

# spgen:

## Creating spatially lagged variables in Stata

Keisuke Kondo<sup>a, b</sup>

<sup>a</sup> Research Institute of Economy, Trade and Industry

<sup>b</sup> Research Institute for Economics and Business Administration, Kobe University

July 20, 2023  
2023 Stata Conference Stanford



# Outline

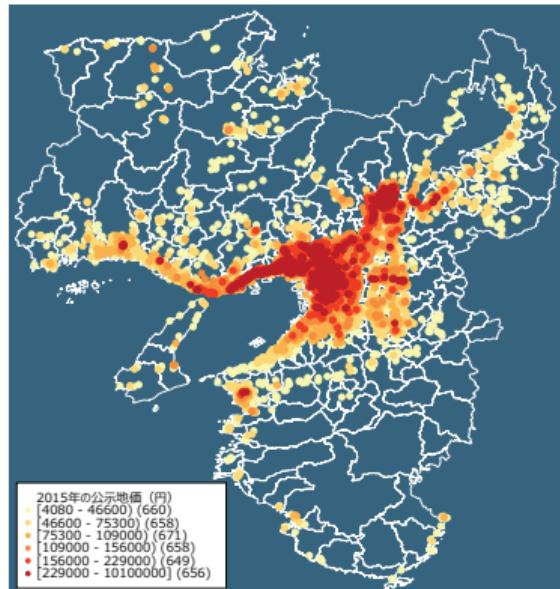
- 1 What is the spatial data?
- 2 How is the spgen command useful?
- 3 Application of the spgen command
- 4 Final remarks

Scan QR code to download this PDF file:



# 1. What is the spatial data?

- Spatial data analysis has gained attention from policy-makers, researchers, and data scientists.
- Spatial data is...(\*)
  - 1 [data structures] Information about the locations and shapes of geographic features and the relationships between them, usually stored as coordinates and topology.
  - 2 [data models] Any data that can be mapped.
- Demand for spatial data analysis is continuously growing among Stata users.



Note: Author's creation based on 2015 land price data in the Osaka metropolitan area (MLIT, Japan)

\* ESRI, spatial data, GIS Dictionary, <https://support.esri.com/en-us/gis-dictionary/spatial-data>  
(accessed July 13, 2023)

## 2. How is the spgen command useful?

In which situation is the spgen command useful?

☞ Examples of research questions:

- How do the neighboring regions affect one another?
- Whether densely populated regions have a higher risk of COVID-19 infection?
- How does local market size affect stores' sales?
- How does local market size affect firms' entry/exit decision?
- How many potential customers are there around stores?
- How many rival stores are there around stores?

## 2. How is the spgen command useful?

### ❖ Sp commands from Stata 15 (since June 2017) ❖

- The Sp commands manage data and fit regressions accounting for spatial relationships (StataCorp, 2023).
- The Sp commands support the estimation of spatial econometric models.
- The `spgenerate` command, which is provided in the Sp commands, provides similar functions to the `spgen` command.
- I would like to explain the motivation why I developed the spgen command and how it is different from the spgenerate command.

Note: The spgen command was first released in 2015 and is not the short abbreviation of the spgenerate command.

## 2. How is the spgen command useful?

### ☞ Comparison between the spgenerate and spgen commands

#### sgenerate

- It calculates the spatially lagged variable.
- It is specialized for the Sp commands (spatial econometrics).
- Users need to prepare spatial weight matrix beforehand.
- It is infeasible for high dimensional spatial weight matrices. High-spec computer is necessary.

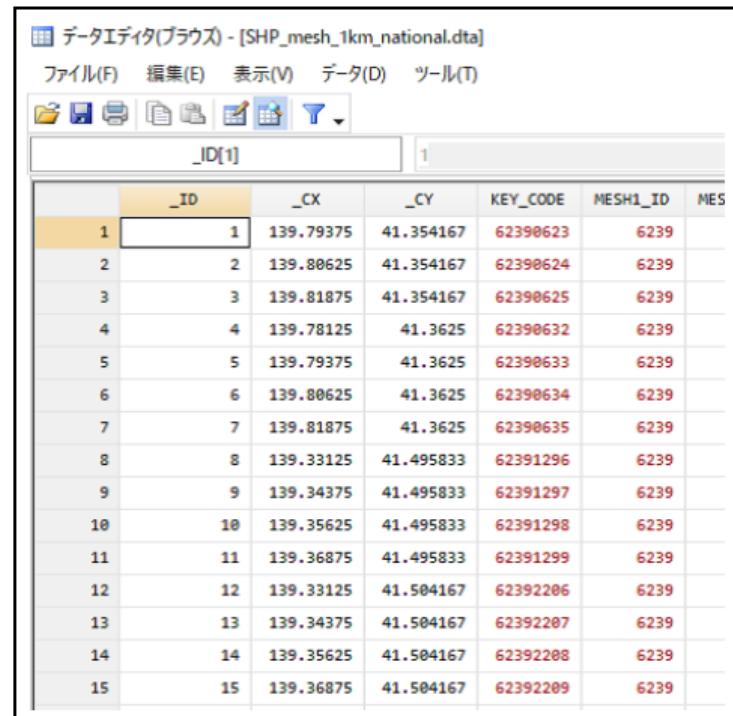
#### spgen

- It calculates the spatially lagged variable.
- It provides more flexible functions for spatial analysis.
- Users do not need to prepare spatial weight matrices beforehand.
- It is feasible for high dimensional spatial weight matrices. It works on low-spec computers.

## 2. How is the spgen command useful?

### How to use the spgen command

- Consider the dataset with location information (e.g., shape file).
- The spshape2dta command converts the shape files to the dta files.
- The variables \_CX and \_CY in this data correspond to the longitude and latitudes, respectively.



The screenshot shows a data editor window with the title "データエディタ(ブラウズ) - [SHP\_mesh\_1km\_national.dta]". The menu bar includes ファイル(F), 編集(E), 表示(V), データ(D), and ツール(T). Below the menu is a toolbar with icons for file operations. The main area displays a table with 15 rows and 6 columns. The columns are labeled \_ID, \_CX, \_CY, KEY\_CODE, MESH1\_ID, and MES. The data consists of coordinates and identifiers, with the first row highlighted in yellow.

	_ID	_CX	_CY	KEY_CODE	MESH1_ID	MES
1	1	139.79375	41.354167	62390623	6239	
2	2	139.80625	41.354167	62390624	6239	
3	3	139.81875	41.354167	62390625	6239	
4	4	139.78125	41.3625	62390632	6239	
5	5	139.79375	41.3625	62390633	6239	
6	6	139.80625	41.3625	62390634	6239	
7	7	139.81875	41.3625	62390635	6239	
8	8	139.33125	41.495833	62391296	6239	
9	9	139.34375	41.495833	62391297	6239	
10	10	139.35625	41.495833	62391298	6239	
11	11	139.36875	41.495833	62391299	6239	
12	12	139.33125	41.504167	62392206	6239	
13	13	139.34375	41.504167	62392207	6239	
14	14	139.35625	41.504167	62392208	6239	
15	15	139.36875	41.504167	62392209	6239	

## 2. How is the spgen command useful?

spgen

```
spgen varlist, lat(varname) lon(varname) swm(swmtype) dist(#) dunit(km|mi)  
          {Latitude}    {Longitude}   {SWM type}  {Dist. threshold} {Dist. unit}  
          [optional settings]
```



- ☞ The spatial lag of varlist is stored in the dataset.

## 2. How is the spgen command useful?

☞ Row-standardized matrix returns the weighted sum of neighbors.

```
. spgen x, lon(_CX) lat(_CY) swm(pow 1) dist(.) dunit(km)
```

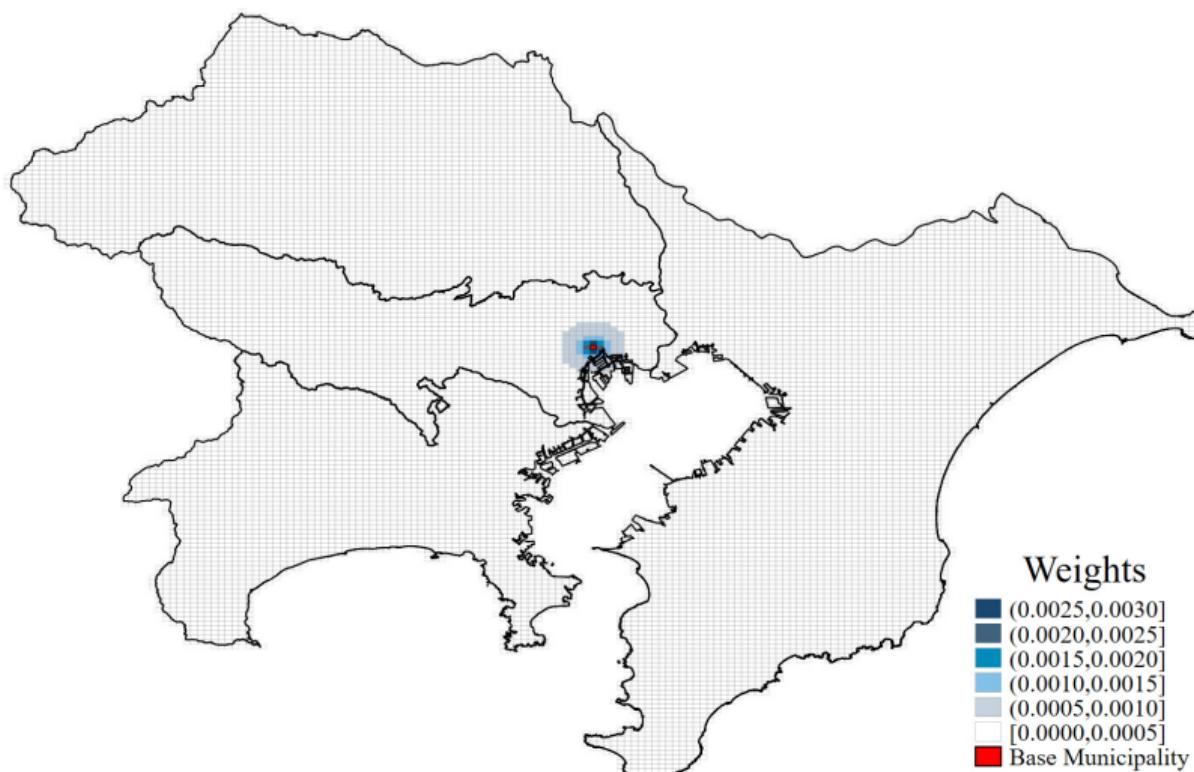
$$\mathbf{W}\mathbf{x} = \underbrace{\begin{pmatrix} 0 & w_{12} & \cdots & w_{1R} \\ w_{21} & 0 & \cdots & w_{2R} \\ \vdots & \vdots & \ddots & \vdots \\ w_{R1} & w_{R2} & \cdots & 0 \end{pmatrix}}_{\text{SWM(pow 1) dist(.)}} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_R \end{pmatrix}}_x = \underbrace{\begin{pmatrix} \sum_{k=1}^R w_{1k}x_k \\ \sum_{k=1}^R w_{2k}x_k \\ \vdots \\ \sum_{k=1}^R w_{Rk}x_k \end{pmatrix}}_{\text{Output of spgen}}$$

power function:  $d_{ij}^{-\delta}$

Note: Spatial weight matrix  $\mathbf{W}$  is often row-standardized (row-sum is equal to one) in the spatial econometrics. Market potential can be calculated with the nostd option (Harris, 1954).

## 2. How is the `spgen` command useful?

- Weights of Neighboring Municipalities (row vector of SWM for base municipality)



## 2. How is the spgen command useful?

- Local sum operator with indicator elements (row is not standardized)

. spgen x, lon(\_CX) lat(\_CY) swm(bin) dist(30) dunit(km) nostd

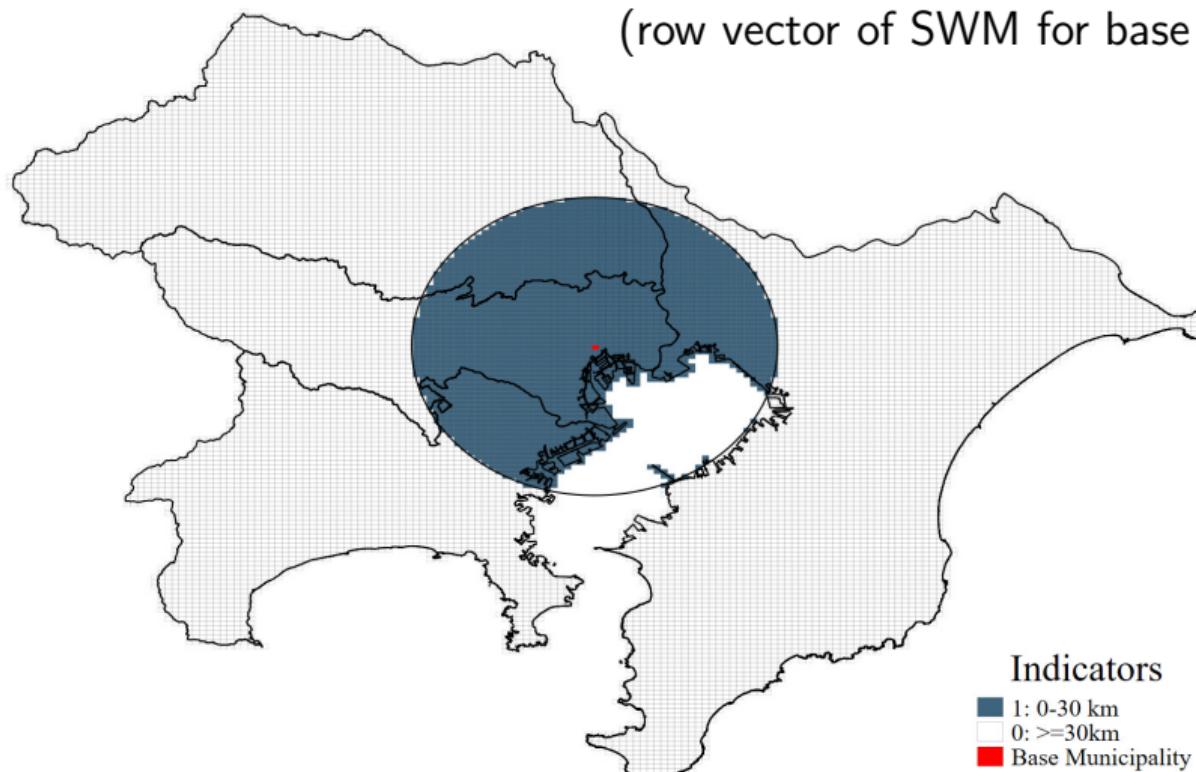
$\overbrace{\quad \quad \quad}$   
indicator function (threshold distance):  $I(d_{ij} < d)$

$$\mathbf{W}\mathbf{x} = \begin{pmatrix} 0 & I(d_{ij} < 30) & \cdots & I(d_{ij} < 30) \\ I(d_{ij} < 30) & 0 & \cdots & I(d_{ij} < 30) \\ \vdots & \vdots & \ddots & \vdots \\ I(d_{ij} < 30) & I(d_{ij} < 30) & \cdots & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_R \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^R I(d_{1k} < 30)x_k \\ \sum_{k=1}^R I(d_{2k} < 30)x_k \\ \vdots \\ \sum_{k=1}^R I(d_{Rk} < 30)x_k \end{pmatrix}$$

SWM(bin) dist(30) nostd                            x                              Output of spgen

## 2. How is the spgen command useful?

- Indicator (1/0) of neighboring municipalities within 30 km  
(row vector of SWM for base municipality)

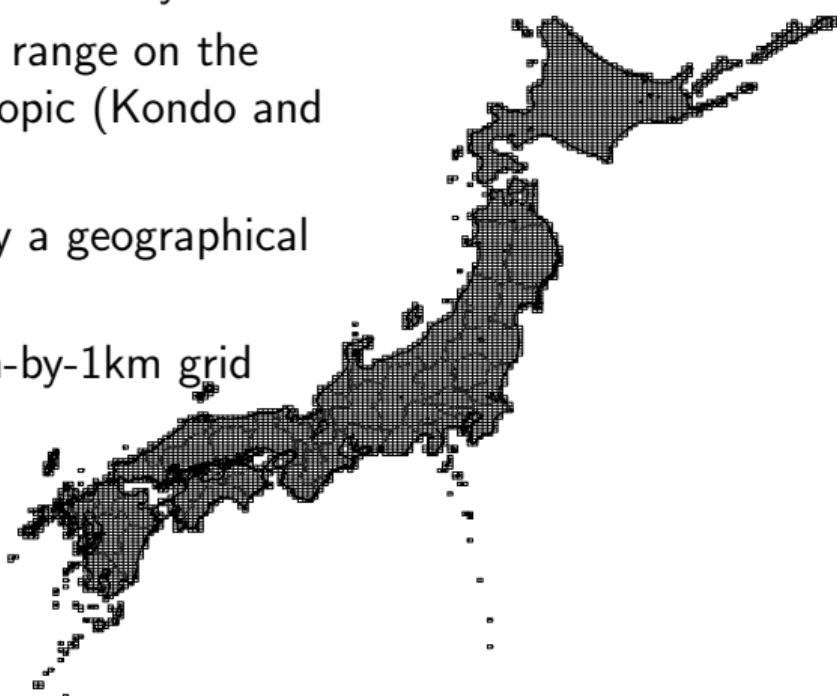


Source: Author's creation using geodist (Picard, 2010) and geocircles (Picard, 2015).

### 3. Application of the spgen command

#### Research Question: Retail Sales and Local Markets Range

- Retail sales are constrained by consumer mobility.
- Analyzing the effect of the local market range on the retail activity is an important research topic (Kondo and Okubo, 2020).
- This simple analysis attempts to identify a geographical range of local markets.
- High-dimensional mesh data at the 1km-by-1km grid level is used (391,451 obs.).

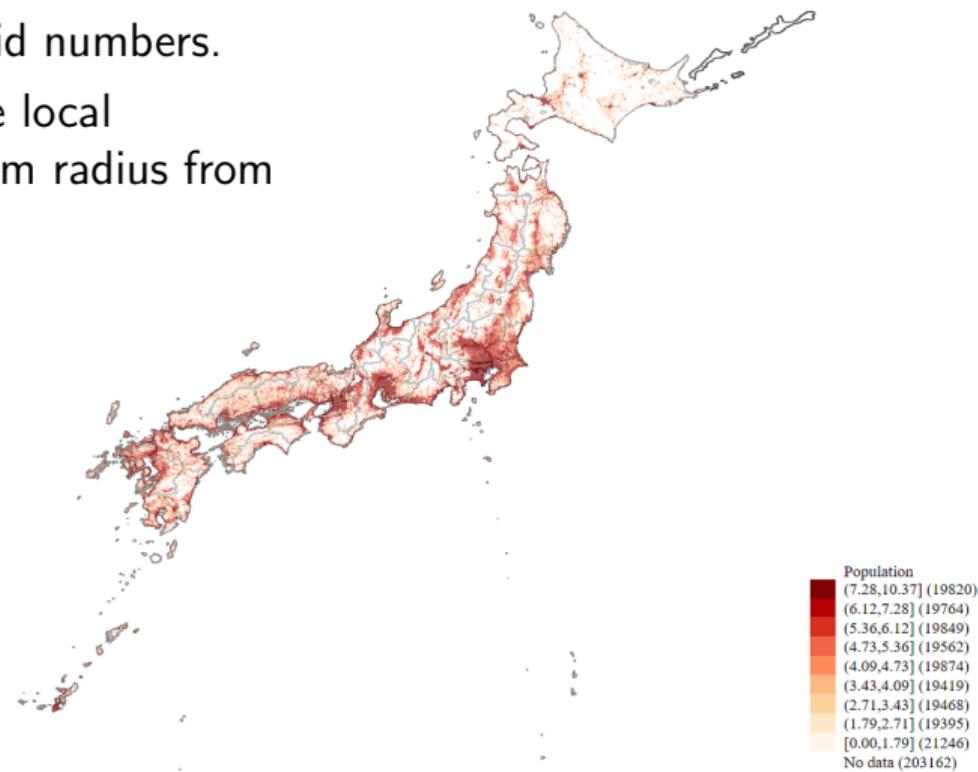


Source: Author's creation based on the shapefiles of mesh data obtained from the e-Stat (Statistical Bureau, MIC).

### 3. Application of the spgen command

#### Mesh Data: Population

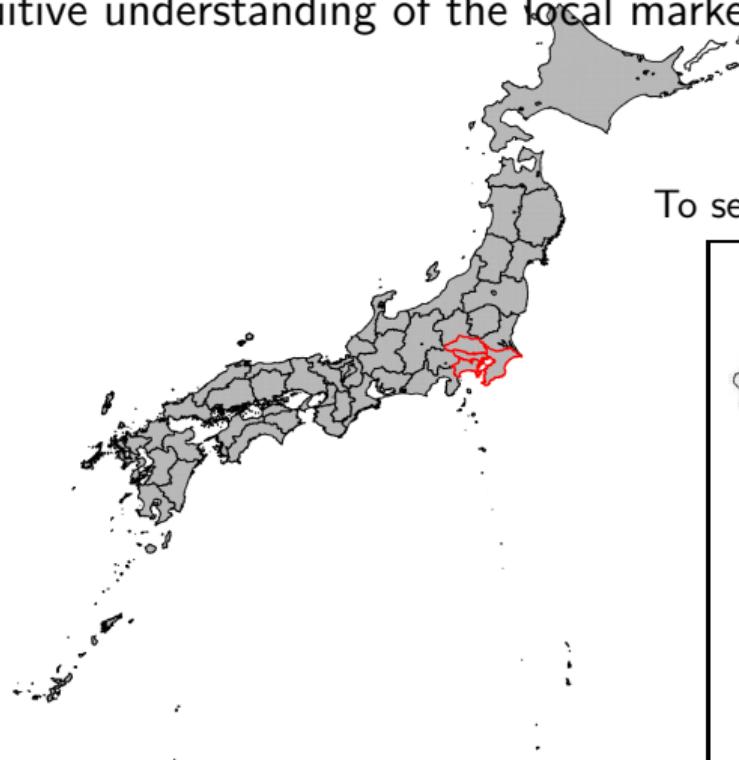
- 178,397 of 391,451 grids have valid numbers.
- Local market size is defined as the local population within the circle of  $k$  km radius from each mesh grid.



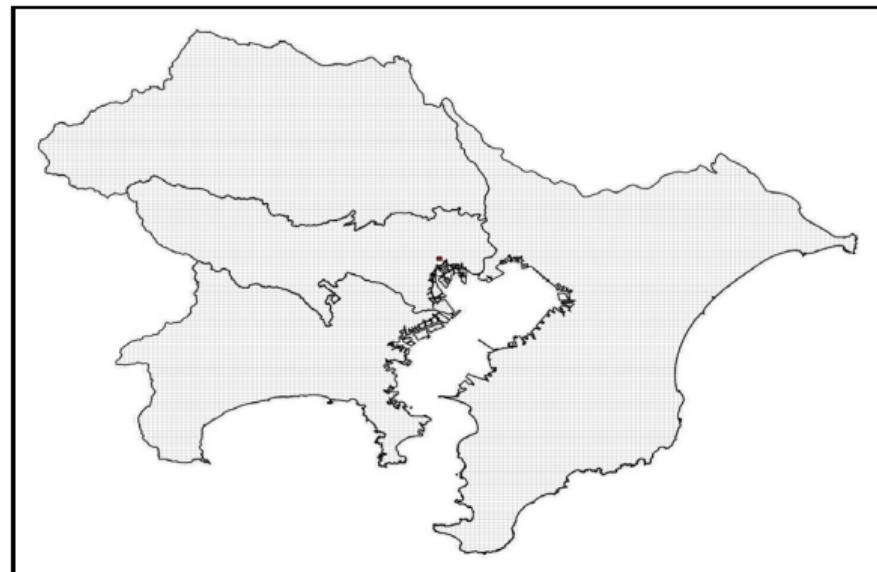
Source: Author's creation based on the 2015 Population Census (MIC).  
The red mesh grid indicates the logarithm of population.

### 3. Application of the spgen command

- ☞ Intuitive understanding of the local market range



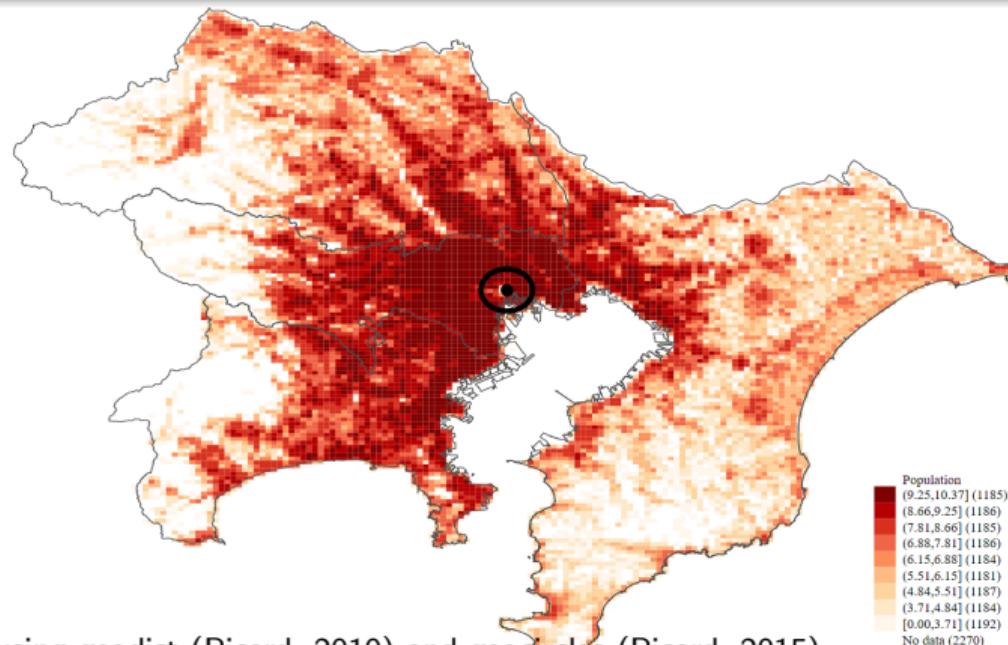
To see grid lines, we zoom in Tokyo metropolitan area.



### 3. Application of the spgen command

→ Local Market Range 0–5 km from each mesh grid

```
. gen rowif_sales_per_area = (sales_per_area > 0 & sales_per_area = .)
. spgen pop_total_all, lon(lon) lat(lat) swm(bin) nostd dist(5) dunit(km) rowif(rowif_sales_per_area)
large size replace
```



Source: Author's creation using geodist (Picard, 2010) and geocircles (Picard, 2015).

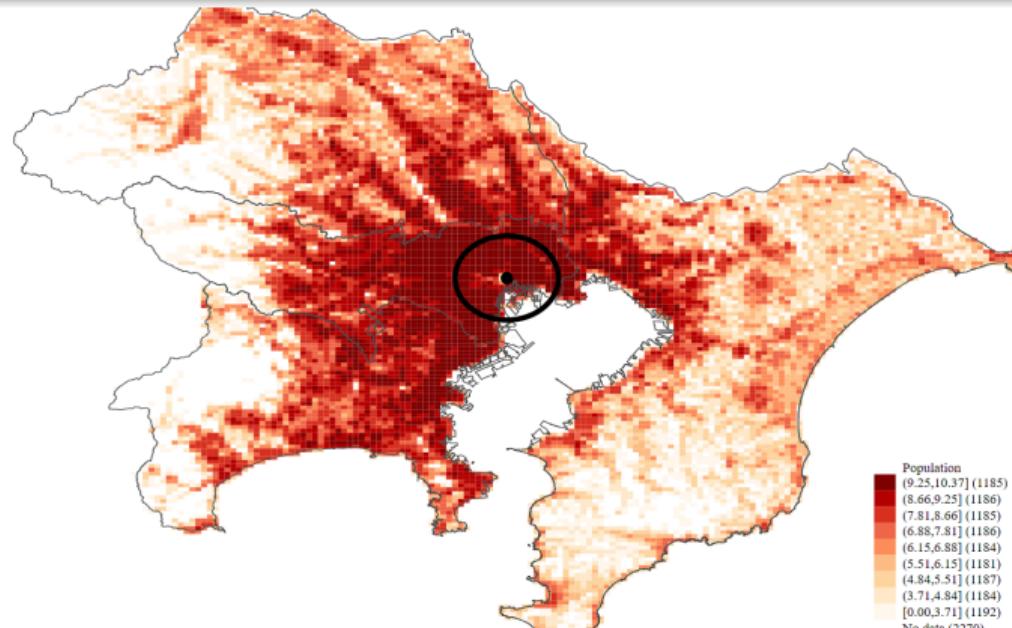
### 3. Application of the spgen command

```
. spgen pop_total_all, ///
>     lon(lon) lat(lat) swm(bin) nostd dist(5) dunit(km) rowif(rowif_sales_per_area) largesize replace
. ROWIF option returns spatial lags for observations with rowif_sales_per_area = 1
. Size of spatial weight matrix: 38334 * 381559
. 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-pop_total_all.b was generated in the dataset.
```

### 3. Application of the spgen command

→ Local Market Range 0–10 km from each mesh grid

```
. gen rowif_sales_per_area = (sales_per_area > 0 & sales_per_area = .)
. spgen pop_total_all, lon(lon) lat(lat) swm(bin) nostd dist(10) dunit(km)
rowif(rowif_sales_per_area) largesize replace
```

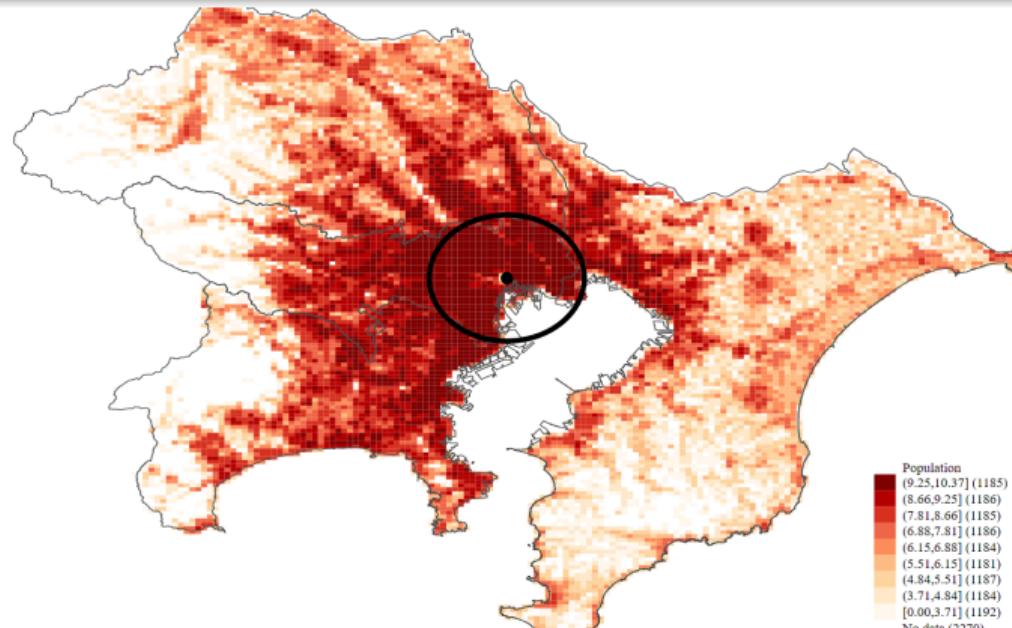


Source: Author's creation using geodist (Picard, 2010) and geocircles (Picard, 2015).

### 3. Application of the spgen command

→ Local Market Range 0–15 km from each mesh grid

```
. gen rowif_sales_per_area = (sales_per_area > 0 & sales_per_area = .)
. spgen pop_total_all, lon(lon) lat(lat) swm(bin) nostd dist(15) dunit(km)
rowif(rowif_sales_per_area) largesize replace
```

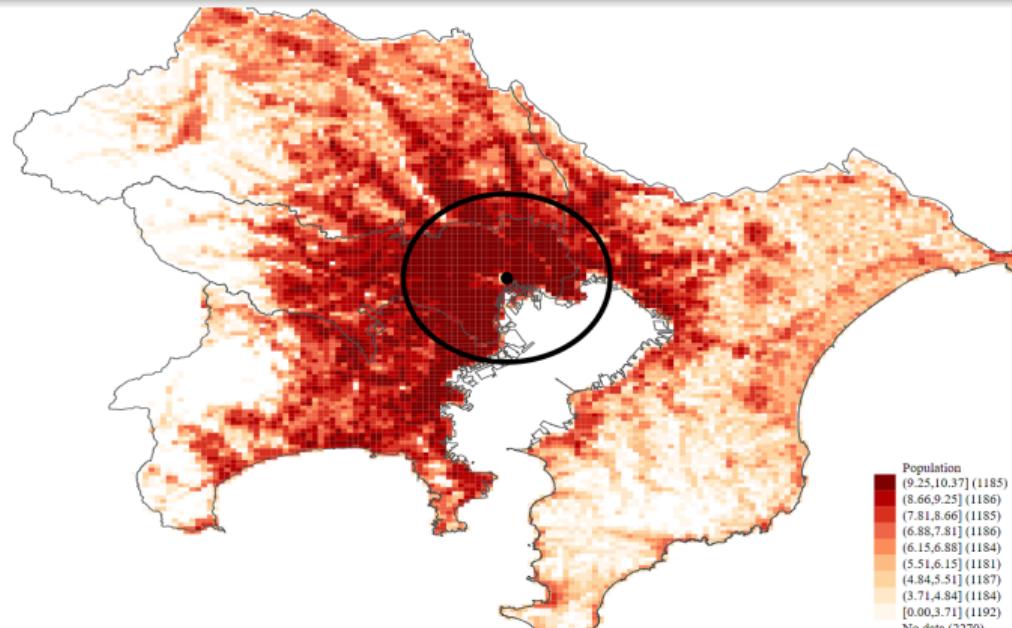


Source: Author's creation using geodist (Picard, 2010) and geocircles (Picard, 2015).

### 3. Application of the spgen command

→ Local Market Range 0–20 km from each mesh grid

```
. gen rowif_sales_per_area = (sales_per_area > 0 & sales_per_area = .)
. spgen pop_total_all, lon(lon) lat(lat) swm(bin) nostd dist(20) dunit(km)
rowif(rowif_sales_per_area) largesize replace
```

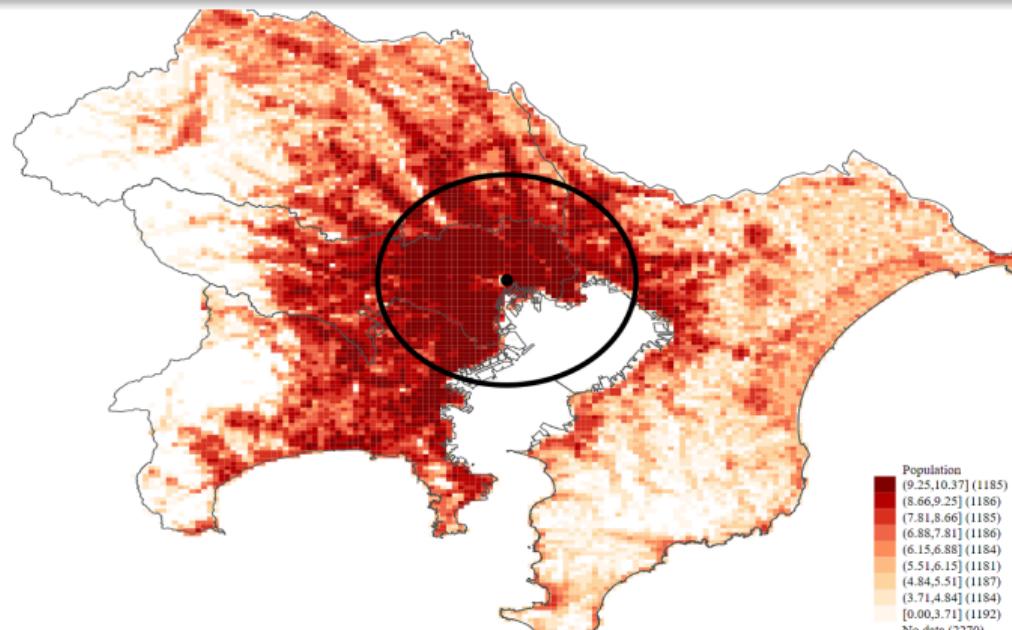


Source: Author's creation using geodist (Picard, 2010) and geocircles (Picard, 2015).

### 3. Application of the spgen command

→ Local Market Range 0–25 km from each mesh grid

```
. gen rowif_sales_per_area = (sales_per_area > 0 & sales_per_area = .)
. spgen pop_total_all, lon(lon) lat(lat) swm(bin) nostd dist(25) dunit(km)
rowif(rowif_sales_per_area) largesize replace
```

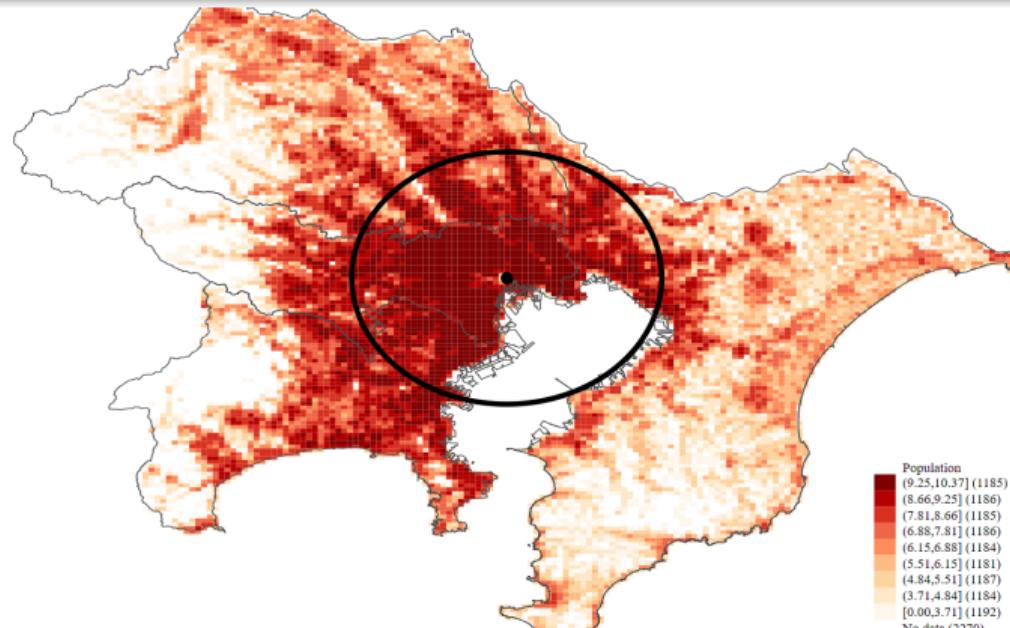


Source: Author's creation using geodist (Picard, 2010) and geocircles (Picard, 2015).

### 3. Application of the spgen command

→ Local Market Range 0–30 km from each mesh grid

```
. gen rowif_sales_per_area = (sales_per_area > 0 & sales_per_area = .)
. spgen pop_total_all, lon(lon) lat(lat) swm(bin) nostd dist(30) dunit(km)
rowif(rowif_sales_per_area) largesize replace
```

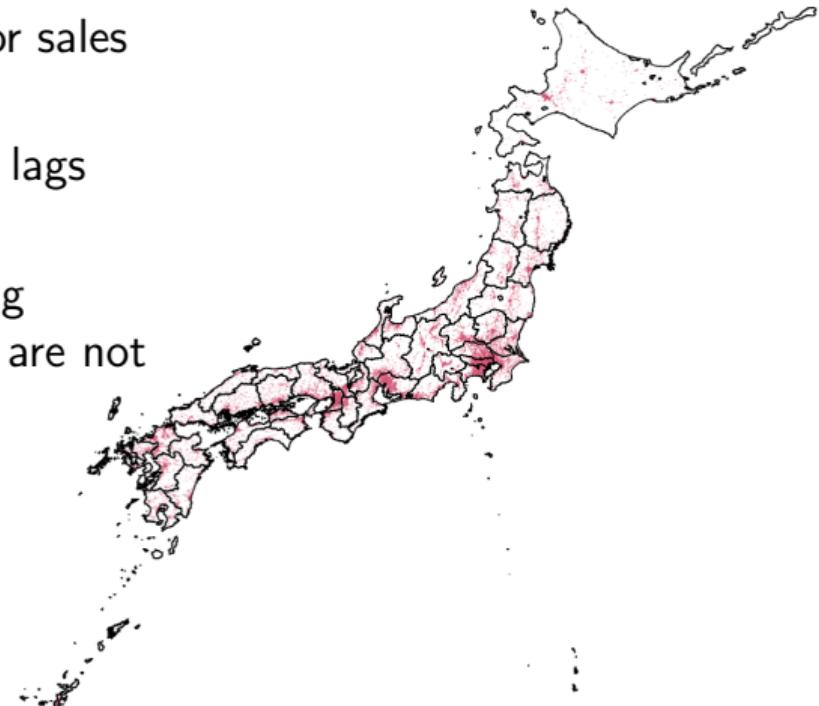


Source: Author's creation using geodist (Picard, 2010) and geocircles (Picard, 2015).

### 3. Application of the spgen command

#### Mesh Data: Sales per floor area

- The 38,334 of 391,451 grids are valid for sales per floor area.
- The `rowif()` option returns the spatial lags only for the selected observations.
- We can save calculation time by skipping observations with missing values, which are not used in the regression.



Source: Author's creation based on the 2014 Census of Commerce (METI). The red mesh grid indicates that the sales per floor area take valid values. The 35,936 mesh grids have anonymized data due to the small number of establishments in the corresponding mesh grids.

### 3. Application of the spgen command

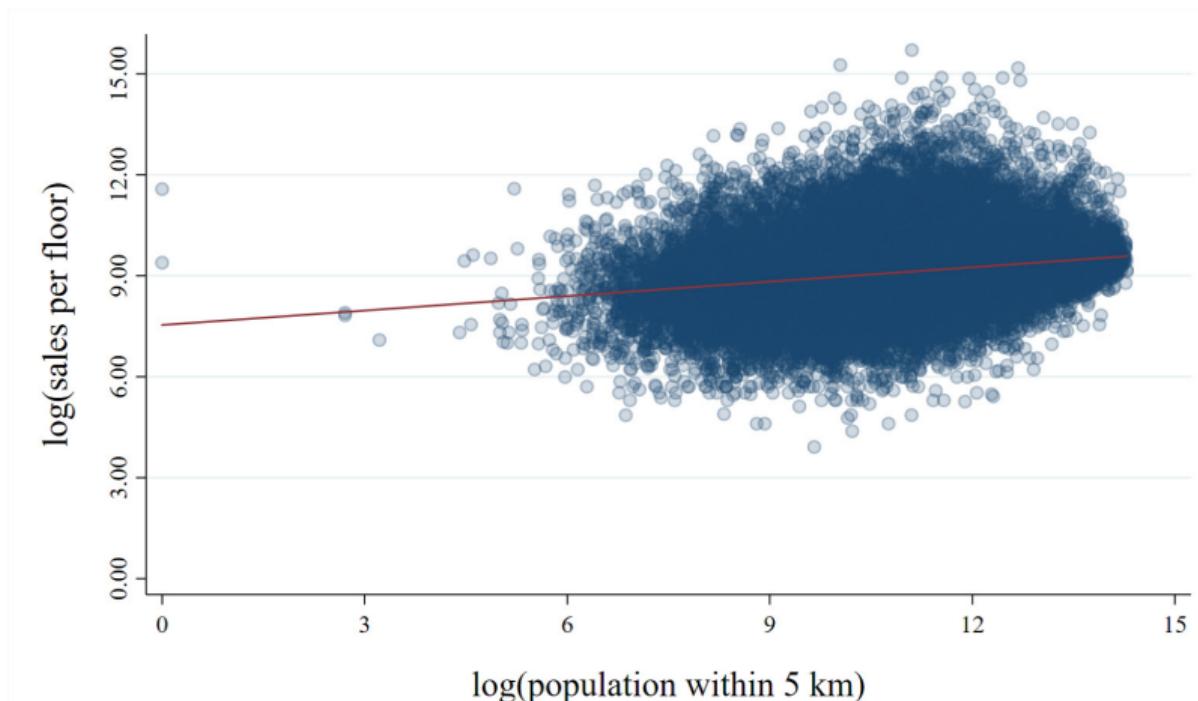


Fig: Scatter plot of sales per floor area and local market size within a 5 km circle

Source: Author's creation based on the 2014 Census of Commerce (METI) and the 2015 Population Census (MIC).

### 3. Application of the spgen command

$$\log(\text{SalesPerFloor}_i) = \beta_{0-kkm} \log(\text{LocalPopulation}_{i,0-kkm}) + \text{Controls}_i + u_i$$

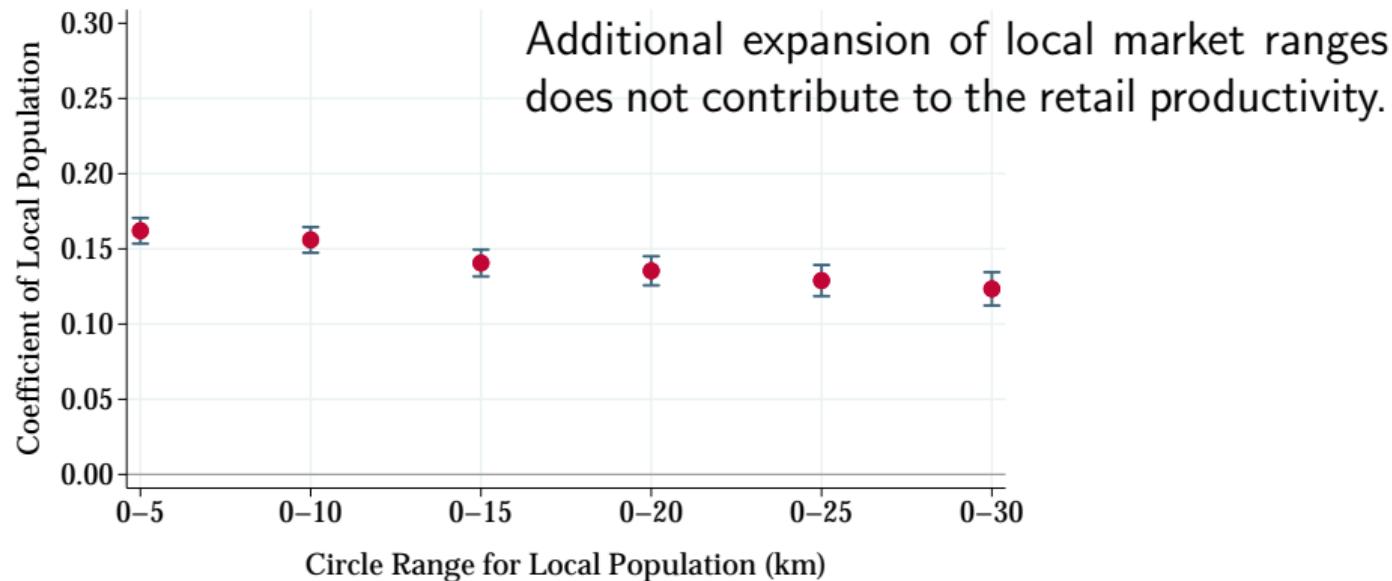


Fig: Coefficient Plot of Local Population

Source: Author's creation. Controls include prefecture dummies.

## 4. Final Remarks

- Beyond the conventional spatial econometrics, the `spgen` command helps researchers who conduct a spatial analysis using geocoded microdata and high-dimensional mesh data (big data).

`spgenerate`

- Spatial econometrics

`spgen`

- Spatial econometrics
- Geocoded microdata + Geospatial data (big data)

- A future challenge is parallel computing for speeding up the calculation.

# References

- Harris, Chauncy 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), pp. 315–348.
- Kondo, Keisuke (2015) "SPGEN: Stata module to generate spatially lagged variables," Statistical Software Components S458105, Boston College Department of Economics, revised 17 Jun 2021.
- Kondo, Keisuke and Toshihiro Okubo (2020) "The impact of market size on firm selection," RIETI Discussion Paper No. 20-E-053.
- Picard, Robert (2010) "GEODIST: Stata module to compute geographical distances," Statistical Software Components S457147, Boston College Department of Economics, revised 24 Jun 2019.
- Picard, Robert (2015) "GEOCIRCLES: Stata module to create circles defined by geographic coordinates," Statistical Software Components S457991, Boston College Department of Economics, revised 16 Aug 2015.
- StataCorp LLC (2023) *Spatial Autoregressive Models Reference Manual*, Release 18, College Station: Stata Press.

# Thank you for listening.

Any comments and suggestions would be greatly appreciated.

To install the spgen command, type:

```
ssc install spgen
```

- ✉ kondo-keisuke@rieti.go.jp
- ⌚ <https://keisukekondokk.github.io/>