Population Research Institute

GIS Population Science UCSB

Center for Spatially Integrated Social Science

Penn State Population Research Institute

Institutional support was provided by The Pennsylvania State University's NICHD Population Research Center Grant 1 R24 HD41025.

GeoDa and Spatial Regression Modeling

June 9, 2006

Stephen A. Matthews
Associate Professor of Sociology & Anthropology, Geography and Demography
Director of the Geographic Information Analysis Core
Population Research Institute

Outline

1. OLS Regression in GeoDa
2. Spatial Regression in GeoDa
3. Examples

This presentation draws on examples and text from both the GeoDa Workbook (0.95i) and the SpaceStat Manual 1.90 (both written by Luc Anselin)

OLS Regression in GeoDa

OLS

The general purpose of linear regression is to find a (linear) relationship between a dependent variable and a set of explanatory variables.

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

OLS

There are usually two objectives:

1. Find a good match (or fit) between predicted values $X\beta$ (sum of the values of explanatory variables, each multiplied by their regression coefficients) and observed values of the explanatory variable y

2. Discover which of the explanatory variables contribute significantly to the linear relationship

OLS

OLS accomplished both stated objectives in an optimal fashion according to some criteria, and is referred to as a Best Linear Unbiased Estimator (BLUE).

OLS estimates for β are found by minimizing the sum of the squared prediction errors (hence least squares).

OLS

In order to obtain the BLUE property and to be able to make statistical inferences about the population regression estimates β by means of your estimates **b**, you need to make certain assumptions about the random part of the regression equation (the random error ε).

Two of these assumptions are crucial to obtain the **unbiasedness** and **efficiency** of the OLS estimates (the U and the E part of BLUE).

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

Assumptions

$$E(\varepsilon) = \varnothing$$

The random error has mean zero (there is no systematic misspecification or bias in the regression equation).

$$E(\varepsilon\varepsilon') = \sigma^2 I$$

The random error terms are uncorrelated and have a constant variance (they are homoskedastic).

OLS - Diagnostics

The assumption of normal, homoskedastic and uncorrelated error terms that lead to the BLUE characteristic of OLS estimators are not necessarily satisfied by the real models and data. Thus, an important part of good practice consists of checking the extent to which these assumptions are violated.

When dealing with spatial data, you must give special attention to the possibility that the errors or the variables in the model show spatial dependence.

Why is spatial autocorrelation important?

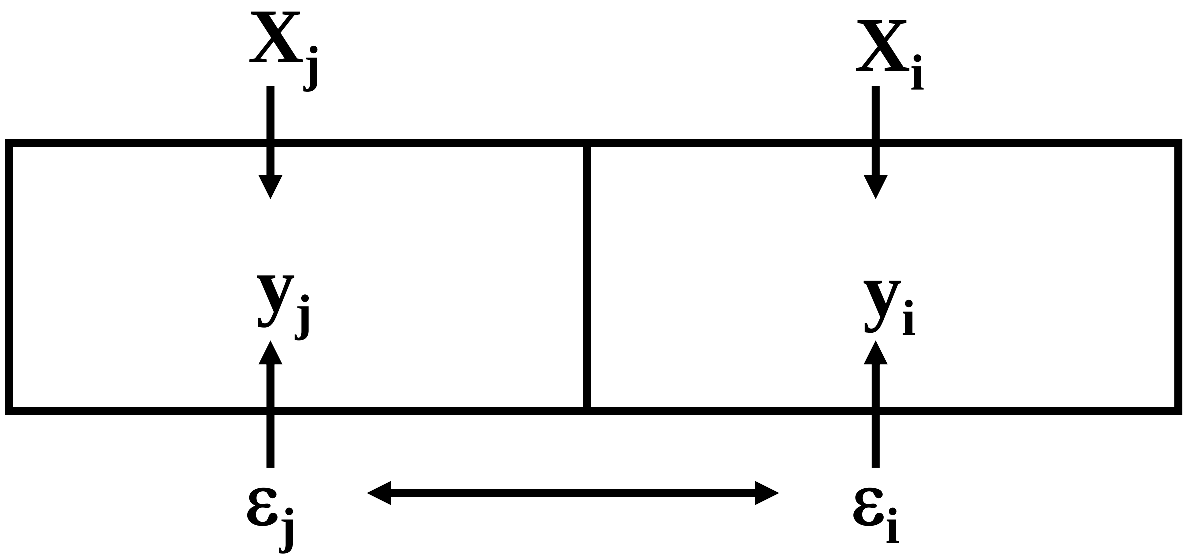
We need to examine the influences of spatial autocorrelation upon the inferences that may be drawn from statistical tests.

As these inferences are based on independence assumptions, then the presence of spatial autocorrelation is likely to bias any resultant inferences.

Spatial Error Effects

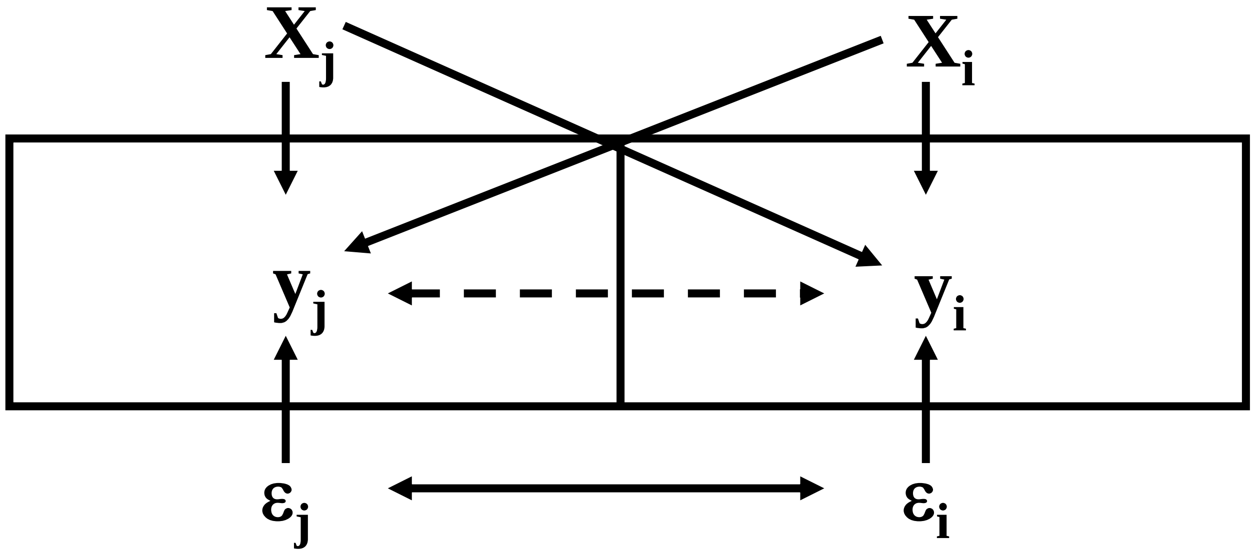
Dependence amongst the errors

OLS estimates become ***inefficient***



Spatial Lag Effects

OLS estimates are ***biased***, and thus inferences based on an OLS model will be incorrect



Spatial Dependence – as a Nuisance

The presence of spatial dependence in cross-sectional georeferenced data has two important consequences.

1. If the interest focuses on obtaining proper statistical inference (estimation, hypothesis tests, predictors) from the dependent data, spatial autocorrelation can be considered a **nuisance**. In such an instance, the main objective is to correct standard statistical procedures for the effect of the spatial dependence, e.g., **by adjustments that incorporate the spatial autocorrelation in a regression error term.**

Spatial Dependence – Substantive

2. When one is intent on discovering the form of the spatial interaction, the precise nature of spatial spillover and the economic and social processes that lie behind it, the spatial dependence can be considered to be **substantive**.

In this case, the focus is on how to incorporate the structure of spatial dependence in to a statistical model and how to estimate and interpret it.

Spatial Dependence

1) **nuisance** involves model residuals only – if this exists it reduces model efficiency and can be corrected by including a **spatial error** specification in the model.

2) **substantive** autocorrelation is where values of Y are systematically related to values of Y in adjacent areas, generating model bias. This can be corrected by including an explicit **spatial lag** term as an explanatory variable in the model.

So why weren't we told about this?

This and the next two slides are taken from a talk by Paul Voss (Wisconsin) presented at a CSISS/PSU GIS workshop in 2003.

Loftin, Colin and Sally K. Ward. 1983. "A Spatial Autocorrelation Model of the Effects of Population Density on Fertility." *American Sociological Review* 48:121-128.

"…[T]he GGM [Galle, Gove, and McPherson, 1972] findings with regard to fertility ***are an artifact of the failure to recognize the presence of disturbance variables which are spatially autocorrelated***…. Our research illustrates the importance of spatial mechanisms in modeling spatial processes. ***The GGM analysis is only one of many examples of studies which use geographically defined areas without due consideration to interactions between units***." (p. 127)

Doreian, Patrick. 1980. "Linear Models with Spatially Distributed Data: Spatial Disturbances or Spatial Effects." *Sociological Methods & Research* 9(1):29-60.

"It is clear that for linear models employing spatially distributed data, **attention must be paid to the spatial characteristics of the phenomena being studied**." (p. 53)

"**The nonspatial model estimated by conventional regression procedures is not a reliable representation and should be avoided when there is a spatial phenomenon to be analyzed.**" (p. 51)

"[T]hese methodological problems ***are not hypothetical ones***." (p. 30, emphasis added)

OLS GeoDa – Diagnostics to detect spatial dependence (and other standard diagnostics)

Multicollinearity

Non-Normal Errors

Heteroskedasticity

Spatial Autocorrelation (Spatial Dependence)

```
REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   6.541828
TEST ON NORMALITY OF ERRORS
TEST                    DF          VALUE            PROB
Jarque-Bera              2           1.835753        0.3993663

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                    DF          VALUE            PROB
Breusch-Pagan test       2           7.900442        0.0192505
Koenker-Bassett test     2           5.694088        0.0580156
SPECIFICATION ROBUST TEST
TEST                    DF          VALUE            PROB
White                    5           19.94601        0.0012792

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : colrook.GAL  (row-standardized weights)
TEST                           MI/DF       VALUE           PROB
Moran's I (error)            0.249862     2.9376173       0.0033076
Lagrange Multiplier (lag)        1        8.7599071       0.0030792
Robust LM (lag)                  1        3.0721737       0.0796429
Lagrange Multiplier (error)      1        5.8148799       0.0158911
Robust LM (error)                1        0.1271465       0.7214092
Lagrange Multiplier (SARMA)      2        8.8870536       0.0117544
========================= END OF REPORT ==============================
```

Multicollinearity

High correlation between independent/explanatory variables (estimates will have very large estimated variances and few coefficients will be found to be significant, even though the regression may be a good fit – High $\mathbf{R}^2$ with low **t** statistics is a good indicator that something is wrong in terms of multicollinearity).

GeoDa's diagnostic that may point to a potential problem is called the "condition number. As a rule of thumb, values of the **condition number > 30 are considered suspect**. A total lack of multicollinearity yields a condition number of 1.

```
REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER    6.541828
```

Non-Normal Errors

Most hypothesis tests and a large number of regression diagnostics assume normal error distributions. It is hard to assess the extent to which this may be violated, since the errors cannot be observed. Instead, tests of non-normal errors must be computed from the regression residuals.

GeoDa reports the Jarque-Bera test. **A low probability indicates a rejection of the null hypotheses of normal error**. If this is the case, the tests for heteroskedasticity and spatial dependence should be interpreted with caution, since they are based on the normal assumption.

TEST ON NORMALITY OF ERRORS

TEST	DF	VALUE	PROB
Jarque-Bera	2	1.835753	0.3993663

Heteroskedasticity

This is the situation where the random regression error does not have a constant variance over all observations (i.e., not homoskedastic).

As a consequence, the indication of precision given by assuming a constant error variance in OLS will be misleading. While the OLS estimates are still unbiased, they will no longer be the most efficient. More importantly, inference based on the usual **t** and **F** statistics will be misleading, and the $\mathbf{R^2}$ measure of the goodness-of-fit will be wrong.

Heteroskedasticity

In spatial data analysis, you will frequently encounter this problem, especially when using data for irregular spatial units (different area), when there are systematic regional differences in the relationships you model (i.e., spatial regimes), or when there is a continuous spatial drift in the parameters in the model (i.e., spatial expansion).

The presence of any of these spatial effects would make a standard regression model that ignores them misspecified. Hence, an indication of heteroskedasticity may point to the need for a more explicit incorporation of spatial effects.

Heteroskedasticity

There are many test for heteroskedasticiy, GeoDa includes a few.

DIAGNOSTICS FOR HETEROSKEDASTICITY

RANDOM COEFFICIENTS

TEST	DF	VALUE	PROB
Breusch-Pagan test	2	7.900442	0.0192505
Koenker-Bassett test	2	5.694088	0.0580156

SPECIFICATION ROBUST TEST

TEST	DF	VALUE	PROB
White	5	19.94601	0.0012792

Both the **BP** and the **KB** test require that you specify the variables to be used in the heteroskedastic specification. When there is little prior information about the form of heteroskedasticity the **White** test is more appropriate, since it has power against any unspecified form of heteroskedasticity.

Heteroskedasticity

DIAGNOSTICS FOR HETEROSKEDASTICITY

RANDOM COEFFICIENTS

TEST	DF	VALUE	PROB
Breusch-Pagan test	2	7.900442	0.0192505
Koenker-Bassett test	2	5.694088	0.0580156

SPECIFICATION ROBUST TEST

TEST	DF	VALUE	PROB
White	5	19.94601	0.0012792

One issue to keep in mind in situations where both heteroskedasticity and spatial dependence may be present is that the tests against heteroskedasticity have been shown to be very sensitive to the presence of spatial dependence.

In other words, while tests may indicate heteroskedasticity, this may not be the problem, but instead spatial dependence may be present (the reverse holds too!).

Spatial Autocorrelation/Dependence

Spatial autocorrelation, or more generally, spatial dependence is the situation where the dependent variable (or the error term) at each location is correlated with observations on the dependent variable (or values for the error term) at other locations.

GeoDa includes many tests

DIAGNOSTICS FOR SPATIAL DEPENDENCE

FOR WEIGHT MATRIX : **colrook.GAL** (row-standardized weights)

TEST	MI/DF	VALUE	PROB
Moran's I (error)	0.249862	2.9376173	0.0033076
Lagrange Multiplier (lag)	1	8.7599071	0.0030792
Robust LM (lag)	1	3.0721737	0.0796429
Lagrange Multiplier (error)	1	5.8148799	0.0158911
Robust LM (error)	1	0.1271465	0.7214092
Lagrange Multiplier (SARMA)	2	8.8870536	0.0117544

========================== END OF REPORT ===============================

Spatial Autocorrelation/Dependence

All these tests …

DIAGNOSTICS FOR SPATIAL DEPENDENCE

FOR WEIGHT MATRIX : **colrook.GAL** (row-standardized weights)

TEST	MI/DF	VALUE	PROB
Moran's I (error)	0.249862	2.9376173	0.0033076
Lagrange Multiplier (lag)	1	8.7599071	0.0030792
Robust LM (lag)	1	3.0721737	0.0796429
Lagrange Multiplier (error)	1	5.8148799	0.0158911
Robust LM (error)	1	0.1271465	0.7214092
Lagrange Multiplier (SARMA)	2	8.8870536	0.0117544

========================= END OF REPORT ==============================

Are based on large sample properties (asymptotics) and their performance in small data sets may be suspect.

Spatial Autocorrelation/Dependence

Moran's I

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : colrook.GAL (row-standardized weights)

TEST	MI/DF	VALUE	PROB
Moran's I (error)	0.249862	2.9376173	0.0033076
Lagrange Multiplier (lag)	1	8.7599071	0.0030792
Robust LM (lag)	1	3.0721737	0.0796429
Lagrange Multiplier (error)	1	5.8148799	0.0158911
Robust LM (error)	1	0.1271465	0.7214092
Lagrange Multiplier (SARMA)	2	8.8870536	0.0117544

========================= END OF REPORT ==============================

This is an extension of Moran's I to measure spatial autocorrelation in regression models. Even though this is the most familiar test it is the most unreliable as it can "pick up" a range of misspecification errors, such as non-normality and heteroscedasticity, as well as spatial lag dependence. Moreover, it does not provide any guidance in terms of which of the substantive (lag of Y) or nuisance (error dependence) is the most likely better alternative model specification.

Spatial Autocorrelation/Dependence

Lagrange Multiplier

Robust LM (lag & error)

DIAGNOSTICS FOR SPATIAL DEPENDENCE

FOR WEIGHT MATRIX : **colrook.GAL** (row-standardized weights)

TEST	**MI/DF**	**VALUE**	**PROB**
Moran's I (error)	0.249862	2.9376173	0.0033076
Lagrange Multiplier (lag)	1	8.7599071	0.0030792
Robust LM (lag)	1	3.0721737	0.0796429
Lagrange Multiplier (error)	1	5.8148799	0.0158911
Robust LM (error)	1	0.1271465	0.7214092
Lagrange Multiplier (SARMA)	2	8.8870536	0.0117544

========================= END OF REPORT ==============================

Based on a number of Monte Carlo simulation experiments the joint use of LM_{LAG} and LM_{ERROR} statistics provides the best guidance with respect to the alternative model specification (alternative to OLS), as long as the assumption of normality is satisfied.

Spatial Autocorrelation/Dependence

Lagrange Multiplier

Robust LM (lag & error)

DIAGNOSTICS FOR SPATIAL DEPENDENCE

FOR WEIGHT MATRIX : **colrook.GAL** (row-standardized weights)

TEST	MI/DF	VALUE	PROB
Moran's I (error)	0.249862	2.9376173	0.0033076
Lagrange Multiplier (lag)	1	8.7599071	0.0030792
Robust LM (lag)	1	3.0721737	0.0796429
Lagrange Multiplier (error)	1	5.8148799	0.0158911
Robust LM (error)	1	0.1271465	0.7214092
Lagrange Multiplier (SARMA)	2	8.8870536	0.0117544

========================= END OF REPORT ==============================

The spatial LM_{LAG} and LM_{ERROR} specifications are highly related, so that tests against one form of dependence will also have power against the other form.

Despite these problems there are a number of practical guidelines that can be followed.

Spatial Autocorrelation/Dependence

Lagrange Multiplier
Robust LM (lag & error)

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : colrook.GAL (row-standardized weights)

TEST	MI/DF	VALUE	PROB
Moran's I (error)	0.249862	2.9376173	0.0033076
Lagrange Multiplier (lag)	1	8.7599071	0.0030792
Robust LM (lag)	1	3.0721737	0.0796429
Lagrange Multiplier (error)	1	5.8148799	0.0158911
Robust LM (error)	1	0.1271465	0.7214092
Lagrange Multiplier (SARMA)	2	8.8870536	0.0117544

========================= END OF REPORT ==============================

The most straightforward testing approach is to use Lagrange Multiplier tests that are based on the residuals of the OLS regression. The separate tests (LM_{LAG} and LM_{ERROR}) are produced, and a simple rule of thumb exists.

Spatial Autocorrelation/Dependence

Lagrange Multiplier

Robust LM (lag & error)

DIAGNOSTICS FOR SPATIAL DEPENDENCE

FOR WEIGHT MATRIX : **colrook.GAL** (row-standardized weights)

TEST	MI/DF	VALUE	PROB
Moran's I (error)	0.249862	2.9376173	0.0033076
Lagrange Multiplier (lag)	1	8.7599071	0.0030792
Robust LM (lag)	1	3.0721737	0.0796429
Lagrange Multiplier (error)	1	5.8148799	0.0158911
Robust LM (error)	1	0.1271465	0.7214092
Lagrange Multiplier (SARMA)	2	8.8870536	0.0117544

========================= END OF REPORT ==============================

Briefly, if **neither** the LM_{LAG} or LM_{ERROR} statistics reject the null hypothesis stick with OLS.

If **one** of the LM statistics rejects the null hypothesis, but the other does not, then the decision is straightforward and you should estimate the alternative "spatial" regression model that matches the test statistic that rejects the null.

Spatial Autocorrelation/Dependence

Lagrange Multiplier

Robust LM (lag & error)

DIAGNOSTICS FOR SPATIAL DEPENDENCE

FOR WEIGHT MATRIX : colrook.GAL (row-standardized weights)

TEST	MI/DF	VALUE	PROB
Moran's I (error)	0.249862	2.9376173	0.0033076
Lagrange Multiplier (lag)	1	8.7599071	0.0030792
Robust LM (lag)	1	3.0721737	0.0796429
Lagrange Multiplier (error)	1	5.8148799	0.0158911
Robust LM (error)	1	0.1271465	0.7214092
Lagrange Multiplier (SARMA)	2	8.8870536	0.0117544

======================== END OF REPORT ==============================

When **both** the LM_{LAG} or LM_{ERROR} statistics reject the null hypothesis focus on the Robust forms of the test statistics. Typically, only one of them will be significant, or one will be more significant than the other. In this case, estimate the spatial regression model matching the (most) significant statistic (above = LAG)

When both are **highly** significant go with the largest value for the test statistic (but there may be other causes of misspecification).

Spatial Autocorrelation/Dependence

Lagrange Multiplier
SARMA statistic

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : **colrook.GAL** (row-standardized weights)

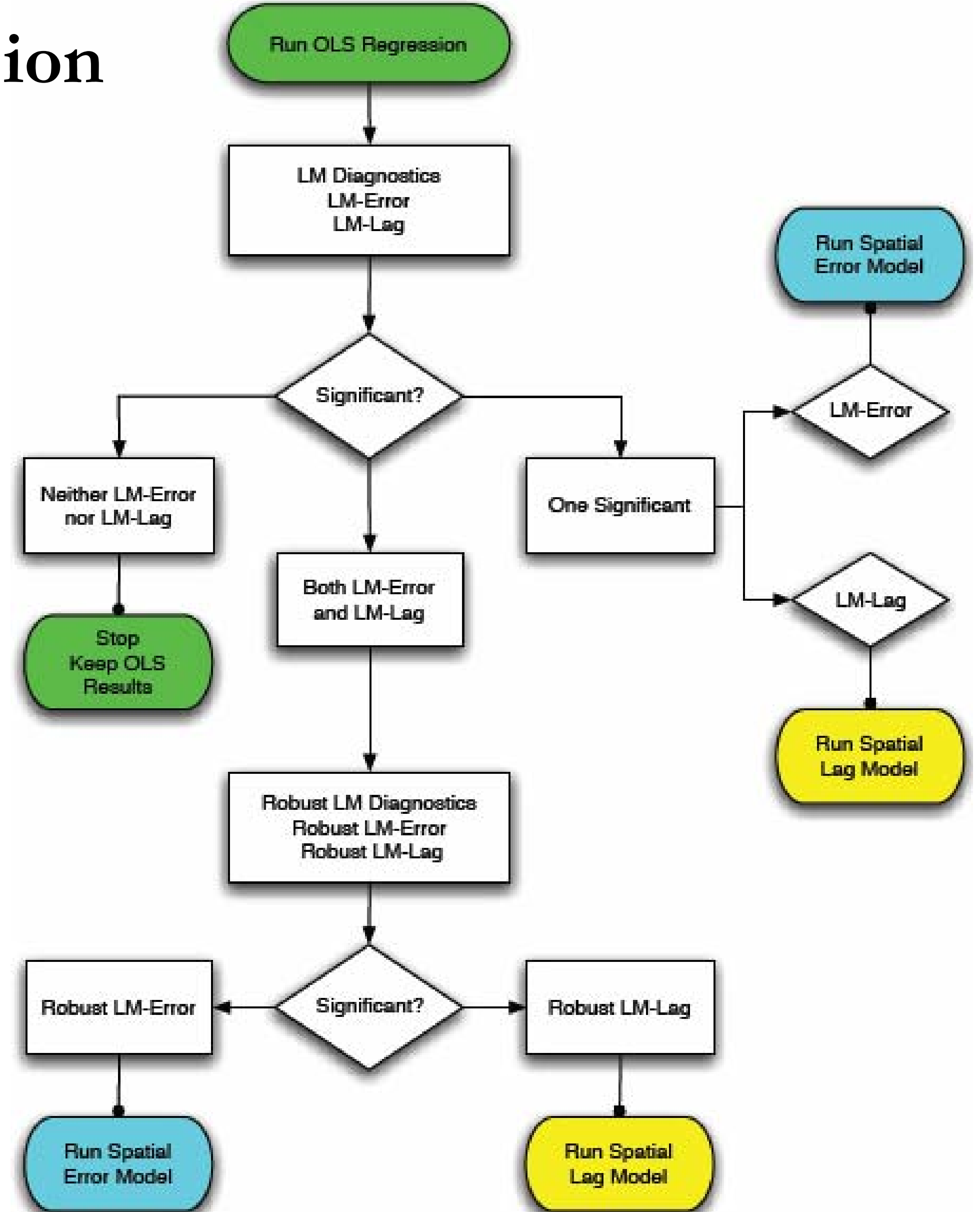
TEST	MI/DF	VALUE	PROB
Moran's I (error)	0.249862	2.9376173	0.0033076
Lagrange Multiplier (lag)	1	8.7599071	0.0030792
Robust LM (lag)	1	3.0721737	0.0796429
Lagrange Multiplier (error)	1	5.8148799	0.0158911
Robust LM (error)	1	0.1271465	0.7214092
Lagrange Multiplier (SARMA)	2	8.8870536	0.0117544

========================= END OF REPORT ==============================

The LM-SARMA will tend to be significant when either the LM_{LAG} or the LM_{ERROR} model are appropriate.

Spatial Regression Decision Tree

(*GeoDa Workbook* p. 199)



Spatial Weights

Remember your results depend on the form of the spatial weights matrix so you may want to look at different forms of the spatial weights matrix.

Spatial Regression in GeoDa

Four steps in Spatial Regression

Model Specification

Model Estimation

Model Diagnostics

Model Prediction

Spatial Regression – Model Specification

The selection of variables to be included in the model and the functional form through which they are related.

When there is no prior theoretical foundations for the choice of model, the indications given by an exploratory analysis of the data (e.g., using LISA statistics) can be very useful.

Spatial Regression – Model Estimation

Typically, a model is first estimated without incorporating spatial effects, but the results of this estimation (and its residuals) form the starting point for the diagnostics for spatial effects.

Spatial Regression – Model Diagnostics

Ideally, diagnostics aid in detecting and distinguishing between substantive (lag) and nuisance (error) spatial autocorrelation.

Spatial Regression – Model Prediction

The use of regression models is often restricted to the interpretation of the significance and magnitude of the coefficients of variables of interest.

In a GIS environment however, the results of spatial regression may also be used to predict values at locations.

Why Spatial Regression?

The concern with accounting for the presence of spatial autocorrelation in a regression model is driven by the fact that the analysis is based on *spatial data* for which the unit of observation is largely arbitrary (such as administrative units).

The methodology focuses on making sure that the estimates and inference from the regression analysis (whether for spatial or a-spatial models) are correct in the presence of spatial autocorrelation.

Why is spatial autocorrelation important?

We need to ascertain whether a spatial distribution is significantly different from the outcome of a random process **so that we do not make the mistake of attributing pattern to what is really a random distribution.**

Therefore any spatial analysis should begin with a test for the presence of spatial autocorrelation in the variables under investigation.

OLS

From an estimation point of view, the problem with an OLS model specification when spatial autocorrelation is present, is that the spatial lag term contains the dependent variables for neighboring observations, which in turn contain the spatial lag for their neighbors, and so on, leading to **simultaneity**. This simultaneity results in a nonzero correlation between the spatial lag and the error term, which violates a standard regression assumption.

OLS

Consequently, ordinary least squares (OLS) estimation will yield inconsistent (and biased) estimates, and inference based on this method will be flawed.

Instead of OLS, specialized estimation methods must be employed that properly account for the spatial simultaneity in the model. These methods are either based on the **maximum likelihood (ML)** principle, or on the application of instrumental variable (IV) estimation in a spatial two-stage, least-squares approach.

Ignoring Spatial Interaction

In practice, the most important aspect of spatial modeling may well be **specification testing**. In fact, even if discovering spatial interaction of some form is not of primary interest, ignoring spatial lag or spatial error dependence when it is present creates serious model misspecification.

Of the two spatial effects, **ignoring lag dependence is the more serious offense**, since, as an omitted variable problem, it results in biased and inconsistent estimates for all the coefficients in the model; and the inference derived from these estimates is flawed.

Ignoring Spatial Interaction

When spatial error dependence is ignored, the resulting OLS estimator remains unbiased, although it is no longer most efficient.

The estimates for the OLS coefficient standard errors will be biased, and, consequently, t-tests and measures of fit will be misleading.

Spatial Error Model

The spatial error model evaluates the extent to which the clustering of an outcome variable not explained by measured independent variables can be accounted for with reference to the clustering of error terms. In this sense, it **captures the influence of unmeasured independent variables**.

Spatial Error Model

The spatial error model takes the form described by two equations:

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

$$\varepsilon = \lambda \mathbf{W} \varepsilon + u$$

Spatial Error Model

$$y = X\beta + \varepsilon$$
$$\varepsilon = \lambda \mathbf{W}\varepsilon + u$$

Where ***y*** is a N×1 vector of observations on the dependent variable, ***X*** is an N×K matrix of observations on the explanatory variables, ***β*** is a K×1 vector of regression coefficients, **ε** in an N×1 vector of spatially autocorrelated error terms, **W**ε is a spatial lag for the errors, **λ** (lambda) is the autoregressive coefficient, and **u** is another error term (independent identically distributed).

Spatial Error Model

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

$$\varepsilon = \lambda \mathbf{W}\varepsilon + u$$

REGRESSION

SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION

Data set	: south		
Spatial Weight	: southrk.GAL		
Dependent Variable	: HR90	Number of Observations	: 1412
Mean dependent var	: 9.549293	Number of Variables	: 6
S.D. dependent var	: 7.036358	Degree of Freedom	: 1406
Lag coeff. (Lambda)	: 0.291609		
R-squared	: 0.345458	R-squared (BUSE)	: -
Sq. Correlation	: -	Log likelihood	: -4471.317119
Sigma-square	: 32.406602	Akaike info criterion	: 8954.63
S.E of regression	: 5.69268	Schwarz criterion	: 8986.150813

Variable	Coefficient	Std.Error	z-value	Probability
CONSTANT	6.693515	1.958045	3.418469	0.0006298
RD90	4.407397	0.237668	18.54434	0.0000000
PS90	1.766328	0.2256524	7.82765	0.0000000
MA90	-0.01663971	0.05298999	-0.3140161	0.7535089
DV90	0.4991464	0.1249123	3.995975	0.0000645
UE90	-0.3878414	0.07847802	-4.942039	0.0000008
LAMBDA	0.2916094	0.03727543	7.823098	0.0000000

Spatial Error Model

A satisfactory spatial error model implies that it is unnecessary to posit distinctive effects of the lagged dependent variable.

The observed spatial clustering in the outcome variable is accounted for simply by the geographic patterning of measured and unmeasured independent variables.

Spatial Lag Model

The spatial lag model in contrast, incorporates the influence of unmeasured independent variables but also stipulates an additional effect of neighboring attribute values, i.e., the lagged dependent variable.

The spatial lag model takes the form:

$$y = \varrho \mathbf{W} y + X\beta + \varepsilon$$

Spatial Lag Model

$$y = \varrho \mathbf{W}y + X\beta + \varepsilon$$

where $\mathbf{W}\boldsymbol{y}$ is an N×1vector of spatial lags for the dependent variables, ϱ (Rho) is spatial autoregressive coefficient, $X\boldsymbol{\beta}$ is an N×K matrix of observations on the exogenous explanatory variables multiplied by a K×1 vector of regression coefficients $\boldsymbol{\beta}$ for each X, and ε is a N×1 vector of normally distributed random error terms.

In the above equation, ϱ (Rho) is a scalar parameter that indicates the **effect of the dependent variable in the neighbors on the dependent variable in the focal area**.

Spatial Lag Model

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

The presence of the spatial lag is similar to the inclusion of endogenous variables on the RHS in simultaneous equations. This model is therefore often referred to as the **simultaneous spatial autoregressive model**.

Spatial Lag Model

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

REGRESSION

SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION

```
Data set                   : south
Spatial Weight             : southrk12.GAL
Dependent Variable         :          HR60  Number of Observations: 1412
Mean dependent var         :       7.29214  Number of Variables   :    7
S.D. dependent var         :       6.41874  Degrees of Freedom    : 1405
Lag coeff.    (Rho)        :      0.532889

R-squared                  :      0.197931  Log likelihood        :    -4488.97
Sq. Correlation            : -              Akaike info criterion :     8991.93
Sigma-square               :       33.0455  Schwarz criterion     :      9028.7
S.E of regression          :       5.74852
```

Variable	Coefficient	Std.Error	z-value	Probability
W_HR60	0.5328888	0.04566825	11.66869	0.0000000
CONSTANT	6.574962	1.172724	5.606573	0.0000000
RD60	1.100473	0.1963386	5.604976	0.0000000
PS60	0.03791171	0.2026779	0.187054	0.8516183
MA60	-0.1752564	0.03671206	-4.773809	0.0000018
DV60	0.9352081	0.2303864	4.059302	0.0000492
UE60	-0.1326599	0.06735334	-1.969612	0.0488827

Spatial Lag Model

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

You can interpret this model specification in two different ways:

1) You consider the inclusion of the $\mathbf{W}y$ in addition to other explanatory variables as a way to assess the degree of spatial dependence, while controlling for the effect of these other variables.

Spatial Lag Model

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

2) Alternatively, the inclusion of $\mathbf{W}y$ allows you to assess the significance of the other (non-spatial) variables, after the spatial dependence is controlled for.

Motivation for Spatial Lag Models

The spatial lag model allows for filtering out the potentially confounding effect of spatial autocorrelation in the variable under consideration. **A motivation for using a spatial lag model is to obtain the proper inference on the coefficients of the other covariates in the model.**

Spatial Lag Models

If the spatial lag model you specified is indeed the correct one, then no spatial dependence should remain in the residuals.

The **Lagrange Multiplier test for spatial error autocorrelation** in the spatial lag model is a diagnostic for this.

Spatial Lag Models

A significant result for the LM test indicates one of two things:

(1) the weights matrix is misspecified - not all spatial dependence has been eliminated (or new, spurious patterns of spatial dependence have been created) which casts doubt on the appropriateness of the spatial weights specification. **The solution** is to try a higher order spatial autoregressive model, a different weights matrix, or a different model specification (e.g., error model).

Spatial Lag Models

A significant result for the LM test indicates one of two things:

(2) It may point to the appropriateness of a mixed autoregressive spatial moving average model (i.e., a model with a spatial lag **and** a spatial moving average process in the error terms. Referred to as a **SARMA** model (Spatial Auto-Regressive Moving Average)

Spatial Lag Model

The spatial lag model is the model most compatible with common notions of **diffusion processes** because it implies an influence of neighboring attribute values that is not simply an artifact of measured or unmeasured independent variables. Rather, the outcome variable in one place actually increase the likelihood of outcome variable in nearby locales.

Notes on Diffusion

It is important to recognize that these models for spatial lag and spatial error processes are designed to yield indirect evidence for diffusion in cross-sectional data.

However, any diffusion process ultimately requires identifiable mechanisms (vectors of transmission) through which events in a given place at a given time influence events in another place at a later time.

Notes on Diffusion

The spatial lag model, as such, is not able to discover these mechanisms. Rather, it depicts a spatial imprint at a given instant of time that would be expected to emerge if the phenomenon under investigation were to be characterized by a diffusion process.

The observation of spatial effects thus indicates that further inquiry into diffusion is warranted, whereas the failure to observe such effects implies that such inquiry is likely to be unfruitful.

Comparing Models

You should not compare the R-squared across the OLS, Spatial Lag, and Spatial Error models since the spatial lag and error models are based on maximum likelihood estimation, not OLS. If you want to compare models use the respective log likelihoods of the maximum likelihood estimations.

Comparing Models

The proper measures for goodness-of-fit are based on the likelihood function and include the value of the maximized likelihood, the Akaike Information Criterion (AIC) and the Schwartz Criterion (SC).

The model with the highest log likelihood, or with the lowest AIC or SC is the best.

Spatial Regression Examples

Retrofitting Context and Integrating Spatial Models

The results that follow originate in an ongoing research project involving my colleagues R. Barry Ruback (Penn State) and Karen L. Hayslett-McCall (UT-Dallas)

OLS Model

Dependent Variable = Residential Burglary

R2-adj	0.5975	LIK	-485.615	AIC	997.231	F-test	15.9710

Variable	Coeff	S.D.	t-value	Prob
Constant	88.3811	17.8773	4.9437	0.0000
Affluence	-3.3846	1.9902	-1.7006	0.0918
Disadvantage	15.3032	2.3422	6.5335	0.0000
Immigration	-4.3356	1.8457	-2.3489	0.0206
Residential Instability	-0.4865	2.2781	-0.2135	0.8312
Population Density	-0.0030	0.0011	-2.6700	0.0087
Bus Rides	-4.1311	2.1230	-1.9458	0.0542
Bars	-2.3364	1.5473	-1.5099	0.1339
Park	0.6238	2.7384	0.2278	0.8202
BarXDisadvantage	-0.1202	1.6915	-0.0710	0.9434
Percent Male	-61.4852	34.1917	-1.7982	0.0749
Percent 18-25	94.1268	16.7951	5.6044	0.0000
Distance from Downtown	-1.5905	0.5443	-2.9219	0.0042

OLS Diagnostics

In our work based on residential burglary the diagnostics (i.e., the Lagrange Multiplier statistic) indicate that a possible alternative model would be one that incorporates a spatial lag of the dependent variable

Spatial Lag Model (MLE)

Dependent Variable = Residential Burglary

R2 0.6779 LIK -474.912 AIC 977.825

Variable	Coeff	S.D.	Z-value	Prob
Spatial Lag Res. Burglary	**0.4642**	**0.0826**	**5.6135**	**0.0000**
Constant	**58.9644**	**15.9686**	**3.6925**	**0.0002**
Affluence	**-3.4375**	**1.6955**	**-2.0273**	**0.0426**
Disadvantage	**10.7493**	**2.0532**	**5.2353**	**0.0000**
Immigration	**-5.5658**	**1.5581**	**-3.5721**	**0.0003**
Residential Instability	2.2048	1.9780	1.1146	0.2650
Population Density	**-0.0036**	**0.0009**	**-3.7237**	**0.0001**
Bus Rides	-1.5791	1.7940	-0.8802	0.3787
Bars	**-2.3192**	**1.3098**	**-1.7705**	**0.0766**
Park	-0.0910	2.3115	-0.0393	0.9685
BarXDisadvantage	1.0312	1.4291	0.7215	0.4705
Percent Male	**-54.9509**	**28.8842**	**-1.9024**	**0.0571**
Percent 18-25	**69.8891**	**14.5786**	**4.7939**	**0.0000**
Distance from Downtown	-0.6099	0.4935	-1.2357	0.2165

Spatial Lag Diagnostics

Dependent Variable = Residential Burglary

SPATIAL LAG MODEL – MLE DIAGNOSTICS

Diagnostics For Heteroskedasticity Random Coefficients

minor problems could exist in model

Lagrange Multiplier Test On Spatial Error Dependence

no spatial dependence in the residuals.

Spatial Regimes: SpaceStat

In many instances, the assumption of a fixed relation between the explanatory variables and the dependent variable that holds over the complete dataset is not tenable.

When different subsets in the data correspond to regions or spatial clusters, Anselin refers to this as a spatial regimes model.

Spatial Regimes: SpaceStat

Spatial Lag Model (ML) for Structural Change using SVS Tract Dummy

R2 0.7282 Sq. Corr. 0.7618 LIK -463.811 AIC 981.622

SC 1057.33 SIG-SQ 110.445 (10.5093)

	SVS = No (0)		SVS = Yes (1)	
Variable	Coeff	t-value	Coeff	t-value
Spatial Lag of Res. Burglary	0.503	6.366***		
Constant	72.906	1.415	44.810	2.601***
Affluence	-6.897	-1.554	-2.188	-1.195
Disadvantage	**-5.780**	**-0.364**	**11.679**	**5.533*****
Immigration	**0.883**	**0.147**	**-4.641**	**-2.892*****
Residential Stability	**7.794**	**1.815***	**1.074**	**0.468**
Population Density	**-0.010**	**-2.667*****	**-0.001**	**-0.700**
Bus Rides	**9.476**	**1.324**	**-5.262**	**-2.224****
Bars	5.823	0.729	-0.476	-0.341
Park	-7.594	-1.143	-0.365	-0.157
BarXDisadvantage	7.704	0.671	1.769	1.098
Percent Male	-142.790	-1.261	-25.563	-0.750
Percent 18-25	22.476	0.467	25.203	0.985
Distance from Downtown	2.378	1.883	-0.598	-1.206

Spatial Regimes Diagnostics

SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
REGRESSION DIAGNOSTICS reveal

Test on structural instability for two regimes (SVS tract dummy)

1. all coefficients jointly	**coefficients not the same in the two regimes significant (0.026)**
2. individually	**There is a significant difference in the relation of the following variables and Residential Burglary in the two Spatial regimes: Pop. Density (0.020) Bus Rides (0.050) & Distance Downtown (0.026)**
Lagrange Multiplier Test On Spatial Error Dependence	**no spatial dependence in the residuals**.

Sampson, Morenoff and Earls (1999)

"Beyond social capital: spatial dynamics of collective efficacy for children" ASR 64, 633-660.

GIS-related techniques used for:

- integrating tract level data to neighborhood
- mapping
- local analysis and spatial modeling

Neighborhood level social organization

Intergenerational closure
– degree to which adults and children are "linked" to one another.

Reciprocated (relatively equal) exchange
- level of interfamily and adult interaction

Informal social control and mutual support of children or "collective efficacy"
- shared values among neighbors and expectations for action within a collectivity

Sampson et al., 1999

Data:

Project on Human Development in Chicago Neighborhoods (PHDCN).

8500 residents in

865 tracts grouped into 343 relatively homogenous "neighborhoods" (race/ethnic mix, SES, housing density, family structure).

Neighborhood measures: concentrated disadvantage, concentrated immigration and residential stability, etc.

Sampson et al., 1999

Method:

Individuals “nested” within neighborhoods therefore used Hierarchical linear models (HLM)

They argue “that the emergence of intergenerational closure, reciprocal exchange, and child-centered social control in a neighborhood benefits not only residents of that area but also others who live nearby. Methodologically this leads to a model of spatial dependence in which neighborhood observations are **interdependent** and are characterized by a **functional relationship** between what happens at one place and what happens elsewhere.” (p.645)

Sampson et al., 1999

Method (continued):

Spatial embeddedness – spatial lag regression models run using SpaceStat.

“we explore a typology of spatial association that decomposes the citywide pattern into its specific local forms. The typology we employ is referred to as a Moran Scatterplot” (p. 649).

Color Versions of Black and White Maps from Sampson, Robert J., Jeffrey Morenoff, and Felton Earls. 1999. "Beyond Social Capital: Spatial Dynamics of Collective Efficacy for Children." *American Sociological Review* 64: 633-660.

Figure 1. Spatial Typology of Adult-Child Exchange: Chicago, 1995

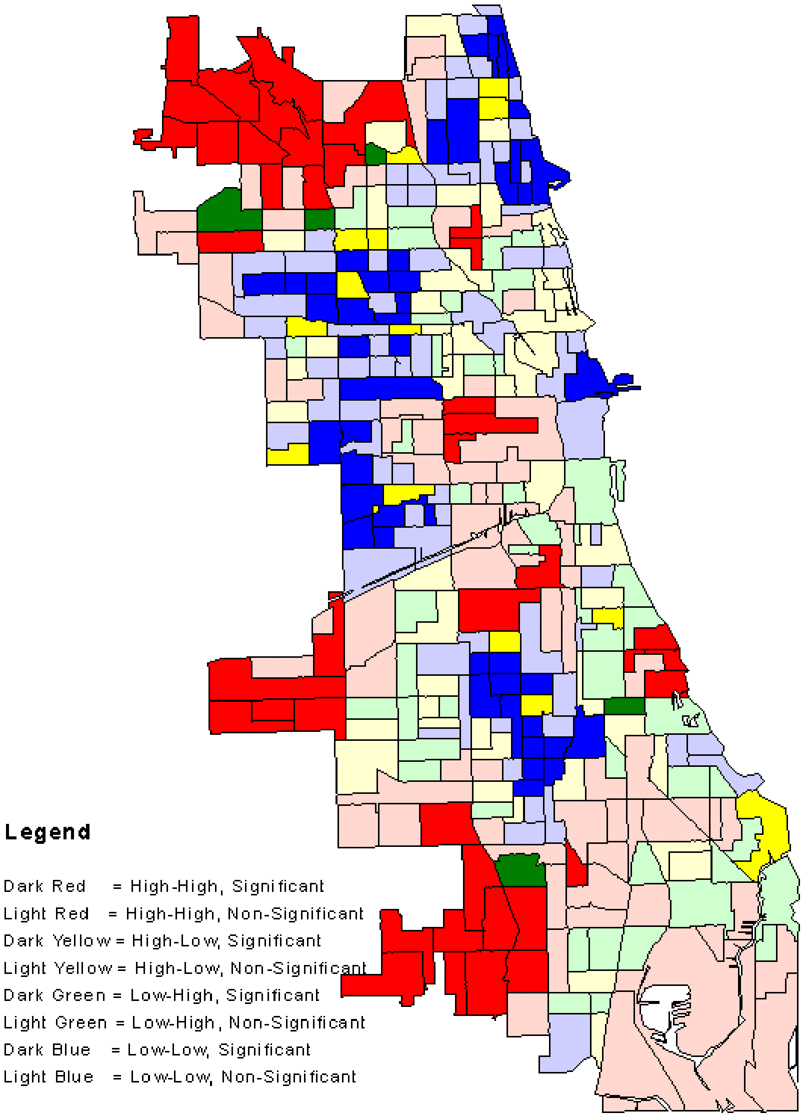
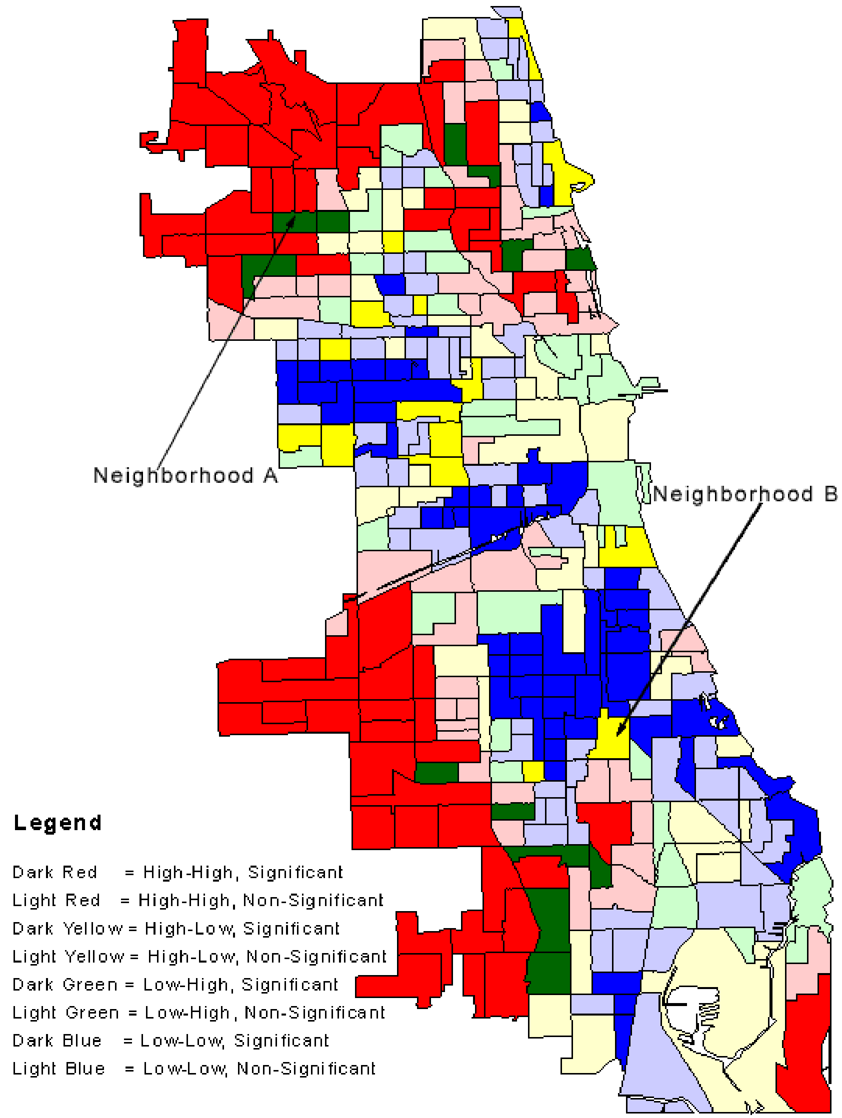


Figure 2. Spatial Typology of Child-Centered Social Control: Chicago, 1995



Sampson et al., 1999

Conclusions:

"The results point to how spatial inequality in a metropolis can translate into local inequalities for children. Above and beyond the internal characteristics of neighborhoods themselves the potential benefits of social capital and collective efficacy for children are linked to a neighborhood's relative position in the larger city… some neighborhoods benefit simply by their proximity to (other) neighborhoods … but white neighborhoods are much more likely to reap the advantages of such spatial proximity.

Sampson et al., 1999

Conclusions (continued):

"Spatial externalities have been overlooked in prior research, but our analysis indicates that social capital and collective efficacy for children are relational in character at a higher level of analysis than the individual or the local neighborhood."

Jeffrey Morenoff (2003)

Neighborhood Mechanisms and the Spatial Dynamics of Birth Weight. *AJS* 108 (5), 976-1017.

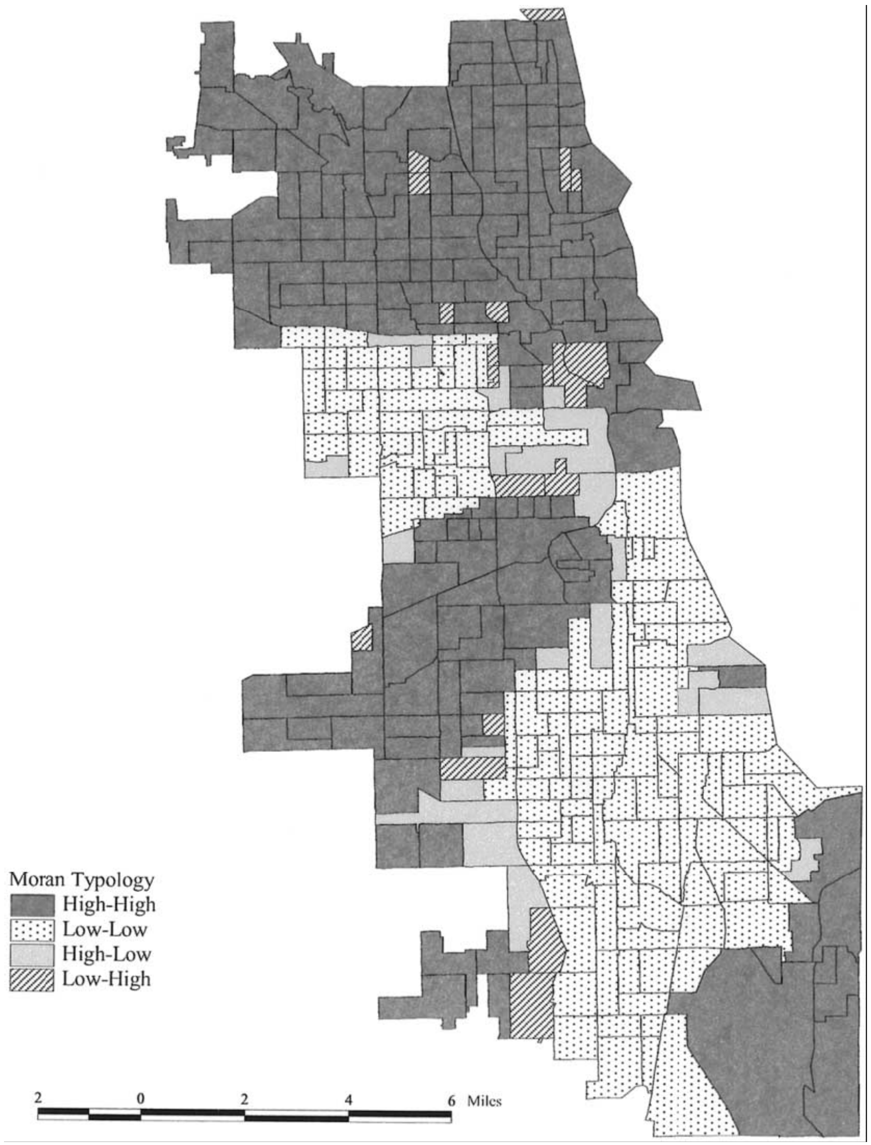
A distinctive methodological feature of this analysis is that it combines multilevel & spatial modeling techniques.

Spatial effects on birth weight are estimated through an autoregressive process in the dependent variable known as a “spatial lag” model.

$$y = \varrho \mathbf{W} y + X\beta + \varepsilon$$

Jeffrey Morenoff (2003)

Local Moran
for Birth Weight



Jeffrey Morenoff (2003)

Local Moran
for the log of the
Violent Crime Rate

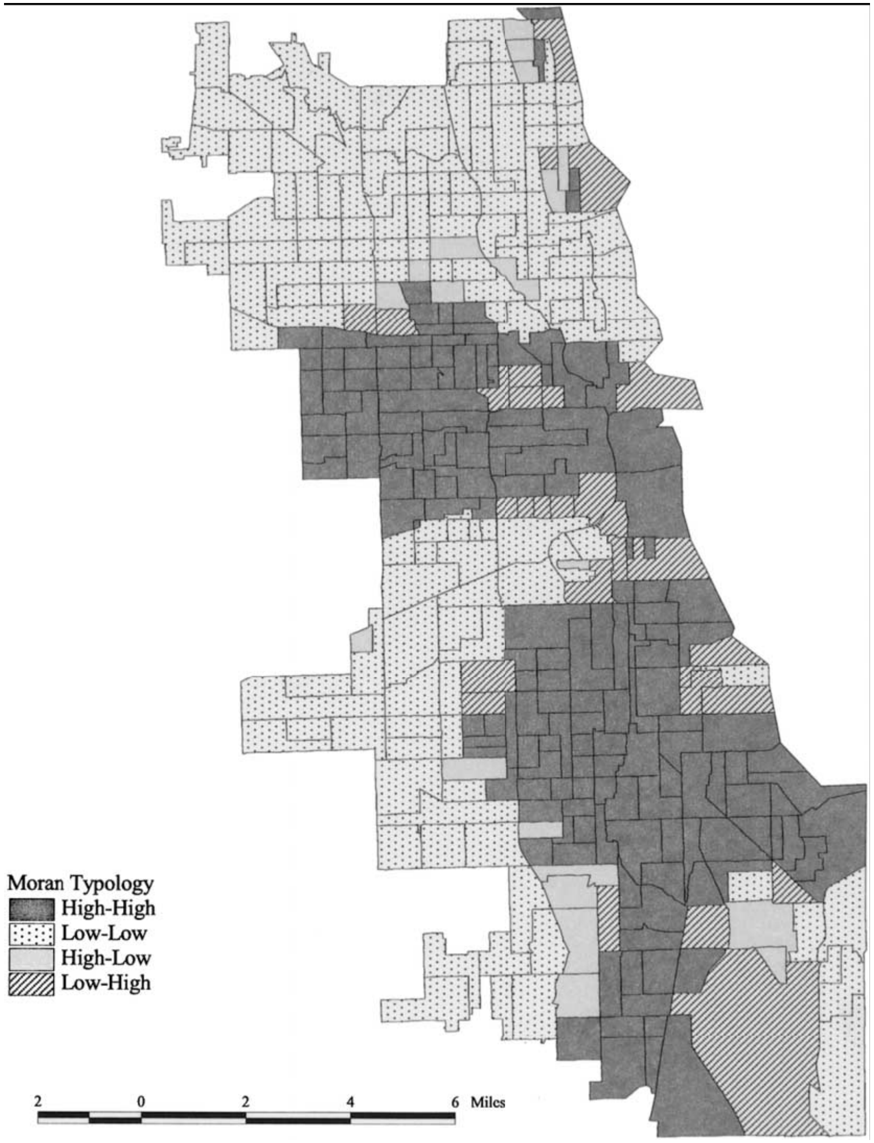


TABLE 4

Coefficients from Spatial Lag Regression Models of Continuous Birth Weight and Low Birth Weight (Adjusted Neighborhood Means)

	Continuous Birth Weight				Low Birth Weight	
	ML (1)		2SLS (2)		2SLS (3)	
Independent Variable	Coefficient	SE	Coefficient	SE	Coefficient	SE
%African-American	2.43	3.36	2.58	3.38	−.02	.03
%Mexican	3.01	2.04	2.89	2.05	.02	.02
%poor families	1.73	2.64	1.02	2.77	−.02	.02
Residential stability	2.37	2.28	1.55	2.43	.00	.02
ln violent crime rate	−9.16	3.57*	−7.50	3.96	.04	.03
Exchange/voluntarism ...	6.30	2.30**	5.46	2.47*	−.02	.02
Spatial lag term (*Wy*)	.33	.07***	.53	.21*	.69	.35*
Intercept	74.69	8.18***	52.91	23.42*	−.08	.09
R^2	.24		.29		.18	

The rho coefficient represents the rate at which spatial externalities—i.e., effects from the observed and unobserved characteristics of adjacent neighborhoods—contribute to birth weight in the focal neighborhood. For continuous birth weight, rho is estimated to be between 0.33 (using ML) and 0.53 (using 2SLS), meaning that the total effects of observed and unobserved neighborhood-level causes of birth weight are about one-third to one-half larger when we take into account the effects of externalities from surrounding areas. For low birth weight, the effects of observed and unobserved causes in adjacent neighborhoods is an astounding 69% as large as it is in the focal neighborhood.

E-mail: matthews@pop.psu.edu