A Brief Introduction to Spatial Regression

Three Important Ways That Spatial Analysis Can Help The Social Scientist

(i) Data integration:
Spatial analysis provides a basis for integration and data collection at different spatial scales and time dimensions. Data integration is a central function of the application of GIS.

(ii) Exploratory spatial data analysis (ESDA):
ESDA is a collection of techniques to describe and visualize spatial distributions, identify atypical locations or spatial outliers, discover patterns of spatial association, clusters or hot spots, and suggest spatial regimes or other forms of spatial heterogeneity (Anselin, 1994, 1999b).

Three Important Ways That Spatial Analysis Can Help The Social Scientist

(iii) Confirmatory spatial data analysis:
Spatial modeling techniques, such as regression analysis can also be implemented to explicitly incorporate the mechanisms underlying the spatial patterns.

Some (but not all) regression assumptions

1. The dependent variable should be normally distributed (i.e., the histogram of the variable should look like a bell curve)
   - Ideally, this will also be true of independent variables, but this is not essential. Independent variables can also be binary (i.e., have two values, such as 1 (yes) and 0 (no))
2. The predictors should not be strongly correlated with each other (i.e., no multicollinearity)
3. Very importantly, the observations should be independent of each other. (The same holds for regression residuals). If this assumption is violated, our coefficient estimates could be wrong!

Cluster analysis

Moran's I: Compares the value of the variable at any one location with the value at all other locations.

$$I = \frac{\sum_i \sum_j w_{i,j}(x_i - \bar{X})(x_j - \bar{X})}{\sum_i (x_i - \bar{X})^2}$$

A focus on *where* the non-randomness may be located, in terms of significant clusters or spatial outliers is provided by an analysis of the local indicators of spatial association (LISA).

$$I_i = \frac{(x_i - \bar{X})\sum_j w_{ij}(x_j - \bar{X})}{\sum_i (x_i - \bar{X})^2}$$

Gobal indicators

- '*Is there spatial autocorrelation*'?
- Global indicators of spatial association provide the answer
- E.g. Moran's I

$$I = \frac{Cov\left(x_i, m\left(x_i\right)\right)}{Var\left(x_i\right)}$$

$$I = \frac{\sum_i \sum_j w_{ij} \tilde{x}_i \tilde{x}_j}{\sum_i \tilde{x}_i^2}$$

- Where "~" to indicate deviations from mean

LISA

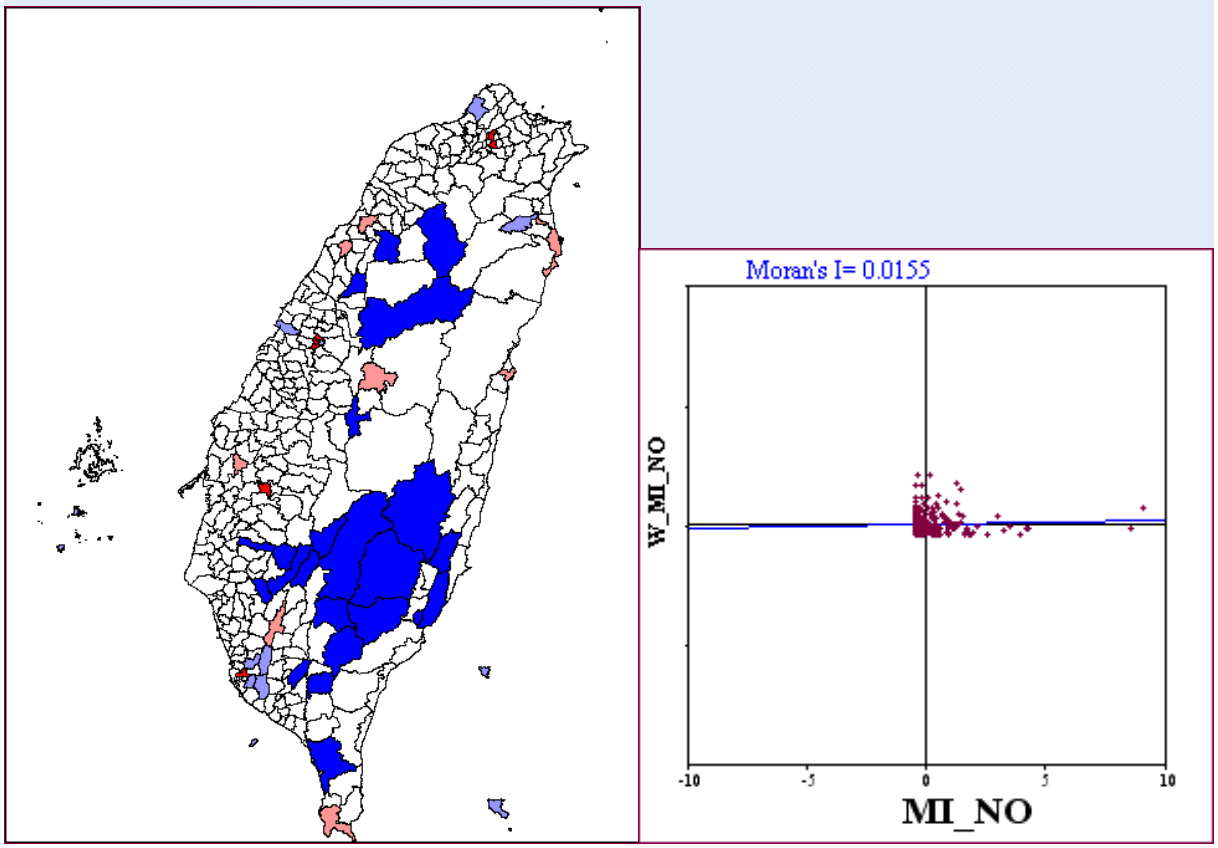
- '*Where is the spatial autocorrelation'*?
- **Local indicators of spatial association** (LISA) provide answer
- Anselin (1995) definition: LISA
  - Indicates spatial clustering of similar values around the observation
  - Sum of LISAs proportional to a Global indicator

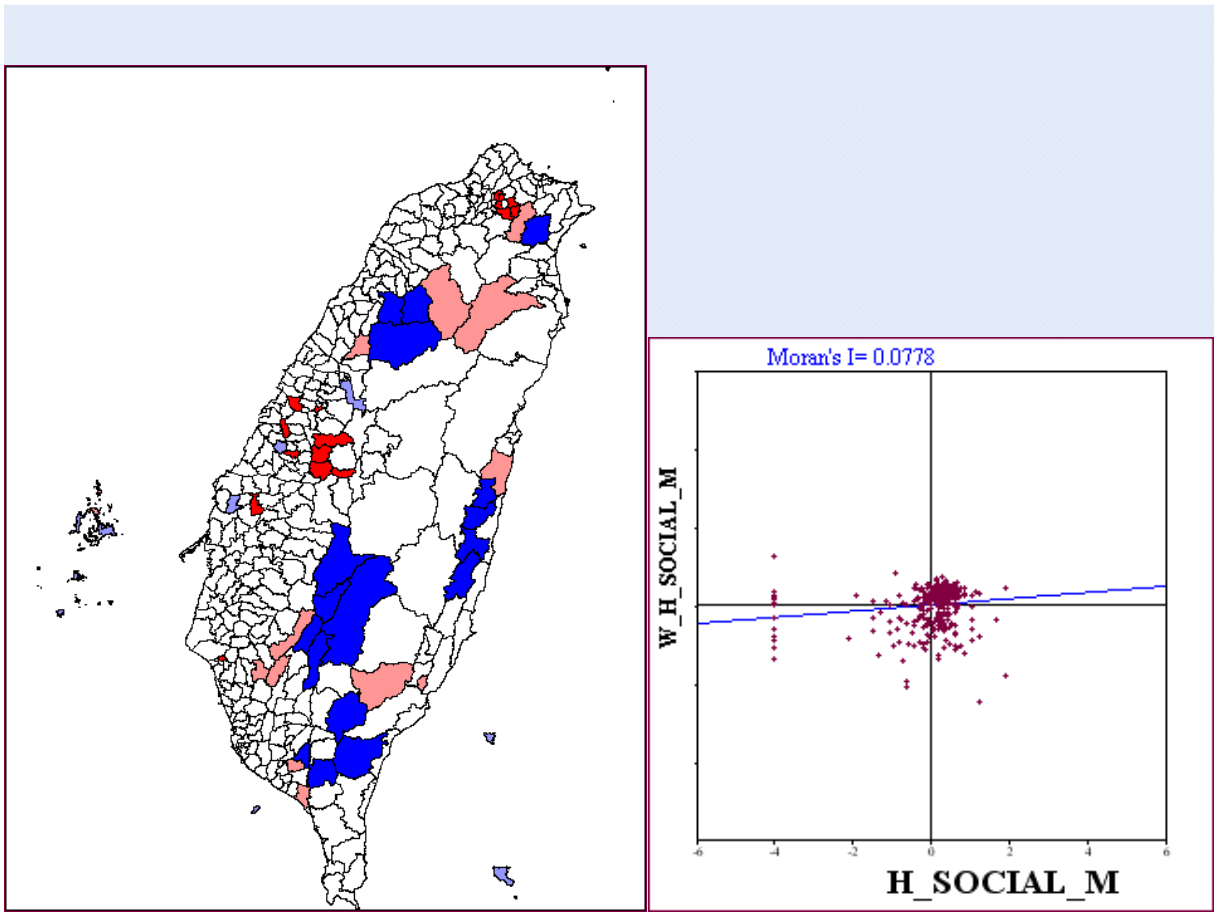
Local Moran I

- Local Indicator (Local Moran I)
  - Product of (centred) x and 'neighbouring' x at place i
  - Divided by the variance of x

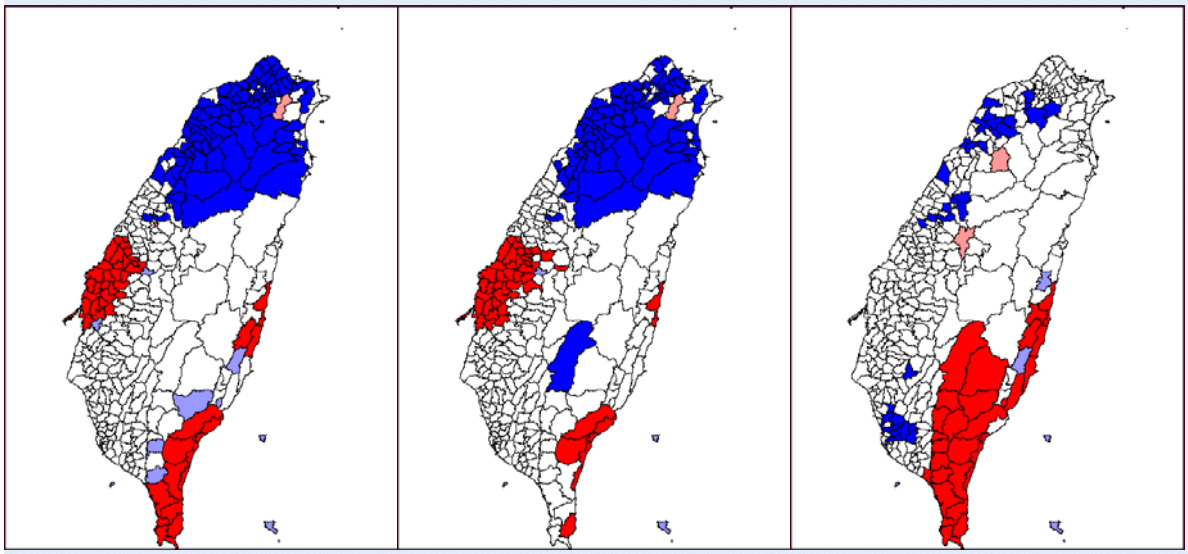
$$I_i = n \frac{x_i \sum_j w_{ij} x_j}{\sum_i x_i^2}$$

- Note: **mean of Local = Global**


Moran's I= 0.0155
W_MI_NO
MI_NO
-10
-5
0
5
10

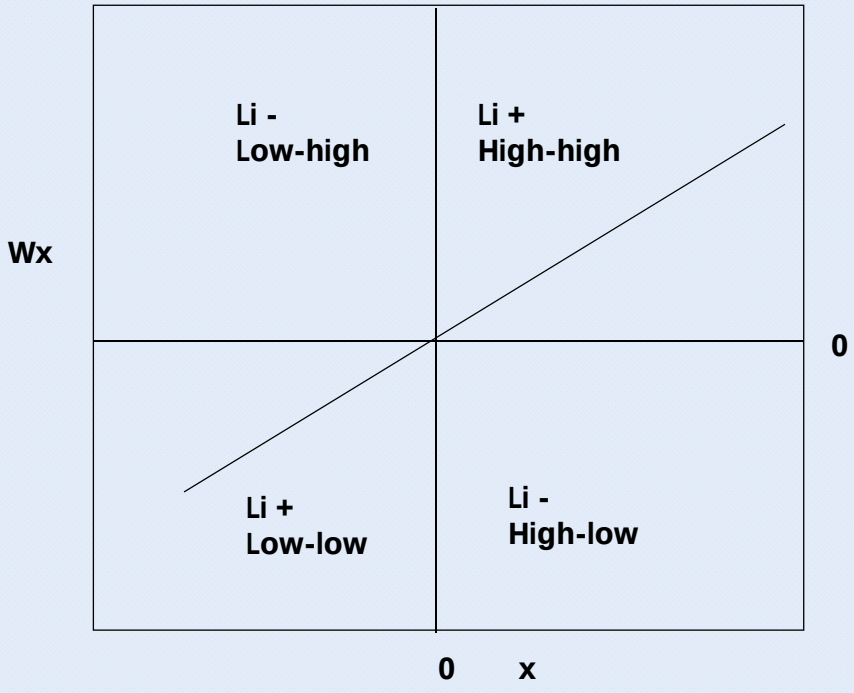

Moran's I= 0.0778
W_H_SOCIAL_M
H_SOCIAL_M
-6
-4
-2
0
2
4
6

Oral Cancer Death Rate

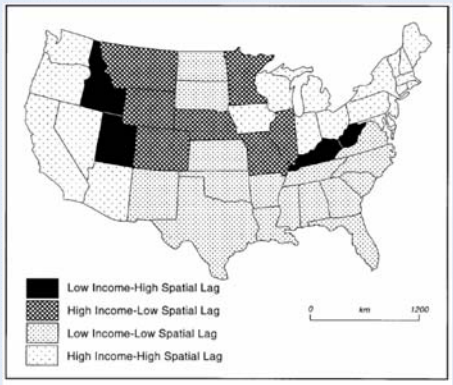


Total Male Female

Moran scatter-plot: components of spatial autocorrelation



US Income Convergence Example



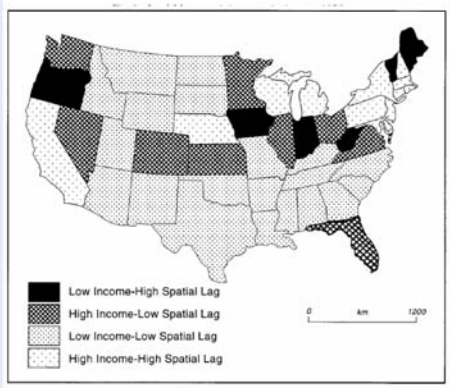
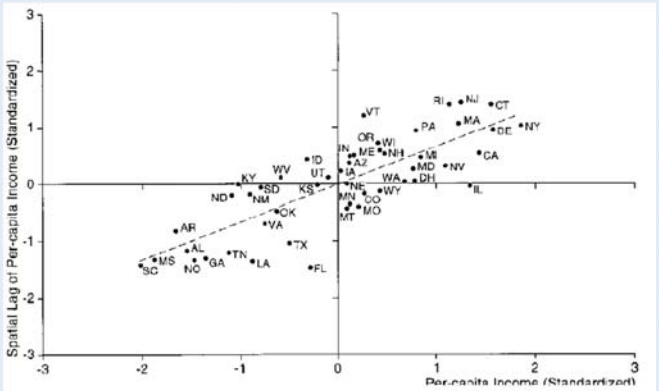
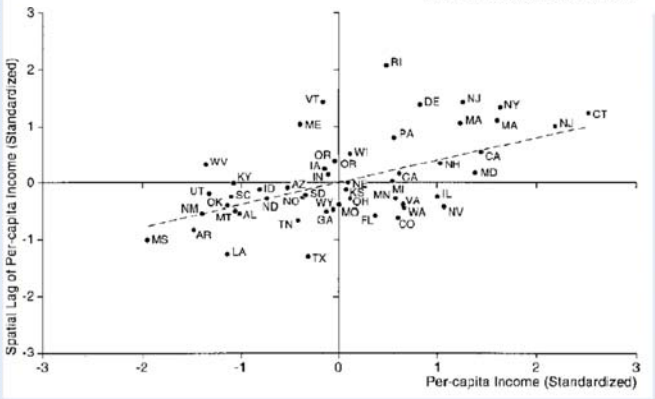


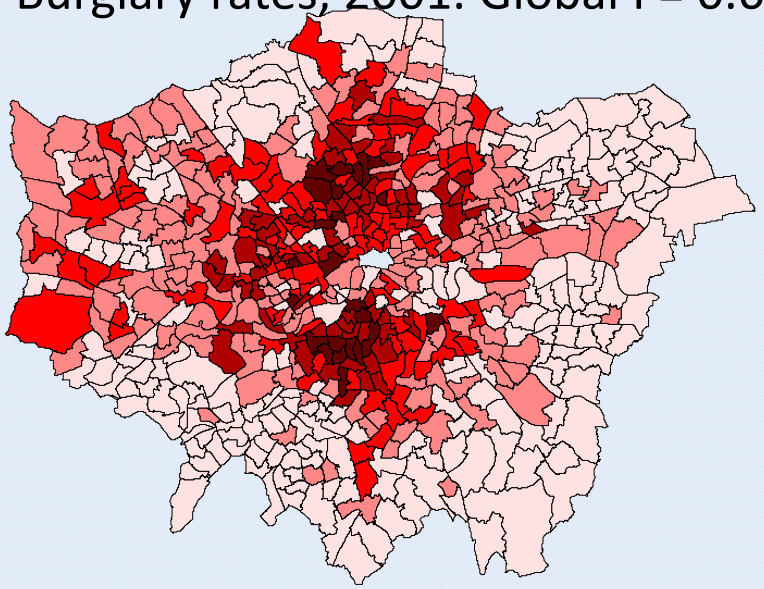
Fig 5. Local Moran statistics per capita income, 1994

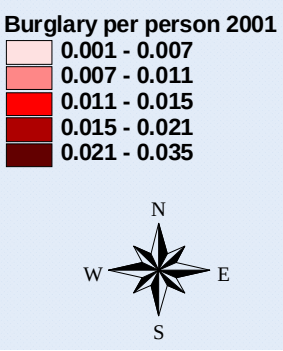




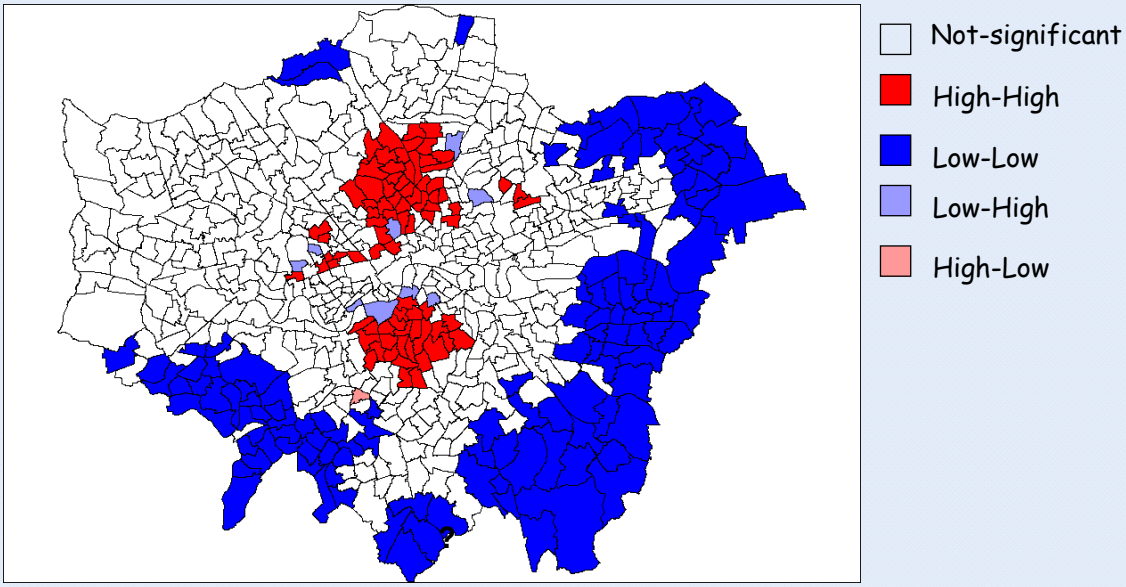
Example: London crime data

- Burglary rates, 2001. Global I = 0.624

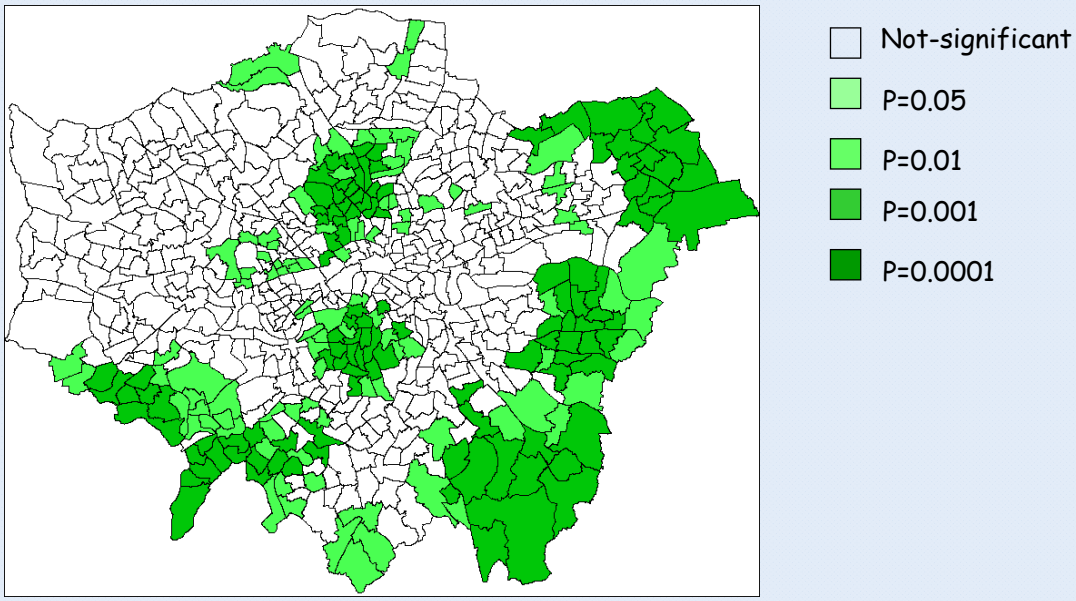




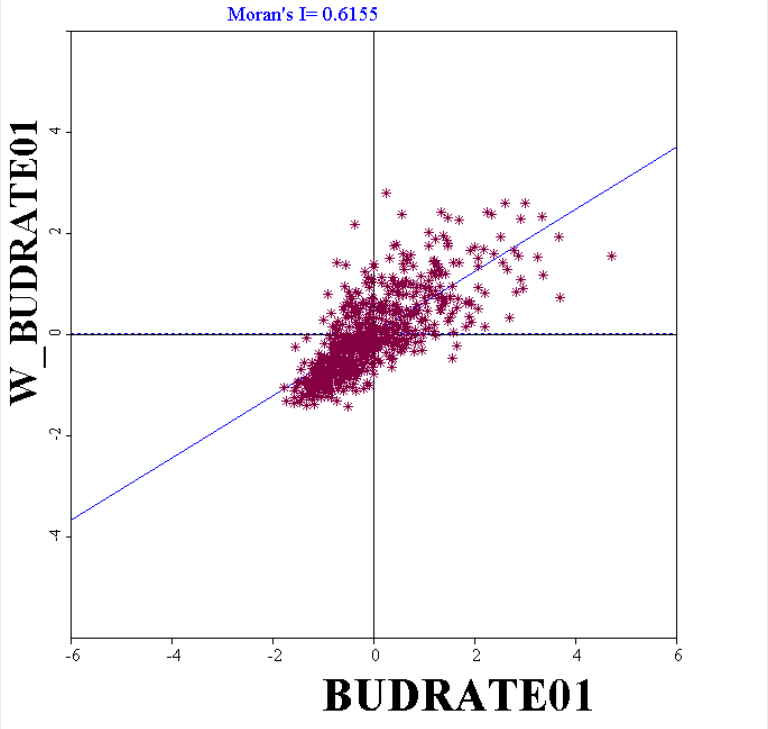
Local Moran I Map



Local Moran Significance Map

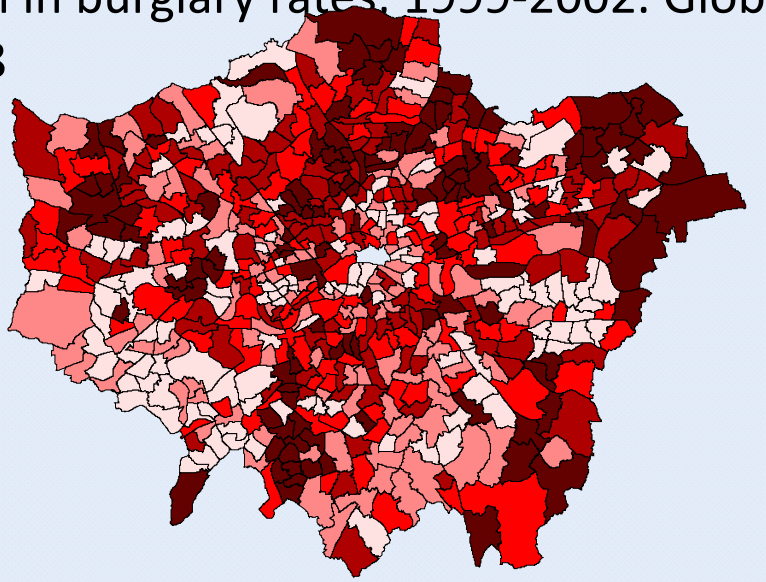


Moran Scatter Plot



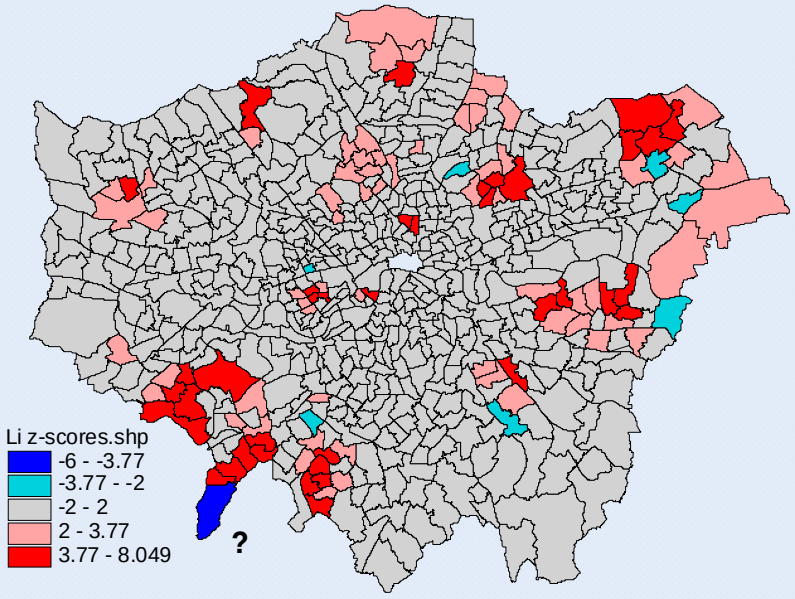
Example: Growth in London crime

- Growth in burglary rates, 1999-2002. Global I = 0.328



Local Moran I z-scores

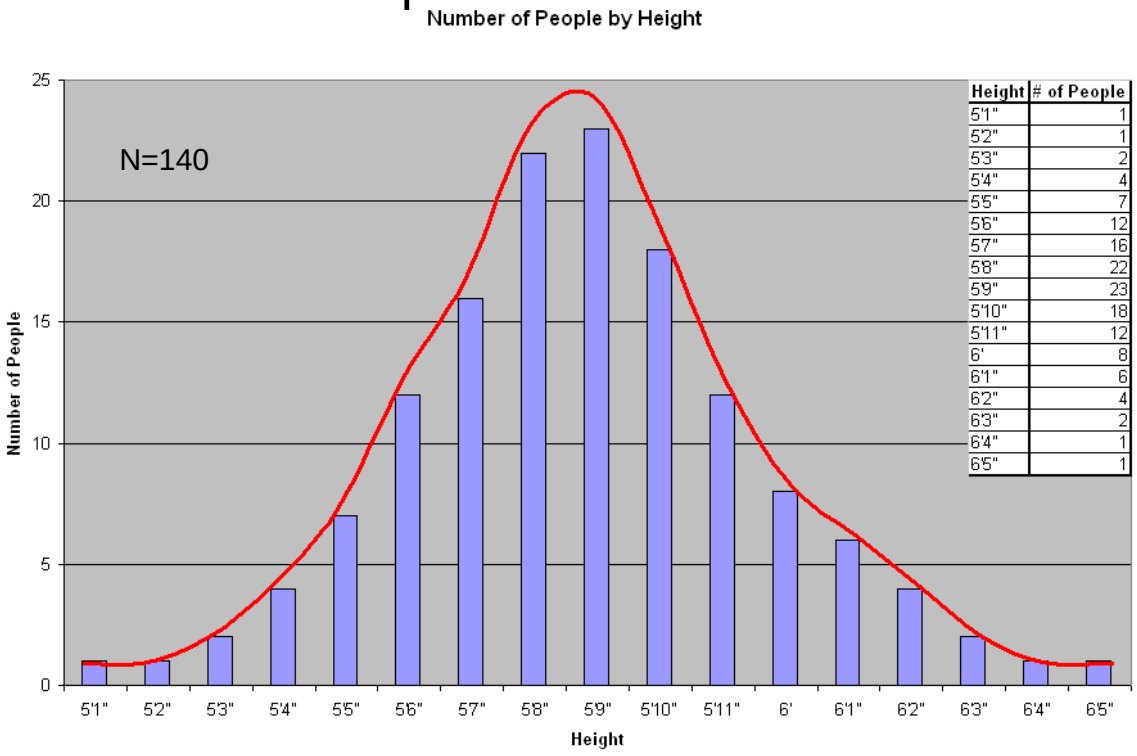
- Z(0.05) = 1.96, no correction
- Z(0.05) = 3.77, bonferroni correction (634 wards)



Conclusions on LISA

- Local Moran's I (and other LISA) useful for showing places where significant spatial autocorrelation exists
- Purely descriptive
- Though potential to combine with regression analysis for further analysis
  - Residuals?
  - Dependent variable?

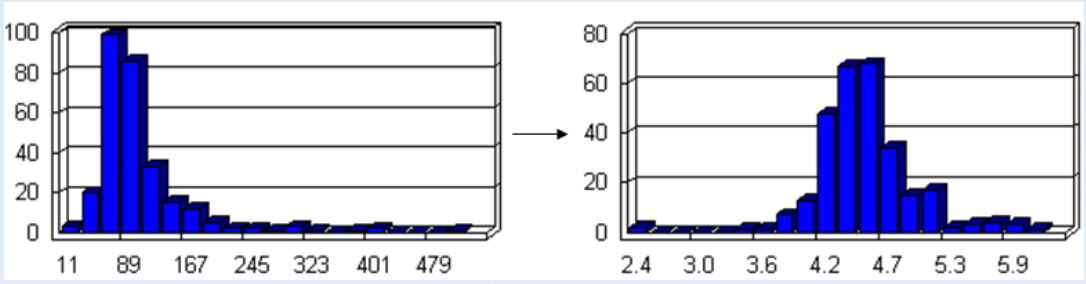
An Example of a Normal Distribution



Height	# of People
5'1"	1
5'2"	1
5'3"	2
5'4"	4
5'5"	7
5'6"	12
5'7"	16
5'8"	22
5'9"	23
5'10"	18
5'11"	12
6'	8
6'1"	6
6'2"	4
6'3"	2
6'4"	1
6'5"	1

Data Transformations

- Sometimes, it is possible to *transform* a variable's distribution by subjecting it to some simple algebraic operation.
  - The logarithmic transformation is the most widely used to achieve normality when the variable is *positively skewed* (as in the image on the left below)
  - Analysis is then performed on the *transformed* variable.



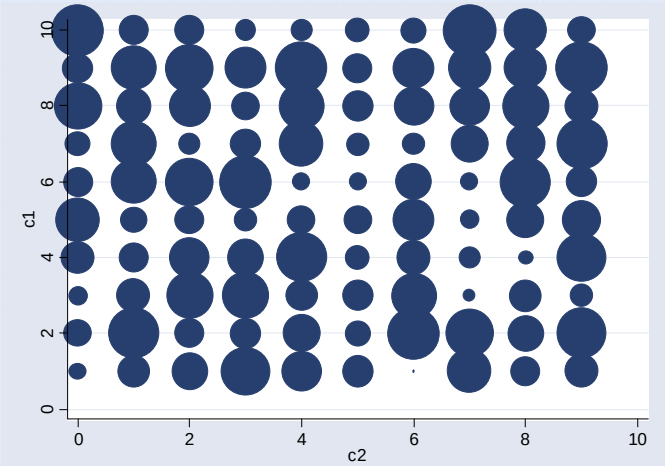
Spatial Autocorrelation

- There is *spatial autocorrelation* in a variable if observations that are closer to each other in space have related values (Tobler's Law)
- One of the regression assumptions is independence of observations. If this doesn't hold, we obtain inaccurate estimates of the β coefficients, and the error term ε contains spatial dependencies (i.e., meaningful information), whereas we want the error to not be distinguishable from random noise.

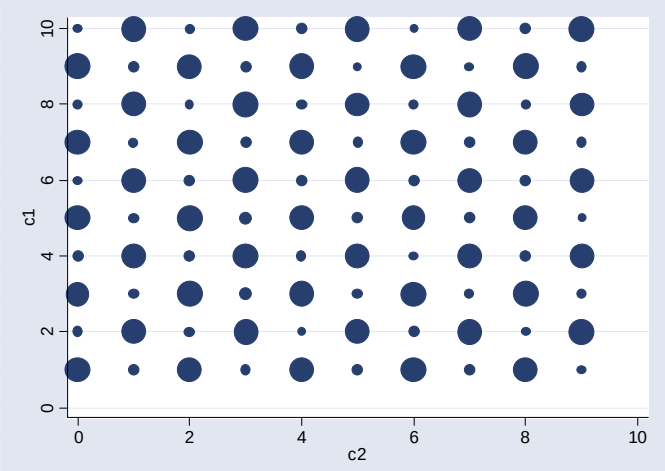
Spatial autocorrelation

- Assume places (regions, districts, firms people etc) are fixed
- Variable (x) recorded at places s
- Is the data x random across space or are there similarities between neighbours?
- Does a high value of x tend to be associated with a high value of x in neighbouring places (and low values with low)?

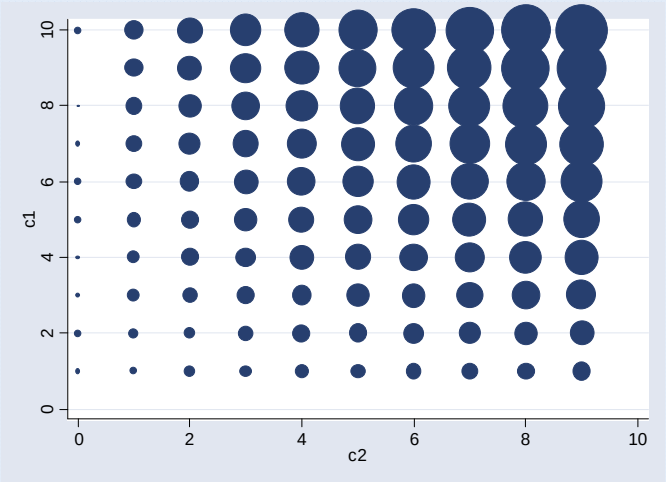
Random - no spatial autocorrelation



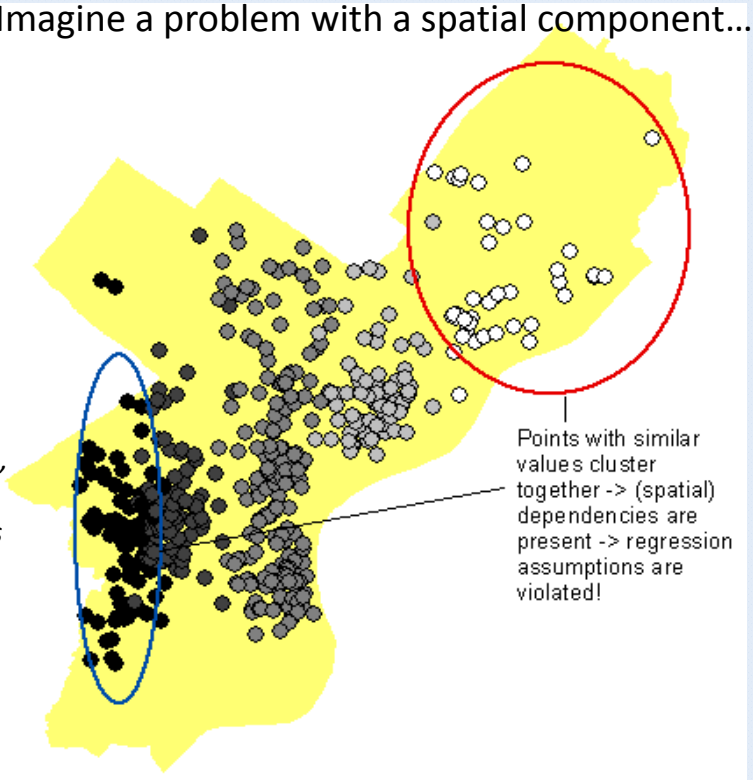
Overly dispersed - negatively autocorrelated



Positive spatial autocorrelation

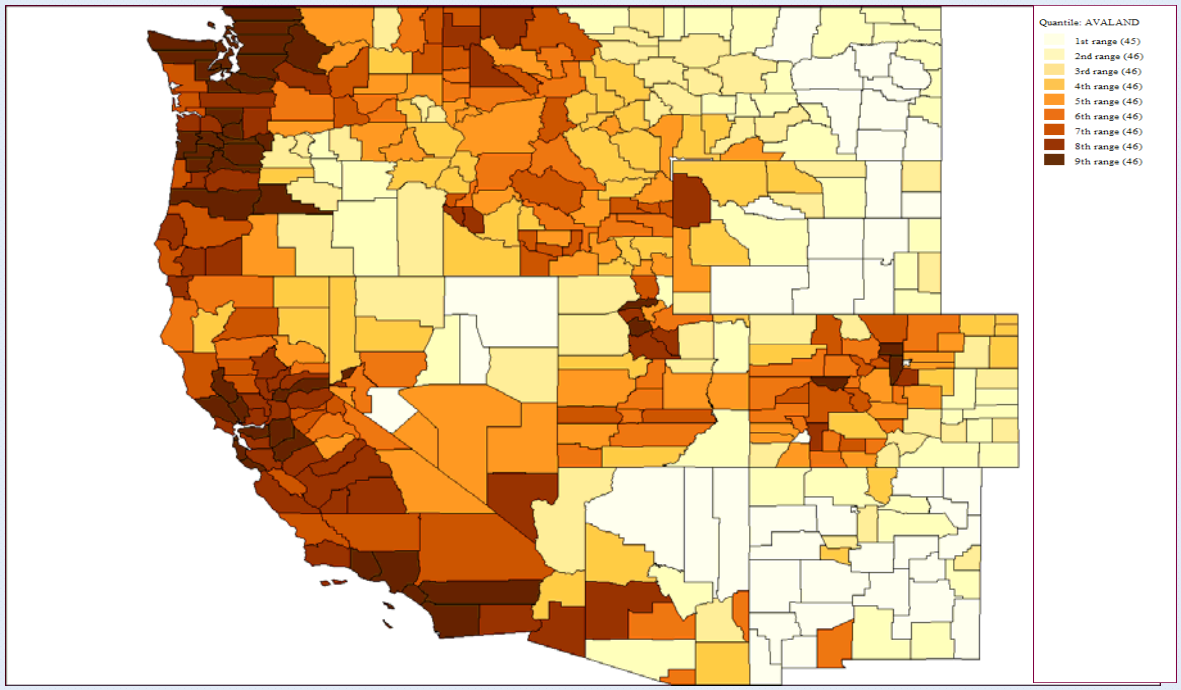


Imagine a problem with a spatial component…



This example is obviously a dramatization, but nonetheless, in many spatial problems points which are close together have similar values

Average Value of Ag Land and Buildings



But how do we know if spatial dependencies exist?

- Moran's I (1950) – a rather old and perhaps the most widely used method of testing for spatial autocorrelation, or spatial dependencies
  - We can determine a p-value for Moran's I (i.e., an indicator of whether spatial autocorrelation is statistically significant).
    - For more on Moran's I, see http://en.wikipedia.org/wiki/Moran%27s_I
  - Just as the non-spatial correlation coefficient, ranges from -1 to 1
  - Can be calculated in ArcGIS
- Other indices of spatial autocorrelation commonly used include:
  - Geary's *c* (1954)
  - Getis and Ord's *G*-statistic (1992)
    - For non-negative values only

So, when a problem has a spatial component, we should:

- Run the ***non-spatial regression***
- Test the *regression residuals* for spatial autocorrelation, using Moran's I or some other index
- If no significant spatial autocorrelation exists, STOP. Otherwise, if the spatial dependencies are significant, use a special model which takes **spatial dependencies** into account.

Types of Models: Spatial Error

- Model
  - Start with basic model
    - $y= \beta x+e \quad e \sim N(0,\sigma^2)$
  - $y= \beta x+e+\lambda we$
    - If $\lambda=0$, reduces to OLS, if $\lambda \neq 0$, OLS is unbiased and consistent, but SE will be wrong and the betas will be *inefficient*

Empirical Model

- Negative binomial (NB) distribution

$$\Pr(Q_{ODi} \mid X_i, \alpha, \lambda_i) = \frac{\Gamma(Q_{ODi} + 1/\alpha)}{Q_{ODi}!\Gamma(1/\alpha)} \left(\frac{1/\alpha}{1/\alpha + \lambda_i} \right)^{1/\alpha} \left(\frac{\lambda_i}{1/\alpha + \lambda_i} \right)^{Q_{ODi}}$$

  - λ is the expected interval-usage count for a given interval (between interchanges) and α is the overdispersion parameter

- Log-likelihood function (L) of the NB regression model

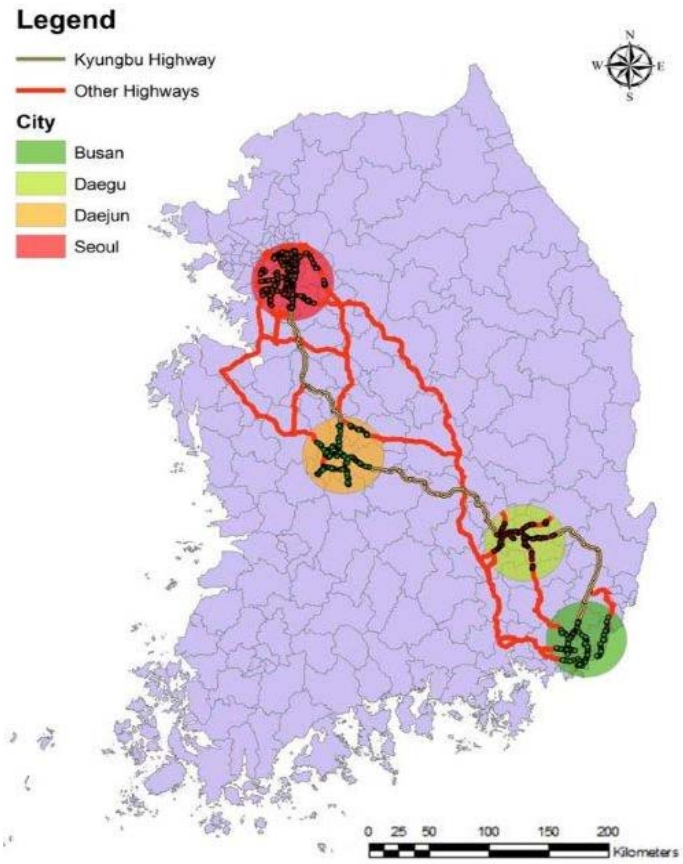
$$L = \sum_{i=1}^{N} \{ \sum_{j=0}^{Q_{ODi}-1} \ln(j + 1/\alpha) - \ln(Q_{ODi}!)$$

$$- (Q_{ODi} + 1/\alpha) \ln(1 + \alpha \exp(X_i \beta)) + Q_{ODi} \ln(\alpha) + Q_{ODi} X_i \beta \}$$

- Spatial heteroskedastic autocorrelation consistent (HAC) estimator

$$u_{ij} = \mathbf{r}_j \varepsilon, \quad \varepsilon \sim iid(0, \sigma^2)$$

  - r_j is the j^{th} row of matrix R.
  - A consistent non-parametric estimator of the asymptotic distribution of the nonstochastic location determinant by Kelejian and Prucha (2007)



(Examples of the highways that are sampled for the case study)

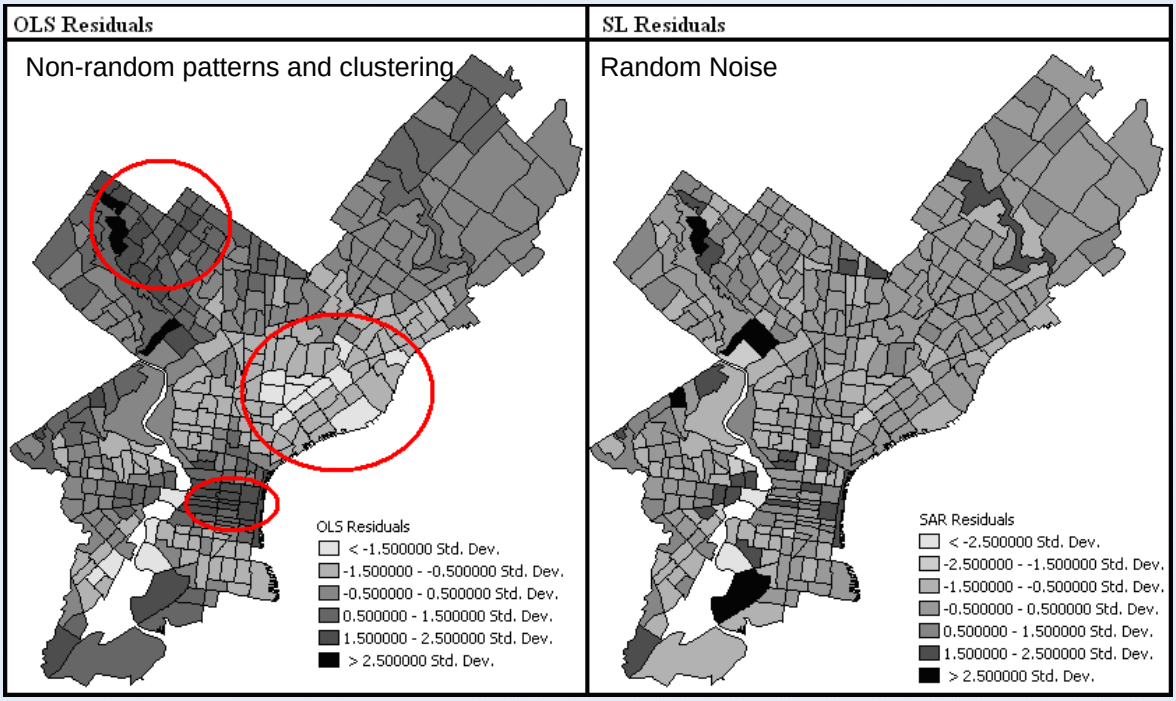
Types of Models: Spatial Lag

- Spatial lag model
  - Dependent variable is affected by the values of the dependent variables in nearby places
    - Land value in a county is a function of land value in nearby counties, not just related to common unmeasured variables

Types of Models: Spatial Lag

- Model
  - $Y = \beta x_i + \phi w_i y + e_i$
    - Can also include $w_i x_i$ term
  - OLS in this case is *biased* and *inconsistent*

OLS Residuals vs. SAR Residuals



Empirical model

General spatial hedonic model

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\boldsymbol{\varepsilon} = (\mathbf{I} - \lambda\mathbf{W})^{-1}\boldsymbol{\mu}$$

- **y** is an n × 1 vector, dependent variable (natural log of the sale price of a single-family house)
- **Wy** is an n × 1 vector, spatial lag of the dependent
- **W** is a spatial weight matrix, neighborhood structure
- ρ, λ are the parameters of the spatially lagged dependent variable and spatial autoregressive structure of the disturbance **ε**
- **X** is an n × (k + 1) matrix, explanatory variables including measures of ambient water quality
- **β** is a vector of parameters,
- **μ** normally distributed error term

Individual parcel data

- Detached single-family houses sold between 2001 and 2004
- Total of 2,135 sales occurred during the 2001–2004 period
  - 1,394 sales in NC and 741 sales in TN
  - Price adjusted to 2001 dollars using the annual housing price index for each state
  - After eliminating missing data, 595 sales from NC and 497 sales from TN used