Correcting for Spatial Effects in Limited Dependent Variable Regression: Assessing the Value of "Ad-Hoc" Techniques[1]

Alessandro De Pinto and Gerald C. Nelson

May 14, 2002

Abstract:
A common test for spatial dependence in regression analysis with continuous dependent variables is the Moran's I. For limited dependent variable models, the standard definition of a residual breaks down because y_i is qualitative. Efforts to correct for potential spatial effects in limited dependent variable models have relied on ad-hoc methods such as including a spatial lag variable or using a regular sample that omits neighboring observations. Kelejian and Prucha have recently developed a version of Moran's I for limited dependent variable models. We present the statistic in a more accessible way and use it to test the value of previously-used ad-hoc techniques with a specific data set.
Keywords: Moran's I, Spatial Autocorrelation, Limited Dependent Variable Models, Land-Use Change, Geographical Information Systems (GIS),

Corresponding author:
Alessandro De Pinto, doctoral candidate, Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign.
Mumford Hall, 1301 W. Gregory Drive, Urbana, IL, 61801
adepinto@uiuc.edu

Gerald Nelson, Associate Professor, Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign.

---

[1] Selected paper prepared for the American Agricultural Economics Association annual meeting, Long Beach, California, July 29-Aug 1, 2002. This material is based upon research supported in part by the Cooperative State Research, Education and Extension Service, U.S. Department of Agriculture, under Project No. ILLU 05-0361 and the Research Board of the University of Illinois, Urbana-Champaign. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors. The authors are solely responsible for the analysis and conclusions presented here.

Introduction

In recent years there has been a rapid increase in the use of spatially-explicit data in economic modeling. Although this type of data can provide unique insights, its use poses conceptual and technical difficulties. In particular, the existence of spatial relationships among observations can result in unreliable estimates and statistical inference of the parameters (Anselin (1988)). Standard econometric techniques often fail in the presence of spatial correlation and heterogeneity. Econometric problems with spatial data are not exclusively a consequence of the interaction among "neighboring" agents. Spatial effects can also arise when data from different sources, different sample designs, or varying aggregation rules is used. The need to integrate data from various sources will tend to result in spatially dependent as well as spatially heterogeneous observations.

Spatial econometrics has relatively well-established procedures for models with continuous LHS variables. However, there are no well-established procedures for testing for the existence of spatial effects or methods to incorporate spatial effects in limited dependent variable models.[2] Some researchers have attempted to reduce the potentially negative effects by including one or more spatial lag variables or using a regular sample

[2] Explicit modeling of spatial effects in limited depend variable models is an extremely active area of research, particularly in the context of spatial probit (Case (1992); McMillen (1992); Bolduc, et al. (1997); Pinkse and Slade (1998), Beron and Vijverberg (Forthcoming), Fleming (2001)). With the exception of an approach proposed by Fleming, all other proposed solutions are either very computationally intensive or adopt ad-hoc solutions to model spatial effects.

that omits neighboring observations but until recently there has been no way of assessing the value of these approaches in reducing the negative consequences of spatial effects. With continuous LHS variables, the statistic known as Moran's I, based on estimates of the residuals ($y - \hat{y}$), provides a common test for spatial dependence. With limited dependent variable models, however, the standard Moran's I approach breaks down because there are no residuals as conventionally defined (no $\hat{y}$). In this article we implement a version of the Moran's I developed by Kelejian and Prucha (2001) that is applicable to limited dependent variable models and use it to evaluate the effectiveness of some of the ad-hoc approaches used in the literature on land use.

Ad-Hoc Corrections for Spatial Effects in Limited Dependent Variable Models

Three types of ad-hoc corrections can be found in the land use literature – regular sampling from a grid, pure spatial lags variables using latitude and longitude index values, and spatial lag variables involving a geophysical variable such as slope or rainfall.

Regular sampling from a grid

Nelson and Hellerstein (1997), following Besag (1974) as described in Haining (1994), applied a "coding" scheme (Besag's terminology) that selects samples over a regular grid in such a way that two observations are not physical neighbors. The rationale for this method is that many spatial relationships between observations decay with Euclidean distance. Observations 'sufficiently' distant do not influence each other. The coding method has been subsequently used by Mertens and Lambin (2000), Munroe, et al. (2001), Nelson, et al. (2001), and Nelson, et al. (forthcoming, 2003).

Latitude and longitude as exogenous variables

Nelson, et al. (2001) corrected for spatial effects using two additional explanatory variable representing latitude and longitude of each observation. This method is likely to be helpful when the spatial effect is caused by an unobserved variable that varies linearly over the area. However, this is a very special case and does not account for all the other possible spatial relationships.

Spatially lagged geophysical variables

Nelson, et al. (2001) and Munroe, et al. (2001) use spatial lags (weighted averages of values in neighboring locations) of key geophysical variables such as soil type and slope as exogenous variables. In essence this approach approximates the use of a (contiguity) spatial weight matrix applied to selected RHS variables**.**

The only attempt to evaluate the effectiveness of one of the ad-hoc methods is Munroe, et al. (2001). After implementing a coding scheme with different sample sizes, the authors use a join count statistic to test the hypothesis of spatial randomness in the land use choices. This test is capable of detecting both positive (i.e. the propensity of similar values to create clusters) and negative (i.e. the propensity of dissimilar values to appear in close association) spatial autocorrelation. However, there is a severe flaw in this method since the test is applied directly to the manifestation of a latent process. Spatial effects could be present in the true underlying process but absent in its realization and vice versa.

Moran's I for Limited Dependent Variables

The standard correlation coefficient, *r*, is a measure of direction and closeness of linear association between 2 variables X and Y. An intuitive explanation of *r* is that if Y and X move together, r is close to 1. If they move in opposite directions, r is close to -1.

Moran's I is similar to a correlation coefficient, except that it refers to correlation of a single variable, x, to itself across space. Moran's I is defined as:

$$I_M = \frac{\mathbf{x'Wx}}{\mathbf{x'x}} \qquad (1)$$

W is a matrix that defines the neighborhood set for each location and can be row-standardized (the original row values are divided by the sum of the original row values). The analogy to a correlation coefficient is that when X_is at neighboring locations are generally greater than $\overline{X}$, I_M approaches 1. When X_i at a location is greater than $\overline{X}$, and its neighbors are generally less than $\overline{X}$, I_M approaches -1. The values in W operationalize our assumptions about what locations are 'neighbors'. Note that W is an n x n matrix (n is number of observations in sample) so operations involving W typically require manipulating a large matrix.

To test for spatial error dependence with a continuous LHS variable, we use Moran's I constructed as follows:

$$I_M = \frac{\hat{\mathbf{e}}'\mathbf{W}\hat{\mathbf{e}}}{\hat{\mathbf{e}}'\hat{\mathbf{e}}} \qquad (2)$$

where I_M is now based on the residuals $\hat{e}_i = y_i - \mathbf{X_i}\hat{\boldsymbol{\beta}}$.

However, in limited dependent variable models the LHS is discrete and the definition of $\hat{e}_i$ above has no meaning.

To follow Kelejian and Prucha's method of estimating a Moran's I for a polychotomous limited dependent variable model, we start with definition of notation:

m – number of choices for the LHS variable y

j – the index (identification number) over the choices; j = 1,m; the assignment of identification numbers to qualitative variables is arbitrary. The only requirement is that they are sequential

n – number of observations

i – index of the number of observations, i = 1,n; also a pointer to a specific location since each observation is at a particular location

x_i – vector of exogenous variables, valued at location i

b_0 – vector of parameters to be estimated.

We start with a function that determines the probability that qualitative variable j is found at location i.

$$\Pr\left(y_i = j\right) = P_j\left(x_i, b_0\right) \qquad (3)$$

P_j is the function that transforms the x's at location i and b_0's into a probability value Pr. Create a new function, f, which is a weighted average of the identification number (index) of the choice with the probability value (P_j) as the weight.

$$f\left(x_i, b_0\right) = \sum_{j=1}^{m} jP_j\left(x_i, b_0\right) \qquad (4)$$

Kelejian and Prucha use this f and an error term as follows

$$y_i = f\left(x_i, b_0\right) + e_i, i = 1,...,n \quad (5)$$

$$E\left(e_i\right) = 0; s_i^2 = E\left(e_i^2\right) = h_i\left(b_0\right) \qquad (6)$$

$$h_i\left(b_0\right) = \sum_{j=1}^{m} j^2 P_j\left(x_i, b_0\right) - \left[\sum_{j=1}^{m} jP_j\left(x_i, b_0\right)\right]^2 \quad (7)$$

Equation (5) says that the correct indicator value (y_i) is the sum of f (a predictor of y_i) and an error term. Expression (6) describes the distribution characteristics. Equation (7) is the expression for the variance of the error term.

The generalized Moran's I test is:

$$I_n = \frac{Q_n^*}{\tilde{s}_{Q_n^*}} \xrightarrow{D} N(0,1) \qquad (8)$$

$$Q_n^* = \hat{\mathbf{e}}' \mathbf{W} \hat{\mathbf{e}} \qquad (9)$$

$$\hat{e}_i = y_i - f\left(x_i, \hat{b}_0\right) \qquad (10)$$

Kelejian and Prucha demonstrate that in absence of spatial autocorrelation the index I has a normal distribution centered on zero and with variance one.

$\tilde{s}_{Q_n^*}$ is a normalization factor, which for limited dependent variable models is:

$$\begin{aligned} \tilde{s}_{Q_n^*} &= \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n}\left(w_{ik,n} + w_{ki,n}\right)^2 s_{i,n}^2 s_{k,n}^2 \\ &= tr\left(W_n \hat{\Sigma}_n W_n \hat{\Sigma}_n + W_n' \hat{\Sigma}_n W_n \hat{\Sigma}_n\right) \\ \hat{\Sigma}_n &= diag\left(\hat{s}_{i,n}^2\right) \end{aligned} \qquad (11)$$

Data Sources and Manipulation

The data set used for this study has been previously used for an investigation on the effects of road construction on land use in an area in Panama. For a more detailed description of this data set, as well as the behavioral model, see Nelson, et al. (2001). Information on the land use choices, the dependent variable, was derived from satellite images. The set of explanatory variables can be divided in two groups: geophysical and socioeconomic. Geophysical variables capture the natural predisposition or physical

limitations of a plot of land to enter in agricultural production. In our case these are temperature, elevation, slope, and soil quality. Socioeconomic variables were chosen to reflect the influence of prices and property rights on land use. For this study, information on land tenure was used as well as cost of access data. Since the objective of this exercise is not the estimation of the regression parameters to provide normative policy recommendation, to speed up the test process we have used only a portion of the original study area. The original data set had a total of about 63,000 observations; in this sub-sample the observations are about 36,000.

Implementation

The pseudo-errors (10), the residuals for the discrete choice model, and the normalization factor $\tilde{s}_{Q_n^*}$ (11) were obtained using Limdep v. 7.0.

The contiguity matrix is constructed using a queen criterion (all 8 neighboring cells) using SpaceStat v. 1.9.1 and is row-standardized. It should be noted that the structure of the contiguity matrix W (i.e. queen, rook, or bishop) reflects the researcher's beliefs regarding the way that the spatial process propagates through space. There are no tests for the specification of W. However, misspecifying W in a test statistic is much less serious than misspecifying the contiguity matrix in a spatial regression model. In a test statistic, misspecification can render the test statistic less powerful; in a regression model it usually causes the estimator to be inconsistent.[3]

[3] In a test statistic, misspecifying W by choosing a simpler structure may increase the power of the test (see e.g. Florax and Rey, (1995)). Stetzer (1982) finds that "in a Monte Carlo study that, although the choice of weight matrix has an effect on the performance of estimators in spatial regression models, other factors, including delineation of the geographical area studied, tend to be more important."

The computation of the Kelejian-Prucha version of the Moran's I was performed using MatLab v. 6.1 mostly because of its ability to handle sparse matrices. As indicated above the size of the contiguity matrix is given by the square of number of observations. Fortunately, most of its elements are zeros and can be easily handled by software that has sparse matrix functionality. For example, with 4,536 observations on a Pentium 4 computer with 256 mb of RAM, MatLab takes about 6.20 minutes to complete the calculation.

Results

The results of our analysis indicate that the common methods used to remove spatial effects have some merit, at least for the data set used here. Although they do not completely remove spatial autocorrelation they greatly reduce it.

Effect of Sampling

Figure 1 shows the results when sampling is used (Besag's coding scheme). With the full data set the value for the Moran's I is 117.8. As the distance between observations increases to 2 Km, the value decreases to 7.1. Although this later value is still above the cutoff level for statistical significance, sampling our data set reduced the magnitude of spatial autocorrelation dramatically.

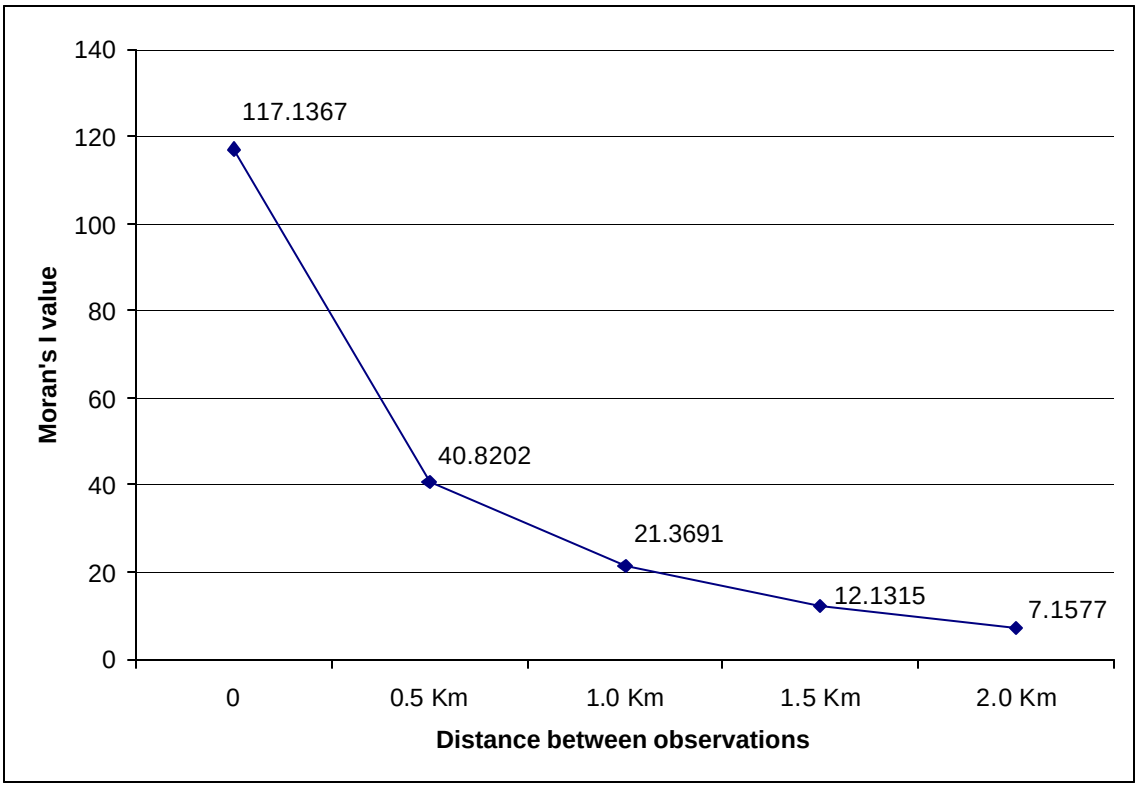

Figure 1: Effect of Sampling on Spatial Effects

Effect of spatial lags

Figure 2 summarizes the effects of various approaches to spatial lags. As it can be observed these ad-hoc methods are again partially effective in reducing the Moran's I value, however they are never successful in completely eliminating spatial autocorrelation.

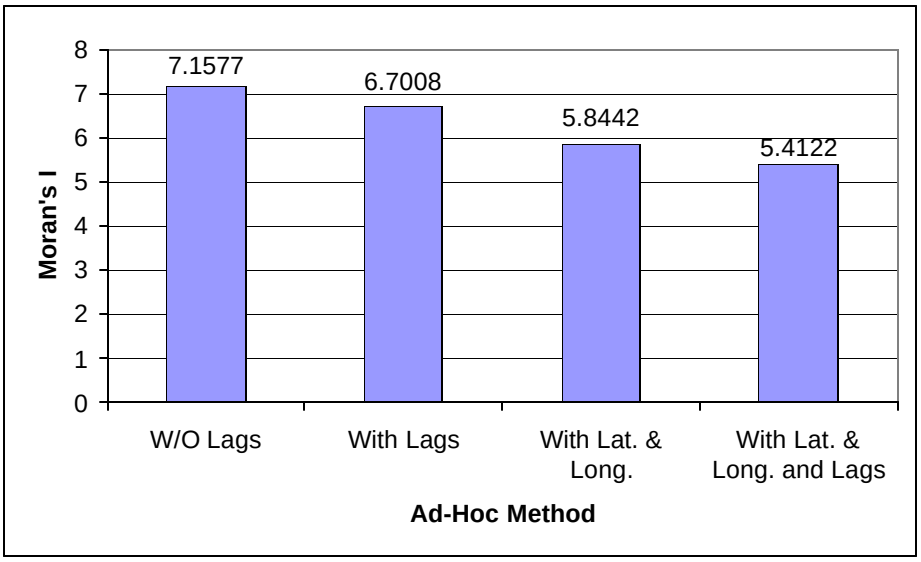

Figure 2: Effect of Spatial Lags on Spatial Effects

The results of this test, however, do not necessarily indicate that the sampling process is ineffective in removing spatial correlation. In fact, not only this test cannot distinguish between error (nuisance) and lag (substantive) spatial dependence but it also detects heteroskedasticity, which in space is also called spatial heterogeneity. The latter is simply structural instability in the form of non-constant error variances (heteroskedasticity) or model coefficients (variable coefficients, spatial regimes). It has been often noted (Anselin 1999, McMillen (1992)) that spatial autocorrelation and spatial heterogeneity may be observationally equivalent. For example, a spatial cluster (i.e., observed in locations that are in close proximity) of extreme residuals may be interpreted as due to spatial heterogeneity (e.g., groupwise heteroskedasticity) or as due to spatial autocorrelation (e.g., a spatial stochastic process yielding clustered values).

That we are in presence of heteroscedastic error terms can be observed in Figure 3 and 4 where we have plotted the pseudo-errors against two explanatory variables, Slope and Soil Type (SoilIndx), after having sampled with a 5x5 scheme.

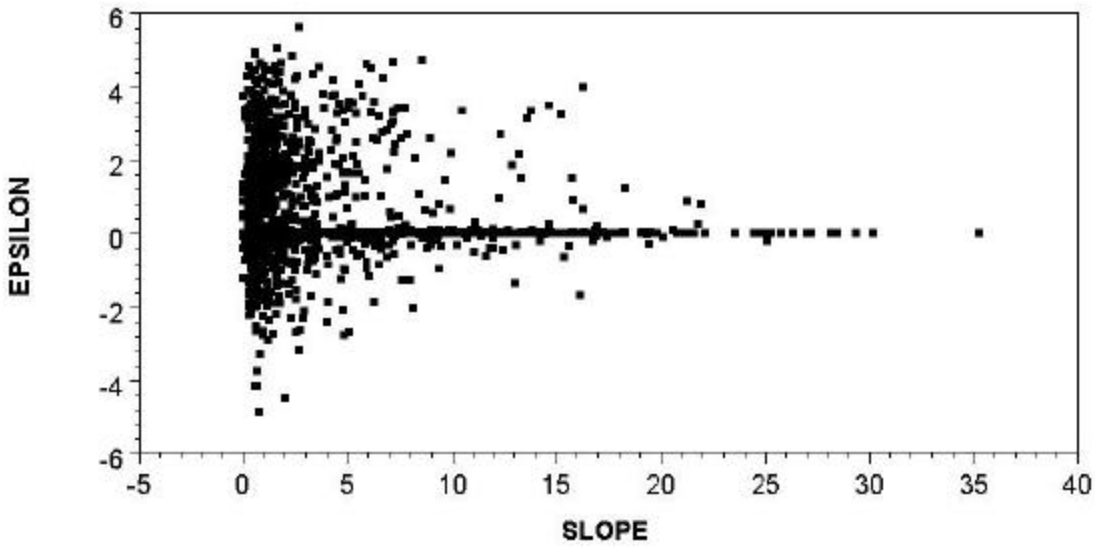

Figure 3: Plot of Pseudo-Errors Against Slope

For the slope variable, the pseudo errors have a much greater range at small values of slope than large values. Since Areas of similar slope are naturally clustered in space, heteroskedasticity in our case is also spatial groupwise heteroscedasticity. For soil type, on the other hand, there is little difference in the range of pseudo error values for the different soil types.

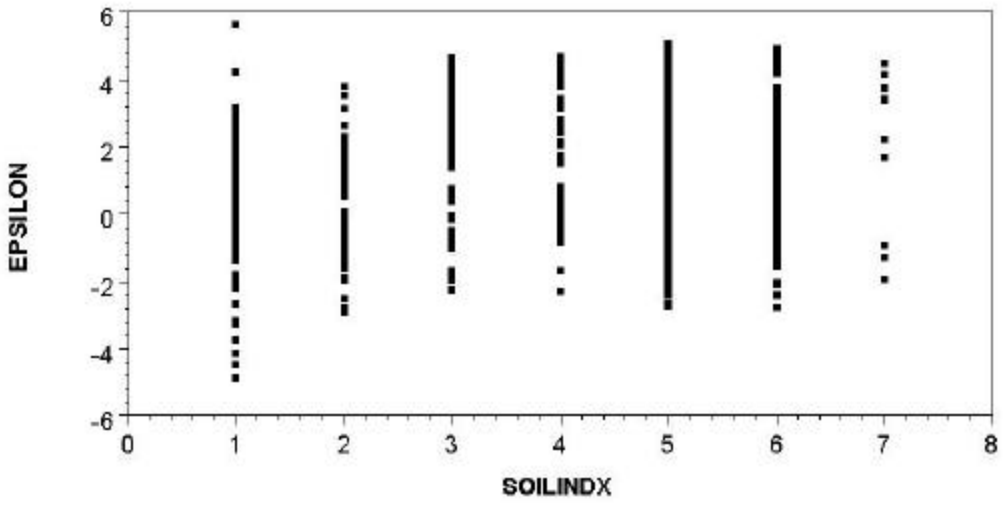

Figure 4: Plot of Pseudo-Errors Against Soil Type

We have created maps of pseudo-errors for a visual inspection of the spatial distribution of the residuals. Clusters of residuals with same sign and similar magnitude can be observed in all three maps and although the clustering is considerably reduced when the original map is sampled with a 5x5 coding scheme, it does not disappear.

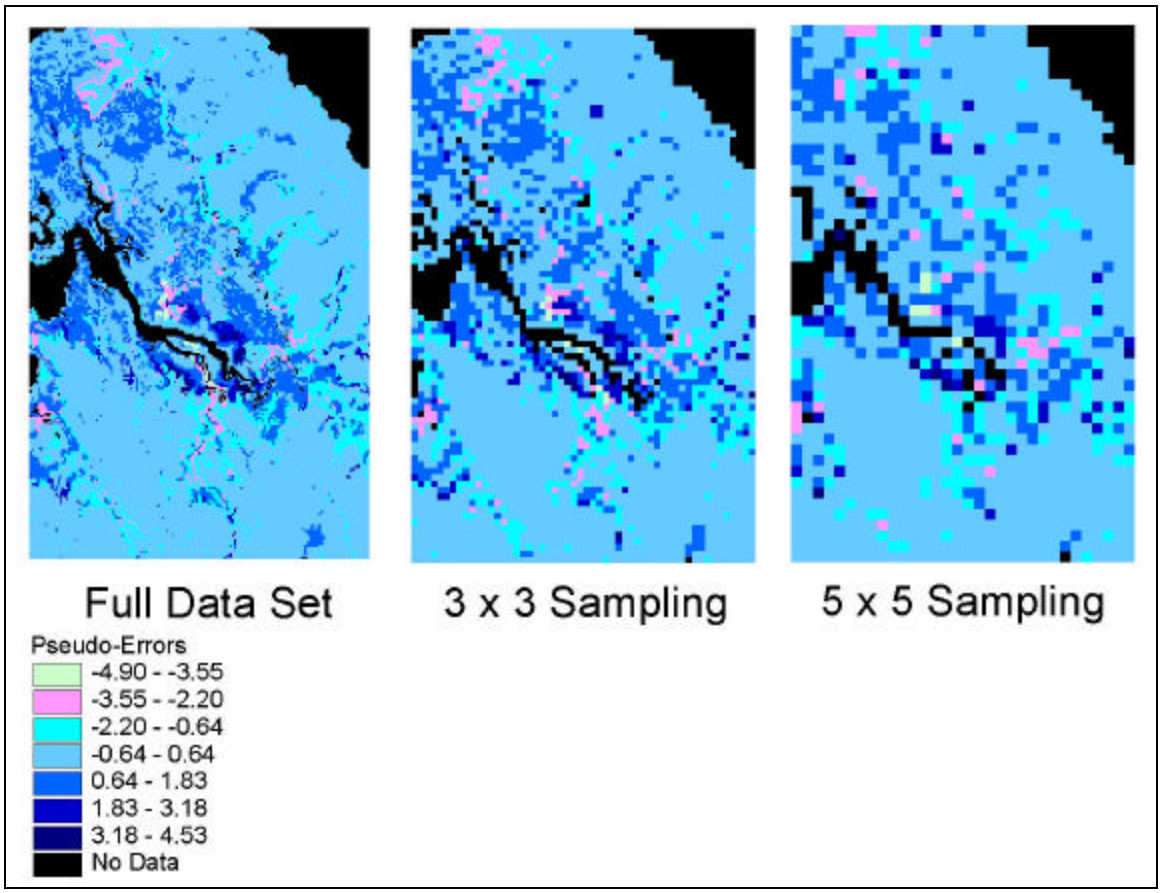

Figure 5: Maps of Pseudo-Errors with Different Sampling Schemes

Conclusions

Researchers have attempted to control for spatial effects in models of land use; however, they did not have a method to assess how successful they were in removing spatial autocorrelation. In this paper we have implemented a version of Moran's I developed by Kelejian and Prucha that applies to limited dependent variable models. Our results show that, at least with for the data set we used, these "ad-hoc" methods are not completely successful in eliminating spatial effects. The value for the Moran's statistic is considerably reduced when the original data set is sampled and a combination of lagged geophysical variables and longitude-latitude variables are introduced in the model

specification. This shows that these techniques have some merit. Sampling the data set is still a viable solution and it possible to select observation further apart than we have attempted in this study. The obvious drawback is that dropping observations decreases the predictive power of the model[4]. As with the regular Moran's I, the test does not discriminate between different kinds of spatial effects or between spatial effects per se and spatial heterogeneity. But for the first time, land use modelers have a statistically valid tool for assessing the presence of spatial effects and the benefit of various kinds of ad-hoc correction procedures.

[4] For an investigation of the effects of sampling on model accuracy see Munroe, Southworth, Tucker (Forthcoming)

Appendix

How to treat missing or rejected observations for the computation of the Moran's I according to K-P method. Take as an example a map with dimensions of 5x4. We select the first row and every other row after that. We take every second column. This means there are 6 observations in our sample.

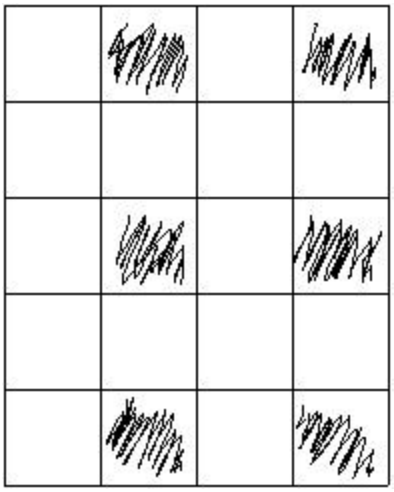

Figure 3: No missing observations

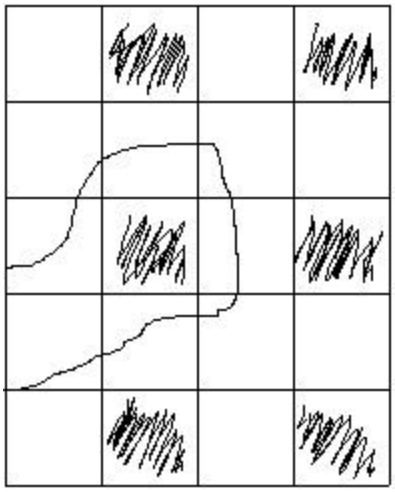

Figure 4: Missing observation 3,2

If there are no missing observations, as in

Figure 3, the corresponding 6x6 queen's case weight matrix is Table 1.

Table 1: 6x6 contiguity matrix

	1,2	1,4	3,2	3,4	5,2	5,4
1,2	0	1	1	1	0	0
1,4	1	0	1	1	0	0
3,2	1	1	0	1	1	1
3,4	1	1	1	0	1	1
5,2	0	0	1	1	0	1
5,4	0	0	1	1	1	0

Suppose now that observation 3,2 is missing as in

Figure 4. The weight matrix that describes the neighbors among the remaining 5 observations is a 5x5 square matrix, eliminating observation 3,2. Note that this matrix is identical to the matrix in Table 1 with the row and column for observation 3,2 missing.

Table 2: 5x5 contiguity matrix (6x6 with obs 3,2 missing)

	1,2	1,4	3,4	5,2	5,4
1,2	0	1	1	0	0
1,4	1	0	1	0	0
3,4	1	1	0	1	1
5,2	0	0	1	0	1
5,4	0	0	1	1	0

From a contiguity perspective, Table 2 is the same as Table 1 with 0s replacing the 1s in the row and column for observation 3,2.

	1,2	1,4	3,2	3,4	5,2	5,4
1,2	0	1	**0**	1	0	0
1,4	1	0	**0**	1	0	0
3,2	**0**	**0**	**0**	**0**	**0**	**0**
3,4	1	1	**0**	0	1	1
5,2	0	0	**0**	1	0	1
5,4	0	0	**0**	1	1	0

The interpretation of the 0s is that the 3,2 observation is no one's neighbor and no observation is the neighbor of observation 3,2. Operationally, this effect can be achieved by injecting a row and a column of 0s in the observation's position.

References

Anselin, L. *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers., 1988.

Beron, K., and W. P. M. Vijverberg (forthcoming). Probit in a Spatial Context: A Monte Carlo Analysis, in *New Advances in Spatial Econometrics*. ed. L. Anselin. Berlin, Springer Verlag, pp.

Besag, J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B* 36(1974): 192 - 236.

Bolduc, D., B. Fortin, and S. Gordon. Multinomial Probit Estimation of Spatially Interdependent Choices: an Empirical Comparison of Two New Techniques. *International Regional Science Review* 20, no. 1-2(1997): 77-101.

Case, A. Neighborhood Influence and Technological Change. *Regional Science and Urban Economics* 22, no. 3(1992): 491-508.

Fleming, M. M.. Development Patterns on the Urban Fringe: A Spatially Correlated Discrete Choice Approach. Paper presented at the Regional Science Association International 48th Annual North American Meeting. Charleston, S.C., November 15-18, 2001.

Haining, R. Diagnostics for Regression Modeling in Spatial Econometrics. *Journal of Regional Science* 34, no. 3(1994): 325-41.

Kelejian, H. H., and I. R. Prucha. On the asymptotic distribution of the Moran I test statistic with applications. *Journal of Econometrics* 104, no. 2(2001): 40.

McMillen, D. P. Probit with Spatial Autocorrelation. *Journal of Regional Science* 32, no. 3(1992): 335-348.

Munroe, D., J. Southworth, and C. M. Tucker. The Dynamics of Land-Cover Change in Western Honduras: Spatial Autocorrelation and Temporal Variation. American Agricultural Economics Association Annual Meetings2001,

Nelson, G. C., V. Harris, and S. W. Stone. Deforestation, Land Use, and Property Rights: Empirical Evidence from Darien, Panama. *Land Economics* 77, no. 2(2001): 187-205.

Nelson, G. C., and D. Hellerstein. Do Roads Cause Deforestation? Using Satellite Images in Econometric Analysis of Land Use. *American Journal of Agricultural Economics* 79, no. 1(1997): 80-88.

Nelson, G. C., De Pinto, A. and S. W. Stone,. Land Use and Road Improvements: a Spatial Econometric Analysis. *International Regional Science Review* (forthcoming, 2003).

Pinkse, J., and M. E. Slade. Contracting in Space: an Application of Spatial Statistics to the Discrete Choice Model. *Journal of Econometrics* 85, no. 1(1998): 125-154.