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

An Liu[a], Henk Folmer[b], Han Oud[c]

[a]Department of Spatial Sciences, University of Groningen, P.O. Box 800, NL-9700AV Groningen, The Netherlands, an.liu@rug.nl, corresponding author
[b]Department of Spatial Sciences, University of Groningen, P.O. Box 800, NL-9700AV Groningen, The Netherlands, and Department of Social Sciences, Wageningen University, PO Box 8130, NL-6700 EW Wageningen, The Netherlands, henk.folmer@wur.nl
[c]Behavioural Science Institute, Radboud University Nijmegen, P.O. Box 9104, NL-6500 HE Nijmegen, The Netherlands, j.oud@pwo.ru.nl

KEY WORDS: spatial autoregressive model, structural equation model, latent variable, Monte Carlo simulation, bias, RMSE

ABSTRACT:

The paper evaluates by means of Monte Carlo simulations the estimator of the regression coefficient obtained by the classical W-based spatial autoregressive model and the structural equations model with latent variables (SEM) on the basis of data sets that contain two types of spatial dependence: spillover from (i) a hotspot and (iia) first order queen contiguity neighbors or (iib) inverse distance related neighbors. The classical models are either correctly specified or ignore (i), as is common in practice. SEM takes spatial dependence into account by means of a fixed number of nearest neighbors as well as the dependent variable in the hotspot weighted by inverse distance. The estimation results are analyzed in terms of bias and root mean squared error (RMSE) for different values of the spatial lag parameters, specifications of weights matrices and sample sizes. The simulation results show that compared to the misspecified models SEM frequently has smaller bias and RMSE and even outperforms the correctly specified models in many cases. These trends increase when the spatial lag parameter for spillover increases. The lead of SEM also increases by sample size. Finally, SEM is more stable in terms of both bias and RMSE over various dimensions.

1. INTRODUCTION

The conventional spatial regression model is based on a spatial weights matrix, usually denoted W, that accounts for spatial dependence and spill-over effects among the spatial units of observation. The latent variables approach (denoted SEM below), introduced by Folmer and Oud (2008), replaces the spatially lagged variables in the structural model by latent variables and models the relationship between latent spatially lagged variables and their observed indicators in the measurement model. SEM not only can produce virtually the same estimates as obtained by the classical approach but also is more general.

In order to gain insight into the characteristics of the estimators of the regression coefficients including the spatial autocorrelation coefficient produced by the classical approaches and SEM, Liu et al (2010a) carried out a series of Monte Carlo simulations on the basis of Anselin's (1988) Columbus, Ohio, crime data set which was also used by Folmer and Oud (2008) for illustrative purposes. The latent spatial lag variable in the SEM model was measured by a number of nearest neighbors. Data was generated on the basis of a first order queen contiguity or an inverse distance weights matrix. The main result was that the classical approach (estimated with weights matrix consistent with the data generation matrix) had lowest bias and RMSE in the majority of cases. SEM outperformed the classical approach for some W matrices, however. Particularly, it had the smallest bias in several cases. Liu et al (2010b) examined the performances of the two approaches in the context of spatial dependence due to spillover from hotspots. In that case spatial lag variable was measured by the values of the dependent variable in hotspots weighted by inverse distance. The simulation results indicated that both approaches performed better for smaller values of the spatial lag parameter and larger sample sizes. SEM tended to outperform the classical approach in term of bias but the classical model based on first order contiguity matrix had lowest RMSE in most cases. Furthermore, SEM was most stable in terms of variations in both bias and RMSE. Globally speaking, the performances of both approaches do not differ much.

In this paper we further evaluate the performances of the classical W-based approach and SEM in a more general setting that combines the two different types of spatial dependence considered in the previous simulations. The remainder of the paper is organized as follows. Section 2 briefly specifies the model structures of the classical W-based approach and SEM. A 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

The classical spatial autoregressive model reads:

$$y = \rho W y + X\beta + \varepsilon \quad (1)$$

$$\varepsilon \sim N(0, \sigma^2 I_n) \quad (2)$$

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 associated coefficient vector β, ε is an $n \times 1$ vector of error terms. W is the $n \times n$ spatial weight matrix, with spatial autoregressive or spatial lag parameter ρ. (For further details see amongst others LeSage and Pace, 2009)

A SEM in general form consists of two basic equations:

$$y = \Lambda\eta + \varepsilon \quad \text{with} \quad \text{cov}(\varepsilon) = \Theta \tag{3}$$

$$\eta = \text{B}\eta + \zeta \quad \text{with} \quad \text{cov}(\zeta) = \Psi \tag{4}$$

Equation (3) is the measurement model with *y* the *p*-vector of observed variables or indicators, Λ the matrix of loadings of the observed variables (indicators) on the *k*-vector of latent variables η [1], and Θ is the $p \times p$ measurement error covariance matrix. In the structural model (4), B specifies the structural relationships among the latent variables and Ψ is the $k \times k$ covariance matrix of the errors in the structural model. The measurement errors ε are assumed to be uncorrelated with the latent variables η as well as with the structural errors ζ who are supposed to be uncorrelated with η. (For details on identification, estimation, testing and respecification of structural equation models see Jöreskog and Sörbom, 1996)

The SEM spatial autoregressive approach replaces the spatially lagged variable *Wy* in the W-based equation (1) by a latent variable η in the structural model. In the measurement model η is measured by a set of observed variables. This model structure implies that both spatially lagged variables related to neighboring regions and hotspots can be indicators of the same latent variable. For detailed model specifications see Folmer and Oud (2008).

3. EXPERIMENTAL DESIGN

The dependent variable (*y*) in each spatial unit is affected by the dependent variable in one or more neighboring units as well as by hotspots. For data generation this implies that besides the spatial structure, the hotspot also needs to be known. However, it is not until the samples are generated that we get to know which region is the hotspot (defined by highest value of *y*). To solve this problem we take a step backward and choose the 'potential' hotspot on the basis of the independent variable *x* instead. That is, we designate the hotspot according to the largest value of *x*.[2]

We consider regular lattice structures of dimensions 7×7 (N=49), 10×10 (N=100) and 15×15 (N=225). The spatial weight matrices are defined on these lattice maps. To generate samples we rewrite equation (1) as:

$$y = \rho_1 W_1 y + \rho_2 W_2 y + x\beta + \varepsilon \tag{5}$$

or

$$y = (I - \rho_1 W_1 - \rho_2 W_2)^{-1}(x\beta + \varepsilon) \tag{6}$$

where W_1 is the weights matrix representing the spatial hotspot structure. Particularly, W_1 is the inverse distance matrix with elements equal to $1/d_{ij}$ for cell *i* and hotspot *j* and 0 elsewhere), W_2 is the conventional first order contiguity or inverse distance matrix, and ρ_1 and ρ_2 are corresponding spatial lag parameters.

Next, *y* is generated as follows:

1. Generate the exogenous variable *x* by drawing from a uniform (0, 10) distribution.
2. Fix the regression coefficient for all simulation runs: $\beta = 1$.
3. The spatial lag parameters ρ_1 and ρ_2 take values 0, 0.1, 0.3, 0.5, 0.7 and 0.9 consecutively[3].
4. Generate values for the error term ε by randomly drawing from a normal distribution with mean zero and variance 2.0.
5. Choose the hotspot according to the values of *x* generated in step 1 and compute *y* according to equation (6).

Both a first order contiguity queen and an inverse distance W_2 is used to generate data.[4] We estimate two types of classical models. One, the TRUE model, estimated with the same W_1 and W_2 as in the model used for data generation. The second is in line with common estimation practice and considers only one overall type of weights matrix, viz. W_2 only. However, we consider both the first order contiguity and inverse distance matrix. Estimation of the SEM model is always based on the first three nearest neighbors and spillover from the hotspot. The estimators are compared in terms of bias and RMSE of β, the coefficient of the regressor *x*.[5] The number of replications is set to 500.

4. SIMULATION RESULTS

In this section we present the main simulation results for TRUE (estimated with W_1 and W_2 used to generate the samples), CONT (estimated with W_2 specified as first order contiguity matrix only), DINV (estimated with W_2 specified as inverse distance matrix only) and SEM (estimated with the first three nearest neighbors and spillover from the hotspot *j* as indicators for cell *i*). Observe that due to the restrictions on ρ_1 and ρ_2, not all parameter combinations are feasible, as explained in the previous section.

Table 1 reports the biases of the estimators of β for samples generated by spillover from hotspot and from first order queen contiguity neighbors for 49 observations. It shows that when $\rho_1 = 0$ and 0.1, CONT has lower biases than SEM in most cases and outperforms TRUE a few times, although the differences are quite small among all models. When $\rho_1 = 0$, $\rho_2 = 0$ or 0.7 and $\rho_1 = 0.1$, $\rho_2 = 0$ or 0.1,

[1] Observe that a SEM will not be identified if the latent variables have not been assigned measurement scales. It is convenient to fix the measurement scale of a latent variable by fixing one λ_i, usually at 1. That is, one often chooses $\lambda_1 = 1$.

[2] Here we only consider one hotspot. However, it is possible to consider several hotspots simultaneously (see Liu et al., 2010b)

[3] Note that in a spatial autoregressive model, the asymptotic properties of the ML estimator require $|I - \rho W| > 0$ (Anselin, 1988). In the present case this constraint is $|I - \rho_1 W_1 - \rho_2 W_2| > 0$.

[4] W_1 is always an inverse distance matrix.

[5] Since they are not directly comparable, we do not compare the spatial autoregressive coefficients (see Liu et al.,2010b).

SEM outperforms TRUE or performs equally well. When $\rho_1 \geq 0.3$, SEM outperforms CONT in almost every case and its dominance becomes more distinct as the value of ρ_1 increases. It is not surprising that CONT is better than SEM for small values of ρ_1, since it is the genuinely true model when $\rho_1 = 0$. Also observe that for each value of ρ_1, the biases of SEM tend to increase as ρ_2 increases. But the increase is not uniform over the interval of ρ_1 for a fixed ρ_2. This is probably due to the limited number of indicators, especially the fixed numbers of neighbors included in SEM estimation.

The RMSE for the model generated on the basis of a first order contiguity matrix is presented in Table 2. The table shows that CONT has smallest RMSE in more than half of the cases. Moreover, the RMSEs of SEM and CONT basically follow the pattern of bias. The lead of CONT diminishes when ρ_1 grows larger and for values of $\rho_1 \geq 0.5$, SEM beats CONT in most cases. Moreover, for the same range of values of ρ_1, SEM even outperforms TRUE in more than half of the cases.

For samples generated with spillover from hotspot and from other regions according to inverse distance, the biases of β are summarized in Table 3. Comparison of this table and Table 1 shows that in the present case both approaches perform worse. It also shows that SEM has lower bias than DINV most of the time. As the value of ρ_1 goes up to 0.9, SEM still remains stable in terms of bias while the estimation results of DINV get extremely biased. Another interesting finding is that SEM outperforms TRUE more often than in the previous case.

Table 4 shows that the RMSEs and bias follow similar patterns in the present case. Moreover, The RMSEs of SEM and DINV tend to diverge more rapidly when ρ_1 increases. Besides, SEM also outperforms TRUE more frequently than in the previous case.

Tables of results for sample sizes 100 and 225 are not presented here due to length limitations. They are available upon request from the first author. The main results are the following. When sample size goes up to 100 and 225, SEM outperforms the classical W-based approaches more in terms of bias, but it does not uniformly outperform them. The comparison in terms of RMSE as a function of the number of observations is very much in line with the bias pattern.

The above analyses of bias and RMSE of the estimator of β show that SEM outperforms the classical W-based models in most cases with more obvious dominance in terms of bias than RMSE. Specifically, it tends to increasingly outperform the classical approach when ρ_1 goes up. These conclusions hold for the correctly specified classical models but even more so for the misspecified models which ignore spillover from hotspots. As far as the type of weights matrix used in sample generation is concerned, both approaches have larger biases and RMSEs for samples generated with spillover from hotspot in combination with inverse distance matrix. Although SEM is also at a disadvantage in the type of samples as it only considers three neighbors in contrast to the correct and much larger number (total sample size) of units that inverse distance matrix model takes into account, DINV produces the most biased results in the majority of cases. SEM makes a winner in terms of stability of bias and RMSE over changing values of the spatial lag parameters, sample sizes and types of weights matrix used for sample generation.

	Hotspot (ρ_1) + Contiguity (ρ_2), Sample size = 49											
	$\rho_2 = 0$		0.1		0.3		0.5		0.7		0.9	
	TRUE CONT	SEM	*TRUE* CONT	SEM	*TRUE* CONT	SEM	*TRUE* CONT	SEM	*TRUE* CONT	SEM	*TRUE* CONT	SEM
$\rho_1 = 0$	*-0.004*		*-0.003*		*-0.001*		*0.001*		*0.004*		*-0.013*	
	-0.006	-0.004	-0.005	0.005	-0.003	0.006	-0.001	0.015	0.000	-0.004	0.003	0.014
0.1	*-0.004*		*-0.003*		*-0.001*		*0.001*		*0.002*		*-0.023*	
	-0.001	0.000	-0.001	-0.003	-0.001	0.004	-0.002	-0.009	-0.011	-0.025	-0.061	0.051
0.3	*-0.004*		*-0.003*		*-0.002*		*0.000*		*0.002*			
	-0.002	0.000	-0.007	-0.004	-0.020	-0.009	-0.028	-0.017	-0.013	-0.023		
0.5	*-0.004*		*-0.003*		*-0.003*		*0.007*					
	-0.010	0.001	-0.017	-0.003	-0.012	-0.010	-0.015	-0.023				
0.7	*-0.004*		*-0.003*		*0.049*							
	0.002	0.000	0.016	-0.001	0.075	-0.037						
0.9	*0.000*		*-0.004*									
	0.051	-0.001	0.060	-0.030								

Table 1. Bias of the estimator of β for spillover from hotspot and first order queen contiguity neighbors

	Hotspot (ρ_1) + Contiguity (ρ_2), Sample size = 49											
	$\rho_2 = 0$		0.1		0.3		0.5		0.7		0.9	
	TRUE CONT	SEM	*TRUE* CONT	SEM	*TRUE* CONT	SEM	*TRUE* CONT	SEM	*TRUE* CONT	SEM	*TRUE* CONT	SEM
$\rho_1 = 0$	*0.061*		*0.060*		*0.060*		*0.060*		*0.061*		*0.081*	
	0.060	0.071	0.060	0.075	0.060	0.075	0.060	0.072	0.061	0.072	0.063	0.138
0.1	*0.061*		*0.061*		*0.060*		*0.061*		*0.063*		*0.079*	
	0.061	0.072	0.061	0.071	0.061	0.072	0.061	0.073	0.063	0.081	0.100	0.078
0.3	*0.061*		*0.061*		*0.061*		*0.061*		*0.060*			
	0.061	0.063	0.061	0.067	0.062	0.062	0.062	0.063	0.100	0.068		
0.5	*0.061*		*0.061*		*0.061*		*0.075*					
	0.061	0.065	0.062	0.063	0.060	0.063	0.127	0.064				
0.7	*0.061*		*0.061*		*0.111*							
	0.062	0.061	0.064	0.061	0.166	0.071						
0.9	*0.064*		*0.060*									
	0.112	0.061	0.220	0.069								

Table 2. RMSE of the estimator of β for spillover from hotspot and first order queen contiguity neighbors

	Hotspot (ρ_1) + Inverse-distance (ρ_2), Sample size = 49											
	$\rho_2 = 0$		0.1		0.3		0.5		0.7		0.9	
	TRUE DINV	SEM	*TRUE* DINV	SEM	*TRUE* DINV	SEM	*TRUE* DINV	SEM	*TRUE* DINV	SEM	*TRUE* DINV	SEM
$\rho_1 = 0$	*-0.012*		*-0.011*		*-0.011*		*-0.010*		*-0.009*		*-0.008*	
	-0.012	-0.004	-0.012	-0.014	-0.011	-0.027	-0.010	-0.060	-0.010	-0.116	-0.009	-0.022
0.1	*-0.012*		*-0.011*		*-0.010*		*-0.009*		*-0.009*		*-0.004*	
	0.000	0.000	0.001	-0.005	0.003	-0.023	0.007	-0.065	0.015	-0.143	0.070	-0.009
0.3	*-0.011*		*-0.010*		*-0.010*		*-0.009*		*-0.008*			
	0.046	0.002	0.049	0.001	0.058	-0.032	0.078	-0.095	0.141	-0.180		
0.5	*-0.010*		*-0.009*		*-0.008*		*-0.008*					
	0.124	0.004	0.134	-0.008	0.168	-0.054	0.256	-0.131				
0.7	*-0.008*		*-0.008*		*-0.007*							
	0.257	0.006	0.296	-0.015	0.506	-0.096						
0.9	*-0.006*		*-0.007*									
	2.518	-0.001	3.244	-0.087								

Table 3. Bias of the estimator of β for spillover from hotspot and inverse distance related neighbors

	Hotspot (ρ_1) + Inverse-distance (ρ_2), Sample size = 49											
	$\rho_2 = 0$		0.1		0.3		0.5		0.7		0.9	
	TRUE DINV	SEM	*TRUE* DINV	SEM	*TRUE* DINV	SEM	*TRUE* DINV	SEM	*TRUE* DINV	SEM	*TRUE* DINV	SEM
$\rho_1 = 0$	*0.066*		*0.066*		*0.066*		*0.065*		*0.065*		*0.064*	
	0.066	0.071	0.065	0.077	0.065	0.084	0.065	0.097	0.064	0.143	0.064	0.127
0.1	*0.066*		*0.066*		*0.066*		*0.065*		*0.065*		*0.063*	
	0.065	0.068	0.065	0.070	0.065	0.079	0.065	0.093	0.065	0.166	0.088	0.144
0.3	*0.066*		*0.066*		*0.066*		*0.065*		*0.065*			
	0.077	0.067	0.078	0.067	0.083	0.067	0.094	0.107	0.144	0.184		
0.5	*0.066*		*0.066*		*0.065*		*0.065*					
	0.130	0.064	0.138	0.062	0.170	0.074	0.256	0.133				
0.7	*0.065*		*0.065*		*0.064*							
	0.258	0.061	0.296	0.061	0.506	0.102						
0.9	*0.064*		*0.063*									
	2.518	0.060	3.244	0.097								

Table 4. RMSE of the estimator of β for spillover from hotspot and inverse distance related neighbors

5. CONCLUSIONS

The paper evaluates by means of Monte Carlo simulations the estimator of the regression coefficient obtained by the classical W-based spatial autoregressive model and the structural equations model with latent variables (SEM) on the basis of data sets that contain two types of spatial dependence: spillover from (i) a hotspot and (iia) first order queen contiguity neighbors or (iib) inverse distance related neighbors. Two types of classical models were considered. SEM takes spatial dependence into account by means of a fixed number of nearest neighbors as well as the dependent variable in the hotspot weighted by inverse distance. The estimation results are analyzed in terms of bias and root mean squared error (RMSE) for different values of the spatial lag parameters, specifications of weights matrices and sample sizes.

The simulation results show that both approaches perform better for samples generated with spillover from the hotspot and from first order queen contiguity neighors. Moreover, compared to the misspecified W-based models, SEM frequently has smaller bias and RMSE and even outperforms the correctly specified models in many cases. These trends increase when the spatial lag parameter for spillover increases. The lead of SEM also increases by sample size. Finally, SEM was found to be more stable in terms of both bias and RMSE over various dimensions.

Finally, note that in the case of SEM not all model search options were exploited. Specifically, the number of observed spatially lagged variables was a priori fixed whereas it offers ample opportunities to search and test the optimal number of observed indicators (see Folmer and Oud, 2008). Another option of SEM that was not exploited was the use of several latent variables to take spatial dependence into account (Folmer and Oud, 2008). Exploitation of this option would have brought SEM closer to the correctly specified model. Full exploitation of all its model search options might improve the performance of SEM in comparison with the classical W-based approaches.

References

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

Florax, R.J.G.M, Folmer, H. and Rey, S.J., 2003. Specification searches in spatial econometrics: the relevance of Hendry's methodology. *Regional Science and Urban Economics, 33, 5,* 557-579.

Florax, R. and Folmer, H., 1992. Specification and estimation of spatial linear regression models: Monte Carlo evaluation of pre-test estimators. *Regional Science and Urban Economics, 22,* 405–432.

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.

Jöreskog K.G. and Sörbom, D., 1996. *Lisrel 8: User's Reference Guide.* Scientific Software International, Chicago, IL.

LeSage, J and Pace, R. K., 2009. *Introduction to Spatial Econometrics*. Chapman & Hall/CRC.

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.

Oud, J. and Folmer, H., 2008. A structural equation approach to models with spatial dependence. *Geographical Analysis, 40,* 152-166.