Week 3_12
Spatial Regressions

CRP 3850/5850

Yujin Hazel Lee
Wenzheng Li

Objectives

- Introduce new spatial terms
- Learn different types of spatial models and their use
- Run regressions for different models

Highly Recommended Readings

- Lesage, J., 2014. What Regional Scientists Need to Know about Spatial Econometrics
- Chi and Zhu, 2008. Spatial Regression Models for Demographic Analysis
- Ruttenauer, T., 2019. Spatial Regression Models: A Systematic Comparison of Different Model Specifications Using Monte Carlo Experiments

(2014) 44, 13–32

The Review of Regional Studies

The Official Journal of the Southern Regional Science Association

2014 SRSA Fellows Address: San Antonio, Texas March 28, 2014
53rd Annual Meetings of the Southern Regional Science Association

What Regional Scientists Need to Know about Spatial Econometrics*

James P. LeSage

Department of Finance and Economcs, Texas State University-San Marcos, USA

Spatial Regression Models for Demographic Analysis

Published: 27 September 2007

Volume 27, pages 17–42, (2008) Cite this article

Spatial Regression Models: A Systematic Comparison of Different Model Specifications using Monte Carlo Experiments

Tobias Rüttenauer

Nuffield College
University of Oxford
New Road
Oxford, OX1 1NF
United kingdom
tobias.ruttenauer@nuffield.ox.ac.uk

RECAP

Spatial Weights – contiguity/ distance-based/block weight/ Kernel matrices

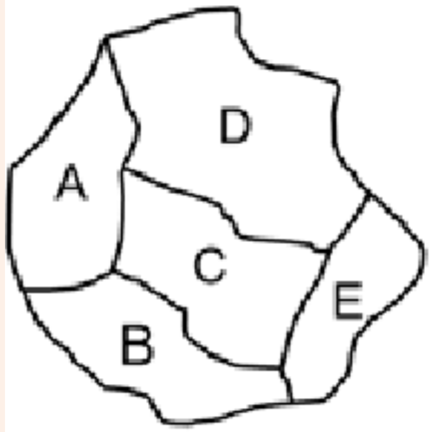
Spatial Autocorrelation – Global Moran's I, LISA map

Linear Regressions - OLS, Interpreting results

Spatial Regressions

What is a Spatial Lag?

- In linear regressions, a ‘**lag**’ shifts the variable by one or more periods in time
- The ‘**spatial lag operator**’ (Wy or Wx or Wu) relates a variable at one point in space to the observation for that variable in other spatial units in the system
- We assume that the values in one unit are indirectly influenced by the values found in the neighbors



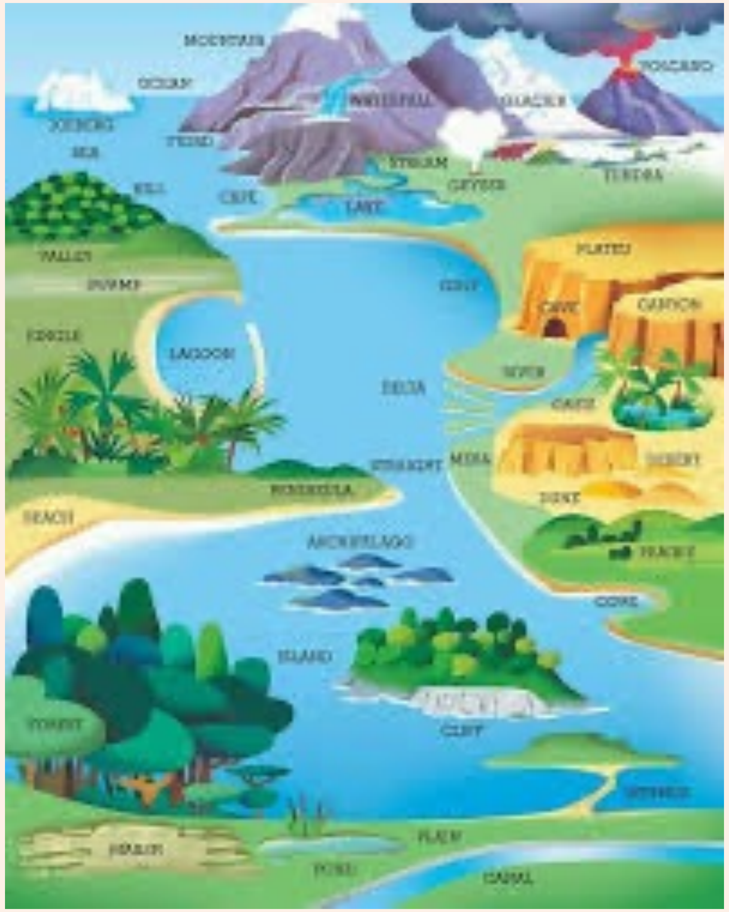
	A	B	C	D	E
A	0	1	1	1	0
B	1	0	1	0	1
C	1	1	0	1	1
D	1	0	1	0	1
E	0	1	1	1	0

Spatial Characteristics and Terminology

- **Spatial Dependence** (= Spatial Autocorrelation) – lack of independence between observations - it relates to the characteristics of the observations in the dataset (i.e. distance)

- **Spatial Heterogeneity** – lack of stability over space and the behavior of other relationships – geographical characteristics are considered as spatial heterogeneity

- **Spatial Regimes** – aggregation of neighboring units that are homogenous in functional terms that “share the same relationship between a dependent variable and some covariates (i.e. urban/rural areas; coastal areas)



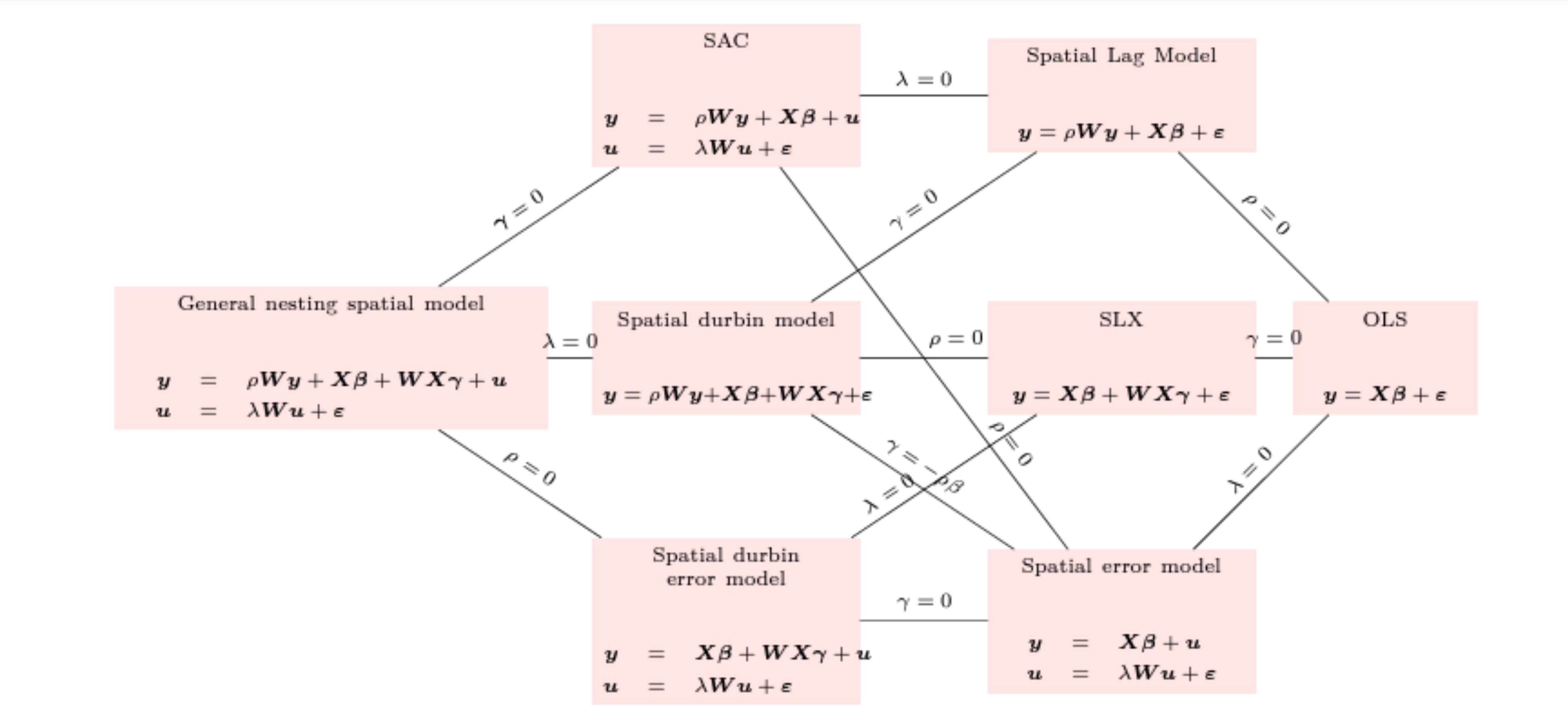
What is a Spatial Regression?

- Assumes that **spatial diffusion** affects the choices and behavioral patterns of regions - adding spatially weighted variables into the regression
- Based on the idea that social events are not simply an artifact of observed and unobserved variables, there’s a spatial component to understand neighbors’ behaviors
- Depending on whether the dependent variable or independent variable is lagged, the name of the models will vary (taxonomically)

- Spatial Regression is still a very new field, and is constantly developing

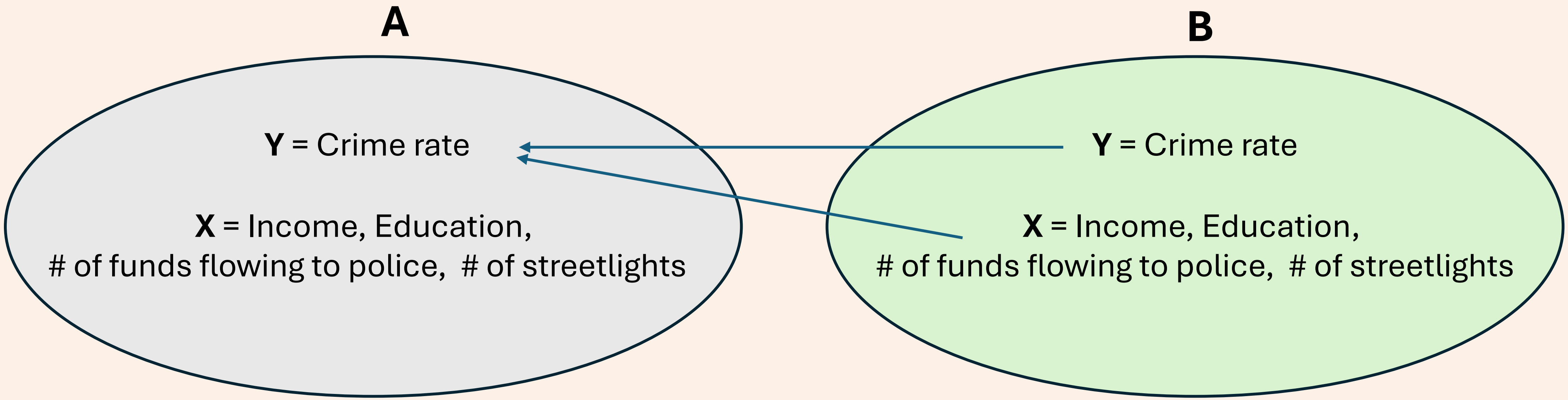
Taxonomy of Spatial Models

W = weights
$\rho = autocorrelation\ parameter\ for\ y$
$\gamma = autocorrelation\ parameter\ for\ x$
$\lambda = autocorrelation\ parameter\ for\ u$



Spatial Spillovers

- **Spatial Spillovers (Diffusion) results in Peer Effects (Indirect Effects)**
- The effect that neighboring regions gives to you, rather than characteristics of your own region



Model Interpretation

- Python results will differ depending on the model you run
- Basic statistics such as p-values, t-stats, and r-squared can be interpreted as the same way as we did in the linear regression
- **Results from the weighted (or lagged) variables will give you the spatial effects**

```
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :     unknown
Weights matrix      :        None
Dependent Variable  :Insurer Nonrenewals                Number of Observations:        1282
Mean dependent var  :    137.9072                Number of Variables   :          19
S.D. dependent var  :    131.3436                Degrees of Freedom    :        1263
R-squared           :      0.1764
Adjusted R-squared  :      0.1647
Sum squared residual: 1.82005e+07                F-statistic           :     15.0284
Sigma-square        :   14410.524                Prob(F-statistic)     :   3.937e-42
S.E. of regression  :     120.044                Log likelihood        :   -7947.541
Sigma-square ML     :   14196.952                Akaike info criterion :   15933.082
S.E of regression ML:    119.1510                Schwarz criterion     :   16031.049

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       -72.97667        26.25543        -2.77949         0.00553
            Avg_Risk         1.82673        10.35052         0.17649         0.85994
            POP_SQMI         0.00075         0.00099         0.75799         0.44860
           Hous_Burd         1.67922         1.03101         1.62872         0.10362
           Education         2.16290         0.74618         2.89864         0.00381
             Poverty        -2.03009         0.66620        -3.04727         0.00236
        Unemployment        -1.07390         1.51413        -0.70925         0.47830
            Ling_Iso        -1.78042         0.91266        -1.95080         0.05130
               Ozone      3563.57676      2245.98469         1.58664         0.11284
                PM25        20.64495         6.10145         3.38361         0.00074
          W_Avg_Risk        26.58061        14.13528         1.88044         0.06028
          W_POP_SQMI        -0.00045         0.00118        -0.38539         0.70002
         W_Hous_Burd         2.44505         1.59971         1.52843         0.12666
         W_Education         3.31552         1.11298         2.97895         0.00295
           W_Poverty        -3.04462         0.96934        -3.14092         0.00172
      W_Unemployment         7.78908         2.77846         2.80338         0.00513
          W_Ling_Iso        -3.88419         1.32753        -2.92587         0.00350
             W_Ozone       936.12278      2287.16269         0.40929         0.68239
              W_PM25       -23.21078         6.57040        -3.53263         0.00043
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           185.697

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2        7802.166           0.0000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test               18         724.080           0.0000
Koenker-Bassett test             18         110.016           0.0000
================================ END OF REPORT =====================================
```

Model Interpretation

- Direct and Indirect Effects are given separately for spatial regressions
- **Direct Effects** (Local effects) - how the independent variables within my area is correlated with the dependent variable within my area
- **Indirect Effects** (Extralocal Effects) – known as the spillover effect. How the independent/dependent variables from my neighbors explains my dependent variable

Table 2. Logistic Coefficients for Regression Analyses of Residential Mobility Out of Census Tract of Origin: White PSID Householders, 1980 to 2003

	Model 1		Model 2	
Independent Variables	b	SE	b	SE
Extralocal Neighborhood Conditions				
Minority concentration in distance-weighted surrounding neighborhoods			–.0092***	.0021
Multiethnic indicator for distance-weighted surrounding neighborhoods			–.0971**	.0350
Change in minority concentration in distance-weighted surrounding neighborhoods			.0445**	.0163
Local Neighborhood Conditions				
Minority concentration	.0159**	.0053	.0231***	.0059
Squared minority concentration	–.0003*	.0001	–.0004**	.0002
Cubed minority concentration[a]	.0002*	.0001	.0002*	.0001
Multiethnic indicator	.0884**	.0332	.0848*	.0335
Change in minority concentration	.0075*	.0038	.0019	.0049
Poverty level	–.0056	.0032	–.0067*	.0033
Level of homeownership	–.0030**	.0011	–.0037***	.0011
Level of single motherhood	.0010	.0023	–.0010	.0023
Micro-Level Characteristics				
Age	–.1347***	.0062	–.1351***	.0063
Age squared	.0010***	.0001	.0010***	.0001
Female	.0634	.0506	.0549	.0506
Education	.0310***	.0067	.0323***	.0068
Family income (in $1,000's)	.0007***	.0002	.0009***	.0002
Employed	–.0486	.0616	–.0545	.0618
Married	–.2298***	.0442	–.2370***	.0442
Children	–.1055***	.0161	–.1069***	.0161
Homeowner	–1.0504***	.0380	–1.0622***	.0381
Household crowding	.0357***	.0071	.0347***	.0071
Long-term resident	–.2442***	.0384	–.2357***	.0383
Year (1980 = 0)	.0419***	.0033	.0427***	.0034
Constant	–80.367***	6.695	–81.687***	6.810
Wald chi-square	4228.38		4290.20	

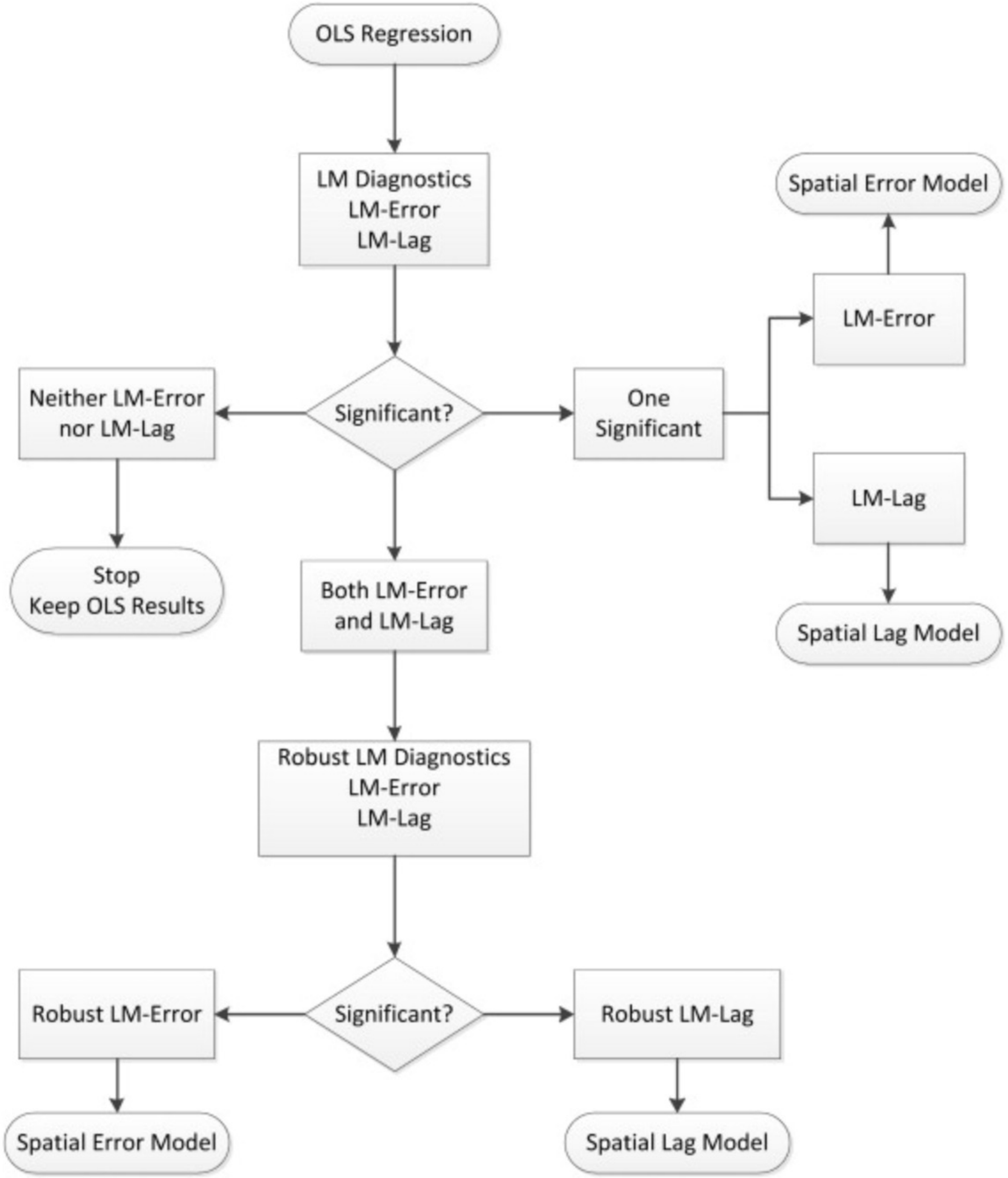
Note: N of observations = 48,508; N of persons = 7,622.
[a] Coefficients and standard errors multiplied by 100.
* $p < .05$; ** $p < .01$; *** $p < .001$ (two-tailed tests).

Model Choice

- Substantive choice
  - i.e. if we're looking at voting results, because elections happen on a single day, it's hard to argue that the results from neighboring regions will spillover
- A lot of the spatial effects are not clear cut though
  - i.e. the feedback loop is unclear (whether there are global or local spillover effects)
  - Sometimes hard to discern to what extent indirect effects would exist – do we assume investment for infrastructure are local or global?

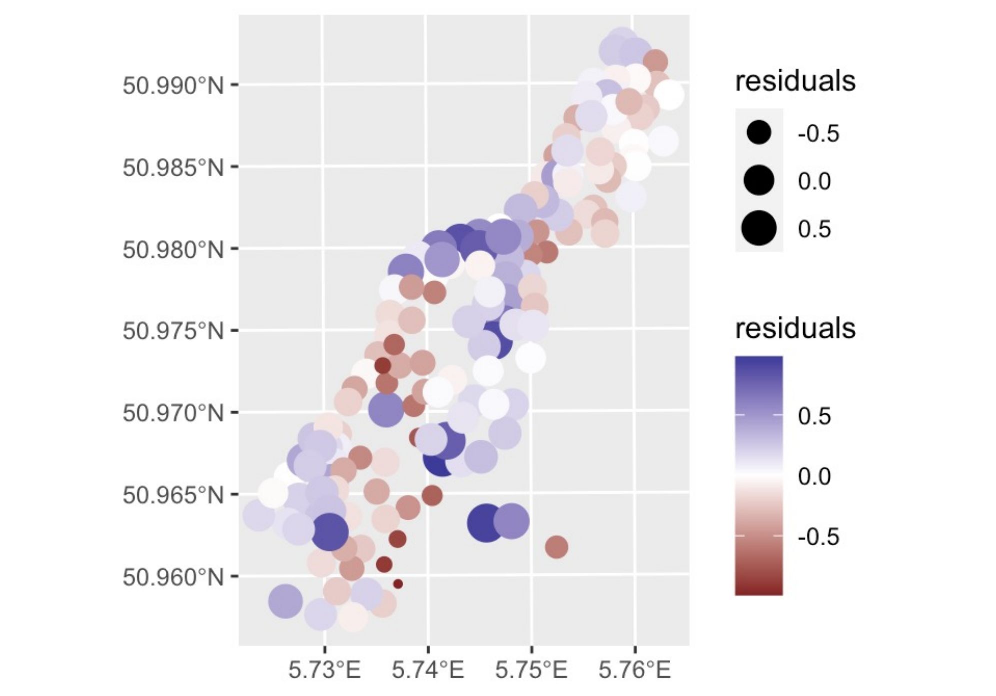
Model Choice

- **Data Driven Approach**
- Series of Lagrange-Multiplier tests that determine whether we would use a spatial lag or spatial error model



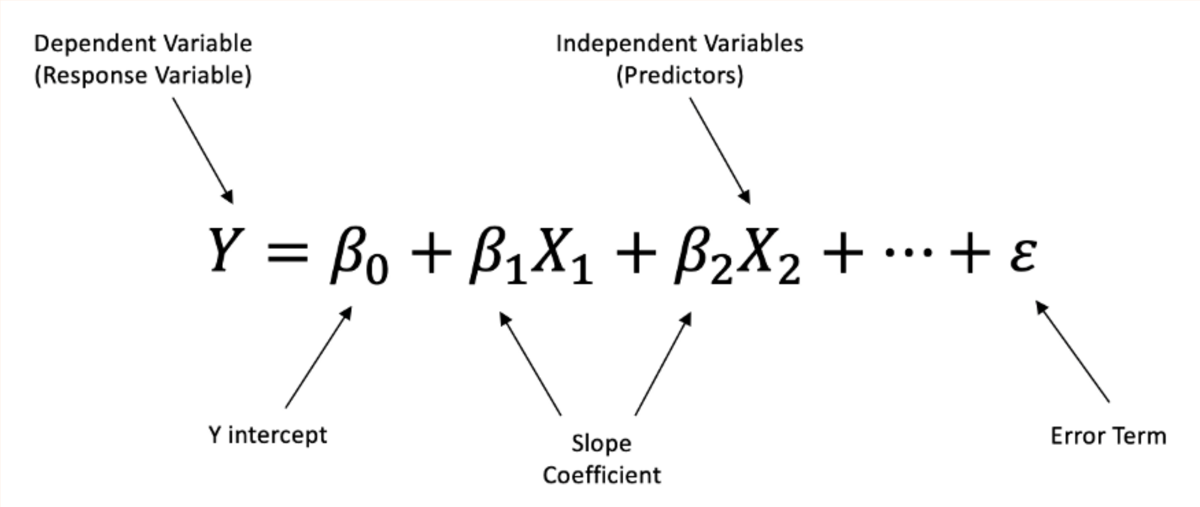
Non-Spatial Regressions

- Assumption: the errors are not correlated (complete spatial randomness of error terms)
- What if we see spatial autocorrelation in the error terms – would indicate that there are issues behind the structural ideas captured in the independent variables that have been unobserved



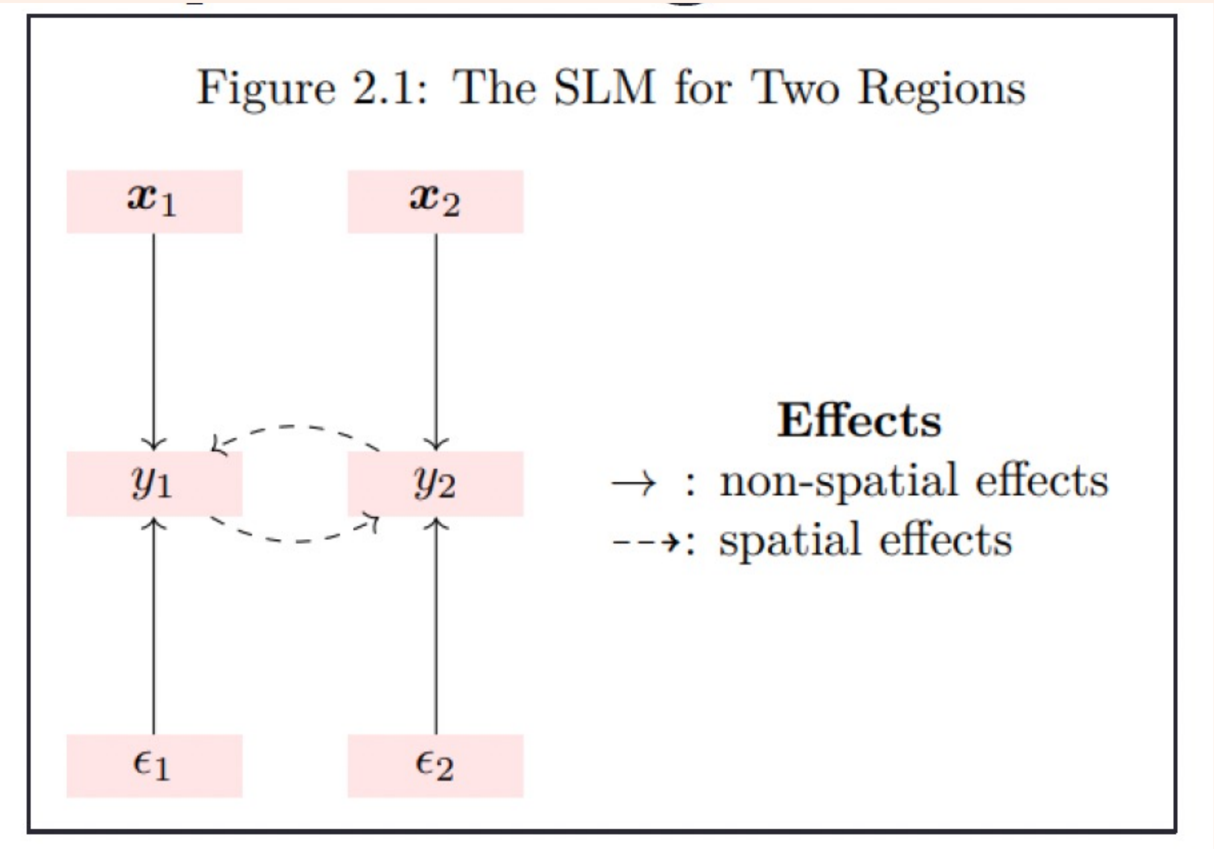
Non-Spatial Regressions

- Why not include regional dummy variable for each state or unit of observation?
  - Less parsimonious (the general rule of thumb is that the less variables are better)
  - Lost opportunity if interested in diffusion



Spatial Lag Model

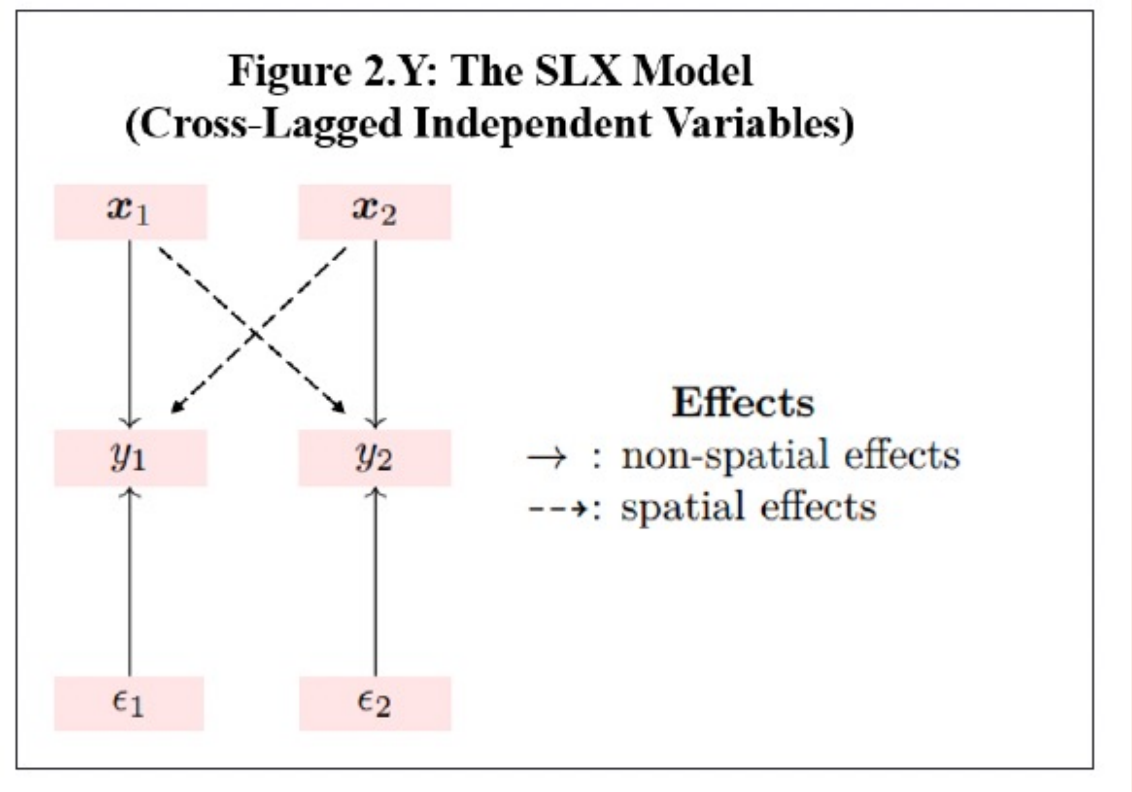
- Global model where we assume that the y values of our neighbors will affect the y values of us
- i.e. adoption of solar panels; air pollution



https://www.msarrias.com/uploads/3/7/7/8/37783629/spatialeconometrics.pdf

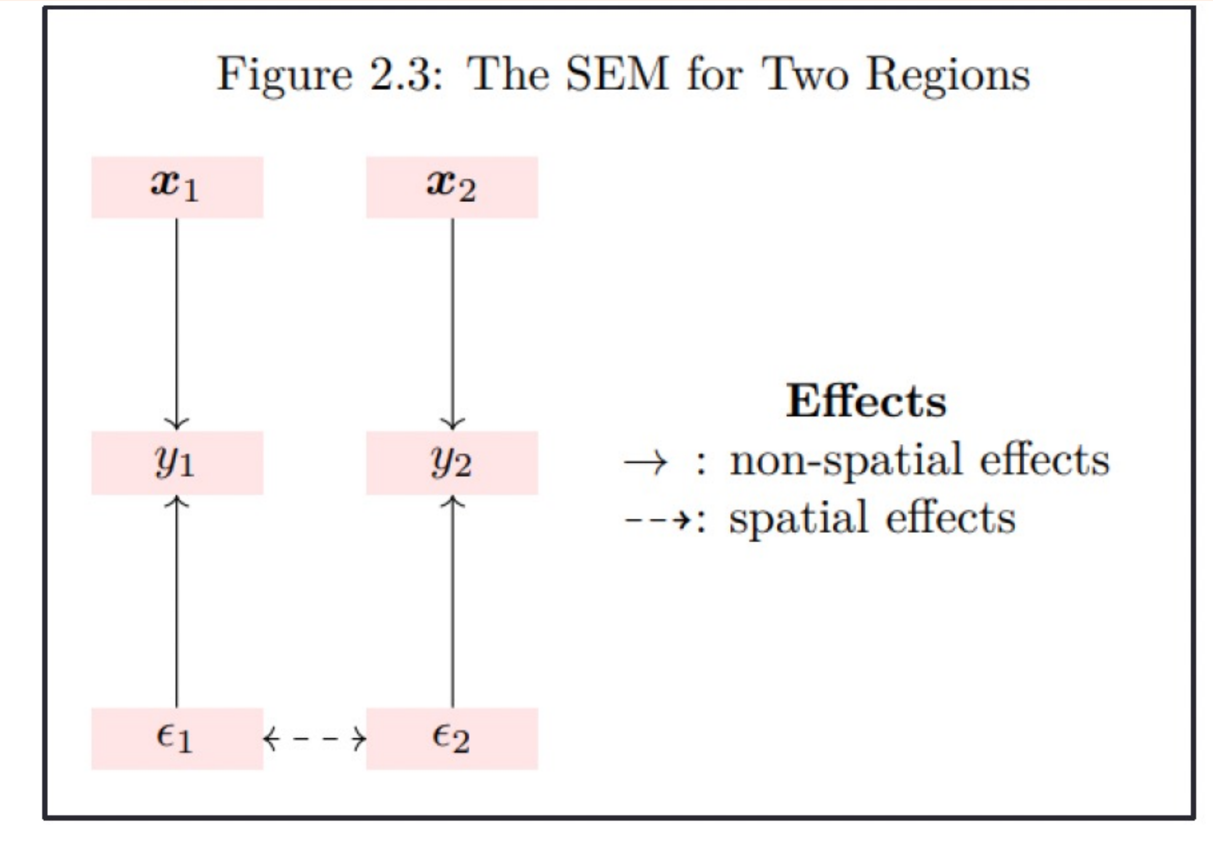
Spatially Lagged X Model (SLX)

- Lagged X variable
- Assuming that the neighbor's x values will spillover to my Y
- If we're looking at the average number of pupils in a school district, the birthrate of my neighbors can also affect the pupils for school districts



https://www.msarrias.com/uploads/3/7/7/8/37783629/spatialeconometrics.pdf

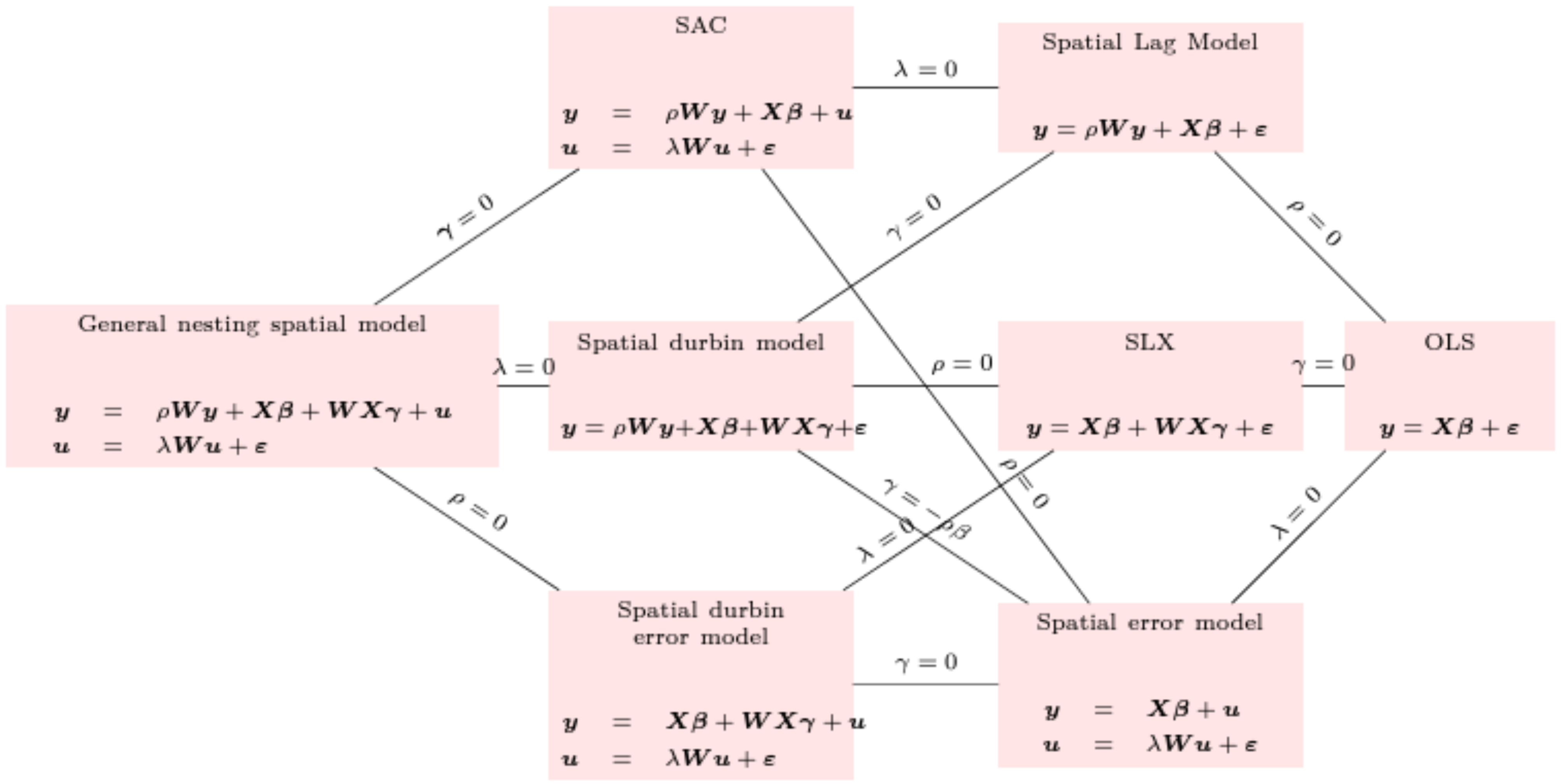
Spatial Error Model



https://www.msarrias.com/uploads/3/7/7/8/37783629/spatialeconometrics.pdf

- Assumes that the y and x-variable is not subject to spatial diffusion
- Autocorrelation parameter would be zero
- No indirect (spillover) effects caused in this model
- There could be spatial issues that are not included in the model – usually justified theoretically.

Taxonomy of Spatial Models

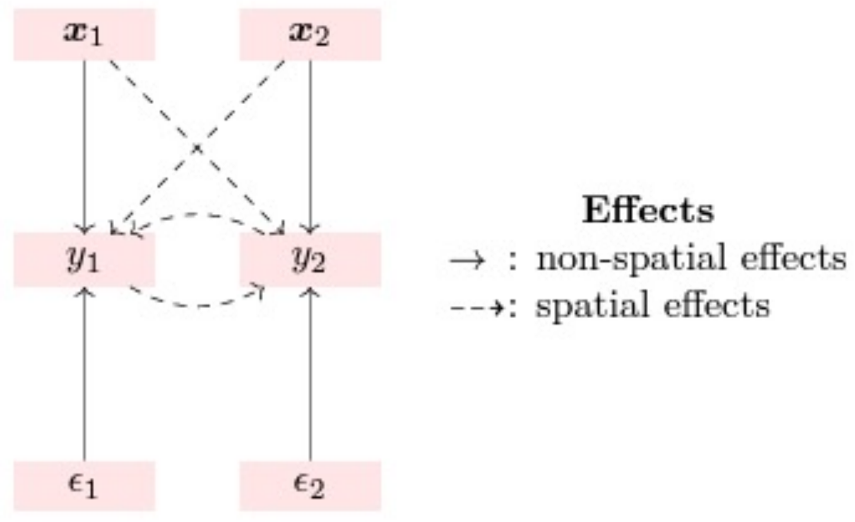


Spatial Durbin Model

$$\boldsymbol{y} = \rho \boldsymbol{W} \boldsymbol{y} + \alpha \boldsymbol{\imath}_n + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\gamma} + \epsilon$$

$$\boldsymbol{y} = \left(\boldsymbol{I}_n - \rho \boldsymbol{W}\right)^{-1} \left(\alpha \boldsymbol{\imath}_n + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}\right)$$

Figure 2.2: The SDM for Two Regions



https://www.msarrias.com/uploads/3/7/7/8/37783629/spatialeconometrics.pdf

- Lagged x and y

Spatial Durbin Error Model

Spatial Durbin Error Models

OLS: $y_i = \beta \boldsymbol{X}_i + \varepsilon_i$

Error: $y_i = \beta \boldsymbol{X}_i + (\boldsymbol{I} - \lambda \boldsymbol{w}_i)^{-1}\varepsilon$

SDEM: $y_i = \beta \boldsymbol{X}_i + \theta(w_i \boldsymbol{X}_i) + (\boldsymbol{I} - \lambda \boldsymbol{w}_i)^{-1}\varepsilon$

y_i	outcome variable for focal unit (i)
$\boldsymbol{X}_i$	vector of covariates
β	coefficients for those covariates
$w_i \boldsymbol{X}_i$	For each x in the vector of covariates, the average value of all neighbors of focal unit (i), based on a specified neighbor weight matrix
θ	Coefficients for each of the spatially lagged covariates
λ	correlation parameter (i.e. correlation coefficient of errors among neighbors). Bounded by -1 and 1. A value of zero indicates no autocorrelation
ε_i	Error (assumed IID and no longer spatially clustered)

- Lagged X and error

The Python Spatial Analysis Library
for open source, cross platform
Geospatial Data Science

Spatial Regression Models (spreg)

spreg, short for “spatial regression,” is a python package to estimate simultaneous autoregressive spatial regression models. These models are useful when modeling processes where observations interact with one another. For more information on these models, consult the Spatial Regression short course by Luc Anselin (Spring, 2017), with the Center for Spatial Data Science at the University of Chicago: