Spatial Regression

Modeling areal data in R

Back to Tobler…

- "All places are related but nearby places are more related than distant places."
- Social and physical phenomena are often highly clustered in space
  - e.g., regional voting patterns, racial segregation, the poverty belt, lung cancer, housing values, crime, farm crops, forest fires, animal habitats, plant species, soil chemistry

Spatial Analysis

- Often these spatial relationships are ignored
  - Weakens our ability generate meaningful inferences about the processes we study
- Spatial regression models include relationships between variables and their neighboring values
  - Include as explanatory variables the values of error terms, x or y values in surrounding regions
- Allows us to examine the impact that one observation has on other proximate observations

Why worry about spatial similarities(1)?

- It tells us something more about what we're studying
  - Is there an unmeasured process that affects the outcome we're interested in?
  - Does this process manifest itself in space?
  - Examples: interaction processes, diffusion, historical or ethnic legacy, programmatic effects

Why worry about spatial similarities(2)?

- Violation of regression assumptions
  - Residuals are uncorrelated with each other
  - Variance is not likely to be constant
- If we ignore the spatial relationships in our data:
  - Our estimated regression coefficients are biased/inconsistant
  - Our R^2 statistic is exaggerated
  - We've made incorrect inferences
  - We'll *never* get it published (or we shouldn't!!!)
- If spatial effects are present, and you don't account for them, your model is not accurate!

If spatial autocorrelation occurs

- There may be unmeasured x's which are causing the failure of independence
  - misspecification error
- There may be a "contagious" process at work (y's in one location may be affecting y's in adjacent locations)
- Value of y may depend on the value of x at the same site as well as nearby sites
- The errors in estimates may be spatially correlated between units

How to treat spatial component?

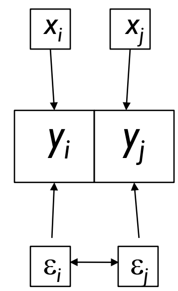
- As a substantive effect of interest
  - build into model/explore
  - e.g. spatial lag, spatial regimes, GWR
- As a nuisance effect due to specification errors
  - eliminate/control
  - e.g. spatial error

Question…

- "A mismatch between the spatial unit of observation and the spatial extent of the phenomena under consideration will result in spatial measurement errors and spatial autocorrelation between these errors in adjoining locations."
  - Anselin & Bera, 1998
- Why?

The spatial error model

- Examines spatial autocorrelation between the residuals of adjacent areas
- Treats spatial correlation primarily as a nuisance
  - Disregards the idea that spatial correlation may reflect some meaningful process



- Positive spatial error may reflect a misspecified model (particularly a omitted variable that is spatially clusters)
- If we ignore spatial error in the residuals:
  - Coefficients unbiased
  - Standard errors are wrong (p-values wrong)

Spatial autocorrelation in residuals Spatial error model

- Incorporates spatial effects through error term

$$y = x\beta + \varepsilon$$
$$\varepsilon = \lambda W\varepsilon + \xi$$

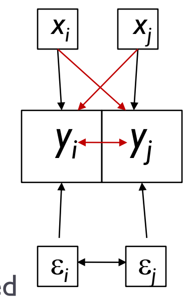
- Where:

ε is the vector of error terms, spatially weighted using the weights matrix (W)
λ is the spatial error coefficient
ξ is a vector of uncorrelated error terms

- If there is no spatial correlation between the errors, then $\lambda = 0$

The spatial lag model

- Incorporates spatial dependence by adding a "spatially lagged" DV (y) on the right-hand side of the regression equation
  - Other, more complex, models may also include spatially lagged IVs (x)
- Treats spatial correlation as a process or effect of interest
  - The values of y in one area are directly influenced by the values of y found in neighboring areas
  - Depends on how to we define neighborhood



Spatial lag model

- Positive spatial lag provides evidence that the y's in adjacent areas covary
- If we ignore the influence of spatially lagged terms:
  - Coefficients will be biased
    - If there is a positive effect of neighboring y's, usually coefficients are biased upward
  - Standard errors are wrong (p-values wrong)

Spatial autocorrelation in DV
Spatial lag model

- Incorporates spatial effects by including a spatially lagged dependent variable as an additional predictor

$$y = \rho W y + x\beta + \varepsilon$$

- Where:

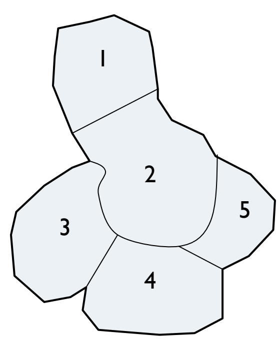
Wy is the spatially lagged DVs for weights matrix W
x is a matrix of observations on the explanatory variables
ε is a vector of error terms
ρ is the spatial coefficient

- If there is no spatial dependence, and y does no depend on neighboring y values, ρ = *0*

How do we calculate that spatial lag term?

- Y is the average of all neighbors

$$W = \begin{matrix} 0 & 1 & 0 & 0 & 0 \\ .25 & 0 & .25 & .25 & .25 \\ 0 & .5 & 0 & .5 & 0 \\ 0 & .33 & .33 & 0 & .33 \\ 0 & .5 & 0 & .5 & 0 \end{matrix}$$



Area	y	*Wy*
1	5	(1*7)=7
2	7	(.25*5)+(.25*9)+(.25*12)+(.25*11)=9.25
3	9	(.5*7)+(.5*12)=9.5
4	12	(.33*7)+(.33*9)+(.33*11)=8.91
5	11	(.5*7)+(.5*12)=9.5

Which type of SR model do we use?

- If residuals are spatial autocorrelated (Moran's I), then use the Langrange Multiplier diagnostic to determine appropriate model
  - Regression residuals (LM-Error)
    - Mis-match of process and spatial units → systematic errors, correlated across spatial units
  - Dependent variable (LM-Lag)
    - Underlying process has led to clustered distribution of variables → influence of neighboring values on unit values
  - Spatial autocorrelation in both

Spatial Regression in R
Example: Housing Prices in Boston

CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 ft^2
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (=1 if tract bounds river; 0 otherwise)
NOX	Nitrogen oxide concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000's

Spatial Regression in R

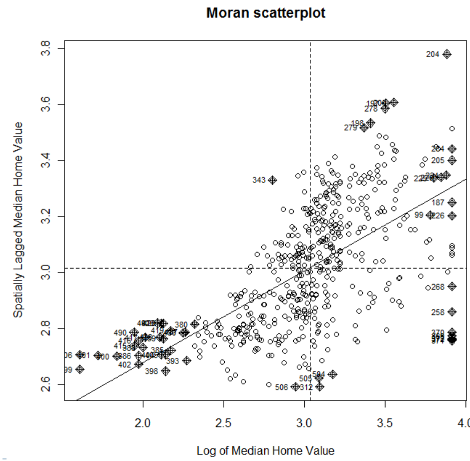
1. Read in boston.shp
2. Define neighbors (k nearest w/point data)
3. Create weights matrix
4. Moran's test of DV, Moran scatterplot
5. Run OLS regression
6. Check residuals for spatial dependence
7. Determine which SR model to use w/LM tests
8. Run spatial regression model

Moran's I on the DV

```
moran.test(boston$LOGMEDV, listw= bost_kd1_w)

 Moran's I test under randomisation

data:  boston$LOGMEDV
weights: bost_kd1_w

Moran I statistic standard deviate = 24.5658, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     0.3273430100     -0.0019801980      0.0001797138
```

Moran Plot for the DV

```
> moran.plot(boston$LOGMEDV, bost_kd1_w,
  labels=as.character(boston$ID))
```



OLS Regression

```
bostlm<-lm(LOGMEDV~RM + LSTAT + CRIM + ZN + CHAS + DIS, data=boston)

Residuals:
     Min       1Q   Median       3Q      Max
-0.71552 -0.11248 -0.02159  0.10678  0.93024

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.8718878  0.1316376  21.817  < 2e-16 ***
RM           0.1153095  0.0172813   6.672 6.70e-11 ***
LSTAT       -0.0345160  0.0019665 -17.552  < 2e-16 ***
CRIM        -0.0115726  0.0012476  -9.276  < 2e-16 ***
ZN           0.0019330  0.0005512   3.507 0.000494 ***
CHAS         0.1342672  0.0370521   3.624 0.000320 ***
DIS         -0.0302262  0.0066230  -4.564 6.33e-06 ***
---
Residual standard error: 0.2081 on 499 degrees of freedom
Multiple R-squared: 0.7433,     Adjusted R-squared: 0.7402
F-statistic: 240.8 on 6 and 499 DF,  p-value: < 2.2e-16
```

Checking residuals for spatial autocorrelation

```
> boston$lmresid<-residuals(bostlm)
> lm.morantest(bostlm,bost_kd1_w)

 Global Moran's I for regression residuals

Moran I statistic standard deviate = 5.8542, p-value =
  2.396e-09
alternative hypothesis: greater
sample estimates:
Observed Moran's I       Expectation          Variance
      0.0700808323     -0.0054856590      0.0001666168
```

Determining the type of dependence

```
> lm.LMtests(bostlm, bost_kd1_w, test="all")

    Lagrange multiplier diagnostics for spatial dependence

LMerr = 26.1243, df = 1, p-value = 3.201e-07
LMlag = 46.7233, df = 1, p-value = 8.175e-12
RLMerr = 5.0497, df = 1, p-value = 0.02463
RLMlag = 25.6486, df = 1, p-value = 4.096e-07
SARMA = 51.773, df = 2, p-value = 5.723e-12
```

- Robust tests used to find a proper alternative
- Only use robust forms when BOTH LMErr and LMLag are significant

One more diagnostic…

```
> library(lmtest)

> bptest(bostlm)

        studentized Breusch-Pagan test

data:  bostlm
BP = 70.9173, df = 6, p-value = 2.651e-13
```

- Indicates errors are heteroskedastic
  - Not surprising since we have spatial dependence

Running a spatial lag model

```
> bostlag<-lagsarlm(LOGMEDV~RM + LSTAT + CRIM + ZN + CHAS + DIS,
  data=boston, bost_kd1_w)
Type: lag
Coefficients: (asymptotic standard errors)
               Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  1.94228260  0.19267675  10.0805 < 2.2e-16
RM           0.10158292  0.01655116   6.1375 8.382e-10
LSTAT       -0.03227679  0.00192717 -16.7483 < 2.2e-16
CRIM        -0.01033127  0.00120283  -8.5891 < 2.2e-16
ZN           0.00166558  0.00052968   3.1445  0.001664
CHAS         0.07238573  0.03608725   2.0059  0.044872
DIS         -0.04285133  0.00655158  -6.5406 6.127e-11

Rho: 0.34416, LR test value:37.426, p-value:9.4936e-10
Asymptotic standard error: 0.051967
    z-value: 6.6226, p-value: 3.5291e-11
Wald statistic: 43.859, p-value: 3.5291e-11

Log likelihood: 98.51632 for lag model
ML residual variance (sigma squared): 0.03944, (sigma: 0.1986)
AIC: -179.03, (AIC for lm: -143.61)
```

A few more diagnostics

```
LM test for residual autocorrelation
test value: 1.9852, p-value: 0.15884

> bptest.sarlm(bostlag)

        studentized Breusch-Pagan test
data:
BP = 60.0237, df = 6, p-value = 4.451e-11
```

- LM test suggests there is no more spatial autocorrelation in the data
- BP test indicates remaining heteroskedasticity in the residuals
  - Most likely due to misspecification

Running a spatial error model

```
> bosterr<-errorsarlm(LOGMEDV~RM + LSTAT + CRIM + ZN + CHAS + DIS,
  data=boston, listw=bost_kd1_w)
Type: error
Coefficients: (asymptotic standard errors)
               Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  2.96330332  0.13381870  22.1442 < 2.2e-16
RM           0.09816980  0.01700824   5.7719 7.838e-09
LSTAT       -0.03413153  0.00194289 -17.5674 < 2.2e-16
CRIM        -0.01055839  0.00125282  -8.4277 < 2.2e-16
ZN           0.00200686  0.00062018   3.2359  0.001212
CHAS         0.06527760  0.03766168   1.7333  0.083049
DIS         -0.02780598  0.01064794  -2.6114  0.009017

Lambda: 0.59085, LR test value: 24.766, p-value: 6.4731e-07
Asymptotic standard error: 0.086787
    z-value: 6.8081, p-value: 9.8916e-12
Wald statistic: 46.35, p-value: 9.8918e-12

Log likelihood: 92.18617 for error model
ML residual variance (sigma squared): 0.03989, (sigma: 0.19972)
AIC: -166.37, (AIC for lm: -143.61)
```

Why we don't use R^2

- R^2 isn't a suitable measure of model fit for spatial regression
- R^2 is calculated based on the ratio between explained and unexplained (residual) variation
  - Requires the residuals are independent of one another
- The reason for using spatial regression is that we found spatial autocorrelation in the residuals
  - e.g., the explained and unexplained variations are not independent in this scenario