

Regression analysis of spatial data

James P. LeSage

University of Toledo

Abstract. Practitioners of regional science often are engaged in statistical analysis of regional data samples collected with reference to points in space. Examples are cross-sectional observations on county-level income, employment or payroll, cross-sectional observations from a group of neighboring states in a region, and firm-level employment or payroll where we know the firm address or an approximate location based on a postal code. Ignoring the spatial configuration of sample observations in regression analysis has been found to produce residuals that vary systematically over space, a phenomenon known as *spatial autocorrelation*. This paper illustrates how to incorporate spatial information in regression relationships that exhibit spatial autocorrelation. I argue that a simple contiguity matrix provides a unified approach that works with cross-sectional continuous linear relationships, as well as with binary and censored dependent variable problems and autoregressive time series relationships.

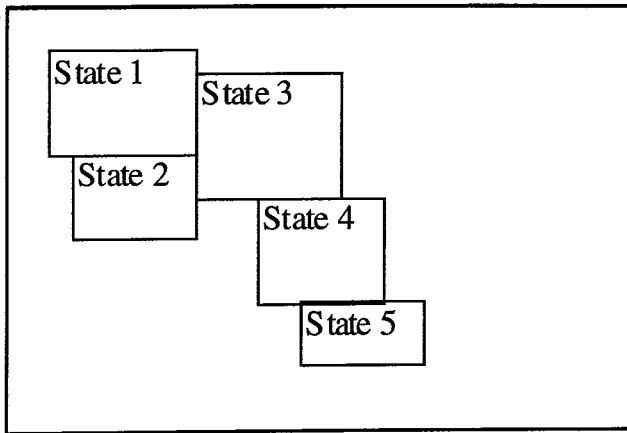
1. Introduction

Practitioners of regional science often are engaged in statistical analysis of regional data samples collected with reference to points in space. Examples are cross-sectional observations on county-level income, employment or payroll, cross-sectional observations from a group of neighboring states in a region, and firm-level employment or payroll where we know the firm address or an approximate location based on a postal code. Ordinary least squares regression methods to analyze this type of sample data have been found to produce residuals that vary systematically over space, a phenomenon known as *spatial autocorrelation*. Anselin (1988) provides a presentation of these ideas aimed at those with econometric expertise, whereas here the presentation is aimed at a larger audience of regional science practitioners and requires only a basic knowledge of ordinary least squares regression methods.

This paper illustrates how ordinary regression estimates of relationships that exhibit spatial autocorrelation may be biased and inconsistent. It also demonstrates estimation methods that correct for these problems in regressions involving continuous, binary, and censored dependent variables and in autoregressive time series.

We begin by introducing the concept of a first order spatial contiguity matrix that provides a unified approach to incorporating the spatial configuration information about the points in space at which our data observations are gathered. Consider the spatial configuration of the five states shown in Figure 1. Figure 1 shows that

Figure 1. Contiguous relationships



states 2 and 3 are first order contiguous to state 1 (that is, they have borders that touch). Similarly, state 2 is first order contiguous to states 1 and 3. In contrast, state 5 has only one neighbor, state 4.

The first order spatial contiguity matrix shown in equation (1) is a convenient way to summarize the spatial configuration of the five states in Figure 1. Information regarding first order contiguity is recorded as ones for states that are neighbors and zeros for those that are not. In the first row we record a value of 1 in the 2nd and 3rd positions to denote that states 2 and 3 border state 1. Similarly, the second row has values of unity in the first and third positions, indicating that state 2 borders states 1 and 3. Each row records the contiguity relationships for each of the five states. By convention, zeros are placed on the main diagonal of the spatial weight matrix.

$$W = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \quad (1)$$

A standardization of the matrix in equation (1) is employed prior to using the matrix in a regression modeling context. The standardization involves normalizing so that row-sums add to unity. A standardized version of W from equation (1) is shown in equation (2) labeled \bar{W} .

$$\tilde{W} = \begin{pmatrix} 0 & 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0.33 & 0.33 & 0 & 0.33 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \quad (2)$$

We can use the standardized first order contiguity matrix to incorporate the spatial configuration of our data in regression relationships. To see how this works, consider the regression relationship in equation (3) which has been labeled a *spatial error model* (SEM) in the spatial econometrics literature (Anselin 1988).

$$\begin{aligned} y &= X\beta + u \\ u &= \rho \tilde{W}u + \varepsilon \end{aligned} \quad (3)$$

The vector y contains cross-sectional observations for states or areas, the matrix X contains a set of explanatory variables for the regression relationship, the vector ε denotes the typical Gaussian disturbance term, and the vector u represents errors in the regression relation that are related to errors from neighboring states or areas. The parameters to be estimated in the model are β and ρ . To see how the standardized first order contiguity matrix works to relate errors from one state to those from neighboring states, consider our five state example from Figure 1 and the associated contiguity matrix \tilde{W} . If we were to expand the vectors and matrix expressions in the second line of equation (3), we would have:

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{pmatrix} = \rho \begin{pmatrix} (1/2)u_2 + (1/2)u_3 \\ (1/2)u_1 + (1/2)u_3 \\ (1/3)u_1 + (1/3)u_2 + (1/3)u_4 \\ (1/2)u_3 + (1/2)u_5 \\ u_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix} \quad (4)$$

Applying the standardized contiguity matrix to the errors in the vector u results in a regression model that relates the errors made by the basic regression relationship in equation (3) from each state to an average of the errors from all neighboring states. The parameter ρ indicates the strength of the correlation between errors across neighbors and is labeled the *spatial autocorrelation parameter* by analogy to the serial correlation problem in time series regression models. The parameters β in the spatial error model can be estimated consistently using least squares, and maximum likelihood estimates for both β and ρ can be obtained easily.

Another way to use the spatial weight matrix \tilde{W} in regression modeling is shown in equation (5), which is labeled a *spatial autoregressive model* (SAR) in the spatial econometrics literature (Anselin 1988).

$$y = \rho \tilde{W}y + X\beta + \varepsilon \quad (5)$$

This model is analogous to the lagged dependent variable model for time series regressions, with the parameter ρ indicating the extent to which variations in the vector of observations y are explained by the average of neighboring observations values. The averaging occurs in a fashion similar to that shown in equation (4), with a partial illustration shown in equation (6) based on the five state example.

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \rho \begin{pmatrix} 0 & 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0.33 & 0.33 & 0 & 0.33 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} + X\beta + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix} \quad (6)$$

Applying ordinary least squares to estimate the parameters β in the spatial autoregressive model produces biased and inconsistent estimates. The problem can be viewed as one similar to simultaneity in least squares. It arises from the introduction of dependence between neighboring observations in the vector y as an explanatory variable on the right side of the regression.

A third more complicated way to introduce spatial configuration information in regression models represents a combination of the spatial error model and the spatial autoregressive model and is shown in equation (7). In this model two spatial weight matrices are usually employed, so we label one as W_1 and the other W_2 . We do not deal with this model, but merely mention it for the sake of completeness.

$$y = \rho W_1 y + X\beta + u \quad (7)$$

$$u = \lambda W_2 u + \varepsilon$$

2. Some examples of the spatial autoregressive and spatial error models

Estimation of the spatial autoregressive and spatial error models proceeds on the basis of an iterative procedure that maximizes the likelihood. A simple search over the spatial correlation parameter ρ is involved. The following steps are used to estimate the parameters β and ρ for the spatial autoregressive model.

1. Compute least squares estimates of $b_0 = (X'X)^{-1} X'y$

2. Determine $bL = (X'X)^{-1} X'Wy$
3. Determine $e_0 = y - Xb_0$ and $eL = Wy - XbL$
4. Search values of ρ to maximize $\log |I_n - \rho W| - (n/2) \log[(e_0 - \rho eL)'(e_0 - \rho eL)]$

A similar iterative procedure can be used to estimate the spatial error model. One can avoid the optimization step by using a grid search over the feasible range of values for ρ between $1/\lambda_1$ and $1/\lambda_2$. Anselin (1988) shows that the parameter ρ falls between $1/\lambda_1$ and $1/\lambda_2$ where $1/\lambda_1$ and $1/\lambda_2$ are the minimum and maximum eigenvalues of the standardized spatial weight matrix W .

Next we turn to an illustration of the problems one might encounter when using least squares estimates to draw inferences regarding cross-sectional observations collected at points in space. A sample of average housing values for each of 88 counties in Ohio will be related to population per square mile, the number of households, and unemployment rates in each county. This regression relationship can be written as:

$$\text{House}_i = \alpha + \beta \text{POP}_i + \gamma \text{Households}_i + \delta \text{Unemploy}_i + \varepsilon_i \quad (8)$$

The motivation for the regression relationship is that population and household density as well as unemployment rates work to determine the house values in each county. The advent of suburban sprawl and the notion of urban rent gradients suggest housing values in contiguous counties should be related. The least squares relationship in equation (8) ignores the spatial contiguity information whereas both the spatial autoregressive and spatial error models would allow for this type of variation in the model. Table 1 shows the OLS, spatial autoregressive, and spatial error model estimates for the relationship in equation (8).

Table 1. Comparison of estimates

	Variable	Estimate	t-statistic	t-probability
OLS	constant	87236.9532	19.822	0.000
OLS	popsqm	25.3791	2.206	0.030
OLS	# households	-0.1146	-1.884	0.062
OLS	unemployment	-4367.4344	-8.401	0.000
	R-squared	0.5552		
SAR	constant	56883.6180	6.311	0.000
SAR	popsqm	13.6897	1.432	0.155
SAR	# households	-0.0711	-1.414	0.161
SAR	unemployment	-3473.4743	-6.950	0.000
SAR	ρ	0.4459	3.834	0.000
	R-squared	0.682		
SEM	constant	86353.0640	18.889	0.000
SEM	popsqm	9.4567	0.895	0.373
SEM	# households	-0.0558	-1.032	0.304
SEM	unemployment	-4067.9448	-7.805	0.000
SEM	ρ	0.526	5.467	0.000
	R-squared	0.508		

In Table 1 the spatial autocorrelation coefficient estimates for the spatial autoregressive and spatial error models are statistically significant, indicating the presence of spatial autocorrelation in the regression relationship. Least squares ignore this type of variation and produces estimates that lead us to conclude all three explanatory variables are significant in explaining housing values across the 88 county sample. In contrast, the spatial autoregressive and spatial error models lead us to conclude that the population density (popsqm) and number of households variables are not statistically significant. The OLS estimates are biased and inconsistent, so the inference of significance from OLS we would draw is likely to be incorrect.

Taking the spatial variation into account improves the fit of the model, raising the R-squared statistic for the spatial autoregressive model and modeling the systematic variation in the residuals in the case of the spatial error model.

Finally, the magnitudes of the OLS parameter estimates indicate that house values are more sensitive to the population density, number of households, and the unemployment rate variables than are the spatial autoregressive and spatial error models. For example, the OLS estimates imply that a one percentage point increase in the unemployment rate leads to a decrease of \$4,367 in house values whereas the spatial autoregressive model places this at \$3,473, and the spatial error model estimates it at \$4,067. Similarly, the OLS estimates for number of households and population density are twice the magnitude of those from the spatial autoregressive and spatial error models.

This illustration shows that ignoring information regarding the spatial configuration of the data observations will produce different inferences that may lead to an inappropriate model specification. Anselin and Griffith (1988) show that traditional specification tests are plagued by the presence of spatial autocorrelation, so we should not rely on these tests in the presence of significant spatial autocorrelation.

3. Other approaches to modeling spatial configuration

Next we discuss a number of other approaches that have been proposed for taking into account spatial location when establishing statistical relationships. Section 3.1 discusses spatial expansion, a method that allows for parameter variation over space, and Section 3.2 points out that binary and censored dependent variable problems involving spatial autoregressive and spatial error model structures also can be estimated. Finally, Section 3.3 demonstrates that the spatial contiguity weight matrix can be used as a Bayesian prior in vector autoregressive time series models.

3.1 Spatial expansion

Another simple approach to incorporating the spatial configuration of the data in regression models is spatial expansion where we use the latitude-longitude coordinates to create multiplicative interactive variables in the model. This model allows variables to have a different impact on the dependent variable based on the point in space from which the sample data are collected.

Casetti (1972, 1992) and Casetti and Jones (1987) propose the spatial expansion model for incorporating spatial influences in regression relationships. To illustrate their approach, let y denote an $n \times 1$ dependent variable vector associated with spatial observations and x_i represent a set of corresponding $k \times 1$ explanatory variable vectors, as shown in equation (9).

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1' & 0 & \cdots & 0 \\ 0 & x_2' & & \\ \vdots & & \ddots & \\ 0 & & & x_n' \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} \tag{9}$$

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = \begin{pmatrix} r_1 \otimes I_k & s_1 \otimes I_k & 0 & \cdots \\ 0 & \ddots & \ddots & \\ \vdots & & r_n \otimes I_k & s_n \otimes I_k \end{pmatrix} \begin{pmatrix} I_k & 0 \\ 0 & I_k \\ \vdots & \\ 0 & I_k \end{pmatrix} \begin{pmatrix} b_r \\ b_s \end{pmatrix}$$

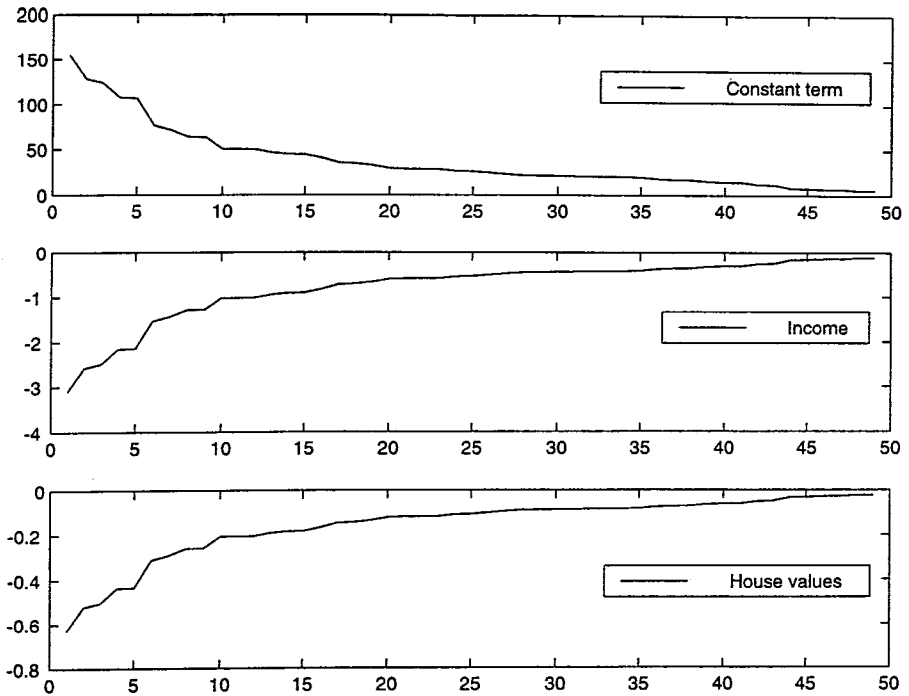
where:

- r_1 and s_1 = The coordinates of observation 1 in the spatial plane (perhaps latitude and longitude centroid coordinates);
- b_r and b_s = Parameters to be estimated; and
- \otimes = The Kronecker product.

This model can be estimated using ordinary least squares to produce a set of estimates for every observation point in space. This allows us to examine the impact of variables on the dependent variable across space. As an example, consider a regression model involving crime as the dependent variable, a constant, household income, and housing values as explanatory variables. The sample represents 49 Columbus neighborhoods (presented in Anselin 1988) that we use to illustrate the spatial expansion model.

Figure 2 shows the variation in the spatial expansion parameter estimates as a function of the distance from a central city neighborhood. The 49 neighborhoods represented on the horizontal axis are ordered by distance from the central city. The vertical axis in the plots shows the magnitude of the spatial expansion estimates—the graph depicts the estimated impact of each explanatory variable as we move from the central city.

We would expect that housing values and household income may have a greater impact on crime in central city neighborhoods and less impact in suburban areas. This is precisely what is shown by the parameter variation as a function of distance from the central city. Least squares estimation of this relationship would assign constant parameter values to all distances with respect to the central city.

Figure 2. Parameters as a function of distance

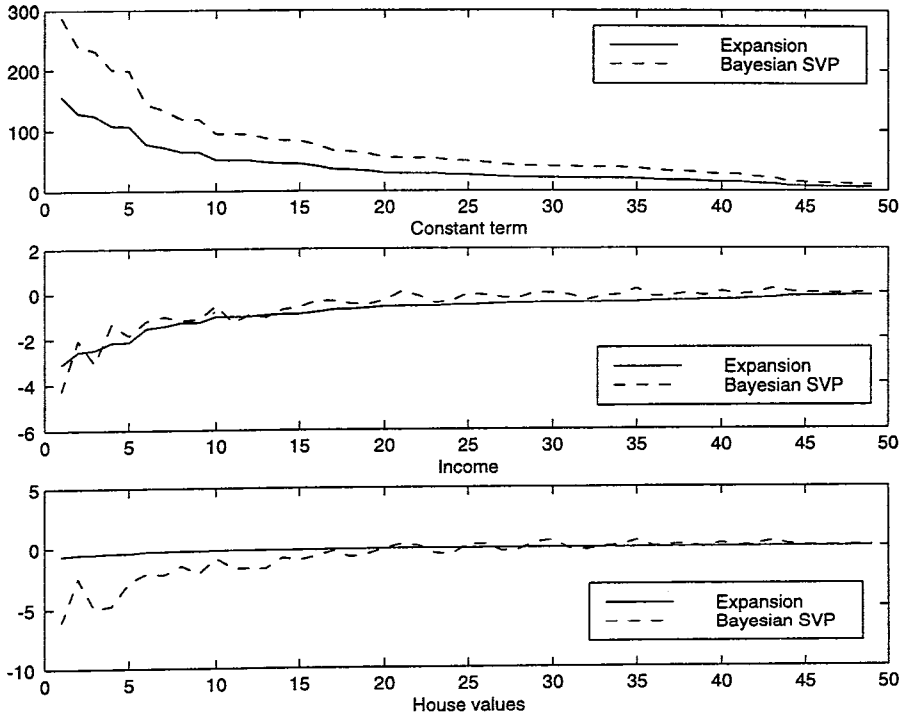
LeSage and Stevens (1994) and LeSage and Muraco (1994) provide a Bayesian extension of the spatial expansion model they label a *Bayesian spatial varying parameter model*. This extension accommodates nonconstant variance and outliers that often are found in spatial data sets. The Bayesian spatial varying parameter estimates are plotted beside the spatial expansion estimates in Figure 3.

The figure shows a greater sensitivity to distance in the Bayesian estimates that is more robust with respect to outlying observations than the least squares based expansion estimates. Nonetheless, the two sets of estimates provide a similar inference that the impact of housing values and household income on crime varies with regard to distance from the central city.

3.2 Limited dependent variable models

With spatial data samples observations often are truncated or represent discrete choice variables from surveys or classifications. McMillen (1992) proposes methods for estimating spatial autoregressive and spatial error probit models containing spatial heteroscedasticity that rely on the EM algorithm. LeSage (1997b) proposes a

Figure 3. Expansion versus Bayesian estimates



Bayesian variant of McMillen’s approach that draws on work by Chib (1992), Albert and Chib (1993), and LeSage (1997a). In addition to extending the class of models beyond probit to include tobit and a family of models that contain probit and logit as special cases, LeSage’s estimation methodology overcomes several drawbacks associated with McMillen’s EM approach.

As an illustration, consider the sample of 49 Columbus neighborhoods used in the previous section. We can artificially generate a binary data sample by setting the dependent variable on crime such that $y_i = 1$ for values of crime greater than 40 and $y_i = 0$ for values of crime less than or equal to 40. The explanatory variables in the model, as in the previous section, are neighborhood housing values and neighborhood income.

Table 2 presents a comparison of EM and Bayesian estimates based on the generated binary data sample. A disadvantage of the EM approach is that no estimates of precision are provided for the spatial autoregressive parameter ρ . The Bayesian estimates are shown in columns labeled *Gibbs*, as this is the method used to produce the estimated parameters. Estimates are shown for two alternative values of a hyper-

Table 2. EM versus Bayesian estimates

	McMillen SAR	Gibbs SAR $q = 7$	Gibbs SAR $q = 5$	McMillen SEM	Gibbs SEM $q = 5$
Constant	2.587	2.768	2.689	2.227	1.896
Standard error	0.888	0.966	0.886	0.715	1.004
t-value	2.912	2.864	3.032	3.115	1.887
Income	-0.128	-0.147	-0.144	-0.123	-0.099
Standard error	0.060	0.071	0.068	0.051	0.070
t-value	-2.137	-2.070	-2.103	-2.422	-2.404
Housing	-0.029	-0.028	-0.025	-0.025	-0.025
Standard error	0.018	0.015	0.014	0.015	0.016
t-value	-1.617	-1.816	-1.793	-1.586	-1.522
ρ	0.429	0.256	0.255	0.279	0.395
Standard error	---	0.270	0.259	---	0.270
t-value	---	0.946	0.984	---	1.463
Median	0.264	0.267		0.428	
Mode		0.353	0.341		0.539

parameter, q , that determines sensitivity to nonconstant variance and outliers that are accommodated by the Bayesian variant of the model.

The same spatial autoregressive and spatial error models introduced for regression relationships can be used when modeling discrete or truncated variable relationships based on sample data collected with reference to points in space.

3.3 A spatial prior for time series models

A popular approach to forecasting regional time series has been to rely on the Bayesian vector autoregressive models introduced by Doan, Litterman, and Sims (1984) for macroeconomic time series forecasting. A principle behind much of the modeling done in regional science is that location in space matters. Despite this, several regional forecasting models have been constructed using vector autoregressions that rely on the Minnesota prior introduced in Doan, Litterman, and Sims (1984). Pan and LeSage (1995) modify the Minnesota prior to better accommodate regional time series models. They draw upon the first order contiguity matrix W to customize the variance of the Minnesota prior, but leave the prior means unaltered.

Krivelyova (1997) introduces a spatial prior that is more appropriate for modeling regional time series variables as it specifies both prior means and variances based on the concept of first order spatial contiguity. To motivate the spatial prior means, consider the spatial autoregressive model above used to model cross-sectional data. Krivelyova uses this model to provide an intuitive motivation for a prior means specification that averages over neighboring time series variables from the previous time period.

She suggests a prior mean for the VAR model coefficients on variables associated with first own-lag spatially contiguous variables equal to $1/c$, where c is the number of spatial entities contiguous to each variable in the model. In the example shown in Figure 1, the prior means for the first own-lag of the contiguous variables

y_{2t-1} and y_{3t-1} in the y_{1t} equation of the VAR would be equal to $1/2$. The prior means for the noncontiguous variables y_{4t-1} and y_{5t-1} as well as the prior mean for the lagged dependent variable y_{1t-1} in this equation would be zero. Her spatial contiguity prior differs from the Minnesota prior in that it downweights the lagged dependent variable y_{1t-1} using a zero prior mean, discounting the autoregressive influence of past values of this variable in the equation describing variation in y_{1t} .

For macroeconomic time series, the Minnesota prior emphasizes a random walk with drift model using prior means centered on a model: $y_{it} = \alpha + y_{it-1}$. The intercept term reflects the drift in the random walk model and is estimated using a diffuse prior. Krivelyova’s spatial prior is centered on a random walk model that averages over contiguous entities and allows for drift:

$$y_{it} = \alpha + \sum_{j=1}^{c_i} (1/c) y_{jt-1} \quad j \in C_i \tag{10}$$

where:

C_i = The set of entities contiguous to entity i .

For example, C_1 equals states 2 and 3 in Figure 1, and c_1 denotes the number of contiguous areas, equal to 2. In addition to using spatial contiguity to specify the prior means in the model, a prior variance is constructed that also draws on spatial contiguity relationships among the variables in the time series model.

Both Pan and LeSage (1995) and Krivelyova (1997) point to improved forecasting performance as a result of incorporating spatial contiguity relationships as prior information in these vector autoregressive forecasting models.

4. Conclusion

Spatial relationships can be represented easily using first order contiguity matrices. For those engaged in modeling sample data collected with reference to points in space, the contiguity matrix can be used to extend regression relationships such that the spatial configuration of the observations is incorporated in the model.

We provide brief illustrations for the case of regression models, including binary dependent variable models. This approach also extends to tobit models where the dependent variables are truncated or censored at some point. Another approach to incorporating spatial information in regression models is the spatial expansion approach that allows variation in the parameters of the relationship over space or distance from some point in space.

Finally, the spatial contiguity matrix has been used to form a spatial prior to replace reliance on the Minnesota prior that often is used in regional time series forecasting. The spatial prior seems more appropriate than the Minnesota because it places the emphasis in regional vector autoregressive modeling where it should be: on what has been happening recently in neighboring areas.

Practitioners engaged in statistical work with regional data samples should attempt to incorporate information on the spatial configuration of the sample data in

their work. Ignoring this information in our models may produce inferences that are qualitatively and quantitatively different from models that contain these relations.

References

- Albert, James H., and Siddhartha Chib, "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, no. 422 (1993), pp. 669-679.
- Anselin, Luc, *Spatial Econometrics: Methods and Models* (Dordrecht: Kluwer Academic Publishers, 1988).
- Anselin, Luc, and D.A. Griffith, "Do Spatial Effects Really Matter in Regression Analysis?" *Papers of the Regional Science Association*, 65 (1988), pp. 11-34.
- Casetti, Emilio, "Generating Models by the Expansion Method: Applications to Geographic Research," *Geographical Analysis*, 4 (1972), pp. 81-91.
- Casetti, Emilio, "Bayesian Regression and the Expansion Method," *Geographical Analysis*, 24 (1992), pp. 58-74.
- Casetti, Emilio, and J.P. Jones, "Spatial Applications of the Expansion Method Paradigm," in C. Dufournaud and D. Dudycha (eds.), *Quantitative Analysis in Geography* (Ontario, Canada: University of Waterloo, 1987), pp. 121-136.
- Chib, Siddhartha, "Bayes Inference in the Tobit Censored Regression Model," *Journal of Econometrics*, 51 (1992), pp. 79-99.
- Doan, T.R., B. Litterman, and C. Sims, "Forecasting and Conditional Projections Using Realistic Prior Distributions," *Econometric Review*, 3 (1984), pp. 1-100.
- Krivelyova, Anna, "A Spatial Prior for Bayesian Vector Autoregressive Models," paper presented at the Mid-Continent Regional Science Association meetings (June 1997).
- LeSage, James P., "Bayesian Estimation of Spatial Autoregressive Models," *International Regional Science Review*, 20, nos. 1 and 2 (1997a), pp. 113-129.
- LeSage, James P., "Bayesian Estimation of Spatial Probit/Tobit models," University of Toledo working paper (February 1997b).
- LeSage, James P., and William A. Muraco, "Spatial Modeling of Gypsy Moth Movements," University of Toledo working paper (September 1994).
- LeSage, James P., and John J. Stevens, "A Bayesian Approach to Spatial Expansion," University of Toledo working paper (September 1994).
- McMillen, Daniel P., "Probit With Spatial Autocorrelation," *Journal of Regional Science*, 32, no. 3 (1992), pp. 335-348.
- Pan, Zheng, and James P. LeSage, "Using Spatial Contiguity as Prior Information in Vector Autoregressive Models," *Economic Letters*, 47 (February 1995), pp. 137-142.