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

Keisuke Kondo[†]

Version of this manual: April 2, 2018
(`spgen`: version 1.32)

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. More importantly, 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 econometrics

1 Introduction

Spatial econometric analysis has attracted recent attention from researchers and policy-makers, and demand for its use is continuously growing among Stata users. The newly developed Stata command, `spgen`, helps them easily perform spatial econometric analysis in Stata (Kondo, 2015).

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.

[*] I thank Christopher Barrie for discovering a crucial typo in the manual. 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 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://sites.google.com/site/keisukekondokk/`).

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

The computation of spatially lagged variables requires a spatial weight matrix, which mathematically describes the spatially dependent structures in the matrix. Some researchers have already developed Stata packages for spatially lagged variables. For example, Drukker et al. (2013) offer the `spmat` command, which constructs spatial weight matrix and calculates spatially lagged variables. Jeanty (2010) also offers the `splagvar` command for spatially lagged variables, However, researchers might have difficulties constructing the spatial weight matrix before computing spatially lagged variables. Although the spatial weight matrix is commonly constructed from the shape file in spatial statistical packages, the shape file of the corresponding study area is not always available. The `spgen` command solves this issue by facilitating a computing procedure of spatial weight matrix.

The `spgen` command provides a good opportunity as a simple introduction to spatial econometric analysis. The `spgen` command computes spatially lagged variables using the geographical information of latitude and longitude. 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.[1] Even if a dataset has no coordinate information (i.e., latitude and longitude), a recent geocoding technique facilitates adding this information to the dataset. The `spgen` command offers flexible extensions for constructing spatial weight matrix based on a distance matrix.

The `spgen` command also contributes to the economic geography literature. 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 empirical analyses in economic geography will advance with Stata.

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

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-section analysis to consider spatial dependence between own region and neighboring 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

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

spatial econometric 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 (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. Various types of spatial weight matrices are proposed in the literature. In this article, three types of spatial weight matrices are considered.[2] The `spgen` command deals with these three types of spatial weight matrices.

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.

[2] 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.

The 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. 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:

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

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.

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. For example, Molho (1995) uses employment size as a weight variable when constructing the spatial weight matrix.

The `spgen` command allows to incorporate a weight variable v into the spatial weight matrix. The power functional type 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 variable of 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:

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

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 varname [if] [in] , lat(varname) lon(varname) swm(swmtype) dist(#) dunit(km|mi)
[ order(#) wvar(varname) nostd nomatsave dms approx detail largesize ]
```

3.2 Options

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

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

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

`dist(#)` specifies the threshold distance # for the spatial weight matrix. The unit of distance is specified by the `dunit(km|mi)` option.

`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.

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.

3.3 Output

3.3.1 Outcome variables

The spgen command creates a spatially lagged variable of *varname* in the dataset.

splag#_*varname*_*swmtype*[_*wvar*] is a #th order spatially lagged variable of *varname*. The value # in order(#) option is inserted after splag. The *varname* is automatically inserted and the suffix b, e, or p is also inserted in accordance with *swmtype*: b for swm(bin), e for swm(exp #), and p for swm(pow #). The weight variable name *wvar* is inserted if used.

3.3.2 Stored results

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

Scalars			
r(N)	number of observations	r(td)	threshold distance
r(dd)	distance decay parameter	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		
Macros			
r(cmd)	spgen	r(varname)	name of variable
r(swm)	type of spatial weight matrix	r(weight)	name of weight variable
r(dist_type)	exact or approximation		

❏ Technical note

When the spatial weight matrix is too large for the computer specs, the spgen command may not calculate spatially lagged variable (The computer may freeze.). I confirm that the spgen command handles about 20,000 × 20,000 spatial weight matrix on the computer (Core-i-5 6500, 16GB Memory, 64bit Windows 10), but it takes approximately one hour. The nomatsave option is recommended to save the memory space in the case of a large-sized spatial weight matrix.

A more useful way is to use the `largesize` option for large-sized data. This option calculates a spatial lagged variable faster when the data is too large. When this option is specified, `nomatsave`, `approx`, and `order(1)` options are automatically applied.

❏

4 Example

4.1 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.[3]

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 type 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
Calculating bilateral distance...
Calculating spatial weight matrix...

Distance by Vincenty formula (unit: km)
splag1_CRIME_p is generated in the dataset.
```

After the implementation of the above command, `splag1_CRIME_p` is generated in the dataset.

❏ Technical note

There are some useful tips for the `spgen` command. Five tips are introduced here. For ease of explanation, consider a dataset that contains four variables: latitude (`y`), longitude (`x`), and two other variables (`var1` and `var2`). First, the `spgen` command calculates the local sum of neighboring regions within a circle of radius d km from the own region using `swm(bin)` and `nostd` options as follows:

```
. spgen var1, lat(y) lon(x) swm(bin) dist(30) dunit(km) nostd
  (output omitted)
```

This command calculates the local sum of neighboring regions located within a circle of radius 30 km.

Second, 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

[3] The Columbus dataset is publicly available from the GeoDa Center for Geospatial Analysis and Computation at Arizona State University (`https://geodacenter.asu.edu/`).

Y_i is income of region i. Thus, the following command computes the market potential variable `mp` in Stata:

```
. spgen var1, lat(y) lon(x) swm(pow 1) dist(.) dunit(km) nostd
  (output omitted)
. gen mp = var1 + spgen1_var1_p
```

In the second line, the diagonal element is added because the `spgen` command use zero for the diagonal elements ($w_{ii} = 0$). In addition, own region's income might be weighted by distance considering differences in area. For example, see Head and Mayer (2010) for further discussion.

Third, the threshold distance d in the spatial weight matrix is not necessarily specified in `dist()` when `swm(pow` *#*`)` or `swm(exp` *#*`)` is used. Therefore, a useful way is to put the dot (`.`) in `dist()` as follows:

```
. spgen var1, lat(y) lon(x) swm(pow 2) dist(.) dunit(km)
  (output omitted)
```

Fourth, the `spgen` command works well by `foreach` loop command when there is a need to compute spatially lagged variables for multiple variables. The example code is given below:

```
. foreach VAR in var1 var2 {
.     spgen `VAR´, lat(y) lon(x) swm(pow 2) dist(.) dunit(km)
. }
  (output omitted)
```

Fifth, the `spgen` command calculate spatially lagged variables using the `forvalues` loop in panel data. Consider a dataset that contains four variables: latitude (`y`), longitude (`x`), some variable (`var`), and year (`year`). The time span ranges from 2010 to 2015. The spatially lagged variable of `var` (`wvar`) can be calculated as follows:

```
gen wvar = .
. forvalues i = 2010(1)2015 {
.     spgen var if year == `i´, lat(y) lon(x) swm(pow 2) dist(.) dunit(km)
.     replace wvar = splag1_var_p if year == `i´
.     drop splag1_var_p
. }
  (output omitted)
```

❏

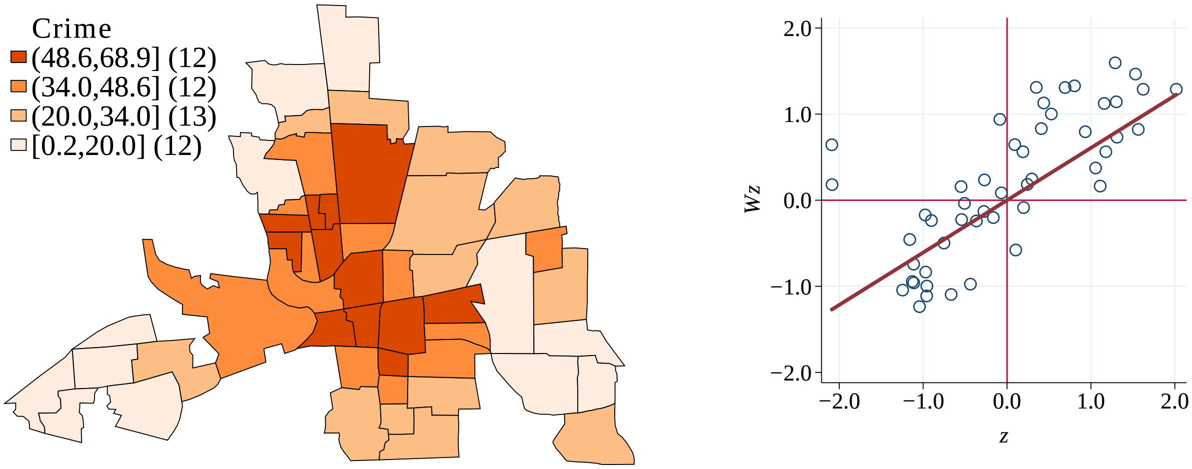
4.2 Moran scatter plot

Visualization is a useful method to foster better understanding of empirical results. 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: $\boldsymbol{W}\boldsymbol{z} = \alpha \boldsymbol{z} +$ residuals, where residuals indicate that any statistical assumption on error terms is not considered. Deriving



(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).

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 $\boldsymbol{Wz}$ 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.[4] 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:

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

. egen std_CRIME = std(CRIME)

. spgen std_CRIME, lat(y_cntrd) lon(x_cntrd) swm(pow 8) dist(.) dunit(km)

. local VAR splag1_std_CRIME_p std_CRIME

. twoway (scatter `VAR´) (lfit `VAR´, est(nocon))

. graph export "FIG_msp.eps", replace
(file FIG_msp.eps written in EPS format)
```

4.3 Empirical application 1: Anselin's Columbus dataset

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

[4] Figure 1 is created by the `shp2dta` that command converts shape files to a DTA file (Crow, 2015) and the `spmap` command that illustrates data on map (Pisati, 2008).

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:

$$\boldsymbol{y} = \rho \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{IID}(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{4}$$

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 (4) 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, the method of instrumental variables (IV), or the generalized method of moments (GMM).[5] In this study, I introduce how the `spgen` helps estimation of the spatial econometric model by IV/GMM.[6]

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

$$\mathrm{CRIME}_i = \rho \mathrm{WCRIME}_i + \beta_1 + \beta_2 \mathrm{INC}_i + \beta_3 \mathrm{HOUSE}_i + u_i, \tag{5}$$

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).

Table 1 presents the estimation results of spatial autoregressive model by OLS and IV/GMM estimations. I follow Anselin and Bera (1998), who compare estimation results between various specifications 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 $\boldsymbol{W}\boldsymbol{y}$. Columns (3)–(4) show similar estimation results each other. For IV/GMM estimation, instead of the spatially lagged variable $\boldsymbol{W}\boldsymbol{y}$, instrumental variables include the first, second, and third orders of spatial lags of explanatory variables ($\boldsymbol{W}\boldsymbol{X}$, $\boldsymbol{W}^2\boldsymbol{X}$, $\boldsymbol{W}^3\boldsymbol{X}$), as suggested by Kelejian and Robinson (1993) and Kelejian and Prucha (1998, 1999).[7]

An advantage is that, 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

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

[6] 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. Further extended Stata modules are needed.

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

Table 1: Estimation Results Using Crime Data in Columbus

	Dependent Variable: Crime Rate (per 1,000 Households)			
	OLS	OLS	IV	GMM
Explanatory Variables	(1)	(2)	(3)	(4)
W Crime Rate		0.594***	0.440***	0.429***
		(0.105)	(0.106)	(0.097)
Income	−1.597***	−0.741	−0.964**	−1.156***
	(0.461)	(0.478)	(0.437)	(0.334)
Housing Value	−0.274*	−0.264	−0.267*	−0.221**
	(0.163)	(0.160)	(0.153)	(0.100)
Constant	68.619***	33.141***	42.386***	43.910***
	(4.233)	(7.375)	(6.585)	(6.259)
Number of Observations	49	49	49	49
Adjusted R^2	0.533	0.685	0.674	0.669
Weak IV			10.914	10.914
Overidentification (p-value)			0.434	0.434

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. Two-step GMM is used with robust weighting matrix in Column (4).

estimation, such as `ivregress` or `ivreg2`.

4.4 Empirical application 2: Japanese municipal data

The spatial analysis of crime data is extended using Japanese municipal data.[8] 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.[9]

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

$$\mathrm{CRIME}_i = \rho \mathrm{WCRIME}_i + \beta_1 + \beta_2 \mathrm{UNEMP}_i + u_i, \tag{6}$$

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

[9] 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.

where UNEMP_i is the unemployment rate of 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 spatial lag of explanatory variable as follows:

$$\text{CRIME}_i = \beta_1 + \beta_2\text{UNEMP}_i + \beta_3\text{WUNEMP}_i + v_i, \tag{7}$$

where WUNEMP_i is the spatial lag of UNEMP_i, v_i is an error term, and β_3 also captures spatial spillover effects on crime rates among municipalities. However, the spatial spillover process differs from the first specification. The unemployment rates in neighboring regions directly affect the own region's crime rate.[10] Unlike the spatial spillover in Model (7), Model (6) 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.

Table 2 presents the estimation results of the spatial econometric Models (6) and (7). 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 (6) and the OLS estimates in Column (2) seem to be inconsistent compared with the IV/GMM estimates in Columns (3)–(4). We can see that there is a significant spatial dependence in crime rates in Columns (3)–(4).

Column (5) shows the estimation results of Model (7). The spatial lag of unemployment rates has a significant positive impact on crime rate, suggesting that the increase in neighboring crime rates also leads to the increase in the own region's crime rate.

If researchers simply estimate standard econometric models without thinking about 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.

5 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.

An advantage of the `spgen` command is that the shape file of the corresponding area is not required to construct the spatial weight matrix. A suitable shape file 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. Furthermore, a possible extension is to examine spatial spillover effects among firms or individuals using microdata with geographic location information. Therefore, the `spgen` command is expected to facilitate spatial econometric analysis using Stata.

[10] Anselin (2003) distinguishes both spillover effects: the first is global spatial spillover and the second is local spatial spillover.

Table 2: Estimation Results Using Japanese Municipal Crime Data

	Dependent Variable: Crime Rate (per 1,000 people)				
	OLS	OLS	IV	GMM	OLS
Explanatory Variables	(1)	(2)	(3)	(4)	(5)
W Crime Rate		0.645***	0.504***	0.478***	
		(0.049)	(0.116)	(0.115)	
Unemployment Rate	0.541***	0.301***	0.353***	0.370***	0.404***
	(0.060)	(0.052)	(0.070)	(0.069)	(0.075)
W Unemployment Rate					0.285***
					(0.083)
Constant	4.262***	0.420	1.086*	1.160*	3.118***
	(0.480)	(0.477)	(0.648)	(0.646)	(0.601)
Region Dummy	Yes	Yes	Yes	Yes	Yes
Number of Observations	1728	1728	1728	1728	1728
Adjusted R^2	0.292	0.483	0.474	0.470	0.297
Weak IV			27.157	27.157	
Overidentification (p-value)			0.278	0.278	

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 6)` 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. Two-step GMM is used with robust weighting matrix in Column (5). Region dummies are classified into eight regions: 1, Hokkaido and Tohoku, 2: Kanto, 3: Hokuriku, 4: Chubu, 5: Kansai, 6: Chugoku, 7: Shikoku, 8: Kyushu and Okinawa.

References

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

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. 2015. SPGEN: Stata module to generate spatially lagged variables.
(URL: `http://econpapers.repec.org/software/bocbocode/S458105.htm`).

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

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

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

Appendix A Japanese municipal data

Japanese municipal data are taken from e-Stat, the portal site of the official statistics of Japan.[11] The municipal unemployment rates are calculated from 2010 population census. The number of criminal cases known to the police in 2009 is taken from the "2011 Statistical Observations of Shi, Ku, Machi, Mura" (Statistical Bureau of Japan). The municipal crime rate (per 1,000 people) is calculated as the number of criminal cases divided by total population. The municipal population is taken from the 2010 population census. Note that the crime rates are 1 year earlier than municipal unemployment rates. Criminal cases for municipalities that merged between 2009 and 2010 are reaggregated by new

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

municipal unit (as of 2010). Table 3 presents the descriptive statistics of Japanese municipal data on crime rates and unemployment rates.

Table 3: Descriptive Statistics of Japanese Municipal Crime Data

Variables	Mean	S.D.	Min	Max
Crime Rate (per 1,000 people)	8.684	5.353	0.000	48.398
W Crime Rate (per 1,000 people)	8.889	4.617	0.005	37.624
Unemployment Rate (%)	6.302	2.175	0.000	22.718
W Unemployment Rate (%)	6.380	1.815	0.628	18.354

Note: The number of observations is 1,728. The prefix W indicates spatial lag of the corresponding variable. The `spgen` command with `swm(pow 4)` and `dist(.)` options is used.