



RIETI Technical Paper Series 16-T-001

Introduction to Spatial Econometric Analysis: Creating spatially lagged variables in Stata

KONDO Keisuke
RIETI



Research Institute of Economy, Trade & Industry, IAA

The Research Institute of Economy, Trade and Industry
<http://www.rieti.go.jp/en/>

Introduction to Spatial Econometric Analysis: Creating spatially lagged variables in Stata*

KONDO Keisuke[†]
RIETI

Abstract

This article introduces the new Stata command `spgen`, which computes spatially lagged variables in Stata. The only additional 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 paper, I offer an interesting illustration for spatial econometric analysis using the `spgen` command.

Keywords: `spgen`, Getis–Ord $G_i^*(d)$, Spatially lagged variable, Spatial econometrics

The views expressed in the papers are solely those of the author(s), and neither represent those of the organization to which the author(s) belong(s) nor the Research Institute of Economy, Trade and Industry.

* 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 (RIETI): 1-3-1 Kasumigaseki, Chiyoda-ku, Tokyo, 100-8901, Japan. (E-mail: kondo-keisuke@rieti.go.jp)

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, 2015b).

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. 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 analysis, 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 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.

¹This method is originally employed by Kondo (2015a).

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 spatial econometric analysis. The spatial weight matrix \mathbf{W} takes the following formula:

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

where diagonal elements take the value of 0 and the sum of each row takes the value of 1 (row-standardization).

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

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

Notice that each element of spatially lagged variable $\mathbf{W}\mathbf{x}$ 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 \mathbf{x} can be defined as

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

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

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.

The case of power functional type is shown below:

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

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:

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

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

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

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

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.

3 Implementation in Stata

3.1 Syntax

```
spgen varname [if] [in] , lat(varname) lon(varname) swm(sumtype) dist(#) [ _order(#)
    dms approx detail ]
```

²A 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.

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.

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

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

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

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

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

3.3 Output

3.3.1 Outcome variables

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

`splag#_varname_swmtype` 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 #)`.

3.3.2 Stored results

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

Scalars

<code>r(N)</code>	number of observations	<code>r(td)</code>	threshold distance
<code>r(dd)</code>	distance decay parameter	<code>r(od)</code>	order of spatial lag
<code>r(dist_mean)</code>	mean of distance	<code>r(dist_sd)</code>	standard deviation of distance
<code>r(dist_min)</code>	minimum value of distance	<code>r(dist_max)</code>	maximum value of distance

Matrices

<code>r(D)</code>	lower triangle distance matrix
-------------------	--------------------------------

Macros

<code>r(cmd)</code>	<code>spgen</code>	<code>r(varname)</code>	name of variable
<code>r(swm)</code>	type of spatial weight matrix	<code>r(dist_type)</code>	exact or approximation

□ Technical note

There are some useful tips for the `spgen` command. Four 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) 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): $MP_i = \sum_{j=1}^n Y_j d_{ij}^{-1}$, for all i, j , where 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(.) 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(.)
(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(.)
. }
(output omitted)
```

□

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(.)
  (output omitted)
```

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

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

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

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

The idea of the Moran scatter plot is as follows. Consider a regression: $\mathbf{W}\mathbf{z} = \alpha\mathbf{z} + \text{residuals}$, 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 $\mathbf{W}\mathbf{z}$ and \mathbf{z} .

Figure 1 presents a Moran scatter plot of CRIME in the Columbus dataset. After implementing the `spgen` command, the `twoway scatter` command can visualize the spatial autocorrelation. As mentioned earlier, the slope through the origin in a Moran scatter plot is equal to the Moran's I . The sample code appears below:

```
. use "columbus-d.dta", clear
. egen std_CRIME = std(CRIME)
. spgen std_CRIME, lat(y_cntrd) lon(x_cntrd) swm(pow 8) dist(.)
. local VAR splag1_std_CRIME_p std_CRIME
. twoway (scatter `VAR') (lfit `VAR', est(nocon))
```

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

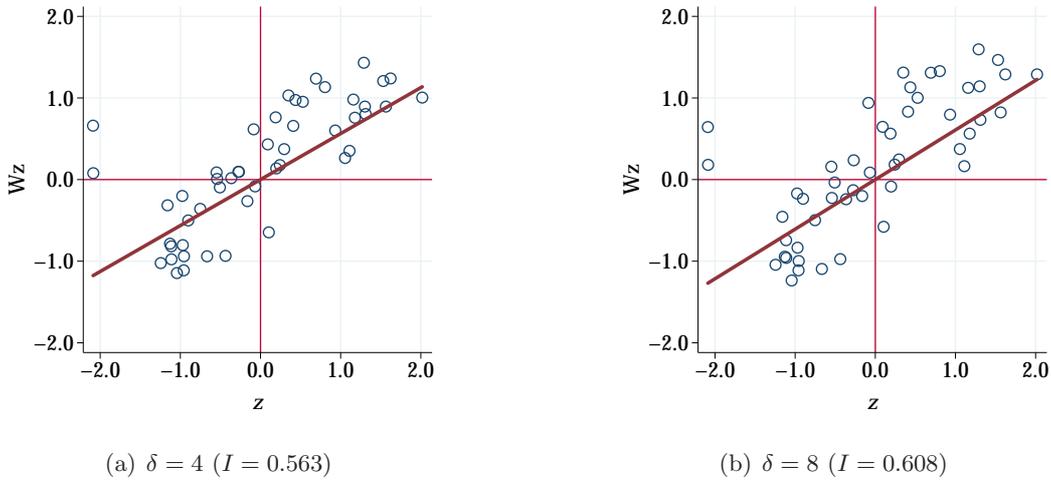


Figure 1: Moran Scatter Plot

Note: Created by the author using the Columbus dataset of Anselin (1988).

```
. 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 is estimated with the `spgen` command.

Let \mathbf{y} , \mathbf{X} , and \mathbf{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:

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

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 $\mathbf{W} \mathbf{y}$ 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).⁴ In this study, I introduce how the `spgen` helps estimation of the spatial econometric model by IV/GMM.⁵

⁴See Anselin (2006) and LeSage and Pace (2009) for further discussion.

⁵The 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.

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

$$\text{CRIME}_i = \rho \text{WCRIME}_i + \beta_1 + \beta_2 \text{INC}_i + \beta_3 \text{HOUSE}_i + u_i, \quad (5)$$

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 $\mathbf{W}\mathbf{y}$. Columns (3)–(4) show similar estimation results each other. For IV/GMM estimation, instead of the spatially lagged variable $\mathbf{W}\mathbf{y}$, instrumental variables include the first and second orders of spatial lags of explanatory variables ($\mathbf{W}\mathbf{X}$, $\mathbf{W}^2\mathbf{X}$), as suggested by Kelejian and Robinson (1993) and Kelejian and Prucha (1998, 1999).⁶

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 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.⁷ 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.⁸

In this empirical illustration, two types of spatial econometric models are estimated. The first

⁶See Anselin (2006) for a review of theoretical background on spatial econometrics.

⁷See Appendix A for more details on the Japanese municipal data.

⁸Ohtake 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

Explanatory Variables	Dependent Variable: Crime Rate (per 1,000 Households)			
	OLS (1)	OLS (2)	IV (3)	GMM (4)
W Crime Rate		0.594*** (0.105)	0.432*** (0.102)	0.448*** (0.100)
Income	-1.597*** (0.461)	-0.741 (0.478)	-0.975** (0.436)	-1.124*** (0.339)
Housing Value	-0.274* (0.163)	-0.264 (0.160)	-0.267* (0.153)	-0.231** (0.102)
Constant	68.619*** (4.233)	33.141*** (7.375)	42.828*** (6.458)	43.550*** (6.255)
Number of Observations	49	49	49	49
Adjusted R^2	0.533	0.685	0.672	0.671
Weak IV			15.930	15.930
Overidentification (p -value)			0.270	0.270

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 is $\mathbf{W}\mathbf{X}$, $\mathbf{W}^2\mathbf{X}$. 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).

specification includes the spatial lag of dependent variable as follows:

$$\text{CRIME}_i = \rho \text{WCRIME}_i + \beta_1 + \beta_2 \text{UNEMP}_i + u_i, \quad (6)$$

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:

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

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.⁹ 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.

⁹Anselin (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

Explanatory Variables	Dependent Variable: Crime Rate (per 1,000 people)				
	OLS (1)	OLS (2)	IV (3)	GMM (4)	OLS (5)
W Crime Rate		0.645*** (0.049)	0.504*** (0.116)	0.483*** (0.115)	
Unemployment Rate	0.541*** (0.060)	0.301*** (0.052)	0.353*** (0.070)	0.367*** (0.069)	0.404*** (0.075)
W Unemployment Rate					0.285*** (0.083)
Constant	4.262*** (0.480)	0.420 (0.477)	1.088* (0.648)	1.135* (0.647)	3.408*** (0.518)
Region Dummy	Yes	Yes	Yes	Yes	Yes
Number of Observations	1728	1728	1728	1728	1728
Adjusted R^2	0.292	0.483	0.474	0.471	0.297
Weak IV			40.202	40.202	
Overidentification (p -value)			0.178	0.178	

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 is $\mathbf{W}\mathbf{X}$, $\mathbf{W}^2\mathbf{X}$. 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.

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 of impact because spatial spillover effects cannot be captured. Therefore, the `spgen` command is expected to facilitate a spatial econometric analysis in Stata.

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 provide 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 information. Therefore, the `spgen` command is expected to facilitate spatial econometric analysis using Stata.

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.
- 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. 2015a. Hot and cold spot analysis using Stata. RIETI Technical Paper 15-T-001.
- . 2015b. SPGEN: Stata module to generate spatially lagged variables. (URL: <http://econpapers.repec.org/software/bocbocode/S458105.htm>).
- LeSage, J., and R. K. Pace. 2009. *Introduction to Spatial Econometrics*. Boca Raton: CRC Press.
- 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).
- Stewart, J. Q. 1947. Empirical mathematical rules concerning the distribution and equilibrium of population. *Geographical Review* 37(3): 461–485.

A Japanese municipal data

Japanese municipal data are taken from e-Stat, the portal site of the official statistics of Japan.¹⁰ 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 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.

¹⁰<https://www.e-stat.go.jp/>