W-BASED VS LATENT VARIABLES SPATIAL AUTOREGRESSIVE MODELS: EVIDENCE FROM MONTE CARLO SIMULATIONS

1. Introduction

When it comes to applying econometric models to analyze georeferenced data, researchers are well aware that ignoring spatial dependencies would lead to inefficient and biased estimates. A lot of theoretical and simulation studies have been undertaken seeking efficient and consistent estimators for spatial data analysis and one major issue concerns the kind of decisions to make regarding the specification of the structure of spatial dependence including the type of spatial weights to be incorporated into the model. Therefore, our study focuses on the evaluation of two approaches that explicitly model spatial autocorrelation: classical W-based regression and the recently proposed latent variables spatial autoregressive model. [1]

In the spatial econometrics literature the W-based spatial regressions have been predominant and most commonly used. The approach is based on one or more spatial structural matrices, usually denoted W, that accounts for spatial dependence and spill-over effects. The selection of spatial weight matrices is a crucial feature of spatial models because they impose a priori a structure of spatial dependence on the models and affect estimates (Bhattacharjee and Jensen-Butler, 2006; Anselin, 2002 and Fingleton, 2003) and substantive interpretation of the research findings (Hepple, 1995).

Autoregressive models can include several types of spatial weight matrices. Most common are the contiguity-based matrices. Two regions are said to be first-order contiguous if they share a common border. Higher-order contiguity is defined in a similar way. The contiguity matrix is based on different forms (whether the units of observation share common boundaries or vertices) or orders of contiguity. The other type of weight matrix is

[1] Another approach to dealing with spatial autocorrelation is spatial filtering. It comes down to the removal of spatial dependence in a spatially autocorrelated variable by partioning it into a filtered nonspatial variable and a residual spatial variable such that conventional regression techniques can be applied to the filtered data (see amongst others Getis and Griffith (2002) and Tiefelsdorf and Griffith (2007). Spatial filtering is not considered in this paper

distance based and calculated using either an algorithm with distance, such as inverse distance or inverse distance squared, or a fixed distance band. Much progress has been made with respect to the comparison of spatial weights matrices (Hepple, 1995), the construction of spatial weights matrices (Getis and Aldstadt, 2003; Aldstadt and Getis (2006) and estimating spatial weights matrices that are consistent with an observed pattern of spatial dependence rather than assuming a priori the nature of spatial interaction. In spite of all these developments the most common procedure is still to assume a priori first order contiguity, as expressed by a matrix W with diagonal elements equal to zero and off-diagonal elements equal to one if two regions are first-order contiguous and zero elsewhere.

The alternative, a latent variables based approach was introduced by Folmer and Oud (2008). It proceeds on the basis of a structural equation model (SEM). SEM allows simultaneous handling of observed and latent variables within one model framework. Latent variables refer to those phenomena that are supposed to exist but cannot be observed directly. One example is socio-economic status. It carries the concept of an individual's standing in the society which cannot be observed or measured directly. However, it can be measured via observable indicators like educational attainment, income, occupational status, etc. The structure of a SEM includes two parts: (i) the structural model presenting the causal relationships between latent variables and (ii) the measurement model showing the relationships between the latent variables and their observable indicators (Oud and Folmer, 2008).

The latent variables approach replaces the spatially lagged variables in the structural model (which is the analogue of the spatial dependence model in the W-based approach) by latent variables and models the relationship between latent spatially lagged variables and its observed indicators in the measurement model. Since one latent variable can be measured by several indicators, this approach allows for the straightforward inclusion of spatial dependence in the model. It is also capable of testing distance decay in a straightforward fashion so as to identify the spatial units that evoke a distance effect and those that do not. Moreover, it is possible to include various different types of contiguity as it allows measurement of spatial dependence by several sets of indicators. In that case more than one latent spatial dependence variable enters the structural model with corresponding indicators in the measurement model.

Folmer and Oud (2008) shows that the latent variables approach can produce the same estimates as obtained by the W-based approach but also that it is more general than the W-based approach. However, further comparison is needed to draw up the pros and cons of each approach. To gain more insight into the quality of the estimators of the regression coefficients including the spatial autocorrelation coefficient in both approaches, we carry out a number of Monte Carlo simulations. The comparison of the two approaches is done by considering Anselin's (1988) Columbus, Ohio, crime dataset. The performance of the approaches will be analyzed in terms of bias and mean squared error for various values of spatial autocorrelation.

The rest of the paper is organized as follows: Section 2 explains the W-based spatial regression approach and the latent variables approach and specifies their model structures. A detailed description of the experimental design is given in section 3. In section 4 we report the simulation results and section 5 concludes the paper.

2. Model Specifications

2.1 The W-based Autoregressive Model

We only consider the spatial lag model which assumes that the dependent variable is a function of exogenous variables and the dependent variable at other locations. We just refer to this model as the 'W-based spatial autoregressive model'. The specification for the model reads:

$$y = \rho W y + X\beta + \varepsilon \tag{1}$$

$$\varepsilon \sim \; N(0, \sigma^2 I_n) \tag{2}$$

where y is an $n \times 1$ vector of observations on the dependent variable, X is an $n \times k$ data matrix of explanatory variables with the associated coefficient vector β, ε is an

$n \times 1$ vector of error terms. W is the row-standardized $n \times n$ spatial weight (contiguity) matrix, ρ is the spatial lag parameter.

Anselin (1988) presents a maximum likelihood method for estimating the parameters of this model that he labels a 'mixed regressive spatial autoregressive model' as it combines the standard regression model with a spatially lagged dependent variable (analogue to the lagged dependent variable model in time-series analysis). Maximum likelihood estimation of this model is based on a concentrated likelihood function in the sense that the maximization can be reduced to an expression in one parameter, conditional upon the values of the other parameters. The corresponding log-likelihood fuction is of the form:

$$L = -\frac{N}{2}\ln \pi - \frac{N}{2}\ln \sigma^2 + \ln|A| - \frac{1}{2\sigma^2}(Ay - X\beta)'(Ay - X\beta) \quad (3)$$

with $A = I - \rho W$. (4)

A few regressions are carried out along with a univariate parameter optimization of the simplified likelihood function over values of the autoregressive parameter ρ. The steps are described in Anselin (1988):

1. Perform OLS of X on y: $y = X\beta_O + \varepsilon_O$

2. Perform OLS of X on Wy: $Wy = X\beta_L + \varepsilon_L$

3. Compute residuals: $e_O = y - X\hat{\beta}_O$, $e_L = Wy - X\hat{\beta}_L$

4. Given e_O and e_L, find ρ that maximizes the concentrated likelihood function:

$L_C = C - \frac{N}{2}\ln\left[\frac{1}{N}\left(e_O - \rho e_L\right)'\left(e_O - \rho e_L\right)\right] + \ln|I - \rho W|$, which is a result of the substitution of estimated β and σ^2 into the log-likelihood, with C as the constant

5. Given $\hat{\rho}$ that maximizes L_C, compute:

$$\hat{\beta} = \hat{\beta}_O - \rho\hat{\beta}_L \ , \ \hat{\sigma}^2 = \frac{1}{N}\left(e_O - \rho e_L\right)'\left(e_O - \rho e_L\right).$$

The likelihood function discussed above contains a Jacobian term $\ln|A|$. Since the matrix A is of dimension equal to the number of observations, its presence in the function to be optimized makes the numerical analysis considerably complex. The determinant and its derivative need to be evaluated at each iteration for a new value of the spatial parameter ρ. Ord (1977) has derived a simplification of determinants such as $|A|$ in terms of its eigenvalues, more specifically:

$$\ln|I - \rho W| = \ln \Pi_i (1 - \rho w_i) = \Sigma_i \ln(1 - \rho w_i) \qquad (5)$$

where the w_i are the eigenvalues of W.

2.2 The Latent Variables Autoregressive Model

A SEM in general form consists of three basic equations:

$$\mathrm{y} = \Lambda_{\mathrm{y}}\eta + \varepsilon \quad \text{with} \quad \mathrm{cov}(\varepsilon) = \Theta_{\varepsilon}, \qquad (6)$$

$$\mathrm{x} = \Lambda_{x}\xi + \delta \quad \text{with} \quad \mathrm{cov}(\delta) = \Theta_{\delta}, \qquad (7)$$

$$\eta = \mathrm{B}\eta + \Gamma\xi + \zeta \text{ with } \mathrm{cov}(\xi) = \Phi, \ \mathrm{cov}(\zeta) = \Psi. \qquad (8)$$

In the measurement models (1) and (2), the vectors y and x are observed endogenous and exogenous variables; the vectors η and ξ contain latent endogenous and exogenous variables; the matrices Λ_{y} and Λ_{x} specify the loadings of the observed variables (indicators) on the vectors of latent variables y and x, and Θ_{ε} and Θ_{δ} are the measurement error covariance matrices. Directly observed variables can be conveniently handled in the structural model by specifying an identity relationship between a given observed variable and the corresponding latent variable in the measurement model.

In the structural model (3) B specifies the structural relationships among the latent endogenous variables mutually and Γ contains the impacts of the exogenous latent variables on the endogenous. Φ is the covariance matrix of latent exogenous variables and Ψ the covariance matrices of the errors in the structural model. The measurement errors in ε and δ are assumed to be uncorrelated with the latent variables in η and ξ as well as with the structural errors in ζ .

In a SEM modelling framework parameter estimation is done by minimizing the distance between the theoretical covariance matrix (based on hypotheses relating to the model structure as specified in the parameter matrices $\Lambda_x, \Lambda_y, \Theta_\varepsilon, \Theta_\delta$, B, Γ, Φ and Ψ) and the observed covariance matrix. Several estimators for SEM have been developed including instrumental variables (IV), two-stage least squares (TSLS), unweighted least squares (ULS), generalized least squares (GLS), fully weighted (WLS) and diagonally weighted least squares (DWLS), and maximum likelihood (ML). ML method is the most commonly used method in estimating SEM models and the default in the statistical packages Mx and LISREL. Below we apply the ML estimation. It maximizes the log-likelihood function:

$$l(\theta \mid Y) = -\frac{N}{2}\ln|\Sigma| - \frac{N}{2}tr\left(S\Sigma^{-1}\right) - \frac{pN}{2}\ln 2\pi \qquad (9)$$

where Σ is the theoretical variance covariance matrix in terms of the free and constrained elements in the 8 parameter matrices:

$$\Sigma = \begin{bmatrix} \Lambda_y(I-B)(\Gamma\Phi\Gamma' + \Psi)(I-B)'\Lambda_y + \Theta_\varepsilon & \Lambda_y(I-B)\Gamma\Phi\Lambda_x' \\ \Lambda_x\Phi\Gamma'(I-B)'\Lambda_y' & \Lambda_x\Phi\Lambda_x' + \Theta_\delta \end{bmatrix} \qquad (10)$$

S is the observed variance covariance matrix for given data Y.

The ML-estimator $\hat{\theta} = \arg\max l(\theta \mid Y)$ chooses that value of θ which maximizes $l(\theta \mid Y)$. Minimizing the fit function:

$$F_{ML} = \ln|\Sigma| + tr\left(S\Sigma^{-1}\right) - \ln|S| - p \qquad (11)$$

gives the same results as maximizing the above likelihood function (Oud and Folmer, 2008). The LISREL software package also contains a variety of test and model evaluation statistics and gives hints about identification problems (see Joreskog and Sorbom, 1996 for details).

On the basis of SEM, we propose a model structure to represent the spatial lag model. We specify the lag model as a SEM and estimate it using the software package Mx.

The model is referred to as 'latent variables spatial autoregressive model' in this paper. We apply the concept of a latent variable to the spatially lagged dependent variable such that in the structural model Wy is replaced by a latent variable while in the measurement model the latent spatially lagged dependent variable is related to its indicators. It replaces the spatially lagged variable Wy in the W-based spatial autoregressive model by a latent variable η:

$$y = \rho\eta + \gamma' x + \zeta \tag{12}$$

and is completed by a measurement equation:

$$y = \Lambda\eta + \varepsilon \tag{13}$$

with

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \end{bmatrix}, \quad \Theta = \begin{bmatrix} \sigma_{\varepsilon_1}^2 & 0 & 0 & 0 \\ 0 & \sigma_{\varepsilon_2}^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_{\varepsilon_m}^2 \end{bmatrix} \tag{14}$$

(Observe that without an appropriate constraint, the loadings λ_i will not be identified. For that reason, one often chooses λ_1 = 1.)

We assume that the spatially lagged observed variables are chosen on the basis of theoretical or ad hoc considerations. Moreover, as pointed out in the introduction more than one spatial feature can be taken into account. For instance, both the relationships to neighbouring regions as relationships to the core regions could be included in the case of the crime model.

We first turn to the measurement model. This model is constructed by means of selection functions or selection matrices S_i which select relevant observations from the vector of observations as follows. In order to distinguish between the standard definition of a SEM in terms of variables and a SEM defined in units of observations, as in standard spatial econometrics, we denote the latter by tilde (~).

$$\begin{aligned} \tilde{y}_1 &= S_1\tilde{y}, \\ \tilde{y}_2 &= S_2\tilde{y}, \\ &\vdots \\ \tilde{y}_m &= S_m\tilde{y}. \end{aligned} \tag{15}$$

That is, S_1 selects the values for the first indicator $\tilde{y}_1$, S_2 for the second indicator $\tilde{y}_2$, etc. For example, S_1 may be defined as the selector of the observations on, say crime, in the nearest contiguous neighbour, S_2 as the selector of the observations on crime in the next nearest contiguous neighbour, S_3 as the spillover of crime from the core, measured, for example, as crime in the core divided by a region's distance from the core, etc. Thus, for the measurement model we obtain:

$$\begin{aligned} \tilde{y}_1 &= S_1\tilde{y} = \lambda_1\tilde{\eta} + \tilde{\varepsilon}_1, \\ \tilde{y}_2 &= S_2\tilde{y} = \lambda_2\tilde{\eta} + \tilde{\varepsilon}_2, \\ &\vdots \\ \tilde{y}_m &= S_m\tilde{y} = \lambda_m\tilde{\eta} + \tilde{\varepsilon}_m. \end{aligned} \tag{16}$$

From the above one observes that spatial dependence is captured by two kinds of parameters, ρ and λ_i, whereas in the standard lag model only the "average" effect ρy_w shows up. This means that a much richer representation and testing of the spatial structure can be obtained than by way of standard spatial econometric approaches. For instance, it allows for determining those units that evoke a distance effect and those that do not, as suggested by Getis and Aldstadt (2004) by testing the significance of the λ_i coefficients.

The standard SEM log-likelihood function needs correction so as to account for the presence of the spatially lagged dependent variable among the explanatory variables. From Folmer and Oud (2008), for

$$A = I - \frac{\rho}{m\lambda_1}S_1 - \frac{\rho}{m\lambda_2}S_2 - \cdots - \frac{\rho}{m\lambda_m}S_m \tag{17}$$

we need to add $\ln|A|$ to the log-likelihood function.

3. Experimental Design

To investigate the performances of the two modelling approaches specified in section 2, we conduct a number of Monte Carlo simulations. Both approaches are applied to Anselin's (1988) Columbus, Ohio, crime dataset. The first step is the design of a data-generating procedure.

We first explain the data generating procedure for the W-based spatial lag model. For this purpose we rewrite equation (1) according to the number of variables and observations involved in Anselin's crime model:

$$y = \rho Wy + \beta_0 + x_1\beta_1 + x_2\beta_2 + \varepsilon \tag{18}$$

Thus the expression of y is:

$$y = (I - \rho W)^{-1}(\beta_0 + x_1\beta_1 + x_2\beta_2 + \varepsilon) \tag{19}$$

The steps for generating samples are:

1. Fix values for the vector $\beta(\beta_0, \beta_1, \beta_2)$ as: $\beta_0 = 45, \beta_1 = -1.0, \beta_2 = -0.3$,

   and vary ρ over the interval [0, 1).
2. Generate samples of error term ε by randomly drawing from a uniform (0, 10) distribution.
3. Compute y according to (19).

Next the spatial lag model is estimated for each sample which gives: ρ, $\beta_0, \beta_1, \beta_2$ and σ^2, on the basis of which we compute bias and mean squared error (MSE). The procedure is similar for the latent variables approach. Specifically, the Wy in standard spatial lag model is replaced by a latent variable while in the measurement model the latent spatially lagged dependent variable is related to its indicators. Here we take the observables for the three contiguous nearest neighbouring regions. Next we estimate the latent spatial lag model for each of the same samples generated from steps 1~3 in the simulation procedure of the standard spatial lag model, which gives ρ (amongst others), $\beta_0, \beta_1, \beta_2$ and σ^2. Then the bias and MSE of the estimators are calculated.

The number of replications is set to 1000 and the main dimensions over which the variation in power of the approaches and the distribution of the pre-test estimators are investigated are the changing values of the spatial lag parameter ρ . We examine the parameter space from [0.0, 0.9] using increments of 0.1. The experiment consists of comparisons in terms of model estimators and their bias and MSE of the two approaches. The results are displayed in tables and graphs in the next section.

4. Simulation Results

Table 1. Mean of estimation results on the Columbus, Ohio, crime dataset by W-based and latent variables spatial autoregressive models

	ρ		β_1		β_2		β_0		σ^2	
Parameter ρ	W-based	Latent	W-based	Latent	W-based	Latent	W-based	Latent	W-based	Latent
0.00	-0.044	-0.009	-1.018	-0.964	-0.297	-0.290	45.907	47.538	91.846	88.168
	(0.169)	(0.243)	(0.305)	(0.424)	(0.093)	(0.131)	(6.082)	(88.589)	(19.345)	(18.718)
0.10	0.055	0.057	-1.021	-1.011	-0.298	-0.295	46.087	45.492	92.013	89.082
	(0.163)	(0.245)	(0.308)	(0.382)	(0.093)	(0.110)	(6.334)	(37.136)	(19.376)	(22.816)
0.20	0.154	0.138	-1.024	-1.024	-0.298	-0.292	46.288	44.078	92.208	90.208
	(0.155)	(0.230)	(0.310)	(0.381)	(0.094)	(0.116)	(6.607)	(11.581)	(19.431)	(19.228)
0.30	0.254	0.245	-1.028	-1.051	-0.298	-0.293	46.516	43.343	92.431	93.163
	(0.146)	(0.254)	(0.313)	(0.394)	(0.094)	(0.103)	(6.908)	(12.441)	(19.510)	(20.453)
0.40	0.354	0.372	-1.032	-1.042	-0.299	-0.286	46.782	41.512	92.681	95.079
	(0.135)	(0.204)	(0.316)	(0.345)	(0.094)	(0.100)	(7.244)	(10.259)	(19.612)	(21.635)
0.50	0.455	0.495	-1.037	-1.047	-0.299	-0.284	47.105	39.595	92.964	96.026
	(0.122)	(0.160)	(0.319)	(0.324)	(0.094)	(0.094)	(7.633)	(10.002)	(19.737)	(22.895)
0.60	0.558	0.599	-1.043	-1.035	-0.299	-0.277	47.520	41.595	93.284	96.168
	(0.108)	(0.124)	(0.322)	(0.322)	(0.094)	(0.094)	(8.106)	(9.065)	(19.886)	(25.045)
0.70	0.662	0.692	-1.049	-1.023	-0.300	-0.270	48.044	41.760	93.678	94.983
	(0.093)	(0.103)	(0.326)	(0.323)	(0.094)	(0.094)	(8.874)	(9.501)	(20.043)	(26.803)
0.80	0.768	0.793	-1.055	-0.969	-0.299	-0.258	48.778	40.759	94.258	90.138
	(0.078)	(0.116)	(0.334)	(0.369)	(0.094)	(0.093)	(10.390)	(14.363)	(20.340)	(31.348)
0.90	0.876	1.140	-1.064	0.110	-0.298	-0.210	50.387	-12.582	95.053	2.491
	(0.061)	(0.269)	(0.353)	(1.433)	(0.094)	(0.201)	(14.803)	(58.167)	(20.605)	(203.505)

Table2. Bias of the estimators for ρ, β_1 and β_2, with 10 different values for the spatial lag parameter

	ρ		β_1		β_2	
Parameter ρ	W-based	Latent	W-based	Latent	W-based	Latent
0.00	0.138	0.140	0.244	0.311	0.074	0.096
0.10	0.133	0.159	0.246	0.285	0.074	0.084
0.20	0.127	0.180	0.248	0.286	0.074	0.085
0.30	0.120	0.178	0.250	0.278	0.075	0.079
0.40	0.111	0.151	0.253	0.266	0.075	0.078
0.50	0.101	0.120	0.255	0.259	0.075	0.076
0.60	0.089	0.094	0.258	0.257	0.075	0.077
0.70	0.076	0.076	0.262	0.257	0.075	0.079
0.80	0.062	0.077	0.269	0.290	0.075	0.082
0.90	0.047	0.251	0.283	1.147	0.075	0.115

Table 3. Mean squared error $(\times 100)$ of the estimators for ρ, β_1 and β_2, with 10 different values for the spatial lag parameter

	ρ		β_1		β_2	
Parameter ρ	W-based	Latent	W-based	Latent	W-based	Latent
0.00	3.032	5.913	9.342	18.047	0.868	1.714
0.10	2.847	6.160	9.499	14.613	0.871	1.215
0.20	2.615	5.682	9.671	14.555	0.874	1.362
0.30	2.341	6.727	9.860	15.743	0.877	1.064
0.40	2.032	4.218	10.067	12.040	0.880	1.019
0.50	1.696	2.562	10.294	10.696	0.883	0.916
0.60	1.344	1.545	10.544	10.468	0.886	0.928
0.70	1.013	1.062	10.865	10.503	0.887	0.973
0.80	0.711	1.349	11.418	13.663	0.892	1.042
0.90	0.427	12.988	12.830	328.359	0.889	4.867

5. Conclusions

Reference

Aldstadt, J. and Getis, A. (2006). Using AMOEBA to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis, 38,* 327-343.

Aldstadt, J. and Getis, A. (2004). Constructing the spatial weights matrix using a local statistic. *Geographical Analysis, 36, 2,* 90-104.

Anselin, L. (2002). Under the hood: Issues in the specification and interpretation of spatial regression models. *Agricultural Economics 27, 3,* 247-267.

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

Bhattacharjee, A. and Jensen-Butler, C. (2006). Estimation of the spatial weights matrix in the spatial error model, with an application to diffusion in housing demand. Available: http://www.econ.cam.ac.uk/panel2006/papers/Bhattacharjeepaper23.pdf.

Fingleton, B. (2003). Externalities, economic geography and spatial econometrics: conceptual and modeling developments, *International Regional Science Review, 26, 2,* 197-207.

Folmer, H. and Oud, J. (2008). How to get rid of W? A latent variables approach to modeling spatially lagged variables. *Environment and Planning A, 40,* 2526-2538.

Hepple, L.W. (1995). Bayesian techniques in spatial and network econometrics:
1. Model comparison and posterior odds. *Environment and Planning A, 27,* 447-469.

Hepple, L.W. (1995). Bayesian techniques in spatial and network econometrics:
2. Computational methods and algorithms. *Environment and planning A, 27,* 615-644.

LeSage, J.P. (1999). *The Theory and Practice of Spatial Econometrics*, 1999. Available: http://www.spatial-econometrics.com/html/sbook.pdf.

Neale M.C., Boker S.M., Xie G., Maes H.H. (2003). *Mx: Statistical Modeling.* VCU Box 900126, Richmond, VA 23298: Department of Psychiatry. 6th Edition.