UNIVERSITÄT LEIPZIG

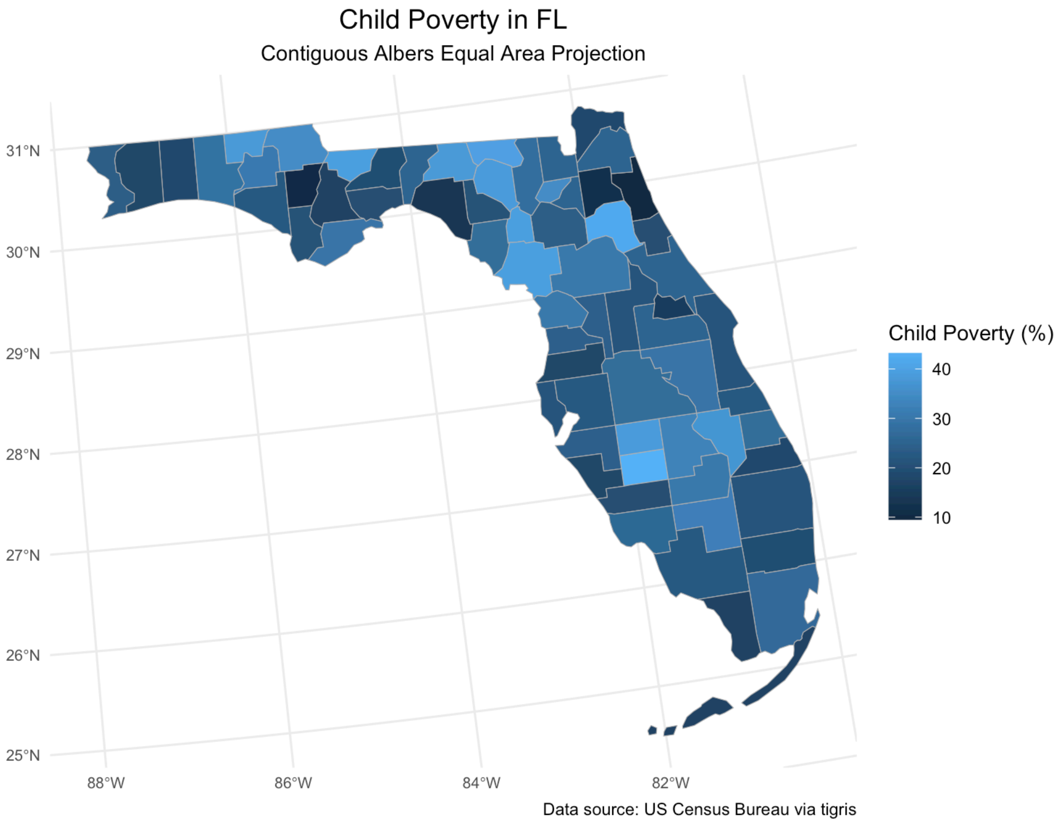
Toolbox CSS – Spatial Data II; inference with spatial data

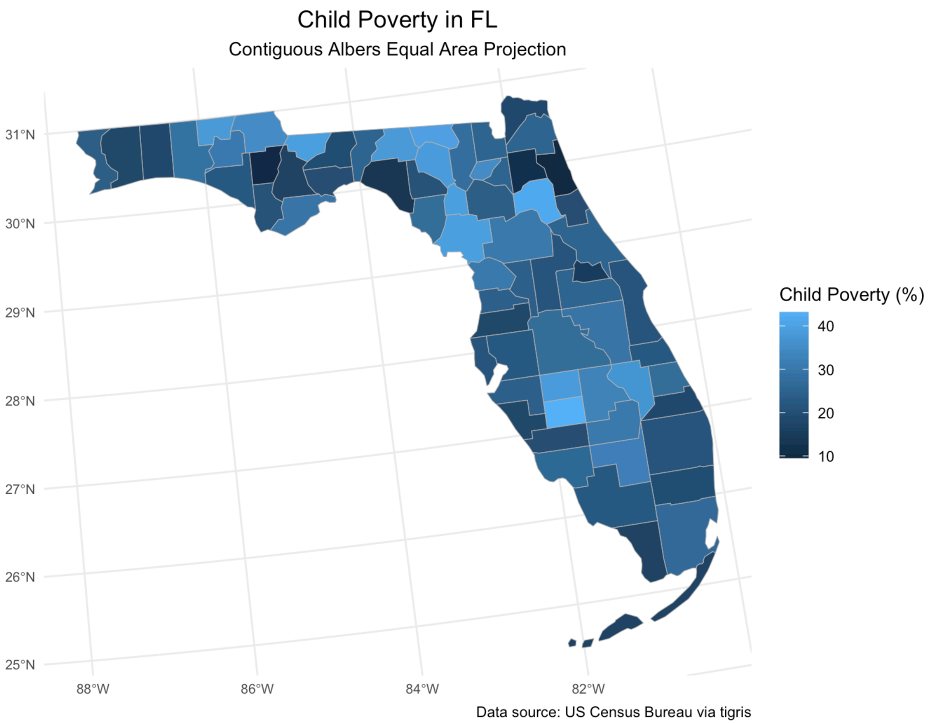
NSG SR 423, 07/01/2025

Felix Lennert, M.Sc.

OUTLINE

- Spatial autocorrelation and how it messes up your regressions
  - Recap: (Local) Moran's I
  - Spillovers
- How to address the problem
  - Lags
  - Errors
- The next weeks


Child Poverty in FL
Contiguous Albers Equal Area Projection
31°N
30°N
29°N
28°N
27°N
26°N
25°N
88°W
86°W
84°W
82°W
Child Poverty (%)
40
30
20
10
Data source: US Census Bureau via tigris



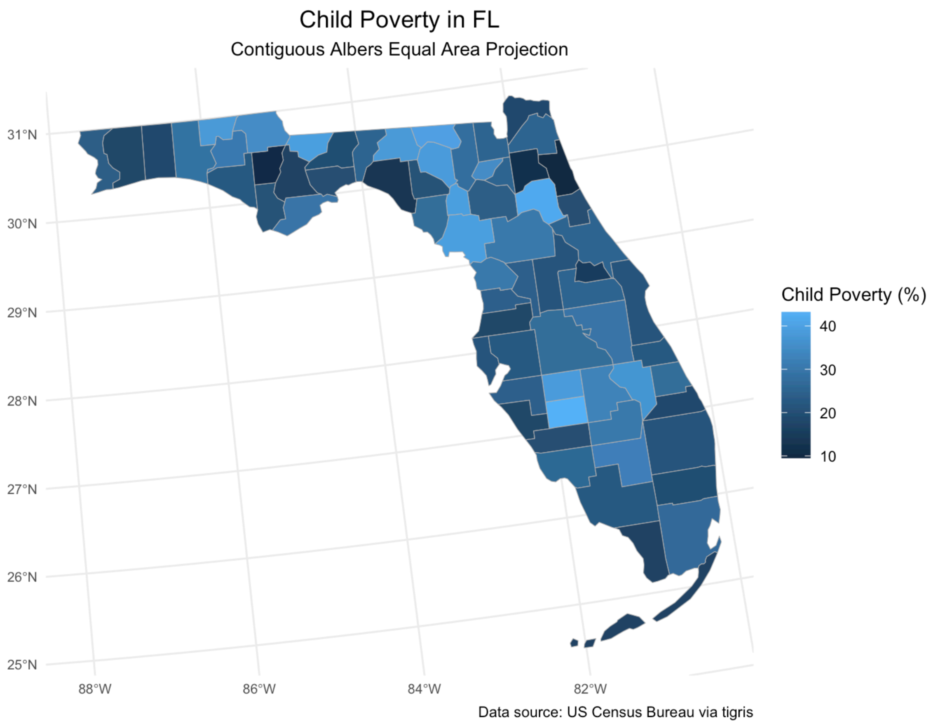
Which (macro-)factors could explain this?

- rural/urban
- race
- marriage
- health insurance
- teenage pregnancies
- job profiles
- income ratios
- something entirely different?

First Law of Geography:
"everything is related, but near things are more related than distant things" (Tobler 1970)

=> Spatial autocorrelation

=> We can measure this using **Moran's I**



AUTOCORRELATION/MORAN'S I

How much do these two points "matter" for each other – distance

$$I = \boxed{\frac{N}{W}} \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Adjustment by number of observations and connectivity

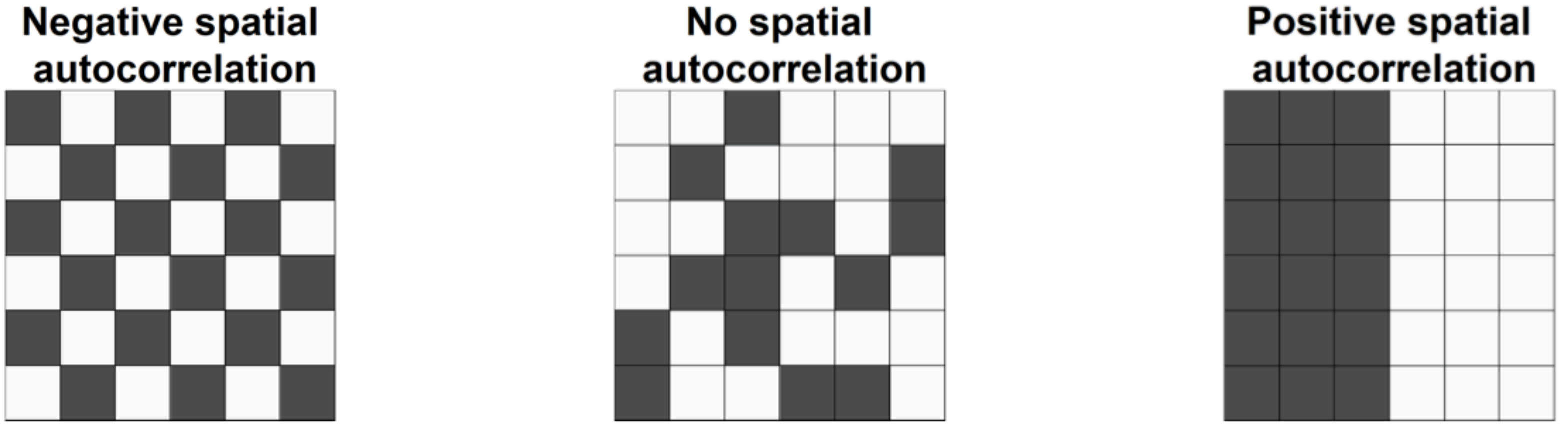
Normalizes by overall variation

How different are these two points from the mean respectively
=> does this for all possible locations
=> gets large if there are more extreme values

Significantly above 0 if…

- areas with similar values are closer to each other
- areas with dissimilar values are further apart from each other

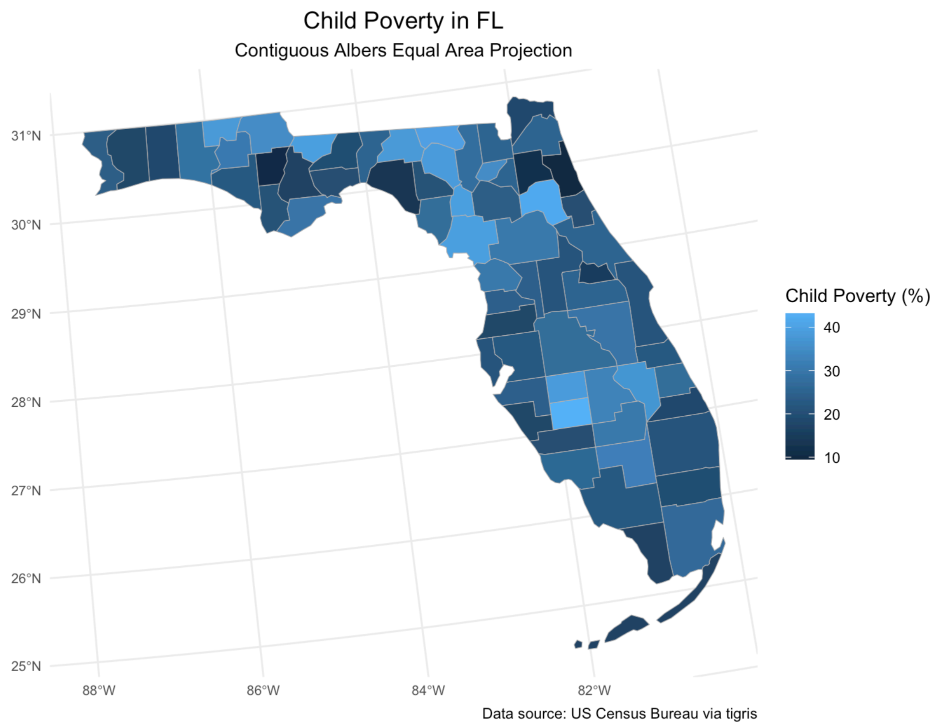
AUTOCORRELATION/MORAN'S I



```
        Moran I test under randomisation

data:  us_counties_child_pov$child_poverty
weights: fl_al_weights

Moran I statistic standard deviate = 1.7891, p-value = 0.0368
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.129925893      -0.015151515       0.006575735
```



Problem: running an OLS regression requires our observations to be independent from each other – spatial data usually violates this

	Dependent variable:
	child poverty
Rural	−10.793
Urban	−2.126
Manufacturing jobs	0.017
Age	1.624
Retail Jobs	4.063
Health Care Jobs	6.916
Construction Jobs	4.219
Less than High School	5.804
Unemployment	4.495
Single Moms	2.164
Share Black	0.727
Share Hispanic	−0.691
Uninsured Ind	9.717
Income Ratio	6.901
Share Teen Births	1.573
Share Unmarried	0.346
Constant	−90.984***
Observations	67
R^2	0.597
Adjusted R^2	0.468
Residual Std. Error	5.917 (df = 50)
F Statistic	4.632*** (df = 16; 50)
Note:	*p<0.1; **p<0.05; ***p<0.01

Do you see any problems here?
(except for the lack of stars)

AUTOCORRELATION IN OLS

Problem: running an OLS regression requires our observations to be independent from each other –
spatial data usually violates this

=> Spillovers

	Dependent variable:
	child poverty
Rural	−10.793
Urban	−2.126
Manufacturing jobs	0.017
Age	1.624
Retail Jobs	4.063
Health Care Jobs	6.916
Construction Jobs	4.219
Less than High School	5.804
Unemployment	4.495
Single Moms	2.164
Share Black	0.727
Share Hispanic	−0.691
Uninsured Ind	9.717
Income Ratio	6.901
Share Teen Births	1.573
Share Unmarried	0.346
Constant	−90.984***
Observations	67
R^2	0.597
Adjusted R^2	0.468
Residual Std. Error	5.917 (df = 50)
F Statistic	4.632*** (df = 16; 50)
Note:	*p<0.1; **p<0.05; ***p<0.01

AUTOCORRELATION AND SPILLOVERS

- Spillovers – neighboring states are somewhat more similar/influence each other
  => we need to control for this – omitted variable bias

- On the right-hand side of the regression formula (independent variables):
  neighborhoods might share local policies, cultural/geographic factors, economic shocks, etc.; also: my neighbor's problems might be similar to my problems
- On the left side (dependent variable):
  Interactions of child poverty between neighboring counties – e.g., higher likelihood of families moving between neighboring counties, shared labor markets (beyond what's covered in our observed data)

- Spillovers can be global or local
  - Local: neighboring states have an immediate, local effect on each other (LeSage 2014: cigarette smuggling across borders)
  - Global: these effects travel through the entire system (LeSage 2014: traffic congestion in county A leading to global effects overall)

AUTOCORRELATION AND SPILLOVERS

Problem: running an OLS regression requires our observations to be independent from each other – spatial data usually violates this

- What does this mean *in data*
  - Predictions are not equally good for all observations
  - Autocorrelation in residuals

    Residual = actual value – predicted value => $r_i = y_i - \hat{y}_i$

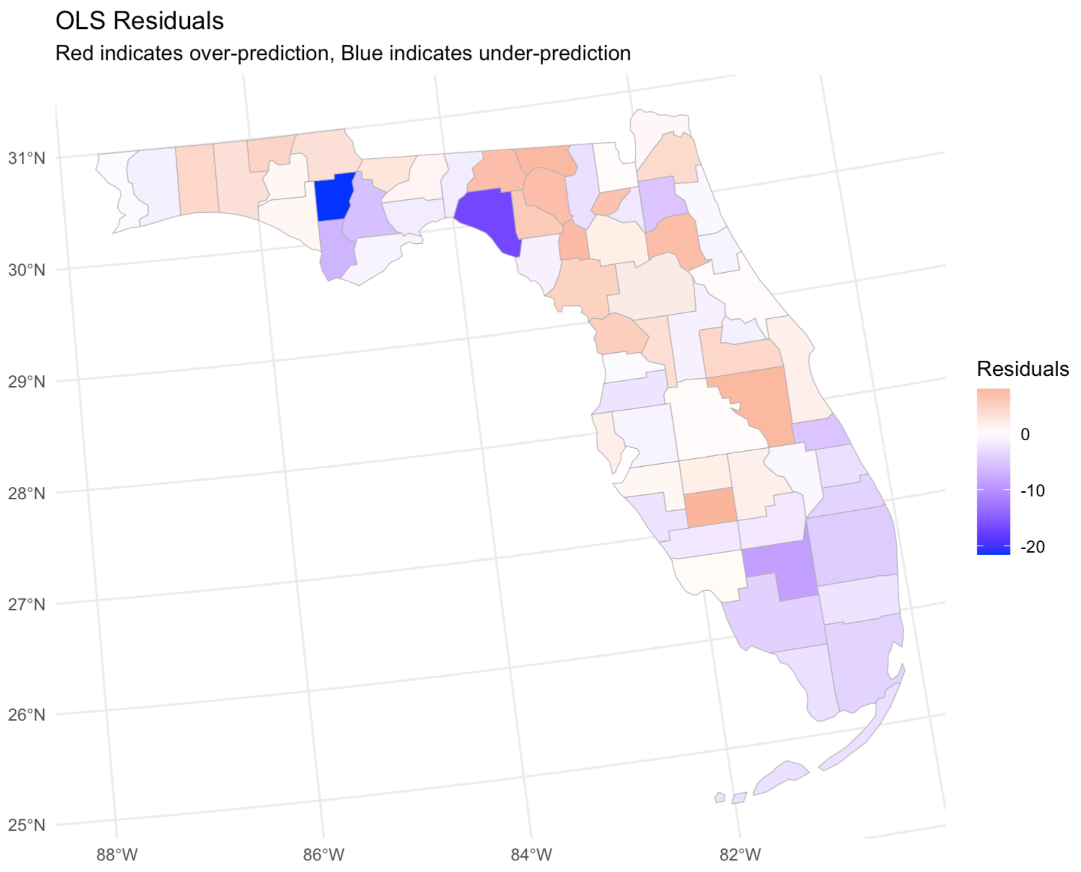
```
        Moran I test under randomisation

data:  us_counties_child_pov_res$residuals
weights: fl_al_weights

Moran I statistic standard deviate = 2.0127, p-value = 0.02207
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.141540396      -0.015151515       0.006060929
```

=> over-/under-prediction seems to be significantly clustered in space

=> including spatial dependencies should enhance model fit



LOCAL MORAN'S I

Global Moran's I gives us a measurement for the entirety of the units

Local Moran's I gives us an estimate *per unit i*

$$I_i = \frac{(x_i - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2/n} \sum_{j \in N_i} w_{ij}(x_j - \bar{x})$$

n = total number of spatial units

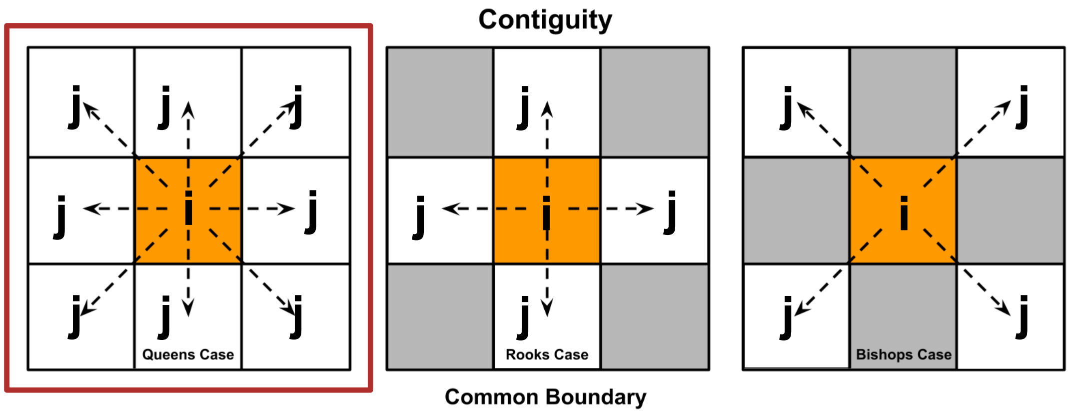
w_{ij} = spatial weight between i and j

x_i = value at location i

x_j = values of all neighboring units

x_k = values of all units

$\bar{x}$ = mean value



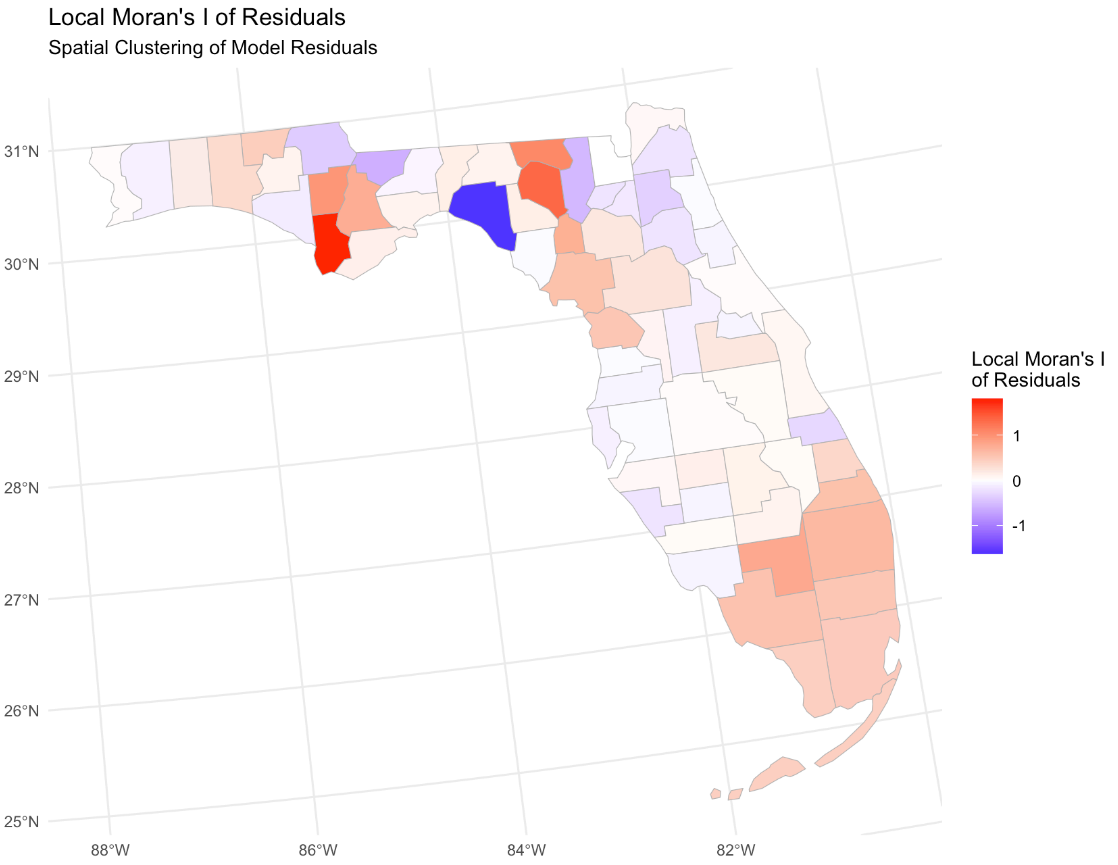
…other specifications exist, e.g., k nearest neighbors

TECHNICAL ASPECTS – LOCAL MORAN'S I

$$I_i = \frac{(x_i - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2/n} \sum_{j \in N_i} w_{ij}(x_j - \bar{x})$$

"how much does x_i differ from the mean" – z-standardized value of x_i

"how much do x_i's neighbors x_j differ from the mean" – z-standardized value of x_j


Local Moran's I of Residuals
Spatial Clustering of Model Residuals
31°N
30°N
29°N
28°N
27°N
26°N
25°N
88°W
86°W
84°W
82°W
Local Moran's I
of Residuals
1
0
-1

SPATIAL LAGS AND ERRORS

Autocorrelation can be modeled in two ways: lags and error

- Lags: something we observe in neighboring entities (here: counties) has an effect on our focal entity
- Error: something we do not observe yet that alters our results is the same for the focal entity and the neighboring ones

=> to get unbiased estimates, we need to include this in our models

SPATIALLY LAGGED X VARIABLES (SLX)

Main idea: how do characteristics of neighboring counties affect the focal county

Solution: include neighboring counties' average values for each independent variable

$$y = X\beta + WX\theta + \varepsilon$$

with $WX\theta$ being the average value of the neighbors independent variables

- Unidirectional: neighboring counties' values can impact focal county

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -114.3738   124.8242  -0.916   0.3660
rural               -5.9267    18.4473  -0.321   0.7500
urban               -1.3428     3.6842  -0.364   0.7178
lnmanufacturing     -2.5830     2.5867  -0.999   0.3251
lnag                 1.8314     1.4228   1.287   0.2067
lnretail             1.2280     5.3221   0.231   0.8189
lnhealthss          -4.2070     7.6829  -0.548   0.5876
lnconstruction       5.0867     4.3053   1.182   0.2456
lnlesshs             1.6336     5.5744   0.293   0.7713
lnunemployment       8.7636    10.1393   0.864   0.3935
lnsinglemom          4.2672     5.2981   0.805   0.4262
lnblack              0.8085     2.3445   0.345   0.7323
lnhispanic           0.3910     3.0728   0.127   0.8995
lnuninsured          1.3458    12.6768   0.106   0.9161
lnincome_ratio       9.7831     9.3062   1.051   0.3006
lnteenbirth          2.5151     2.5712   0.978   0.3349
lnunmarried          1.3596     4.6326   0.293   0.7709
W_rural            -11.3333    34.0115  -0.333   0.7410
W_urban             -8.9307     8.1773  -1.092   0.2825
W_lnmanufacturing    3.2345     6.1164   0.529   0.6004
W_lnag               2.5738     3.0888   0.833   0.4105
W_lnretail          25.5169    12.2309   2.086   0.0445 *
W_lnhealthss        31.7171    16.7790   1.890   0.0673 .
W_lnconstruction    -2.2581    11.5896  -0.195   0.8467
W_lnlesshs           2.5208    13.2305   0.191   0.8500
W_lnunemployment   -22.4139    21.1983  -1.057   0.2978
W_lnsinglemom       -4.8413    12.7953  -0.378   0.7075
W_lnblack           -0.7456     5.5445  -0.134   0.8938
W_lnhispanic         3.1679     6.4761   0.489   0.6279
W_lnuninsured      -12.6350    31.7693  -0.398   0.6933
W_lnincome_ratio    -9.3658    22.5827  -0.415   0.6809
W_lnteenbirth        4.3760     6.0091   0.728   0.4715
W_lnunmarried       -0.4253     9.1812  -0.046   0.9633

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.662 on 34 degrees of freedom
Multiple R-squared:  0.7491,    Adjusted R-squared:  0.513
F-statistic: 3.173 on 32 and 34 DF,  p-value: 0.0006266
```

“Normal” OLS coefficients

coefficients of independent variables of neighbors
=> if x increases in neighboring regions by 1, y in focal region increases by coefficient

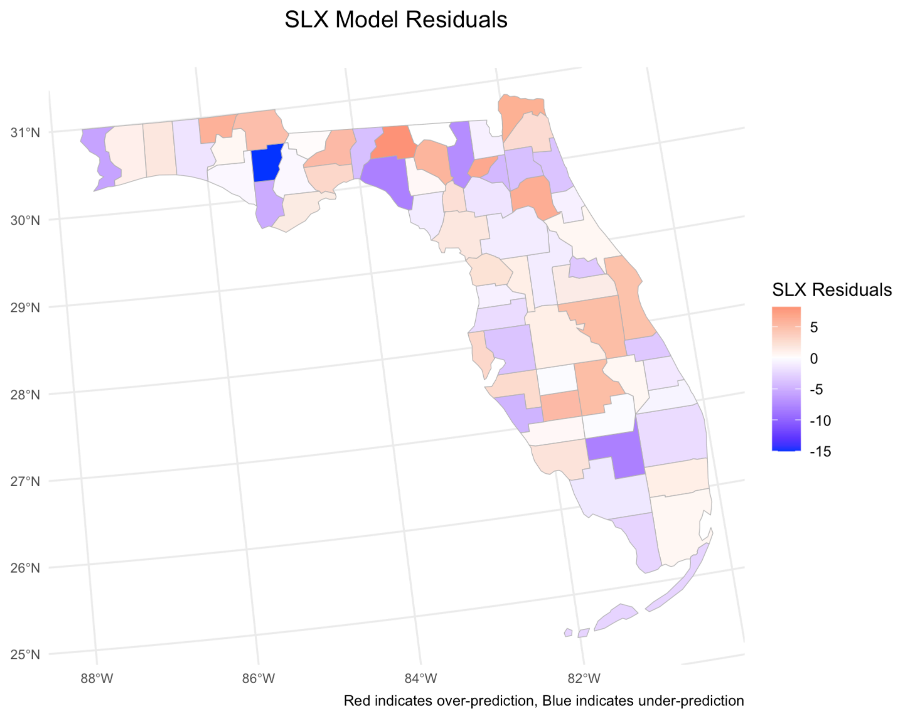
fit has improved – R^2 of OLS was 0.468

```
        Moran I test under randomisation

data:  us_counties_child_pov$slx_residuals
weights: fl_al_weights

Moran I statistic standard deviate = -0.83399, p-value = 0.7979
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     -0.081608083      -0.015151515       0.006349685
```

=> over-/under-prediction not significantly clustered in space anymore



SPATIALLY LAGGED Y VARIABLES (SAR)

Main idea: how do characteristics of neighboring counties' outcome variable affect the focal county's outcome variable

Solution: include neighboring counties' outcome values

$$y = X\beta + \rho Wy + \varepsilon$$

with Wy being the weighted average value of the neighbors outcome variable

- Global spillover: effect ripples across neighbors and to focal unit – effect is not limited to the focal area
- Should be theoretically justified

```
Type: lag
Coefficients: (asymptotic standard errors)
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -86.50330   28.97481 -2.9855 0.002831
rural            -9.49166   11.60343 -0.8180 0.413354
urban            -1.59776    2.36777 -0.6748 0.499807
lnmanufacturing  -0.42812    1.65571 -0.2586 0.795965
lnag              1.29667    1.06902  1.2130 0.225147
lnretail          2.53321    4.03190  0.6283 0.529812
lnhealthss        4.43444    5.31895  0.8337 0.404447
lnconstruction    4.25652    3.04978  1.3957 0.162812
lnlesshs          5.91324    3.79790  1.5570 0.119476
lnunemployment    6.42765    6.31359  1.0181 0.308647
lnsinglemom       2.87869    3.47234  0.8290 0.407084
lnblack           0.85718    1.61375  0.5312 0.595300
lnhispanic       -0.73967    1.83809 -0.4024 0.687382
lnuninsured       8.35100    8.51561  0.9807 0.326755
lnincome_ratio    6.89367    6.77731  1.0172 0.309073
lnteenbirth       1.41859    1.76632  0.8031 0.421899
lnunmarried       0.57856    2.81877  0.2053 0.837375

Rho: 0.20847, LR test value: 1.5856, p-value: 0.20796
Asymptotic standard error: 0.13068
    z-value: 1.5953, p-value: 0.11065
Wald statistic: 2.5449, p-value: 0.11065

Log likelihood: -203.5836 for lag model
ML residual variance (sigma squared): 25.256, (sigma: 5.0256)
Number of observations: 67
Number of parameters estimated: 19
AIC: 445.17, (AIC for lm: 444.75)
LM test for residual autocorrelation
test value: 4.3808, p-value: 0.036347
```

Coefficients not directly interpretable
=> direction and significance – a bit like Pearson's r
=> to get effect size: impact – see this week's lab

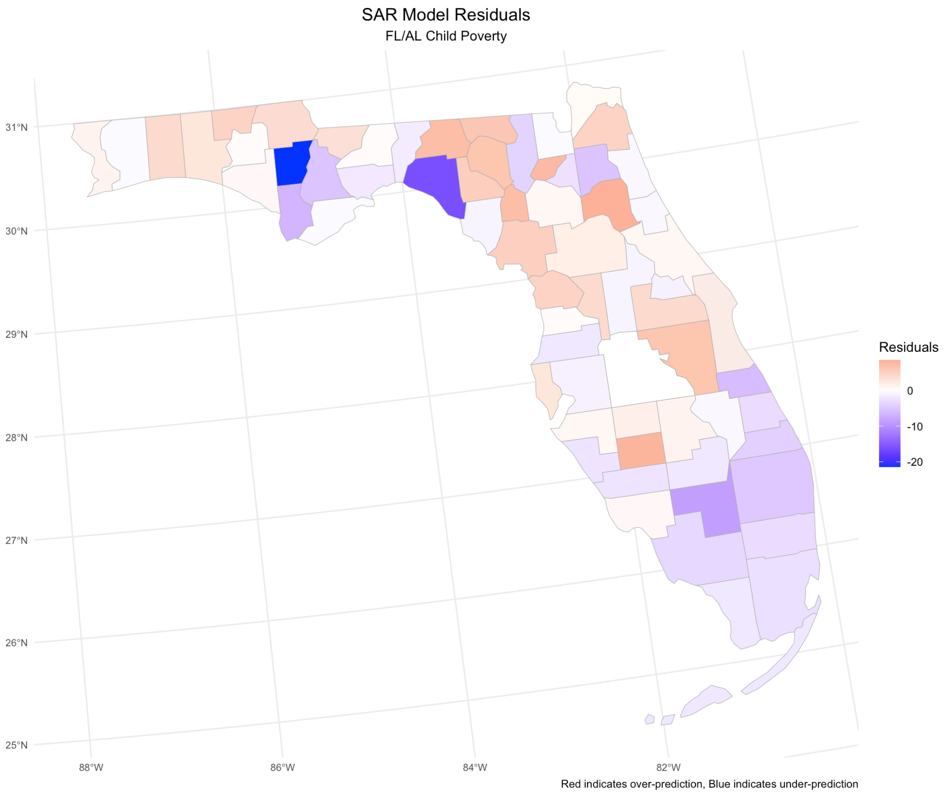
Insignificant – including lagged y does not tell us anything

AIC is higher (== worse) than basic OLS

```
        Moran I test under randomisation

data:  us_counties_child_pov$sar_residuals
weights: fl_al_weights

Moran I statistic standard deviate = 1.5529, p-value = 0.06023
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.105699393      -0.015151515       0.006056513
```



=> over-/under-prediction not significantly clustered in space anymore

=> however, still more clustered in space than with SLX model

SPATIAL ERROR MODEL (SEM)

Main idea: there might be things we can’t measure that affect the focal county and the neighboring ones

Solution: include an error term for the focal county and the neighboring counties

$y = X\beta + u$

with $u = \lambda W u + \varepsilon$, the function of our unexplained error (ϵ) and our neighbors residual values

- Assumption: some clustered residuals are higher than expected and, therefore, there needs to be another missing variable that we cannot account for with our data

```
Call:errorsarlm(formula = model_formula, data = us_counties_child_pov,
    listw = fl_al_weights, zero.policy = TRUE)

Residuals:
      Min        1Q    Median        3Q       Max
-18.95014  -1.97514   0.25588   2.61763   9.27680

Type: error
Coefficients: (asymptotic standard errors)
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -75.20687   29.23447 -2.5725  0.01010
rural            -4.50894   10.81857 -0.4168  0.67684
urban            -0.59601    2.06935 -0.2880  0.77333
lnmanufacturing  -2.38896    1.62885 -1.4667  0.14247
lnag              0.92031    0.99846  0.9217  0.35667
lnretail         -0.37313    3.53117 -0.1057  0.91585
lnhealthss        0.50771    5.38915  0.0942  0.92494
lnconstruction    4.64469    2.71901  1.7082  0.08759
lnlesshs          6.07122    3.55716  1.7068  0.08787
lnunemployment    9.08747    6.34399  1.4325  0.15201
lnsinglemom       2.54767    3.28203  0.7762  0.43760
lnblack           1.86149    1.45370  1.2805  0.20036
lnhispanic       -0.65779    1.96083 -0.3355  0.73728
lnuninsured       9.57106    8.37092  1.1434  0.25289
lnincome_ratio    8.55170    6.11118  1.3994  0.16171
lnteenbirth       1.26952    1.51936  0.8356  0.40340
lnunmarried       1.16675    2.49139  0.4683  0.63956

Lambda: 0.56704, LR test value: 7.6788, p-value: 0.0055872
Asymptotic standard error: 0.11623
    z-value: 4.8785, p-value: 1.0691e-06
Wald statistic: 23.799, p-value: 1.0691e-06

Log likelihood: -200.537 for error model
ML residual variance (sigma squared): 21.322, (sigma: 4.6176)
Number of observations: 67
Number of parameters estimated: 19
AIC: 439.07, (AIC for lm: 444.75)
```

Interpretation similar to OLS coefficients; but: need to take into account errors => here: positive lambda, hence positive spatial correlation in errors – unobserved factors influence neighboring counties

Significant Lambda – error term is spatially autoregressive

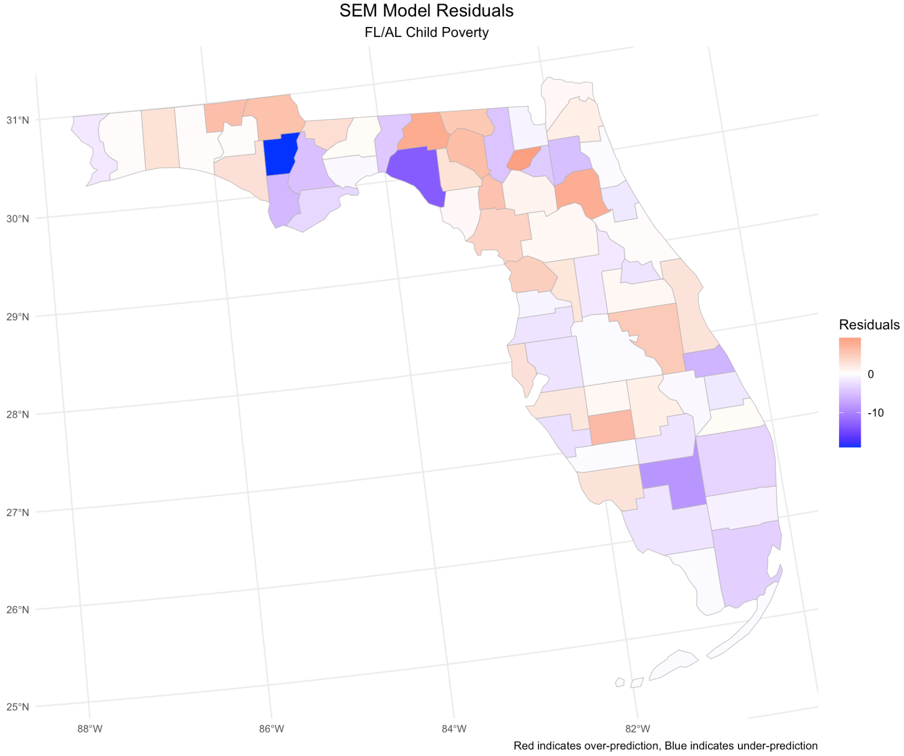
AIC is lower (== better) than basic OLS

```
        Moran I test under randomisation

data:  us_counties_child_pov$sem_residuals
weights: fl_al_weights

Moran I statistic standard deviate = 0.29554, p-value = 0.3838
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.008036192      -0.015151515       0.006155852
```

=> over-/under-prediction not significantly clustered in space anymore



HERE

SEM shows best performance (lowest AIC, removes Spatial autocorrelation in residuals)

More on this: this week's lab

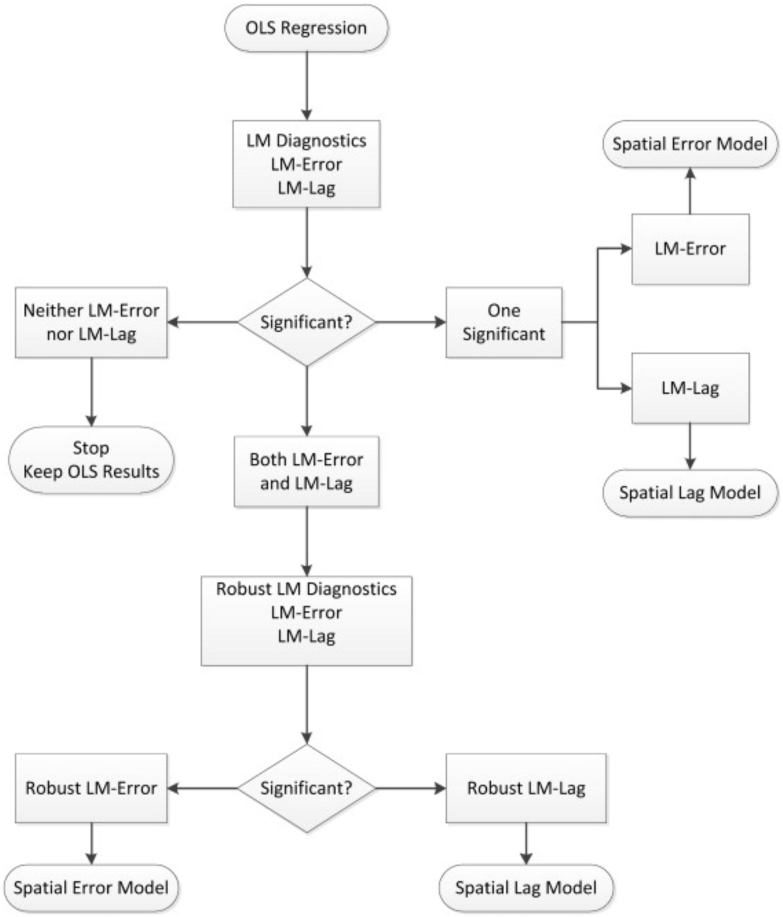
SO, WHAT TO DO IN PRACTICE?

Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity

Luc Anselin

Abstract

Several diagnostics for the assessment of model misspecification due to spatial dependence and spatial heterogeneity are developed as an application of the Lagrange Multiplier principle. The starting point is a general model which incorporates spatially lagged dependent variables, spatial residual autocorrelation and heteroskedasticity. Particular attention is given to tests for spatial residual autocorrelation in the presence of spatially lagged dependent variables and in the presence of heteroskedasticity. The tests are formally derived and illustrated in a number of simple empirical examples.


OLS Regression
LM Diagnostics
LM-Error
LM-Lag
Significant?
Neither LM-Error nor LM-Lag
Stop
Keep OLS Results
One Significant
LM-Error
Spatial Error Model
LM-Lag
Spatial Lag Model
Both LM-Error and LM-Lag
Robust LM Diagnostics
LM-Error
LM-Lag
Significant?
Robust LM-Error
Spatial Error Model
Robust LM-Lag
Spatial Lag Model

ANOTHER OPTION: SDM (SPATIAL DURBIN MODEL) AND SDEM (SPATIAL DURBIN ERROR MODEL)

SDM: includes lagged x and y of neighbors – can be simplified to SLX, SAR OLS

$$y = \rho Wy + X\beta + WX\theta + \varepsilon$$

SDEM: does only include lagged x of neighbors + error term – can be simplified to SEM, SLX, OLS

$$y = X\beta + WX\theta + u, \quad u = \lambda Wu + \varepsilon$$

LeSage 2014: Durbin Models should be used at all times – all other models are just social cases

Start with theory – global or local; then use appropriate model (local = SDEM, global = SDM)

More on this, including decision criteria: in this week's script

THE NEXT WEEKS

- Next two weeks: ABMs

- Then: 1 week sans class – work on your projects, prepare presentation
  => Presentation should include: motivation (w/ some theory/prior research), research question, data source, method, perhaps first results

- Deadline for presentation: January 29, 6PM, via email

- Will forward them to one of your peers who will serve as an opponent

- Now: feel free to stick around and ask questions

MERCI

Felix Lennert
Institut für Soziologie

felix.lennert@uni-leipzig.de
www.uni-leipzig.de