows benchmark estimation results by OLS estimation. Although the unemployment rate significantly increases crime rate, the magnitude differs from those of the spatial econometric models. Columns (2)–(4) show estimation results of the regression (8) and the OLS estimates in Column (2) seem to be inconsistent compared with the IV/GMM estimates in Columns (3)–(4). Column (5) shows the ML estimation results, which are similar to the IV/GMM estimation results. We can see that there is a significant spatial dependence in crime rates in Columns (3)–(5). Column (6) shows the estimation results of Model (9), and the spatially lagged unemployment rates and income per capita are positively correlated with crime rate.

Summing up, if researchers simply estimate standard econometric models without considering spatial dependence, then they might miss an important channel because spatial spillover effects cannot be captured. The `spgen` command is expected to enable us to easily examine spatial dependence in the data.

6 Concluding remarks

I have introduced the new command `spgen`, which easily computes spatially lagged variables in Stata using geographical information of latitude and longitude. In this article, I have provided interesting examples of spatial econometric analysis with the `spgen` command.

Table 2: Estimation Results Using Japanese Municipal Crime Data

	Dependent Variable: Crime Rate (per 1,000 people)					
	OLS	OLS	IV	GMM	ML	OLS
Explanatory Variables	(1)	(2)	(3)	(4)	(5)	(6)
W Crime Rate		0.534***	0.365***	0.359***	0.345***	
		(0.033)	(0.079)	(0.078)	(0.021)	
Unemployment Rate	0.987***	0.675***	0.773***	0.764***	0.785***	0.850***
	(0.086)	(0.072)	(0.083)	(0.082)	(0.066)	(0.100)
log(Income per Capita)	23.726***	15.761***	18.276***	18.534***	18.578***	19.860***
	(1.639)	(1.570)	(1.927)	(1.889)	(1.226)	(2.003)
W Unemployment Rate						0.307**
						(0.126)
W log(Income per Capita)						8.106***
						(2.265)
Constant	−186.687***	−125.612***	−144.895***	−146.812***	−147.211***	−221.326***
	(13.076)	(12.465)	(15.122)	(14.838)	(9.796)	(15.328)
Prefecture Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Number of Observations	1719	1719	1719	1719	1719	1719
Adjusted R^2	0.562	0.645	0.637	0.636		0.567
Weak IV			36.524	36.524		
Overidentification (*p*-value)			0.841	0.841		

Note: Heteroskedasticity-consistent standard errors are in parentheses. * denotes statistical significance at the 10% level, ** at the 5% level, and *** at the 1% level. The prefix W indicates spatial lag of the corresponding variable. The `spgen` command with `swm(pow 4)` and `dist(.)` options is used. The instruments for spatial lag of dependent variable are $\boldsymbol{WX}$, $\boldsymbol{W}^2\boldsymbol{X}$, and $\boldsymbol{W}^3\boldsymbol{X}$ except prefectural dummies. Weak IV is the robust Kleinbergen–Paap rk Wald F statistic for test of weak instruments. Overidentification shows p-value of Sagan–Hansen J test. Iterative GMM is used with robust weighting matrix in Column (4). ML estimates are obtained by the `spregress` command in Column (5).

An advantage of the `spgen` command is that the shapefile of the corresponding area is not required to construct the spatial weight matrix. A suitable shapefile is not available in some situations. Instead, the geographical information on latitude and longitude is the only requirement; it is easily added into a dataset by the geocoding technique. Another advantage of the `spgen` is to be designed for large-sized dataset. The Sp commands on Stata 15 or later rely on the matrix manipulation, which is intuitive and convenient in programming. However, the inverse of high dimensional matrix is not easy to calculate on low spec computer. Thus, the `spgen` command facilitates spatial econometric analysis for large-sized datasets.

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.

———. 2003. Spatial externalities, spatial multipliers, and spatial econometrics. *International Regional Science Review* 26(2): 153–166.

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

Anselin, L., and A. K. Bera. 1998. Spatial dependence in linear regression models with an introduction to spatial econometrics. In *Handbook of Applied Economic Statistics*, ed. A. Ullah and D. E. Giles, chap. 7, 237–289. New York: Marcel Dekkar.

Crow, K. 2015. SHP2DTA: Stata module to converts shape boundary files to Stata datasets. `https://ideas.repec.org/c/boc/bocode/s456718.html`.

Drukker, D. M., H. Peng, I. R. Prucha, and R. Raciborski. 2013. Creating and managing spatial-weighting matrices with the spmat command. *Stata Journal* 13(2): 242–286.

Harris, C. D. 1954. The market as a factor in the localization of industry in the United States. *Annals of the Association of American Geographers* 44(4): 315–348.

Head, K., and T. Mayer. 2010. Illusory border effects : Distance mismeasurement inflates estimates of home bias in trade. In *The Gravity Model in International Trade: Advances and Applications*, ed. S. Brakman and P. van Bergeijk, chap. 6, 165–192. New York: Cambridge University Press.

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 Department of Economics, revised 09 Aug 2012.

Kelejian, H. H., and I. R. Prucha. 1998. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics* 17(1): 99–121.

———. 1999. A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review* 40(2): 509–533.

Kelejian, H. H., and D. P. Robinson. 1993. A suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a county expenditure model. *Papers in Regional Science* 72(3): 297–312.

Kondo, K. 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 Department of Economics, revised 14 Jun 2021.

———. 2019. Municipality-level panel data and municipal mergers in Japan. RIETI Technical Paper No. 19-T-001.

LeSage, J., and R. K. Pace. 2009. *Introduction to Spatial Econometrics*. Boca Raton: CRC Press.

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

Ohtake, F., and M. Kohara. 2010. The relationship between unemployment and crime: Evidence from time-series data and prefectural panel data. *Japanese Journal of Sociological Criminology* 35: 54–71. (in Japanese).

Pisati, M. 2008. SPMAP: Stata module to visualize spatial data. Statistical Software Components S456812, Boston College Department of Economics, revised 18 Jan 2018.

Stewart, J. Q. 1947. Empirical mathematical rules concerning the distribution and equilibrium of population. *Geographical Review* 37(3): 461–485.

Appendix A Data

A.1 Columbus dataset

Table 3 presents the summary statistics of the data used in Anselin (1988).

Table 3: Descriptive Statistics of Crime Data in Columbus

Variables	Mean	S.D.	Min	Max
Crime Rate (per 1,000 people)	35.129	16.732	0.178	68.892
Income	14.375	5.703	4.477	31.070
Housing Value	38.436	18.466	17.900	96.400
W Crime Rate (per 1,000 people)	38.323	13.802	14.445	61.835
W Income	13.195	4.440	4.642	24.908
W Housing Value	35.915	12.383	19.340	75.999

Note: The number of observations is 49. The prefix W indicates spatial lag of the corresponding variable. The `spgen` command with `swm(pow 8)` and `dist(.)` options is used.

A.2 Japanese municipal data

Table 4 presents the descriptive statistics of Japanese municipal data on crime rates and unemployment rates. Japanese municipal data are taken from e-Stat, the portal site of the official statistics of Japan.[14] The municipal unemployment rates are calculated from 2005 population census. The number of criminal cases known to the police in 2006 is taken from the "Statistical Observations of Shi, Ku, Machi, Mura" (Statistical Bureau of Japan). The municipal crime rate (per 1,000 people aged 15 or above) is calculated as the number of criminal cases divided by total population aged 15 or above. The municipal population is taken from the 2005 population census. Note that the crime rates are 1 year later than municipal unemployment rates. All municipal variables in 2005 and 2006 are reaggregated by municipal unit defined as of Octover 1, 2015 (Kondo, 2019).

Table 4: Descriptive Statistics of Japanese Municipal Panel Data on Crime

Variables	Mean	S.D.	Min	Max
Crime Rate (per 1,000 people)	12.204	7.291	0.000	49.579
Unemployment Rate (%)	5.546	2.073	0.000	19.406
log(Income per Capita)	8.009	0.160	7.654	11.618
W Crime Rate (per 1,000 people)	12.523	6.458	0.310	43.1 49
W Unemployment Rate (%)	5.630	1.815	0.650	16.370
W log(Income per Capita)	8.010	0.124	7.729	8.570

Note: The number of observations is 1,719. The prefix W indicates spatial lag of the corresponding variable. The `spgen` command with `swm(pow 4)` and `dist(.)` options is used.

[14] URL: `https://www.e-stat.go.jp/`