Spatial Dependence with Polygons

Ignacio Sarmiento-Barbieri

Universidad de los Andes

June 25, 2022

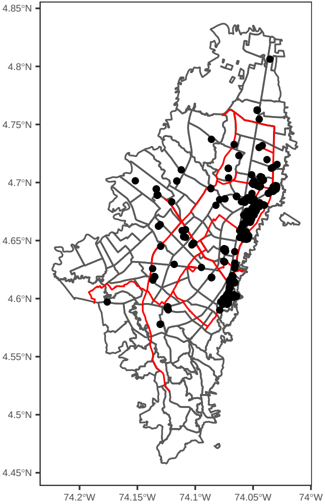
Recap

- Types of Spatial Data
- Reading and Mapping spatial data in R
- Projections
- Creating Spatial Objects
- Measuring Distances

Agenda

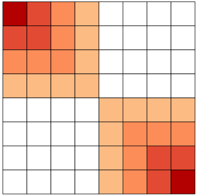
Motivation

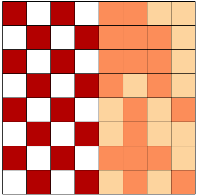
- Independence assumption between observation is no longer valid
- Attributes of observation i may influence the attributes of observation j.
- We will consider various alternatives to model spatial dependence
- Think as a way to model $f(X)$



Motivation

- Independence assumption between observation is no longer valid
- Attributes of observation i may influence the attributes of observation j.
- Positive Spatial correlation arises when units that are *close* to one another are more similar than units that are far apart
- Similarly spatial heterogeneity arises when some areas present more variability than others





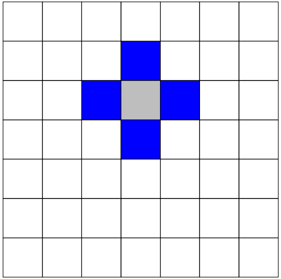
Closeness

"Everything is related to everything else, but close things are more related than things that are far apart" (Tobler, 1979).

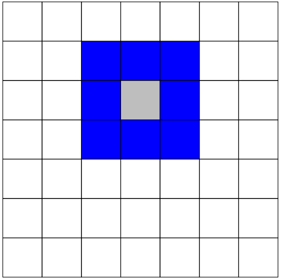
- One of the major differences between standard econometrics and standard spatial econometrics lies, in the fact that, in order to treat spatial data, we need to use two different sets of information
  1. Observed values of the economic variables
  2. Particular location where those variables are observed and to the various links of proximity between all spatial observations

Closeness

Rook criterion: two units are close to one another if they share a side



Queen criterion: two units are close if they share a side or an edge.



Weights Matrix

- At the heart of traditional spatial econometrics is the definition of the *weights matrix*:

$$W = \begin{pmatrix} w_{11} & \dots & \dots & w_{n1} \\ \vdots & w_{ij} & & \vdots \\ \vdots & & \ddots & \vdots \\ w_{n_1} & \dots & \dots & w_{nn} \end{pmatrix}_{n \times n} \tag{1}$$

with generic element:

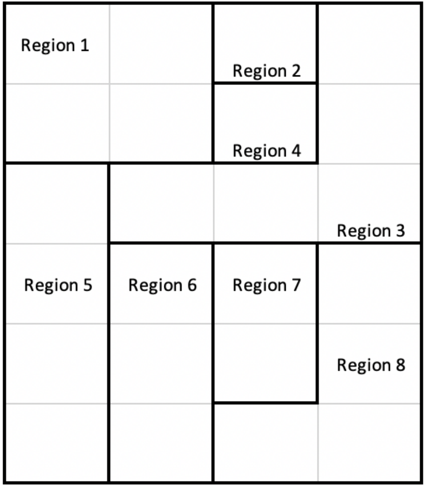
$$w_{ij} = \begin{cases} 1 & if\, j \in N(i) \\ 0 & o.w \end{cases} \tag{2}$$

$N(i)$ being the set of neighbors of location j. By convention, the diagonal elements are set to zero, i.e. $w_{ii} = 0$.

Weights Matrix

- The specification of the neighboring set ($N(i)$) is quite arbitrary and there's a wide range of suggestions in the literature.
  - Rook criterion
  - Queen criterion
  - Two observations are neighbors if they are within a certain distance, i.e., $j \in N(j)$ if $d_{ij} < d_{max}$ where d is the distance between location i and j.
  - Closest neighbor, ties can be solved randomly
  - More general matrices can also be specified by considering entries of w_{ij} as functions of geographical, economic or social distances between areas rather than simply characterized by dichotomous entries

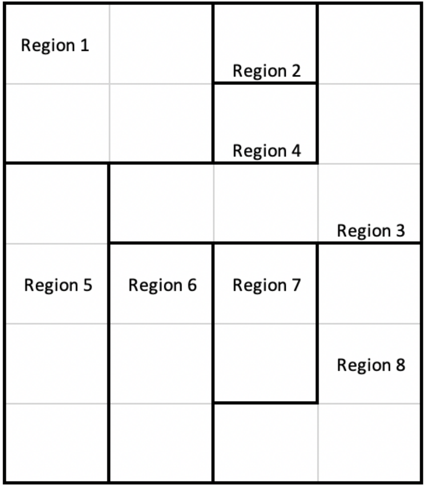
Some Examples of Weights Matrices



Adjacency Criterion

$$W = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}_{8\times 8}$$

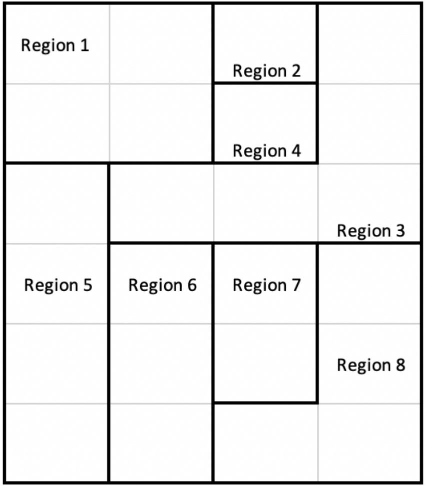
Some Examples of Weights Matrices



Nearest Neighbor

$$W = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}_{8\times 8}$$

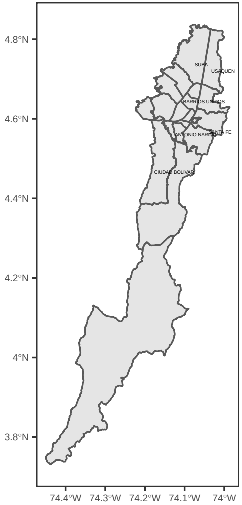
Some Examples of Weights Matrices



Distance < 2

$$W = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}_{8\times 8}$$

Some Examples of Weights Matrices



Some Examples of Weights Matrices

	ANTONIO NARIÑO	TUNJUELITO	RAFAEL URIBE URIBE	CANDELARIA	BARRIOS UNIDOS	TEUSAQUILLO	PUENTE ARANDA	LOS MARTIRES	SUMAPAZ	USAQUEN	CHAPINERO	SANTA FE	SAN CRISTOBAL	USME	CIUDAD BOLIVAR	BOSA	KENNEDY	FONTIBON	ENGATIVA	SUBA
ANTONIO NARIÑO	0	1	1	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0
TUNJUELITO	1	0	1	0	0	0	1	0	0	0	0	0	0	1	1	0	1	0	0	0
RAFAEL URIBE URIBE	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
CANDELARIA	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
BARRIOS UNIDOS	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	1	1
TEUSAQUILLO	0	0	0	0	1	0	1	1	0	0	1	1	0	0	0	0	0	1	1	0
PUENTE ARANDA	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	0	0
LOS MARTIRES	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0
SUMAPAZ	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
USAQUEN	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
CHAPINERO	0	0	0	0	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	1
SANTA FE	1	0	0	1	0	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0
SAN CRISTOBAL	1	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
USME	0	1	1	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0
CIUDAD BOLIVAR	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0
BOSA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
KENNEDY	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	1	0	0
FONTIBON	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	1	0
ENGATIVA	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1
SUBA	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0

Some Examples of Weights Matrices

Quite often the W matrices are standardized to sum to one in each row

$$w^*_{ij} = \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}} \tag{3}$$

This can be quite useful since

$$L(y) = W^* y \tag{4}$$

in which each single element is equal to

$$\begin{aligned} L(y_i) &= \sum_{j=1}^{n} w^*_{ij} y_j \\ &= \sum_{j=1}^{n} \frac{w_{ij} y_j}{\sum_{j=1}^{n} w_{ij}} \end{aligned} \tag{5}$$

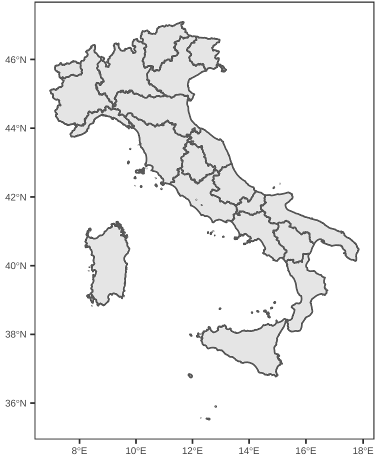
$$\sum_{j \in N(i)} y_j$$

Some Examples of Weights Matrices

	ANTONIO NARIÑO	TUNJUELITO	RAFAEL URIBE URIBE	CANDELARIA	BARRIOS UNIDOS	TEUSAQUILLO	PUENTE ARANDA	LOS MARTIRES
ANTONIO NARIÑO	0.0000000	0.1666667	0.1666667	0.0000000	0.0000000	0.0000000	0.1666667	0.1666667
TUNJUELITO	0.1666667	0.0000000	0.1666667	0.0000000	0.0000000	0.0000000	0.1666667	0.0000000
RAFAEL URIBE URIBE	0.2500000	0.2500000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
CANDELARIA	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
BARRIOS UNIDOS	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.2000000	0.0000000	0.0000000
TEUSAQUILLO	0.0000000	0.0000000	0.0000000	0.0000000	0.1428571	0.0000000	0.1428571	0.1428571
PUENTE ARANDA	0.1666667	0.1666667	0.0000000	0.0000000	0.0000000	0.1666667	0.0000000	0.1666667
LOS MARTIRES	0.2500000	0.0000000	0.0000000	0.0000000	0.0000000	0.2500000	0.2500000	0.0000000
SUMAPAZ	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
USAQUEN	0.0000000	0.0000000	0.0000000	0.0000000	0.3333333	0.0000000	0.0000000	0.0000000
CHAPINERO	0.0000000	0.0000000	0.0000000	0.0000000	0.2000000	0.2000000	0.0000000	0.0000000
SANTA FE	0.1666667	0.0000000	0.0000000	0.1666667	0.0000000	0.1666667	0.0000000	0.1666667
SAN CRISTOBAL	0.2500000	0.0000000	0.2500000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
USME	0.0000000	0.2000000	0.2000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
CIUDAD BOLIVAR	0.0000000	0.2500000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
BOSA	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
KENNEDY	0.0000000	0.2000000	0.0000000	0.0000000	0.0000000	0.0000000	0.2000000	0.0000000
FONTIBON	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.2500000	0.2500000	0.0000000
ENGATIVA	0.0000000	0.0000000	0.0000000	0.0000000	0.2500000	0.2500000	0.0000000	0.0000000
SUBA	0.0000000	0.0000000	0.0000000	0.0000000	0.2500000	0.0000000	0.0000000	0.0000000

	SUMAPAZ	USAQUEN	CHAPINERO	SANTA FE	SAN CRISTOBAL	USME	CIUDAD BOLIVAR	BOSA	KENNEDY	FONTIBON	ENGATIVA	SUBA
ANTONIO NARIÑO	0.0	0.00	0.0000000	0.1666667	0.1666667	0.0000000	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000
TUNJUELITO	0.0	0.00	0.0000000	0.0000000	0.0000000	0.1666667	0.1666667	0.00	0.1666667	0.0000000	0.0000000	0.0000000
RAFAEL URIBE URIBE	0.0	0.00	0.0000000	0.0000000	0.2500000	0.2500000	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000
CANDELARIA	0.0	0.00	0.0000000	1.0000000	0.0000000	0.0000000	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000
BARRIOS UNIDOS	0.0	0.20	0.2000000	0.0000000	0.0000000	0.0000000	0.0000000	0.00	0.0000000	0.0000000	0.2000000	0.2000000
TEUSAQUILLO	0.0	0.00	0.1428571	0.1428571	0.0000000	0.0000000	0.0000000	0.00	0.0000000	0.1428571	0.1428571	0.0000000
PUENTE ARANDA	0.0	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.00	0.1666667	0.1666667	0.0000000	0.0000000
LOS MARTIRES	0.0	0.00	0.0000000	0.2500000	0.0000000	0.0000000	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000
SUMAPAZ	0.0	0.00	0.0000000	0.0000000	0.0000000	1.0000000	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000
USAQUEN	0.0	0.00	0.3333333	0.0000000	0.0000000	0.0000000	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.3333333
CHAPINERO	0.0	0.20	0.0000000	0.2000000	0.0000000	0.0000000	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.2000000
SANTA FE	0.0	0.00	0.1666667	0.0000000	0.1666667	0.0000000	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000
SAN CRISTOBAL	0.0	0.00	0.0000000	0.2500000	0.0000000	0.2500000	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000
USME	0.2	0.00	0.0000000	0.0000000	0.2000000	0.0000000	0.2000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000
CIUDAD BOLIVAR	0.0	0.00	0.0000000	0.0000000	0.0000000	0.2500000	0.0000000	0.25	0.2500000	0.0000000	0.0000000	0.0000000
BOSA	0.0	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.5000000	0.00	0.5000000	0.0000000	0.0000000	0.0000000
KENNEDY	0.0	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.2000000	0.20	0.0000000	0.2000000	0.0000000	0.0000000
FONTIBON	0.0	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.00	0.2500000	0.0000000	0.2500000	0.0000000
ENGATIVA	0.0	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.00	0.0000000	0.2500000	0.0000000	0.2500000
SUBA	0.0	0.25	0.2500000	0.0000000	0.0000000	0.0000000	0.0000000	0.00	0.0000000	0.0000000	0.2500000	0.0000000

Some Examples of Weights Matrices



Some Examples of Weights Matrices

	Piemonte	Valle D'Aosta	Lombardia	Trentino-Alto Adige	Veneto	Friuli Venezia Giulia	Liguria	Emilia-Romagna	Toscana	Umbria	Marche
Piemonte	0.0000000	0.25	0.2500000	0.00	0.0000000	0.00	0.2500000	0.2500000	0.0000000	0.0000000	0.0000000
Valle D'Aosta	1.0000000	0.00	0.0000000	0.00	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Lombardia	0.2500000	0.00	0.0000000	0.25	0.2500000	0.00	0.0000000	0.2500000	0.0000000	0.0000000	0.0000000
Trentino-Alto Adige	0.0000000	0.00	0.5000000	0.00	0.5000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Veneto	0.0000000	0.00	0.2500000	0.25	0.0000000	0.25	0.0000000	0.2500000	0.0000000	0.0000000	0.0000000
Friuli Venezia Giulia	0.0000000	0.00	0.0000000	0.00	1.0000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Liguria	0.3333333	0.00	0.0000000	0.00	0.0000000	0.00	0.0000000	0.3333333	0.3333333	0.0000000	0.0000000
Emilia-Romagna	0.1666667	0.00	0.1666667	0.00	0.1666667	0.00	0.1666667	0.0000000	0.1666667	0.0000000	0.1666667
Toscana	0.0000000	0.00	0.0000000	0.00	0.0000000	0.00	0.2000000	0.2000000	0.0000000	0.2000000	0.2000000
Umbria	0.0000000	0.00	0.0000000	0.00	0.0000000	0.00	0.0000000	0.0000000	0.3333333	0.0000000	0.3333333
Marche	0.0000000	0.00	0.0000000	0.00	0.0000000	0.00	0.0000000	0.2000000	0.2000000	0.2000000	0.0000000
Lazio	0.0000000	0.00	0.0000000	0.00	0.0000000	0.00	0.0000000	0.0000000	0.1666667	0.1666667	0.1666667
Abruzzo	0.0000000	0.00	0.0000000	0.00	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.3333333
Molise	0.0000000	0.00	0.0000000	0.00	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Campania	0.0000000	0.00	0.0000000	0.00	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Puglia	0.0000000	0.00	0.0000000	0.00	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Basilicata	0.0000000	0.00	0.0000000	0.00	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Calabria	0.0000000	0.00	0.0000000	0.00	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Sicilia	0.0000000	0.00	0.0000000	0.00	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
Sardegna	0.0000000	0.00	0.0000000	0.00	0.0000000	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000

	Lazio	Abruzzo	Molise	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna
Piemonte	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0	0
Valle D'Aosta	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0	0
Lombardia	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0	0
Trentino-Alto Adige	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0	0
Veneto	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0	0
Friuli Venezia Giulia	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0	0
Liguria	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0	0
Emilia-Romagna	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0	0
Toscana	0.2000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0	0
Umbria	0.3333333	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0	0
Marche	0.2000000	0.2000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0	0
Lazio	0.0000000	0.1666667	0.1666667	0.1666667	0.0000000	0.0000000	0.0000000	0	0
Abruzzo	0.3333333	0.0000000	0.3333333	0.0000000	0.0000000	0.0000000	0.0000000	0	0
Molise	0.2500000	0.2500000	0.0000000	0.2500000	0.2500000	0.0000000	0.0000000	0	0
Campania	0.2500000	0.0000000	0.2500000	0.0000000	0.2500000	0.2500000	0.0000000	0	0
Puglia	0.0000000	0.0000000	0.3333333	0.3333333	0.0000000	0.3333333	0.0000000	0	0
Basilicata	0.0000000	0.0000000	0.0000000	0.3333333	0.3333333	0.0000000	0.3333333	0	0
Calabria	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	1.0000000	0.0000000	0	0
Sicilia	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0	0
Sardegna	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0	0

Weights Matrix in R

```r
require("sf")
require("spdep")
require("dplyr")
```

```r
chi.poly<-read_sf("foreclosures/foreclosures.shp")
st_crs(chi.poly) #doesn't have a projection
```

```
## Coordinate Reference System: NA
```

```r
st_crs(chi.poly)<-4326 #WGS84 set it in the map
```

Weights Matrix in R

```r
chi.poly<-st_transform(chi.poly,26916) #reproject planarly
#NAD83 UTM Zone 16N
st_crs(chi.poly)
```

```
## Coordinate Reference System:
##   User input: EPSG:26916
##   wkt:
## PROJCS["NAD83 / UTM zone 16N",
##     GEOGCS["NAD83",
##         DATUM["North_American_Datum_1983",
##             SPHEROID["GRS 1980",6378137,298.257222101,
##                 AUTHORITY["EPSG","7019"]],
##             TOWGS84[0,0,0,0,0,0,0],
##             AUTHORITY["EPSG","6269"]],
##         PRIMEM["Greenwich",0,
##             AUTHORITY["EPSG","8901"]],
##         UNIT["degree",0.0174532925199433,
##             AUTHORITY["EPSG","9122"]],
##         AUTHORITY["EPSG","4269"]],
##     PROJECTION["Transverse_Mercator"],
##     PARAMETER["latitude_of_origin",0],
##     PARAMETER["central_meridian",-87],
##     PARAMETER["scale_factor",0.9996],
##     PARAMETER["false_easting",500000],
##     PARAMETER["false_northing",0],
##     UNIT["metre",1,
##         AUTHORITY["EPSG","9001"]],
##     AXIS["Easting",EAST],
##     AXIS["Northing",NORTH],
##     AUTHORITY["EPSG","26916"]]
```

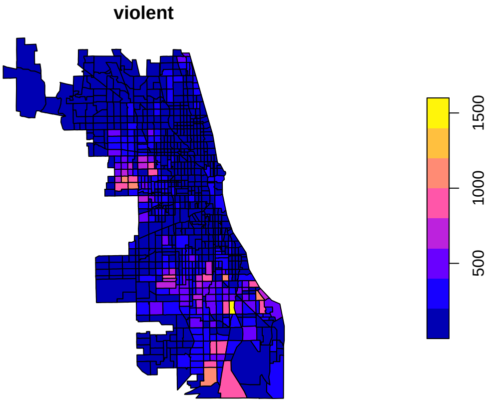
Weights Matrix in R

```r
str(chi.poly)
```

```
## tibble [897 x 17] (S3: sf/tbl_df/tbl/data.frame)
##  $ SP_ID     : chr [1:897] "1" "2" "3" "4" ...
##  $ fips      : chr [1:897] "17031010100" "17031010200" "17031010300" "17031010400" ...
##  $ est_fcs   : int [1:897] 43 129 55 21 64 56 107 43 7 51 ...
##  $ est_mtgs  : int [1:897] 904 2122 1151 574 1427 1241 1959 830 208 928 ...
##  $ est_fcs_rt: num [1:897] 4.76 6.08 4.78 3.66 4.48 4.51 5.46 5.18 3.37 5.5 ...
##  $ res_addr  : int [1:897] 2530 3947 3204 2306 5485 2994 3701 1694 443 1552 ...
##  $ est_90d_va: num [1:897] 12.61 12.36 10.46 5.03 8.44 ...
##  $ bls_unemp : num [1:897] 8.16 8.16 8.16 8.16 8.16 8.16 8.16 8.16 8.16 8.16 ...
##  $ county    : chr [1:897] "Cook County" "Cook County" "Cook County" "Cook County" ...
##  $ fips_num  : num [1:897] 1.7e+10 1.7e+10 1.7e+10 1.7e+10 1.7e+10 ...
##  $ totpop    : int [1:897] 5391 10706 6649 5325 10944 7178 10799 5403 1089 3634 ...
##  $ tothu     : int [1:897] 2557 3981 3281 2464 5843 3136 3875 1768 453 1555 ...
##  $ huage     : int [1:897] 61 53 56 60 54 58 48 57 61 48 ...
##  $ oomedval  : int [1:897] 169900 147000 119800 151500 143600 145900 153400 170500 215900 114700 ...
##  $ property  : num [1:897] 646 914 478 509 641 612 678 332 147 351 ...
##  $ violent   : num [1:897] 433 421 235 159 240 266 272 146 78 84 ...
##  $ geometry  :sfc_POLYGON of length 897; first list element: List of 1
##   ..$ : num [1:15, 1:2] 443923 444329 444814 444839 444935 ...
##   ..- attr(*, "class")= chr [1:3] "XY" "POLYGON" "sfg"
##  - attr(*, "sf_column")= chr "geometry"
##  - attr(*, "agr")= Factor w/ 3 levels "constant","aggregate",..: NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "names")= chr [1:16] "SP_ID" "fips" "est_fcs" "est_mtgs" ...
```

Weights Matrix in R

```
plot(chi.poly['violent'])
```



Weights Matrix in R

```r
list.queen<-poly2nb(chi.poly, queen=TRUE)
W<-nb2listw(list.queen, style="W", zero.policy=TRUE)
W
```

```
## Characteristics of weights list object:
## Neighbour list object:
## Number of regions: 897
## Number of nonzero links: 6140
## Percentage nonzero weights: 0.7631036
## Average number of links: 6.845039
##
## Weights style: W
## Weights constants summary:
##     n     nn  S0       S1       S2
## W 897 804609 897 274.4893 3640.864
```

Weights Matrix in R

```
plot(W,st_geometry(st_centroid(chi.poly)))
```

Weights Matrix in R

```
coords <- st_centroid(st_geometry(chi.poly), of_largest_polygon=TRUE)
W_dist<-dnearneigh(coords,0,1000)
W_dist
```

```
## Neighbour list object:
## Number of regions: 897
## Number of nonzero links: 5448
## Percentage nonzero weights: 0.6770991
## Average number of links: 6.073579
## 55 regions with no links:
## 141 142 143 145 153 154 155 158 462 631 637 638 642 643 644 645 655 656 657 658 659 758 759 769 820 821 822 823 824 855 856 857 861 862 8
```

```
plot(W_dist, coords)
```

Weights Matrix in R

```r
W_dist<-dnearneigh(coords,0,4300)
W_dist
```

```
## Neighbour list object:
## Number of regions: 897
## Number of nonzero links: 87988
## Percentage nonzero weights: 10.9355
## Average number of links: 98.09142
```

```r
plot(W_dist, coords)
```

Traditional Spatial Econometrics

Spatial Autoregressive (SAR) Models

- Spatial lag dependence in a regression setting can be modeled similar to an autoregressive process in time series. Formally,

$$y = \rho W y + X\beta + \epsilon$$

- Wy induces a nonzero correlation with the error term, similar to the presence of an endogenous variable.
- Unlike to time series, Wy_i is always correlated with ϵ_i
- OLS estimates in the non spatial model will be biased and inconsistent. (Anselin and Bera, 1998)
- The estimation of the SAR model can be approached in two ways.
  1. Assume normality of the error term and use maximum likelihood.
  2. Use 2SLS
- In `R` the function `lagsarlm` uses MLE

Prediction with SAR Models

- The usual *prolegomena*

```
set.seed(101010) #sets a seed
#70% train
indic<-sample(1:nrow(chi.poly),floor(.7*nrow(chi.poly)))

#Partition the sample
train<-chi.poly[indic,]
test<-chi.poly[-indic,]


ols<-lm(violent~est_fcs_rt+bls_unemp, data=train)
test$yhat<-predict(ols,newdata=test)
mean((test$violent-test$yhat)^2)
```

```
## [1] 29773.64
```

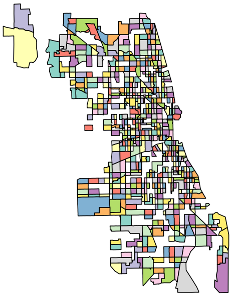
Prediction with SAR Models

- Modeling the spatial structure with a SAR Model

```
list.queen_train<-poly2nb(train, queen=TRUE)
W_train<-nb2listw(list.queen_train, style="W", zero.policy=TRUE)
W_train

Error in print.listw(x) : regions with no neighbours found, use zero.policy=TRUE

plot(train["fips"])
```



Prediction with SAR Models

- Use distance instead

```
coords <- st_centroid(st_geometry(train), of_largest_polygon=TRUE)
W_train<-dnearneigh(coords,0,4300)
W_train<-nb2listw(W_train, style="W", zero.policy=TRUE)



coords <- st_centroid(st_geometry(test), of_largest_polygon=TRUE)
W_test<-dnearneigh(coords,0,4300)
W_test<-nb2listw(W_test, style="W", zero.policy=TRUE)



require("spatialreg")

sar.chi<-lagsarlm(violent~est_fcs_rt+bls_unemp, data=train, W_train)

test$yhat_sar<-predict(sar.chi,newdata=test,listw=W_test)
```

Prediction with SAR Models

- Comparing to OLS

```
mean((test$violent-test$yhat)^2)
```

```
## [1] 29773.64
```

```
mean((test$violent-test$yhat_sar)^2)
```

```
## [1] 28662.23
```

Spatial Lag Model

Let's consider the following model:

$$y = \lambda Wy + X\beta + u$$

with $|\lambda| < 1$, we also assume that W is exogenous

If W is row standardized:

- Guarantees $|\lambda| < 1$ (Anselin, 1982)
- $[0, 1]$ Weights
- Wy Average of neighboring values
- W is no longer symmetric $\sum_j w_{ij} \neq \sum_i w_{ji}$ (complicates computation)

Spatial Lag Model
Maximum Likelihood Estimator

Note that we can write

$$(I - \lambda W)y = X\beta + u$$

- We can think this model as a way to correct for loss of information coming from spatial dependence.
- $(1 - \lambda W)y$ is a spatially filtered dependent variable, i.e., the effect of spatial autocorrelation taken out

Spatial Lag Model

In this case, endogeneity emerges because the spatially lagged value of y is correlated with the stochastic disturbance.

$$y = (I - \lambda W)^{-1} X\beta + (I - \lambda W)^{-1} u$$

$$E((Wy)u') = E(W(I - \lambda W)^{-1} X\beta u' + W(I - \lambda W)^{-1} uu')$$

$$= W(I - \lambda W)^{-1} X\beta E(u') + W(I - \lambda W)^{-1} E(uu')$$

$$= W(I - \lambda W)^{-1} E(uu')$$

$$= \sigma^2 W(I - \lambda W)^{-1} \neq 0$$

Spatial Lag Model

Maximum Likelihood Estimator

- One solution that emerged in the literature is MLE
- We need an extra assumption, i.e., $u \sim_{iid} N(0, \sigma^2 I)$.

$$y = (I - \lambda W)^{-1} X\beta + (I - \lambda W)^{-1} u$$

note that

$$E(y) = (I - \lambda W)^{-1} X\beta + (I - \lambda W)^{-1} u$$

$$= (I - \lambda W)^{-1} X\beta + (I - \lambda W)^{-1} E(u)$$

$$= (I - \lambda W)^{-1} X\beta$$

Spatial Lag Model
Maximum Likelihood Estimator

$$
\begin{aligned}
E(yy') &= ((I-\lambda W)^{-1}X\beta + (I-\lambda W)^{-1}u)((I-\lambda W)^{-1}X\beta + (I-\lambda W)^{-1}u)' \\
&= (I-\lambda W)^{-1}X\beta\beta' X'(I-\lambda W')^{-1} + (I-\lambda W)^{-1}u\beta' X'(I-\lambda W')^{-1} \\
&+ (I-\lambda W)^{-1}X\beta u'(I-\lambda W')^{-1} + (I-\lambda W)^{-1}uu'(I-\lambda W')^{-1} \\
&= (I-\lambda W)^{-1}X\beta\beta' X'(I-\lambda W')^{-1} + (I-\lambda W)^{-1}uu'(I-\lambda W')^{-1} \\
&= (I-\lambda W)^{-1}X\beta\beta' X'(I-\lambda W')^{-1} + (I-\lambda W)^{-1}(I-\lambda W')^{-1}\sigma^2
\end{aligned}
$$

then

$$
\begin{aligned}
V(y) &= E(yy') - E(y) \\
&= [(I-\lambda W)'(I-\lambda W)]^{-1}\sigma^2 \\
&= \Omega\sigma^2
\end{aligned}
\tag{7}
$$

The associated likelihood function is then

$$\mathcal{L}\left(\sigma^2,\lambda,y\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^n |\sigma^2\Omega|^{-\frac{1}{2}} exp\left\{-\frac{1}{2\sigma^2}(y-(I-\lambda W)^{-1}X\beta)'\Omega^{-1}(y-(I-\lambda W)^{-1}X\beta)\right\}$$

the log likelihood

$$l\left(\sigma^2,\lambda,y\right) = constant - \frac{1}{2}ln|\sigma^2\Omega| - \frac{1}{2\sigma^2}(y-(I-\lambda W)^{-1}X\beta)'\Omega^{-1}(y-(I-\lambda W)^{-1}X\beta)$$

note that $|\sigma^2\Omega| = \sigma^{2n}|\Omega|$, and that

$$\begin{aligned} |\Omega| &= |[(I-\lambda W)'(I-\lambda W)]^{-1}| \\ &= |(I-\lambda W)^{-1}(I-\lambda W')^{-1}| \\ &= |(I-\lambda W)^{-1}||(I-\lambda W')^{-1}| \\ &= |(I-\lambda W)|^{-2} \end{aligned} \tag{8}$$

so returning to the log likelihood we have that the log likelihood is

$$
\begin{aligned}
l\left(\sigma^2,\lambda,y\right) = constant &- \frac{n}{2} ln\left(\sigma^2\right) + ln\left(\left|(I-\lambda W)\right|\right) \\
&- \frac{1}{2\sigma^2}(y-(I-\lambda W)^{-1}X\beta)'(I-\lambda W)'(I-\lambda W)(y-(I-\lambda W)^{-1}X\beta)
\end{aligned} \tag{9}
$$

then

$$
\begin{aligned}
l\left(\sigma^2,\lambda,y\right) = constant &- \frac{n}{2} ln\left(\sigma^2\right) \\
&- \frac{1}{2\sigma^2}((I-\lambda W)y-X\beta)'((I-\lambda W)-X\beta) \\
&+ ln\left(\left|(I-\lambda W)\right|\right)
\end{aligned} \tag{10}
$$

Spatial Lag Model

Maximum Likelihood Estimator

- The determinant $|(I - \lambda W)|$ is quite complicated because in contrast to the time series, where it is a triangular matrix, here it is a full matrix.
- However, Ord (1975) showed that it can be expressed as a function of the eigenvalues ω_i

$$|(I - \lambda W)| = \Pi_{i=1}^{n}(1 - \lambda\omega_i)$$

So the log likelihood is simplified to

$$\begin{aligned} l\left(\sigma^2, \lambda, y\right) = & \; constant - \frac{n}{2} ln\left(\sigma^2\right) \\ & - \frac{1}{2\sigma^2}((I - \lambda W)y - X\beta)'((I - \lambda W) - X\beta) \\ & + \sum ln(1 - \lambda\omega_i) \end{aligned} \tag{11}$$

Spatial Lag Model

Maximum Likelihood Estimator

Applying FOC, the ML estimates for β and σ^2 are:

$$\beta_{MLE} = (X'X)^{-1}X'(I - \lambda W)y$$

$$\sigma^2_{MLE} = \frac{1}{n}(y - \lambda Xy - X\beta_{MLE})'(y - \lambda Xy - X\beta_{MLE})$$

- Conditional on λ these estimates are simply OLS applied to the spatially filtered dependent variable and explanatory variables X.

Spatial Lag Model

Maximum Likelihood Estimator

- Substituting these in the log likelihood we have a concentrated log-likelihood as a nonlinear function of a single parameter λ

$$l(\lambda) = -\frac{n}{2} ln\left(\frac{1}{n}(e_0 - \lambda e_L)'(e_0 - \lambda e_L)\right) + \sum ln(1 - \lambda\omega_i) \quad (12)$$

- where e_0 are the residuals in a regression of y on X and
- e_L of a regression of Wy on X.

- This expression can be maximized numerically to obtain the estimators for the unknown parameters λ.

Spatial Lag Model

Maximum Likelihood Estimator

The asymptotic variance follows as the inverse of the information matrix

$$AsyVarl\left(\lambda, \beta, \sigma^2\right) = \begin{pmatrix} tr(W_A)^2 + tr(W_A' W_A) + \frac{(W_A X\beta)'(W_A X\beta)}{\sigma^2} & \frac{(X' W_A X\beta)'}{\sigma^2} & \frac{tr(W_A)'}{\sigma^2} \\ \frac{(X' W_A X\beta)'}{\sigma^2} & \frac{(X'X)}{\sigma^2} & 0 \\ \frac{tr(W_A)'}{\sigma^2} & 0 & \frac{n}{2\sigma^4} \end{pmatrix}^{-1} \quad (13)$$

- where $W_A = W(I - \lambda W)^{-1}$.
- Note that
  - the covariance between β and σ^2 is zero, as in the standard regression model,
  - this is not the case for λ and σ^2.

Spatial Lag Model

Two-Stage Least Squares estimators

- An alternative to MLE we can us 2SLS to eliminate endogeneity.
- Key is to identify proper instruments
  - Need to be uncorrelated with the error term
  - Correlated with WY

Spatial Lag Model

Two-Stage Least Squares estimators

Consider the following

$$E(y) = (I - \lambda W)^{-1} X\beta$$

now, since $|\lambda| < 1$ we can use Neumann series property to expand the inverse matrix as

$$(I - \lambda W)^{-1} = I + \lambda W + \lambda^2 W^2 + \lambda^3 W^3 + ...$$

hence

$$E(y) = (I + \lambda W + \lambda^2 W^2 + \lambda^3 W^3 + ...) X\beta$$

$$= X\beta + \lambda W X\beta + \lambda^2 W^2 X\beta + \lambda^3 W^3 X\beta + ...$$

so we can express $E(y)$ as a function of X, WX, W^2X,...

Spatial Lag Model

Two-Stage Least Squares estimators

We can use the first three elements of the expansion as instruments. Let's define H as the matrix with our instruments

$$H = [X, WX, W^2X]$$

Now,

$$y = \lambda Wy + X\beta + u$$

$$= M\theta + u$$

where $M = [Wy, X]$ and $\theta = [\lambda, \beta]$.

Spatial Lag Model

Two-Stage Least Squares estimators

The first stage is

$$M = H\gamma + \eta$$

and

$$\hat{\gamma} = (H'H)^{-1}H'M$$

$$\hat{M} = H\hat{\gamma} = P_H M$$

and the second stage is

$$y = \hat{M}\theta + u \tag{14}$$

and

$$\hat{\theta}_{2SLS} = (\hat{M}'\hat{M})^{-1}\hat{M}'y$$

Interpretation of Parameters

- Consider the following model for the $i - th$ observation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_r x_{ir} + \cdots + \beta_k x_{ik} \;\; i = 1, \ldots, n$$

- Recall that in OLS we have

$$\beta_1 = \frac{\partial y_i}{\partial x_{i1}}$$

or generically

$$\beta_r = \frac{\partial y_i}{\partial x_{ir}} \;\; \forall i = 1, \ldots, n \;\&\; r = 1, \ldots, k$$
$$\beta_r = \frac{\partial y_i}{\partial x_{jr}} \;\; \forall j \neq i \;\&\; \forall r = 1, \ldots, k$$

- Interpretation is straight forward as long as we take into account units
- In spatial models the interpretation is less immediate and require some clarification

Interpretation of Parameters

- Lets consider the case of a simple Spatial Lag model with a single regressor

$$y_i = \alpha + \beta x_i + \lambda \sum w_{ij} y_j + \epsilon_i \tag{16}$$

with $|\lambda| < 1$, and

$$\beta \neq \frac{\partial y_i}{\partial x_i}$$

$$\frac{\partial y_i}{\partial x_i} = diag(I - \lambda W)^{-1} \beta$$

- The impact depends also on the parameter λ
- The impact is different in each location

Interpretation of Parameters

More generally consider

$$
\begin{aligned}
y &= \lambda W y + X\beta + u \\
&= (I - \lambda W)^{-1} X\beta + (I - \lambda W)^{-1} u
\end{aligned}
$$

Then

$$E(y) = (I - \lambda W)^{-1} X\beta \tag{17}$$

we define

$$S(W) = (I - \lambda W)^{-1} \beta \tag{18}$$

Interpretation of Parameters

Therefore the impact of *each variable* x on y can be described through the partial derivatives $\frac{\partial E(y)}{\partial x}$ which can be arranged in the following matrix:

$$S(W) = \frac{\partial E(y)}{\partial x} = \begin{pmatrix} \frac{\partial E(y_1)}{\partial x_1} & \dots & \frac{\partial E(y_1)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial E(y_i)}{\partial x_1} & \dots & \frac{\partial E(y_i)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial E(y_n)}{\partial x_1} & \dots & \frac{\partial E(y_n)}{\partial x_n} \end{pmatrix} \tag{19}$$

Interpretation of Parameters

On this basis, LeSage and Pace (2009) suggested the following impact measures that can be calculated for each independent variable X_i included in the model

- *Average Direct Impact*: this measure refers to the impact of changes in the $i-th$ observation of x, which we denote x_i, on y_i. This is the average of all diagonal entries in S

$$
\begin{aligned}
ADI &= \frac{tr(S(W))}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n} S(W)_{ii}
\end{aligned} \tag{20}
$$

Interpretation of Parameters

- *Average Total Impact To* an observation: this measure is related to the impact produced on one single observation y_i. For each observation this is calculated as the sum of the $i-th$ row of matrix S

$$
\begin{aligned}
ATIT_j &= \frac{\iota' S(W)}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n} S(W)_{ij}
\end{aligned}
\tag{21}
$$

Interpretation of Parameters

- *Average Total Impact From* an observation: this measure is related to the total impact on all other observations y_i . For each observation this is calculated as the sum of the $j-th$ column of matrix S

$$
\begin{aligned}
ATIF_i &= \frac{1}{n} S(W) \iota \\
&= \frac{\sum_{j=1}^{n} S(W)_{ij}}{n}
\end{aligned} \tag{22}
$$

Interpretation of Parameters

- A Global measure of the average impact obtained from the two previous measures.
- It is simply the average of all entries of matrix S

$$ATI = \frac{1}{n}\iota' S(W)\iota = \frac{1}{n}\sum_{i=1}^{n} ATIT_i = \frac{1}{n}\sum_{j=1}^{n} ATIF_i \tag{23}$$

- The numerical values of the summary measures for the two forms of average total impacts are equal.
- The ATIF relates how changes in a single observation j influences all observations.
- In contrast, the ATIT considers how changes in all observations influence a single observation i.

Interpretation of Parameters

- *Average Indirect Impact* obtained as the difference between ATI and ADI

$$AII = ATI - ADI \tag{24}$$

- It is simply the average of all off-diagonal entries of matrix S

Interpretation of Parameters: Example

- We have data on 20 Italian regions on GDP and unemployment.
- We want to estimate the effect of GDP on Unemployment (Okun's Law)

	OLS	Spatial Lag Model
Intercept	10.971***	3.12275***
GDP	-3.326***	-1.13532***
λ	-	0.7476***
ADI	-	-1.542448
AII	-	-2.95571
ATI	-	-4.498159

Further Readings

- Arbia, G. (2014). A primer for spatial econometrics with applications in R. Palgrave Macmillan. (Chapter 2 and 3)
- Anselin, Luc, & Anil K Bera. 1998. “Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics.” Statistics Textbooks and Monographs 155. MARCEL DEKKER AG: 237–90.
- Sarmiento-Barbieri, I. (2016). An Introduction to Spatial Econometrics in R. http://www.econ.uiuc.edu/~lab/workshop/Spatial_in_R.html
- Tobler, WR. 1979. “Cellular Geography.” In Philosophy in Geography, 379–86. Springer.