콜로키움자료집 14-S09

청소년이 행복한 지역사회 지표개발 및 조성사업 연구II

콜로키움 I :
STATA를 이용한 공간분석의 기초

- 일시 : 2014년 3월 27일 10:30~17:30
- 장소 : 한국청소년정책연구원 9층 회의실
- 주최 : 한국청소년정책연구원

한국청소년정책연구원
National Youth Policy Institute

■ 제목

STATA를 이용한 공간분석의 기초

■ 발표자: 신인철(서울시립대 도시사회학과 강사)

청 소 년 정 책 연 구 원
공 간 분 석 특 강 자 료

STATA를 이용한 공간분석의 기초

2014. 3. 27

강사: 신인철
(incholshin@gmail.com/010-7455-9597)

※ 무단 복사 및 대외인용을 삼가주시기 바랍니다.

1 ArcGIS의 shape boundary files의 변환

□ ESRI의 ArcGIS 파일 소개

❍ 공간분석 연구에 이용되는 자료의 형태는 ESRI shapefile과 MapInfo Interchange 포맷이 가장 많이 통용되나, 국내에서는 전자가 많이 이용되고 있으며 통계청에서도 동 포맷 으로 자료를 제공해주고 있음

❍ ESRI 파일의 구성

- shape file인 .shp, dBASE file인 .dbf, index file인 .shx 등 3개의 파일로 구성
- 이들 파일 중에서 Stata용 자료변환에 실제로 이용되는 것은 지도상의 좌표와 관련된 .shp와 속성자료가 포함되어 있는 .dbf임

□ STATA module을 이용한 ESRI 파일의 변환

❍ STATA에서 ESRI 파일들을 직접적으로 이용할 수 없으며, 이를 위해서는 관련 자료들을 STATA용으로 변환해야 함

❍ 이를 위해 Stata 회사의 Kevin Crow(2006)이 개발한 **'shp2dta'** 모듈을 이용해야 함

- 동 모듈을 이용하기 위해서는 Stata 9.1 이상의 버전이어야 하며, .shp와 .dbf파일의 파일명들이 동일해야 함
- 개발 당시에는 폴리콘 shapefiles에만 적용되었으나, 2008년 업데이트를 통해 최근에는 점(point), 선(line) shapefiles에도 적용할 수 있음
- 동 모듈을 이용하기 위해서는 명령어창에 **"ssc install shp2dta"** 를 입력하여 다운받아야 함

❍ 'shp2dta' 명령문의 구성

```
shp2dta using shpfilename , database(filename) coordinates(filename) ///
                            replace genid(newvarname) gencentroids(stub)
```

- database( )는 자료의 속성정보가 포함되어 있는 .dbf를, coordinates( )는 좌표정보가 포함되어 있는 .shp 파일을 각각 새로운 Stata용 자료명을 지정하기 위한 것으로 반드시 포함되어야하는 옵션임
- genid( )는 각 개별 지역을 나타내는 새로운 id변수를 생성하고자 할 때 이용되는 옵션이고, gencentroids( )는 각 지역의 중심지 좌표를 생성하여 x_와 y_로 저장하도록 하기 위한 옵션

※ 자세한 내용은 도움말(help shp2dta)을 통해 확인 바람

□ STATA의 shp2dta module을 이용한 ESRI 파일의 변환의 예

❍ 자료: spatial econometrics의 대가인 Anselin(1988)이 이용한 자료로, Ohio주의 주도인 columbus의 49개 지역들의 1,000가구당 주거침입 및 차량도난 범죄율(변수명 CRIME), 가구소득(변수명 INC), 주택가치(변수명 HOVAL) 등이 포함된 자료

❍ 자료 변환의 예

- working directory에 변환하고자 하는 ESRI Map 파일이 있고, 새로운 자료도 working directory에 저장하고자 하는 경우

```
· cd "H:₩spatial analysis₩shp2dta"
· shp2dta using columbus, database(columbusdb) coordinates(columbusxy) ///
                                gencentroids(centroid) genid(nid)
```

- 특정 directory에 있는 ESRI Map 파일들을 이용해서 새로운 자료들을 새로운 directory에 저장하고자 하는 경우

```
· shp2dta using "H:₩spatial analysis₩shp2dta₩columbus", ///
                database("H:₩spatial analysis₩spmap₩columbusdb") ///
                coordinates("H:₩spatial analysis₩spmap₩columbusxy") ///
                gencentroids(centroid) genid(nid)
```

- 새로 생성된 Database 데이터의 탐색

```
· describe using "H:₩spatial analysis₩shp2dta₩columbusdb"

Contains data
  obs:            49
 vars:            21
 size:         4,753
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
AREA            double %10.0g                 AREA
PERIMETER       double %10.0g                 PERIMETER
COLUMBUS_       byte   %10.0g                 COLUMBUS_
COLUMBUS_I      byte   %10.0g                 COLUMBUS_I
POLYID          byte   %10.0g                 POLYID
NEIG            byte   %10.0g                 NEIG
HOVAL           double %10.0g                 HOVAL
INC             double %10.0g                 INC
CRIME           double %10.0g                 CRIME
OPEN            double %10.0g                 OPEN
PLUMB           double %10.0g                 PLUMB
DISCBD          double %10.0g                 DISCBD
X               double %10.0g                 X
Y               double %10.0g                 Y
NSA             byte   %10.0g                 NSA
NSB             byte   %10.0g                 NSB
EW              byte   %10.0g                 EW
CP              byte   %10.0g                 CP
THOUS           int    %10.0g                 THOUS
NEIGNO          int    %10.0g                 NEIGNO
_ID             byte   %12.0g
-------------------------------------------------------------------------------
Sorted by:  _ID
```

- 새로 생성된 coordinate 데이터의 탐색

```
· describe using "H:₩spatial analysis₩shp2dta₩columbusxy"

Contains data
  obs:         1,245
 vars:             3
 size:        26,145
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
_ID             byte   %12.0g
_X              double %10.0g
_Y              double %10.0g
-------------------------------------------------------------------------------
Sorted by:  _ID
```

<참고문헌>

Crow, Kevin. 2006. “SHP2DTA: Stata module to converts shape boundary files to Stata datasets” Stata Help Files. URL: http://ideas.repec.org/c/boc/bocode/s456718.html.

2 Visualization of spatial data using STATA

□ STATA의 **'spmap'** module 소개

❍ shp2dta module을 이용해 생성된 Stata용 자료를 이용하여 Stata에서 도면을 구현하는 방법은 크게 두 가지로 **spmap** module과 **tmap** module을 이용하는 방법임

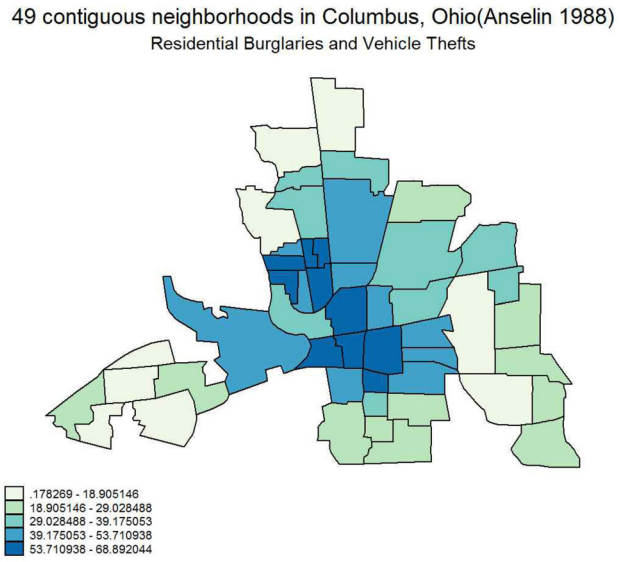
- user-wrriten 모듈인 spmap은 Maurizio Pisati(2007)에 의해 개발된 것으로, 이를 이용하기 위해서는 Stata 9.2 버전 이상이어야 함
- 만약 Stata 8.2 ~ 9.1 버전을 가진 사용자라면 Maurizio Pisati(2005)이 이전에 개발한 " tmap"이라는 모듈을 이용하여 유사한 도면구현을 실행할 수 있음

□ 단계구분도(Choropleth maps)의 예

❍ quantiles method로 나누어 구현하기

- 동 방법은 default로 clmethod()옵션을 생략해도 됨
- **clnumber** 옵션은 나눌 quantiles의 수를 지정하고, **fcolor**는 도면의 색, **ocolor**는 지역경계의 색 등을 지정하는 것임

```
· spmap CRIME using "columbuscoor.dta", id(id)              ///
      clnumber(5) fcolor(GnBu) ocolor(black) osize(medthin)      ///
      title("49 contiguous neighborhoods in Columbus, Ohio(Anselin 1988)", size(*0.8)) ///
      subtitle("Residential Burglaries and Vehicle Thefts" " ", size(*0.8))  ///
      legorder(lohi) legstyle(2) legend(ring(1))
```

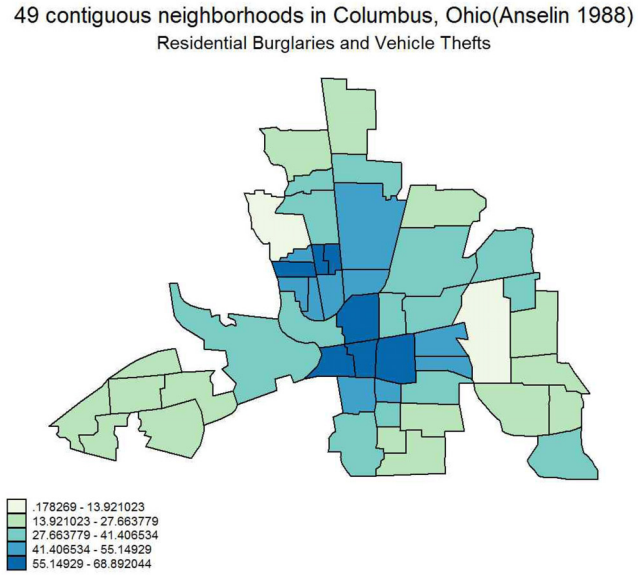


❍ equal intervals method로 나누어 구현하기

- 위와 동일한 명령문에 clmethod(eqint)옵션을 추가로 지정하면 됨
- 이전 결과를 비교해보면, 5개 구역으로 나뉘지만 범례의 범위가 다름을 확인할 수 있음

※ 이 외에도 boxplot법(boxcox), 표준편차법(stdev), k-means법(kmeans), 사용자가 범위를 지정하는 방법(custom), 값들을 모두 제시하는 방법(unique) 등이 있음

```
· spmap CRIME using "columbuscoor.dta", id(id)                ///
      clmethod(eqint) clnumber(5) fcolor(GnBu) ocolor(black) osize(medthin)      ///
      title("49 contiguous neighborhoods in Columbus, Ohio(Anselin 1988)", size(*0.8)) ///
      subtitle("Residential Burglaries and Vehicle Thefts" " ", size(*0.8))   ///
      legorder(lohi) legstyle(2) legend(ring(1))
```

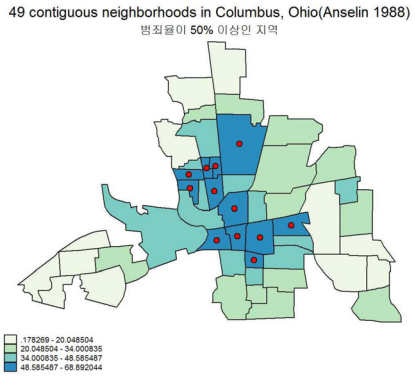


□ Dot Maps의 예

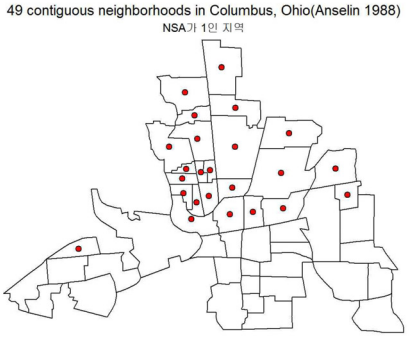
❍ point 옵션을 이용하여 point의 특성들을 지정함

- point 옵션의 하위 옵션인 x()와 y()를 통해 점의 좌표를 지정하는데, 아래의 예에서는 shp2dta module을 이용할 때 생성된 지역중심 좌표를 지정하였음
- point 옵션의 하위 옵션인 select()옵션에서는 제시하고자 하는 점들의 범위를 지정할 수 있음

```
· spmap CRIME using "columbusxy.dta", id(nid)  fcolor(GnBu) ///
      point(x(x_centroid) y(y_centroid) select(keep if inrange(CRIME, 50, 70)) ///
      fcolor(red) ocolor(black) osize(medthin))     ///
      title("49 contiguous neighborhoods in Columbus, Ohio(Anselin 1988)", size(*0.8)) ///
      subtitle("범죄율이 50% 이상인 지역", size(*0.8)) ///
      legorder(lohi) legstyle(2) legend(ring(1))
```




```
· spmap NSA using "columbusxy.dta", id(nid)  fcolor(white) ///
      point(x(x_centroid) y(y_centroid) select(keep if NSA==1) ///
      fcolor(red) ocolor(black) osize(medthin))     ///
      title("49 contiguous neighborhoods in Columbus, Ohio(Anselin 1988)", size(*0.8)) ///
      subtitle("NSA가 1인 지역", size(*0.8)) ///
      legorder(lohi) legstyle(2) legend(ring(1))
```

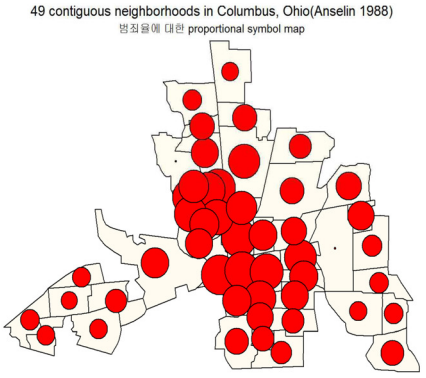


□ Proportional Symbol Maps의 예

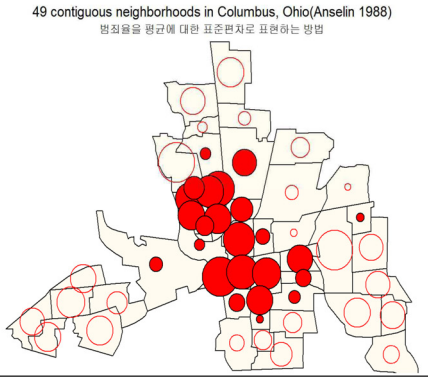
❍ point 옵션을 이용하여 point의 특성들을 지정함

- 나머지는 dot map을 구현하는 방법과 동일하나, proportional( ) 옵션이나 deviation( ) 옵션을 통해 비례적으로 symobol 크기를 정하고자 하는 변수를 지정하는 것만 다름

```
· sspmap using "columbusxy.dta", id(nid)  fcolor(eggshell) ///
      point(x(x_centroid) y(y_centroid) proportional(CRIME) ///
            fcolor(red) ocolor(black) osize(medthin))      ///
      title("49 contiguous neighborhoods in Columbus, Ohio(Anselin 1988)", size(*0.8)) ///
      subtitle("범죄율에 대한 proportional symbol map", size(*0.8))
```



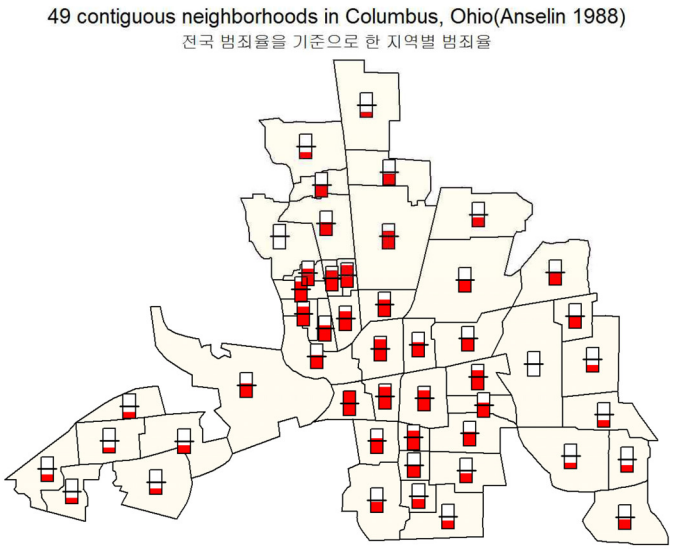

```
· spmap using "columbusxy.dta", id(nid)  fcolor(eggshell) ///
      point(x(x_centroid) y(y_centroid) deviation(CRIME) ///
          fcolor(red) ocolor(black) osize(medthin))      ///
      title("49 contiguous neighborhoods in Columbus, Ohio(Anselin 1988)", size(*0.8)) ///
          subtitle("범죄율이 50% 이상인 지역", size(*0.8))
```



□ Diagram Maps의 예

- diagram( ) 옵션을 통해 평균을 기준으로 해당 값들을 나타내고자 하는 변수를 지정하면 됨

```
· spmap using "columbusxy.dta", id(nid)   fcolor(eggshell) ///
      diagram(var(CRIME) x(x_centroid) y(y_centroid) ///
            fcolor(red) ocolor(black) osize(medthin))      ///
      title("49 contiguous neighborhoods in Columbus, Ohio(Anselin 1988)", size(*0.8)) ///
          subtitle("전국 범죄율을 기준으로 한 지역별 범죄율", size(*0.8))
```



<참고문헌>

Crow, Kevin and William Gould, "How do I graph data onto a map with spmap?" FAQ page (http://www.stata.com/support/faqs/graphics/spmap.html)

Pisati, Maurizio. 2004. "Simple Thematic Mapping," The Stata Journal 4(4): 361-378.

_____. 2008. "SPMAP: Stata Module to Visualize Spatial Data." Stata Help Files . URL: http://econpapers.repec.org/software/bocbocode/s456812.htm

_____. 2012. "Exploratory Spatial Data Analysis Using Stata." 2012 German Stata Users Group meeting, WZB social Science Research Center, Berlin.

3 Spatial Weighting Matrix의 생성과 이용

□ 공간가중행렬의 출발과 목적

❍ Whittle(1954)의 주장을 기반으로, Cliff와 Ord(1973, 1981)는 분석단위들 간의 상관관계뿐만 아니라, 소위 spatial lags라는 공간적 상호작용을 다룰 수 있는 모형을 제안함

- 이러한 형태의 모형들을 'Cliff-Ord 모형' 라고 부르는데, 여기서 spatial lag는 인접 단위들(지역들)의 가중평균(weighted average)로 정의됨

❍ 일반적인 공간자기회귀모형(spatial autoregressive model)

$$y_i = \delta \sum_{j=1}^{n} w_{ij} y_j + \varepsilon_i, \qquad i = 1, \ldots, n$$

- 여기서 y_i는 단위(지역) i의 종속변수, w_{ij}는 비확률 가중치들(nonstochastic weights)로 공간가중치라고 하며, $\sum_{i=1}^{n} w_{ij} y_j$를 spatial lag라 함

- 위의 식을 행렬형태로 나타낼 경우 아래와 같음

$$\mathrm{y} = \delta W \mathrm{y} + \varepsilon$$

$$\mathrm{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad W = \begin{bmatrix} 0 & w_{12} & \ldots & w_{1,n-1} & w_{1n} \\ w_{21} & 0 & \ldots & w_{2,n-1} & w_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ w_{n-1,1} & w_{n-1,2} & \ldots & 0 & w_{n-1,n} \\ w_{n1} & w_{n2} & \ldots & w_{n,n-1} & 0 \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- 이러한 공간가중행렬은 Tobler(1970)의 지리학 제1법칙인 "everything is related to everything else, but near things are more related than distant things"을 반영한 것이라 할 수 있음

□ 공간가중행렬 생성을 위한 STATA module들의 비교

❍ 횡단자료에 한하여 공간가중행렬을 생성할 수 있는 STATA의 module들은 크게 **spatwmat**(Pisati, 2001), **spwmatrix**(Jeanty, 2010), **sppack**의 하위 module인 **spmat**(Drukker, Peng, Prucha and Raciborski, 2011), 그리고 **spweightcs**(Shehata, 2011) 등 4가지가 개발되어 있음

- 위의 각 개발자들은 아래와 같이 STATA에서 이용 가능한 다양한 공간분석기법들을 module화하였는데, 사실 위의 module들을 이용하여 공간가중행렬을 생성할 경우에만 각 개발자들의 분석기법들을 실행할 수 있는 한계가 있음

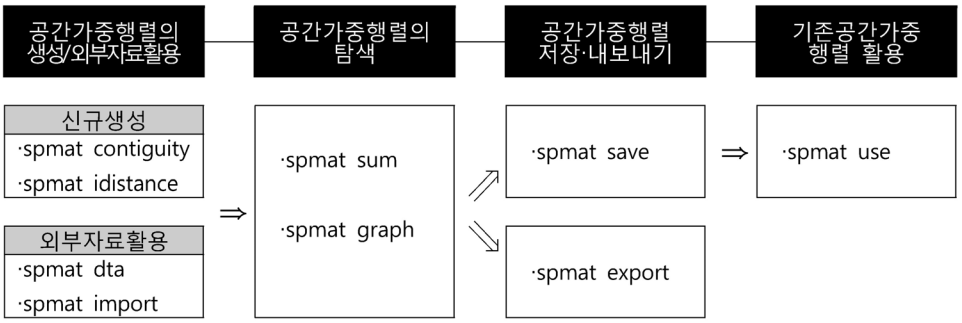
- 그럼에도 불구하고 다른 module들과의 연계성 측면에서 'spmat' module이 좀 더 유용하다고 판단됨

[횡단자료의 공간분석을 위한 STATA Module들의 비교]

	spatial weighting matrix	measures of spatial autocorrelation	Diagnostic tests for spatial dependence in OLS Regression	spatial model by ML	spatial model by IV/GMM
Pisati(2001)	.spatwmat	.spatgsa .spatlsa	.spatdiag	.spatreg	-
Jeanty(2010)	.spwmatrix	.splagvar	.anketest	.spmlreg	-
Drukker, Peng, Prucha, Raciborski (2011)	.sppack(하위모듈들의 통합패키지명)				
	.spmat	-		.spreg	.spivreg
Shehata (2011)	.spweightcs	.spautoreg			

□ 공간가중행렬 생성과 활용을 위한 spmat module의 하위 명령어 개요

❍ spmat module에 포함되어 공간가중행렬의 생성과 활용을 위해 직접적으로 관련이 있는 하위 명령어들은 크게 4가지로 구분할 수 있음



□ ‘spmat contiguity’ 명령어를 이용한 인접 공간가중행렬 생성

❍ 인접 공간가중행렬을 생성하기 위해서는 (1)neighbour criterion을 선택하고 이에 부합하는 지역들 간의 spatial neighbours 또는 공간적 인접성을 정의하고, (2)정의된 neighbour links에 가중치를 부여하는 2단계 과정을 거침(Bivand, Pebesma and Gomez-Rubio, 2008)

❍ 공간적 인접성을 나타내는 가장 기본적인 형태는 이진 연결성(binary connectivity)으로, 구성원리는 지역간의 인접성 여부에 따라 1과 0을 부여하는 형태로 되어 있는데, 마치 체스판과 동일한 구성으로 보고 가중치를 부여

- Linear Contiguity는 특정 지역의 경계선에 공유된 다른 지역의 경계선이 완전히 공유되는 경우 행렬의 원소 $W_{ij}=1$을, 다른 경우에는 $W_{ij}=0$을 부여
- Rook Contiguity는 두 지역의 경계선이 공유되기만 하면 $W_{ij}=1$을, 그렇지 않으면 $W_{ij}=0$을 부여
- Bishop Contiguity는 두 지역이 한 모서리를 공유할 경우에는 $W_{ij}=1$를 부여
- Queen Contiguity는 Rook Contiguity과 Bishop Contiguity의 합집합 형태로 두 지역이 한 면이나 모서리를 공유하면 $W_{ij}=1$, 그렇지 않으면 $W_{ij}=0$이 부여

< spatial binary contiguity matrix 생성법의 비교 >

<Linear>

0	0	0
1	0	1
0	0	0

<Rook>

0	1	0
1	0	1
0	1	0

<Bishop>

1	0	1
0	0	0
1	0	1

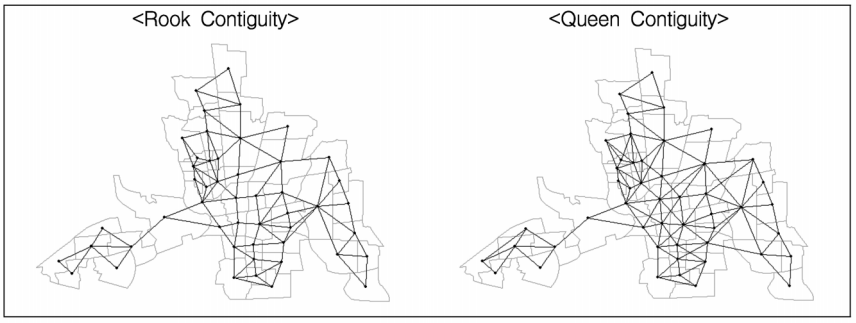
<Queen>

1	1	1
1	0	1
1	1	1

출처: 최명섭, 김의준, 박정욱(2003), p.69.

❍ 일반적으로 Rook Contiguity와 Queen Contiguity 방법이 가장 많이 활용되며, ‘spmat contiguity’ 명령어에서는 후자의 방법이 default임

< 이진연결성의 차이에 따른 공간인접행렬의 비교: Rook과 Queen >



❍ 이러한 과정을 통해 생성된 binary contiguity 공간가중행렬은 그대로 이용하기보다는 분석결과의 해석상의 용이성을 위해 정규화(normalization)하여 이용하는데, 크게 3가지 방법이 가장 많이 활용됨(Kelejian and Prucha, 2010)

유형	방법	특징
row-normalization	-공간가중행렬의 각 행의 합으로 각 행렬의 원소들을 나누어 정규화 시키는 방법 -정규화 $\widetilde{W}$의 (i,j)번째 원소 $\widetilde{w_{ij}} = w_{ij}/r_i$ * r_i: W의 i번째 행의 합	-binary contiguity 행렬은 대칭 행렬이지만 변환된 정규행렬은 비대칭행렬이 됨 -공간계량분석에서 가장 많이 활용되는 방법
minmax-normalization	-공간가중행렬의 각 행의 합들 중 최대값과 각 열의 합들 중 최대값에서 최소값으로 나누어 정규화 시키는 방법 -정규화 $\widetilde{W}$의 (i,j)번째 원소 $\widetilde{w_{ij}} = w_{ij}/m$ *$m = \min(\max_i(r_i), \max_i(c_i))$	-binary contiguity 행렬은 대칭 행렬처럼 변환된 정규행렬 역시 대칭행렬이 됨
spectral-normalization	-공간가중행렬의 가장 큰 고유값으로 나누어 정규화시키는 방법 -정규화 $\widetilde{W}$의 (i,j)번째 원소 $\widetilde{w_{ij}} = w_{ij}/v$ *v: 행렬 W의 가장 큰 고유값	-binary contiguity 행렬은 대칭 행렬처럼 변환된 정규행렬 역시 대칭행렬이 됨

- 공간가중행렬의 표준화 방법으로 row-normalization이 가장 많이 활용되고 있으나, 정규화행렬이 비대칭이 됨에 따라 단위(지역) i가 단위(지역) j에 미치는 영향과 단위(지역) j가 단위(지역) i에 미치는 영향이 달라질 뿐만 아니라, 각 지역이 점유하고 있는 상대적 위치를 무시하게 됨(Elhorst, 2013)

- 이로 인해 Kelejian and Prucha(2010)는 row-normalization 방법을 이용하기 위해서는 이론적 뒷받침이 필요하며, 이것이 없을 경우에는 minmax-normalization이나 spectral-normalization 방법을 이용할 것을 권함

❍ 인접 공간가중행렬의 생성 및 결과 비교

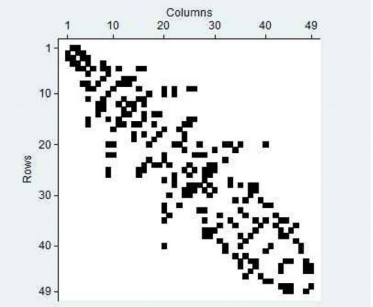
- 명령문-(1)은 binary contiguity 공간가중행렬을 생성하라는 것으로, 동 명령어를 사용하기에 앞서 stata memory에 ‘columbusdb.dta’라는 속성데이터를 먼저 불러와야 하며, 동 데이터에는 id( ) 옵션에 지정한 지역을 나타내는 변수(아래 분석에서는 ‘nid’로 지정)가 포함되어 있어야 함
- 또한, 명령문-(1)에서 ‘using’ 다음에 지정한 데이터는 앞서 ‘shp2dta’ 모듈을 통해 생성한 각 지역별 좌표가 기록된 자료이며, 속성데이터와 좌표데이터 모두 같은 working directory에 들어있기 때문에 별도의 폴더 경로를 지정하지 않음
- 명령문-(2)는 새로 생성된 공간가중행렬인 ‘spweight1.spmat’의 범위, 평균 등의 요약 통계량을 제시하도록 한 것이고, 명령문-(3)는 ‘spweight1.spmat’ 공간가중행렬의 강도 도표를 보여주도록 한 것
- 명령문-(4)는 메모리에 임시로 저장된 공간가중행렬을 .spmat이라는 확장자로 disck에 저장하도록 한 것으로 using 이후에 저장할 파일명을 지정하였으며, 향후 동 행렬은 ‘spmat use weight using nspweight1.spmat’ 등과 같은 명령문을 통해 다시 메모리로 불러 올 수 있음
- 명령문-(5)는 메모리에 임시로 저장된 공간가중행렬을 disck에 .txt 파일로 저장하도록 한 것으로 using 이후에 저장할 파일명을 지정함

```
· use columbusdb.dta, clear
· spmat contiguity spweight1 using columbusxy, id(nid) replace        .............(1)
· spmat summarize spweight1                                           ............(2)
```

Summary of spatial-weighting object spweight1

Matrix	Description
Dimensions	49 x 49
Stored as	49 x 49
Values	
min	0
min>0	1
mean	.0982924
max	1

```
· spmat graph spweight1                                               .............(3)
```




```
· spmat save spweight1 using nspweight.spmat                          .............(4)
· spmat export spweight1 ///
    using "H:₩spatial analysis₩spatial weighting matrix₩spweight_1.txt", replace  ............(5)
```

- 명령문-(6)부터 명령문-(8)은 각각 binary contiguity 공간가중행렬이 아닌 row-normalization, minmax-normalization, 그리고 spectral-normalization 방법으로 정규화시킨 후 각각 .txt 파일로 저장하도록 한 것임

- disk에 저장된 .txt파일을 열어 확인해보면 각각 정규화 방법에 따라 다른 공간가중행렬이 생성된 것을 확인할 수 있음

· * row-normalized matrix

· spmat contiguity spweight2 using columbusxy, id(nid) replace normalize(row)

· spmat export spweight2 using spweight_2.txt, replace .....(6)

```
49
1 0 .5 .5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 .333333333333333 0 .333333333333333 .333333333333333 0 0 0 0 (
3 .25 .25 0 .25 .25 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4 0 .25 .25 0 .25 0 0 .25 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5 0 0 .125 .125 0 .125 0 .125 .125 0 .125 0 0 0 .125 .125 0 0 0
6 0 0 0 0 .5 0 0 0 .5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
7 0 0 0 0 0 0 0 0 .25 0 0 0 .25 .25 .25 0 0 0 0 0 0 0 0 0 0 0 0
8 0 0 0 .166666666666667 .166666666666667 0 .166666666666667 0 (
9 0 0 0 0 .125 .125 0 0 0 .125 0 0 0 0 .125 0 0 0 0 .125 0 .125
```

* min-max matrix

· spmat contiguity spweight3 using columbusxy, id(nid) replace normalize(minmax)

· spmat export spweight3 using spweight_3.txt, replace .....(7)

```
49
1 0 .1 .1 0 0 0 0 0 0 0 0 0 0 0
2 .1 0 .1 .1 0 0 0 0 0 0 0 0 0 0 1
3 .1 .1 0 .1 .1 0 0 0 0 0 0 0 0 0
4 0 .1 .1 0 .1 0 0 .1 0 0 0 0 0 0
5 0 0 .1 .1 0 .1 0 .1 .1 0 .1 0
6 0 0 0 0 .1 0 0 0 .1 0 0 0 0 0
7 0 0 0 0 0 0 0 .1 0 0 0 .1 .1
8 0 0 0 .1 .1 0 .1 0 0 0 .1 .1
9 0 0 0 0 .1 .1 0 0 0 .1 0 0 0 1
```

* spectral-normalied matrix

· spmat contiguity spweight4 using columbusxy, id(nid) replace normalize(spectral)

· spmat export spweight4 using spweight_4.txt, replace .....(8)

```
49
1 0 .163297772449161 .163297772449161 0 0 0
2 .163297772449161 0 .163297772449161 .1632
3 .163297772449161 .163297772449161 0 .1632
4 0 .163297772449161 .163297772449161 0 .163
5 0 0 .163297772449161 .163297772449161 0 .1
6 0 0 0 0 .163297772449161 0 0 0 .163297772
7 0 0 0 0 0 0 0 .163297772449161 0 0 0 .163
8 0 0 0 .163297772449161 .163297772449161 0
9 0 0 0 0 .163297772449161 .163297772449161
```

- 다음에 제시된 명령문들은 명령문-(8)을 통해 생성된 spectral-normalized 공간가중행렬이 생성되는 과정을 검토해보기 위한 것들임
- 이를 위해 먼저 binary contiguity 공간가중행렬을 stata 데이터로 불러와야 하는데, 이전에 spweight_1.txt 파일에서 지역수를 나타내는 첫번째 행인 49 삭제 후 nspweight_1.txt로 저장 후 사용하였음

· infile a1-a49 using nspweight_1.txt, clear // text 파일을 stata에서 불러오기

· mkmat a1-a49, matrix(spectral) // stata 데이터를 matrix로 만들기

· matrix symeigen X lambda = spectral //matrix의 고유값을 만들기

· matrix list lambda // matrix의 고유값 확인하기

```
lambda[1,49]
              e1          e2          e3          e4          e5          e6          e7          e8          e9         e10         e11         e12
r1     6.1237822   5.0695208   4.3436946   3.9672266   3.4259231   3.2691997   2.8101323   2.6677192   2.3095614   2.0068857   1.6370489   1.3774978

             e13         e14         e15         e16         e17         e18         e19         e20         e21         e22         e23         e24
r1     1.221683   1.0466855    .90281145    .62557238    .50045296    .1621447    .08892515    .08698807   -.20031203   -.2626006   -.41511209   -.54050958

             e25         e26         e27         e28         e29         e30         e31         e32         e33         e34         e35         e36
r1   -.70032595  -.74009963  -.84673902          -1  -1.0187739  -1.0426076  -1.0809601  -1.1303779   -1.281979  -1.3823748  -1.5393355  -1.6573691

             e37         e38         e39         e40         e41         e42         e43         e44         e45         e46         e47         e48
r1   -1.6631141  -1.7304448  -1.8393856  -1.9195996          -2  -2.0232245  -2.0902508  -2.1948291  -2.3339655   -2.431431   -2.555971  -2.8958342

             e49
r1   -3.1259285
```

* 가장 큰 고유값인 6.1237822로 1을 나누어 그것이 명령문-(8)의 결과에 동일한지 확인

· display 1/6.1237822

```
.16329777

49
1 0 .16329777244916 1 .16329777244916 1 0 0 0
2 .16329777244916 1 0 .16329777244916 1 .1632
3 .16329777244916 1 .16329777244916 1 0 .1632
4 0 .16329777244916 1 .16329777244916 1 0 .163
5 0 0 .16329777244916 1 .16329777244916 1 0 .1
6 0 0 0 0 .16329777244916 1 0 0 0 .163297772
7 0 0 0 0 0 0 0 .16329777244916 1 0 0 0 .163
8 0 0 0 .16329777244916 1 .16329777244916 1 0
9 0 0 0 0 .16329777244916 1 .16329777244916 1
```

4 Measures of Global Spatial Autocorrelation

□ 전역적 공간자기상관(Global Spatial Autocorrelation)

- ⓐ전역적이라 함은 하나의 분석으로부터 일련의 결과들이 산출되고, 그 결과들은 연구지역 전체에 대해 동일하게 적용되는 것으로 간주됨을 의미하고, ⓑ공간적 자기상관은 한 지역이 가지는 속성값은 그 지역과 이웃하는 다른 지역간의 상호작용으로 인해 영향을 받는다는 개념임(Pfeiffer et al., 2008)
- 따라서 전역적 공간자기상관 지수는 연구지역내 단위구역들(areal units)의 속성 값이 그 근린(neighborhood)의 값들과 유사한 평균 정도를 일련의 표준 수식에 의해 정량적으로 측정하는 것

□ Moran의 전역적 공간자기상관지수 I

- 전역적 공간자기상관을 측정하는 대표적인 지수인 Moran의 I는 공분산 개념을 사용하여 속성값 사이의 유사성을 측정(Moran, 1948)

- 산출식 및 추론

$$I = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} Z_i Z}{S_0 m_2} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\text{(식-1)}$$

  - 여기에서 w_{ij}는 위치쌍 (i,j)에 대한 공간가중행렬 W의 원소이고, $Z_i = Y_i - \overline{Y}$로서 Y_i는 지역 i의 관심변수 Y를, $\overline{Y}$는 전체 지역에 대한 Y의 평균을 나타냄. 그리고 $S_0 = \sum_i \sum_j w_{ij}$와 $m_2 = \sum_i Z_i^2 / N$으로 정의됨
  - 전역적 공간자기상관이 없다는 귀무가설 하에서 I에 대한 기대값은 아래의 (식-2)와 같음

$$E(I) = -1/(N-1) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\text{(식-2)}$$

  - 만약 I가 기대값보다 크다면[$I > E(I)$], 변수Y의 전체 분포는 정적인 공간자기상관을 갖는 것으로 각 지역 i에서 Y값은 그곳과 공간적으로 인접한 지역의 Y값과 유사한 경향이 높다는 것을 의미함. 반대로 I가 기대값보다 작다면[$I < E(I)$], 변수Y의 전체 분포는 부적인 공간자기상관을 갖는 것으로 각 지역 i에서 Y값은 그곳과 공간적으로 인접한 지역의 Y값과 다른 경향이 높다는 것
  - 추론은 아래의 (식-3)과 같이 z값에 기초하는데, I와 그 기대값 $E(I)$의 차이를 I의 표준편차인 $sd(I)$로 나눈 것으로, $sd(I)$는 확률화 가정(randomization assumption)하에서 계산되기 때문에, 이러한 가정을 통해 z_I는 점근적으로 정규분포를 따름(Anselin, 1995).

$$z_I = \frac{I - E(I)}{sd(I)} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\text{(식-3)}$$

□ Geary의 전역적 공간자기상관지수 c

○ Geary(1954)가 제안한 것으로 다음의 (식-4)을 따름

$$c = (N-1)\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(Z_i - Z_j)^2}{2NS_0 m_2} \quad \cdots\cdots(식-4)$$

- 전역적 공간자기상관이 없다는 귀무가설 하에서 c의 기대값은 1
- Moran의 I와는 달리 c가 1보다 크면($c > 1$), 변수Y의 전체 분포는 부적인 공간자기상관을 갖는 것으로 각 지역 i에서 Y값은 그곳과 공간적으로 인접한 지역의 Y값과 다른 경향이 높다는 것을 의미
- 반대로 1보다 작다면($c < 1$), 변수Y의 전체 분포는 정적인 공간자기상관을 갖는 것으로 각 지역 i에서 Y값은 그곳과 공간적으로 인접한 지역의 Y값과 유사한 경향이 높다는 것을 의미

○ 하지만 c 역시 추론은 아래의 (식-5)과 같이 z값에 기초하며 c와 그 기대값 1의 차이를 c의 표준편차인 $sd(c)$로 나눈 것으로, $sd(c)$도 확률화 가정(randomization assumption)하에서 계산되기 때문에, 이러한 가정을 통해 z_c는 점근적으로 정규분포를 따름(Anselin, 1995)

$$z_I = \frac{I - E(I)}{sd(I)} \quad \cdots\cdots(식-5)$$

□ Geary와 Ord의 전역적 공간자기상관지수 G

○ 위의 두 지수와는 다소 다른 전역적 공간자기상관으로 Getis와 Ord(1992)가 제안한 G는 아래의 (식-6)을 통해 산출됨

$$G = \frac{\sum_{i \neq j} w_{ij} Y_i Y_j}{\sum_{i \neq j} Y_i Y_j} \quad \cdots\cdots(식-6)$$

- **Y는 양의 값만을 가질 수 있고, w_{ij}는 $w_{ii} = 0$인 비표준화된 대칭이산가중행렬(nonstandardized symmetric binary weights matrix)의 원소**를 나타냄

○ 전역적 공간자기상관이 없다는 귀무가설 하에서 G의 기대값은 아래의 (식-7)과 같은데, G는 Y값들의 전체적인 공간군집성(spatial clustering)을 측정하는 것으로 앞서 살펴본 I와 c와는 다소 다른 방식으로 해석됨

$$E(G) = \frac{S_0}{N(N-1)} \quad \cdots\cdots(식-7)$$

- G가 기대값보다 크다면, 값이 높은 군집들(high-values clusters)이 많은 정적인 공간자기상관을 갖음을 나타냄
- 반면, 기대값보다 작다면, 정적인 공간자기상관을 갖고는 있으나, 값이 낮은 군집들(low-values clusters)이 많은 경우를 말함

❍ 하지만 G 역시 추론은 아래의 (식-8)과 같이 z값에 기초하며, G와 그 기대값의 차이를 표준편차 $sd(G)$로 나눈 것으로, z_G도 점근적으로 정규분포를 따름

$$z_G = \frac{G - E(G)}{sd(G)} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\text{(식-8)}$$

□ STATA의 'spatgsa' module을 이용한 전역적 공간자기상관지수 산출

❍ 적용되는 공간가중행렬에 따른 공간자기상관계수의 차이 비교

- Pisati(2001)가 개발한 'spatgsa' module을 이용하여 전역적 공간자기상관지수 산출하기 위해서는 번거롭지만, 'spatwmat' module을 통해 공간가중행렬을 생성하고 이것이 memory에 임시로 저장되어 있어야 함

※ 분석에 앞서 해당 모듈을 다운받아 이용해야 함

- 이에 앞서 'spmat' module을 통해 생성했던 공간가중행렬 중 binary contiguity matrix와 minmax-normalized matrix에서 첫 행인 지역수를 나타내는 수를 삭제하고 전자는 swm_nst.dta, 후자는 swm_st.dta 등 STATA 자료로 변환한 후 사용함
- 특히, 'spatwmat' module이 가지고 있는 한계점은 중심좌표값을 이용하지 않을 경우에는 폴리곤의 좌표값을 이용하여 자체적으로 인접행렬을 산출할 수 없다는 것임

```
/* spatwmat module을 이용하여 3개의 공간자기 상관행렬 생성 */
* binary contiguity matrix
· spatwmat using swm_nst.dta, name(BW)

* row-normalized spatial weighting matrix
· spatwmat using swm_nst.dta, name(RSW) standardize

* spectral-normalized spatial weighting matrix
· spatwmat using swm_st.dta, name(MMSW)
```

- 전역적 공간자기상관계수를 산출하기 위해서는 분석하고자 하는 속성데이터를 불어와야 하는데, 본 분석에서는 앞서 이용한 'columbusdb.dta'를 활용함
- 아래의 명령문들은 CRIME, INC, HOVAL 등 3개의 변수에 대해 전역적 공간자기상관계수를 산출토록 한 것으로, 여기서 유의해야 할 것은 다른 지수들과는 달리 Geary와 Ord의 G는 정규화시키지 않은 binary matrix여야 한다는 것임

```
· use columbusdb, clear
· spatgsa CRIME INC HOVAL, weight(BW) moran geary go
```

<위에서 계속>

Measures of global spatial autocorrelation

Weights matrix

Name: BW
Type: Imported (binary)
Row-standardized: No

Moran's I

Variables	I	E(I)	sd(I)	z	p-value*
CRIME	0.515	-0.021	0.086	6.212	0.000
INC	0.412	-0.021	0.085	5.102	0.000
HOVAL	0.221	-0.021	0.084	2.870	0.002

Geary's c

Variables	c	E(c)	sd(c)	z	p-value*
CRIME	0.592	1.000	0.108	-3.795	0.000
INC	0.714	1.000	0.130	-2.200	0.014
HOVAL	0.789	1.000	0.137	-1.539	0.062

Getis & Ord's G

Variables	G	E(G)	sd(G)	z	p-value*
CRIME	0.128	0.100	0.006	4.638	0.000
INC	0.100	0.100	0.005	-0.067	0.473
HOVAL	0.099	0.100	0.006	-0.311	0.378

*1-tail test

- 만약 Geary와 Ord의 G를 산출하는데 있어 정규화된 공간가중행렬을 설정할 경우 다음과 같은 error 메시지를 받게 됨

· spatgsa CRIME INC HOVAL, weight(MMSW) go

```
Spatial weights in matrix MMSW are not binary.
To compute Getis and Ord's G you must use a
non-standardized binary weights matrix
```

- 다음은 공간가중행렬의 유형에 따른 전역적 공간자기상관계수의 값의 차이를 비교한 것인데, row-normalized matrix를 적용했을 때만 제외하고 모두 동일한 값을 산출하고 있음을 알 수 있음

<binary contiguity matrix 적용시>

Moran's I

Variables	I
CRIME	0.515
INC	0.412
HOVAL	0.221

Geary's c

Variables	c
CRIME	0.592
INC	0.714
HOVAL	0.789

<spectral-normalized matrix 적용시>

Moran's I

Variables	I
CRIME	0.515
INC	0.412
HOVAL	0.221

Geary's c

Variables	c
CRIME	0.592
INC	0.714
HOVAL	0.789

Measures of global spatial autocorrelation

Weights matrix

Name: MMSW
Type: Imported (non-binary)
Row-standardized: No

<mimmax-normalized matrix 적용시>

Moran's I

Variables	I	E(I)	sd(I)	z	p-value*
CRIME	0.515	-0.021	0.086	6.212	0.000
INC	0.412	-0.021	0.085	5.102	0.000
HOVAL	0.221	-0.021	0.084	2.870	0.002

Geary's c

Variables	c	E(c)	sd(c)	z	p-value*
CRIME	0.592	1.000	0.108	-3.795	0.000
INC	0.714	1.000	0.130	-2.200	0.014
HOVAL	0.789	1.000	0.137	-1.539	0.062

*1-tail test

<row-normalized matrix 적용시>

Moran's I

Variables	I
CRIME	0.500
INC	0.416
HOVAL	0.180

Geary's c

Variables	c
CRIME	0.541
INC	0.676
HOVAL	0.806

<참고문헌>

Anselin, L. 1995. “Local indicators of spatial association-LISA.” *Geographical Analysis* 27: 93-115.

Pfeier, D., Robinson, T., Stevenson, M., Stevens, K., Rogers, D. and A. Clements. 2008. *Spatial Analysis in Epidemiology*. Oxford: Oxford University Press.

Pisati, M. 2001. sg162: Tools for spatial data analysis. Stata Technical Bulletin 60: 21-7. College Station, TX: Stata Press.

_____. 2012. “Exploratory Spatial Data Analysis Using Stata.” 2012 German Stata Users Group meeting, WZB social Science Research Center, Berlin.

5 Measures of Local Spatial Autocorrelation

□ 지역적 공간자기상관(Local Spatial Autocorrelation)의 유용성

○ 전역적 공간자기상관계수는 관심있는 변수의 평균적인 공간 분포를 나타내주는 것이기 때문에, 각 지역들 간에 유사성 또는 이질성을 보이고 있는가를 확인할 수 없다는 한계가 있었으며 이를 보완한 방법이 지역적 공간자기상관계수임

□ 지역적 공간자기상관계수들의 비교

○ Moral의 지역적 공간자기상관계수 I_i

$$I_i = \sum_{j=1}^{N} w_{ij}^{std} \left(\frac{y_i - \bar{y}}{\sigma_y} \right) \left(\frac{y_j - \bar{y}}{\sigma_y} \right)$$

- 위의 식에서 σ_y는 분석하고자 하는 변수 Y의 표준편차이며, 특히 여기서 유의해야 할 것은 전역적 공간자기상관계수와 달리 **"row-normalized spatial weights matrix"**가 사용된다는 것임

- 만약 지역 r_i의 지역적 공간자기상관계수가 기댓값보다 크다면[$I_i > E(I_i)$], 이것은 지역 r_i가 변수Y에 대해 r_i와 유사한 지역들로 둘러싸여 있음을 나타냄(**positive autocorrelation**)
  a) 만약 $(y - \bar{y}) > 0$이면, r_i를 "hot spot"이라고 함
  b) 만약 $(y - \bar{y}) < 0$이면, r_i를 "cold spot"이라고 함

- 만약 지역 r_i의 지역적 공간자기상관계수가 기댓값보다 작다면[$I_i < E(I_i)$], 이것은 지역 r_i가 변수Y에 대해 r_i와 다른 지역들로 둘러싸여 있음을 나타냄(**negative spatial autocorrelation**)

○ Getis와 Ord의 지역적 공간자기상관계수 G_i와 G_i^*

$$G_i = \frac{\sum_{j=1}^{N} w_{ij} y_j}{\sum_{j=1}^{N} y_j} \quad (i \neq j)$$

$$G_i^* = \frac{\sum_{j=1}^{N} w_{ij} y_j}{\sum_{j=1}^{N} y_j}$$

- 위 두 공식에서는 차이는 공간가중행렬의 차이로 G_i에서는 $w_{ii} = 0$인데 반해, G_i^*는 $w_{ii} = 1$임에 유의할 필요가 있음

- 하지만, 양자 모두 앞서 전역적 공간자기상관계수인 G와 마찬가지로 정적 공간자기상관관계(positive spatial autocorrelation)만을 나타냄
  a) $G_i > E(G_i)$ 또는 $G_i^* > E(G_i^*)$일 경우, 지역r_i는 "hot spot"을 나타냄
  b) $G_i < E(G_i)$ 또는 $G_i^* < E(G_i^*)$일 경우, 지역r_i는 "cold spot"을 나타냄

□ STATA의 'spatlsa' module을 이용한 지역적 공간자기상관지수 산출

❍ Moran I_i는 row-normalized matrix를 이용한다는 점을 유의해서 지역적 공간자기상관계수를 산출함

- Z-test(단측검정임에 유의) 결과에서 유의미한 지역들을 선택하여 해당 지역들이 positive spatial autocorrelation을 보이고 있는지를 확인한 후, 지역별 CRIME 변수의 값과 지역평균값을 비교해서 hot spot 또는 cold spot 여부를 결정

· spatlsa CRIME, weight(RSW) moran

```
Weights matrix
-------------------------------------------------------------
Name: RSW
Type: Imported (binary)
Row-standardized: Yes
-------------------------------------------------------------

Moran's Ii (CRIME)
```

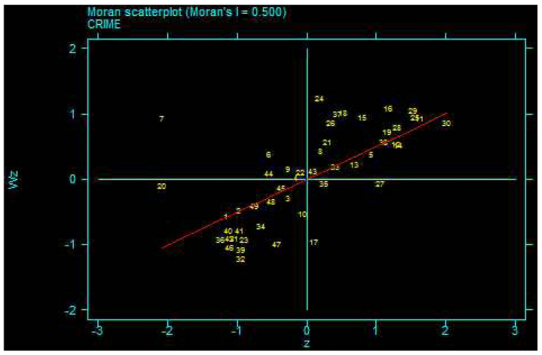
Location	Ii	E(Ii)	sd(Ii)	z	p-value*	
1	0.737	-0.021	0.691	1.097	0.136	
2	0.529	-0.021	0.558	0.985	0.162	
3	0.094	-0.021	0.478	0.240	0.405	
4	0.005	-0.021	0.478	0.054	0.479	
5	0.304	-0.021	0.323	1.006	0.157	negative
6	-0.182	-0.021	0.691	-0.233	0.408	spatial autocorrelation
7	-1.861	-0.021	0.478	-3.850	0.000	
8	0.077	-0.021	0.381	0.255	0.399	
9	-0.031	-0.021	0.323	-0.030	0.488	positive
10	0.040	-0.021	0.478	0.128	0.449	spatial autocorrelation
11	1.458	-0.021	0.423	3.498	0.000	
12	0.639	-0.021	0.381	1.730	0.042	$y_{11}(62.3)$가 $\bar{y}(35.1)$보다
13	0.125	-0.021	0.478	0.306	0.380	크기 때문에 Hot Spot
14	0.615	-0.021	0.381	1.667	0.048	
15	0.735	-0.021	0.381	1.981	0.024	
16	1.239	-0.021	0.323	3.904	0.000	positive
17	-0.108	-0.021	0.558	-0.156	0.438	spatial autocorrelation
18	0.517	-0.021	0.478	1.126	0.130	
19	0.800	-0.021	0.558	1.471	0.071	$y_{16}(54.8)$가 $\bar{y}(35.1)$보다
20	0.341	-0.021	0.281	1.287	0.099	크기 때문에 Hot Spot

- 이하 생략 -

○ 이러한 결과를 좀 더 가시적으로 제시하기 위해 graph 옵션을 통해 Moran Scatterplot을 제시함

- 아래의 명령문에서 symbol( ) 옵션은 각 지역들을 나타내는 방법으로 n이 아닌, id( ) 옵션과 함께 id로 바꿀 경우 지역 id들이 scatterplot에 나타남
- 동 분석에서는 지역순서와 id가 동일해서 둘 다 동일한 결과를 보여줌

```
· spatlsa CRIME, weight(RSW) moran graph(moran) symbol(n)
```



○ 하지만 유의결과는 어떤 지역이 통계적으로 유의미한 지역들인지 알 수 없기 때문에 Pisati(2012)가 제공한 STATA용 'genmsp' program을 이용하면 아래와 같음

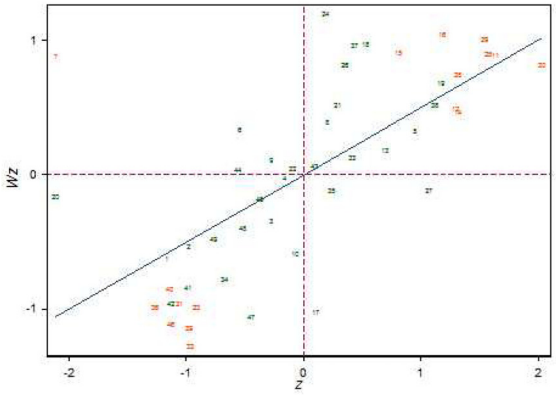
- genmsp 프로그램을 실행시키면, 현재 활성화된 'columbusdb.dta' 데이터의 마지막 부분에 std_CRIME, Wstd_CRIME, zval_CRIME, pval_CRIME, msp_CRIME 등 5개의 새로운 변수가 생성됨을 확인할 수 있음
- 이 변수들을 이용해서 STATA의 scatterplot을 그리는 명령어를 이용해 나타낸 것임

```
· do "genmsp.do"
  - 프로그램 실행 결과 생략 -

· genmsp CRIME, w(RSW)

· graph twoway ///
        (scatter Wstd_CRIME std_CRIME if pval_CRIME >=0.05, ///
                msymbol(i) mlabel(nid) mlabsize(*0.6) mlabpos(c)) ///
        (scatter Wstd_CRIME std_CRIME if pval_CRIME <0.05, ///
                msymbol(i) mlabel(nid) mlabsize(*0.6) mlabpos(c)) ///
        (lfit Wstd_CRIME std_CRIME), ///
        yline(0, lpattern(--)) xline(0, lpattern(--)) ///
        xlabel(-2(1)2, labsize(*0.8)) xtitle("{it:z}") ///
        ylabel(-1(1)1, labsize(*0.8) angle(0)) ytitle("{it:Wz}") legend(off) scheme(s1color)
```

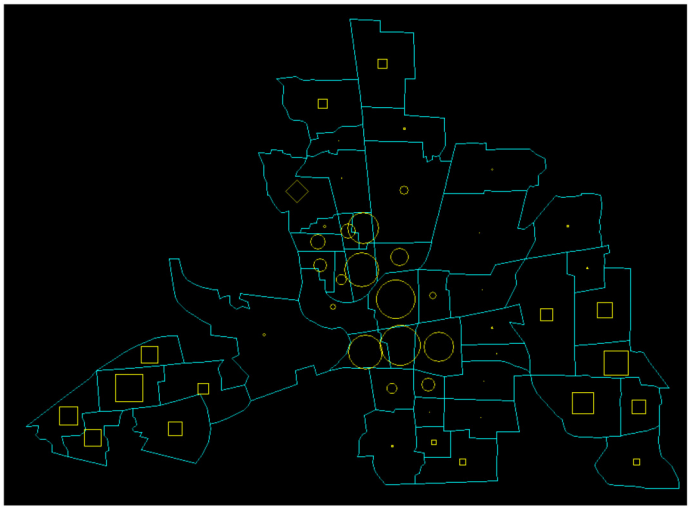
<위의 결과 계속>



❍ 다음은 위의 결과들을 map으로 나타낸 것으로, map은 앞서 shp2dta에서 생성된 지역별 좌표 데이터인 ‘columbusxy.dta’를 이용함

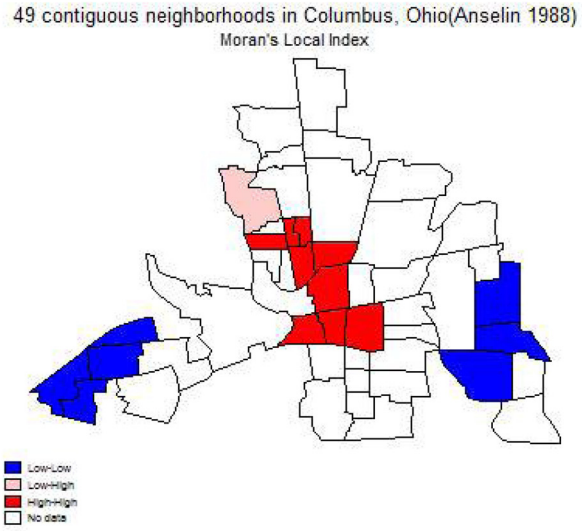
- 아래의 명령문에서 옵션 x( )와 y( )는 columbusxy.dta가 아닌 columbusdb.dta에 포함되어 있는 중심점 좌표를 나타냄

```
· spatlsa CRIME, weight(RSW) moran graph(moran) ///
        map(columbusxy.dta) x(x_centroid) y(y_centroid)
```



❍ 이것을 다시 spmap 모듈을 이용해서 cluster map으로 나타내면 아래와 같음

```
· spmap msp_CRIME using "columbusxy.dta", id(nid)              ///
      clmethod(unique) fcolor(blue red*0.2 red) ocolor(black) osize(medthin)      ///
      title("49 contiguous neighborhoods in Columbus, Ohio(Anselin 1988)", size(*0.8)) ///
          subtitle("Moran's Local Index", size(*0.8))   ///
      legorder(lohi) legstyle(2) legend(ring(1))
```



6 Linear Spatial Dependence Models for Cross-Section Data

□ 공간계량경제학 모형의 발전

❍ 일반적으로 공간계량경제학 모형에서는 표준적인 선형 회귀모형에 1)종속변수의 내생적인 상호작용효과, 2)독립변수의 외생적인 상호작용 효과, 3)오차항의 상호작용 효과 등 크게 3가지 형태의 상호작용효과를 고려하고 있음

❍ 초기에는 Anselin이 1988년에 발간한 "spatial econometrics: methods and models" 이라는 저서에서 제시한 spatial lag 모형 또는 spatial error 모형 중심의 분석과 이후 Anselin과 동료들(1996)에 의해 발전된 robust LM(Lagrange Multipiler) 검정에 의해 spatial lag 모형 또는 spatial error 모형에 대한 검정과정 등을 중심으로 연구되어 왔음

❍ 이후 2007년에 Harry Kelejian이 제1차 세계공간계량경제학회(World Conference of the Spatial Econometrics Association)에서 내생적 상호작용효과와 오차항의 상호작용 효과를 모두 포함하는 모형을 제안하였고, 이후 여러 이름으로 불리며 최근까지 발전되어 왔음

□ 공간계량경제학 모형의 종류

❍ 일반 선형회귀모형의 식

$$Y = \alpha_{\iota N} + X\beta + \varepsilon$$

❍ 일반 선형회귀모형에 포함시킬 수 있는 3가지의 상호작용항

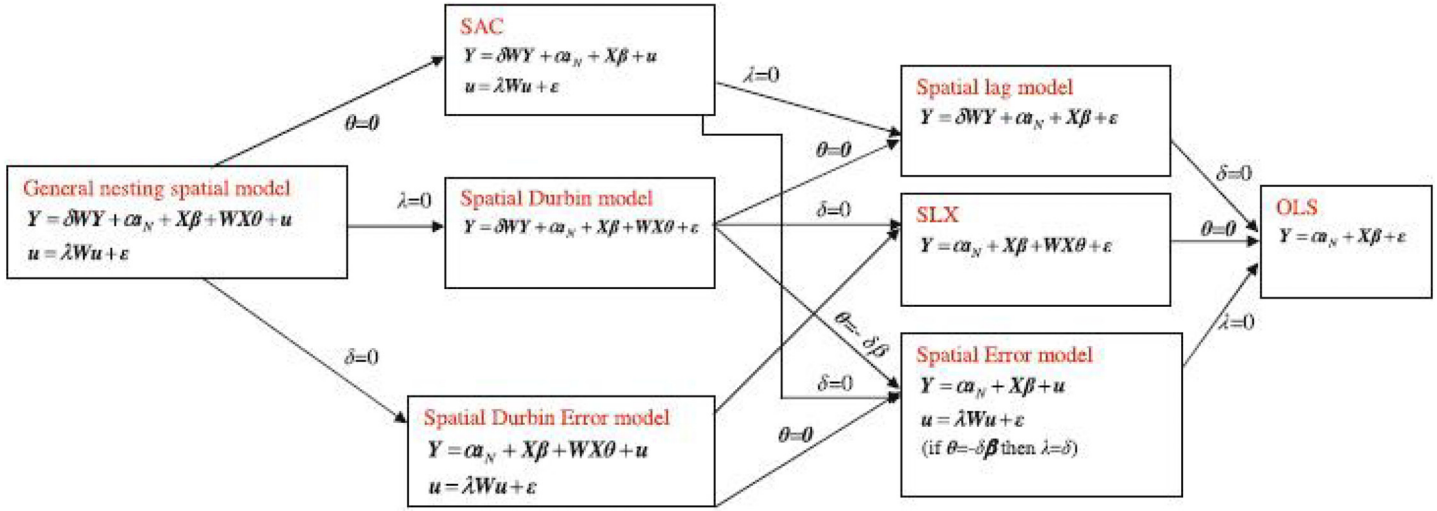
- Dependent variable y of unit A ↔ Dependent variable y of unit B
- Independent variable x of unit B ↔ Dependent variable y of unit A
- Error term u of unit A ↔ Error term u of unit B

❍ 3가지의 상호작용항이 포함된 완전 모형: General nesting spatial model

$$Y = \delta WY + \alpha_{\iota N} + X\beta + WX\theta + u$$

$$u = \lambda Wu + \varepsilon$$

- WY는 종속변수의 내생적인 상호작용효과, WX는 독립변수의 외생적인 상호작용효과, Wu는 다른 분석단위의 교란항 내의 상호작용효과를 나타냄
- 그리고 이러한 모형은 이러한 상호작용이 어떻게 포함되느냐에 따라 아래와 같이 나눌 수 있음


General nesting spatial model
$Y = \delta WY + \alpha\iota_N + X\beta + WX\theta + u$
$u = \lambda Wu + \varepsilon$
$\theta=0$
$\lambda=0$
$\delta=0$
SAC
$Y = \delta WY + \alpha\iota_N + X\beta + u$
$u = \lambda Wu + \varepsilon$
Spatial Durbin model
$Y = \delta WY + \alpha\iota_N + X\beta + WX\theta + \varepsilon$
Spatial Durbin Error model
$Y = \alpha\iota_N + X\beta + WX\theta + u$
$u = \lambda Wu + \varepsilon$
$\lambda=0$
$\theta=0$
$\delta=0$
$\theta=-\delta\beta$
$\delta=0$
$\theta=0$
Spatial lag model
$Y = \delta WY + \alpha\iota_N + X\beta + \varepsilon$
SLX
$Y = \alpha\iota_N + X\beta + WX\theta + \varepsilon$
Spatial Error model
$Y = \alpha\iota_N + X\beta + u$
$u = \lambda Wu + \varepsilon$
(if $\theta=-\delta\boldsymbol{\beta}$ then $\lambda=\delta$)
$\delta=0$
$\theta=0$
$\lambda=0$
OLS
$Y = \alpha\iota_N + X\beta + \varepsilon$

□ 직접효과(direct effect)와 간접효과(indirect or spillover effect)의 구분

❍ 공간계량경제학모형을 적용한 많은 연구들은 앞서 살펴본 여러 상호작용항을 고려해 왔으며, 그 결과를 가지고 공간적인 spillover effect가 존재하는가를 논의해왔으나 최근 이것이 잘못된 결과일 수 있다는 LeSage가 2009년 저서인 "Introduction to Spatial Econometrics"에서 지적한 바 있음

❍ 위의 general nesting spatial model(이하 GNS)은 아래와 같이 재표현할 수 있음

$$Y = (I - \delta W)^{-1}(X\beta + WX\theta) + R$$

- 여기에서 R은 상수항과 오차항을 모두 나타냄
- k번째 독립변수의 단위 1부터 단위 N까지 Y의 기댓값의 편도함수행렬(matrix of partial derivatives)은 다음과 같이 나타낼 수 있음(Elhorst, 2013)

$$\left[\frac{\partial E(Y)}{\partial x_{1k}} \cdot \frac{\partial E(Y)}{\partial x_{Nk}}\right] = \begin{bmatrix} \frac{\partial E(y_1)}{\partial x_{1k}} \cdot \frac{\partial E(y_1)}{\partial x_{Nk}} \\ \cdot \quad \cdot \quad \cdot \\ \frac{\partial E(y_N)}{\partial x_{1k}} \cdot \frac{\partial E(y_N)}{\partial x_{Nk}} \end{bmatrix}$$

$$= (I - \delta W)^{-1} \begin{bmatrix} \beta_k & w_{12}\theta_k & \cdot & w_{1N}\theta_k \\ w_{21}\theta_k & \beta_k & \cdot & w_{2N}\theta_k \\ \cdot & \cdot & \cdot & \cdot \\ w_{N1}\theta_k & w_{N2}\theta_k & \cdot & \beta_k \end{bmatrix}$$

- 위의 식을 토대로 다음과 같은 3가지의 특성을 요약할 수 있음

1) 특정 단위에서의 설명변수 변화는 해당 단위에서의 종속변수뿐만 아니라 다른 단위의 종속변수의 변화를 가져온다는 것으로, 전자는 direct effect라고 하고 후자를 indirect effect라고 함
⇨ 편도함수행렬의 대각선의 모든 원소는 direct effect를 비대각선의 모든 원소들은 indirect effect를 나타냄 ⇨ 따라서 모든 비대각선 원소들이 0일 경우, 즉 $\delta = 0$이고 $\theta_k = 0$일 때 indirect effect는 존재하지 않음

2) 표본의 서로 다른 분석단위별로 direct effect와 indirect effect가 다름
⇨ $\delta \neq 0$일 때 행렬$(I_N - \delta W)^{-1}$의 대각원소들이 분석단위별로 다르기 때문에 direct effect들이 분석단위에 따라 다름 ⇨ $\delta \neq 0$ 그리고(또는) $\theta_k \neq 0$일 때, 행렬$(I_N - \delta W)^{-1}$의 비대각원소들이 분석단위별로 다를 뿐만 아니라 행렬W의 비대각원소들이 분석단위별로 다르기 때문에 indirect effect는 분석단위에 따라 다름

3) $\theta_k \neq 0$일 때 발생하는 간접효과를 local effects라고 하며, $\delta \neq 0$일 때 발생하는 간접효과를 global effects라고 함
⇨ 공간가중행렬의 원소 w_{ij}가 0이 아니면 y_i에 대해 한 단위의 이웃단위들에 의한 x_{jk}의 효과가 존재하고, 공간가중행렬의 원소 w_{ij}가 0이면 y_i에 대해 한 단위의 이웃단위들에 의한 x_{jk}의 효과가 존재하지 않기 때문에 local effects라고 함 ⇨ 공간가중행렬 W과 달리 행렬$(I_N - \delta W)^{-1}$는 $\delta \neq 0$일 때 0인 원소들을 포함하지 않기 때문에 한 단위의 이웃단위집합에 속하지 않은 단위에서 그 효과가 발생하는 것이기 때문에 global effects라고 함 ⇨ 단, $\delta \neq 0$이고 $\theta_k \neq 0$일 때 global effect와 local effect를 구분할 수 없음

- 위와 같은 특성들로 인해 표본의 분석단위별로 직접효과와 간접효과가 달라서 이를 제시하는 방법이 문제가 될 수 있음. 즉, 분석단위지역이 N개이고, 독립변수가 k개일 때 k개의 $N \times N$ 직간접효과의 행렬을 얻게 되어 이를 보고하는 것이 어려울 수 있음

- 이에 Lesage와 Pace(2009)는 위 행렬에서 대각원소들의 평균을 direct effect로, 비대각 원소들의 행 또는 열의 합들에 대한 평균을 indirect effect로 보고할 것을 제안함

□ 공간계량경제학 모형 분석을 위한 STATA Module의 비교

❍ 앞서 공간가중행렬 생성을 위한 module 소개에서 살펴본 바와 같이 현재까지 다양한 연구자들에 의해 제안된 공간계량경제학 모형 분석을 위한 module은 아래와 같음

	spatial weighting matrix	measures of spatial autocorrelation	Diagnostic tests for spatial dependence in OLS Regression	spatial model by ML	spatial model by IV/GMM
Pisati(2001)	.spatwmat	.spatgsa .spatlsa	.spatdiag	.spatreg	-
Jeanty(2010)	.spwmatrix	.splagvar	.anketest	.spmlreg	-
Drukker, Peng, Prucha, Raciborski (2011)	.sppack(하위모듈들의 통합패키지명)				
	.spmat	-		.spreg	.spivreg
Shehata (2011)	.spweightcs	.spautoreg			

- 아래의 모형들은 개발된지 10년이 훨씬 지난 module부터 최근에 계속해서 업데이트되는 module까지 매우 다양함
- 하지만 앞서 간략한 발전과정에서 소개된 바와 같이 최근의 상호작용항을 모두 포함하는 모형까지 분석할 수 있는 모듈은 이집트의 농경제학자인 Emad Abd Elmessih Shehata(2011)이 만든 'spautoreg' module이 유일하며, 직간접효과를 산출하여 제시해주는 module 역시 이것이 유일하여 편의성이 높음
- 특히 Shehata는 공간계량경제학모형의 발전과정을 지속적으로 반영하여 module을 2011년부터 현재까지 계속 수정/보완하고 있기 때문에 유용성이 더 크다 하겠음

□ STATA의 'spautoreg' module을 이용한 공간계량경제학모형 분석

❍ Shehata(2011)의 'spautoreg' module을 이용함에 있어 두 가지 방법이 있는데, 그가 개발한 공간가중행렬 생성 module인 'spweightcs'을 이용해서 공간가중행렬을 새로 생성하거나 기존에 생성된 공간가중행렬을 이용하는 방법임

- 하지만, 전자의 방법을 이용하기 위해서는 아래의 그림과 같은 각 분석단위(지역)별 인접지역에 대한 리스트가 포함된 별도의 파일이 필요한데, 물론 이것은 앞서 살펴보았던 spmat module이나 GeoDa라는 무료 공간분석프로그램을 통해 얻을 수 있음
- 이러한 불편으로 인해 spautoreg' module을 이용하고자 한다면 spmat module 등을 통해 기존에 생성한 공간가중행렬을 이용하는 것이 좀 더 용이할 것으로 판단됨

```
· ssc install spautoreg, replace

· use columbusdb.dta, clear

* MLE Spatial Lag Model (SAR)
spautoreg CRIME INC HOVAL, wmfile(swm_nst) model(sar) mfx(lin)          //binary contiguity matrix
spautoreg CRIME INC HOVAL, wmfile(swm_nst) model(sar) mfx(lin) stand   //row-normalized matrix
spautoreg CRIME INC HOVAL, wmfile(swm_st) model(sar) mfx(lin)           //minmax-normalized matrix

* MLE Spatial Error Model (SEM)
spautoreg CRIME INC HOVAL, wmfile(swm_nst) model(sem) mfx(lin)          //binary contiguity matrix
spautoreg CRIME INC HOVAL, wmfile(swm_nst) model(sem) mfx(lin) stand   //row-normalized matrix
spautoreg CRIME INC HOVAL, wmfile(swm_st) model(sem) mfx(lin)           //minmax-normalized matrix

* MLE Spatial Durbin Model (SDM):
spautoreg CRIME INC HOVAL, wmfile(swm_nst) model(sdm) mfx(lin)          //binary contiguity matrix
spautoreg CRIME INC HOVAL, wmfile(swm_nst) model(sdm) mfx(lin) stand   //row-normalized matrix
spautoreg CRIME INC HOVAL, wmfile(swm_st) model(sdm) mfx(lin)           //minmax-normalized matrix
```

```
* MLE Spatial AutoCorrelation Model (SAC):
spautoreg CRIME INC HOVAL, wmfile(swm_nst) model(sac) mfx(lin)          //binary contiguity matrix
spautoreg CRIME INC HOVAL, wmfile(swm_nst) model(sac) mfx(lin) stand   //row-normalized matrix
spautoreg CRIME INC HOVAL, wmfile(swm_st) model(sac mfx(lin)            //minmax-normalized matrix
```


<참고문헌>

Anselin L (1988). *Spatial econometrics: methods and models*. Kluwer, Dordrecht.

Anselin L, Bera AK, Florax R, Yoon MJ (1996) Simple diagnostic tests for spatial dependence. *Reg Sci Urban Econ* 26(1):77-104.

Elhorst, J.P. (2013). *Spatial Econometrics: from cross-sectional data to spatial panels*. Springer.

LeSage J.P., Pace RK (2009) *Introduction to spatial econometrics*. CRC Press, Taylor & Francis Group, Boca Raton

LeSage JP, Pace RK (2011) Pitfalls in higher order model extensions of basic spatial regression methodology. *Rev Reg Stud* 41(1):13-26.

콜로키움자료집 14-S09

STATA를 이용한 공간분석의 기초

인 쇄 2014년 3월 24일
발 행 2014년 3월 26일

발행처 **한국청소년정책연구원**
서울특별시 서초구 태봉로 114
발행인 이 재 연

등 록 1993. 10. 23 제 21-500호
인쇄처 (주)아르빛 전화 02)503-3223 대표 유채란

사전 승인없이 보고서 내용의 무단전재 · 복제를 금함.
구독문의 : (02) 2188-8844(정보자료 · 전산보안팀)