Lett Spat Resour Sci (2011) 4: 71–80
DOI 10.1007/s12076-010-0047-3

ORIGINAL PAPER

Estimating regression coefficients by W-based and latent variables spatial autoregressive models in the presence of spillovers from hotspots: evidence from Monte Carlo simulations

An Liu · Henk Folmer · Johan H.L. Oud

Received: 17 May 2010 / Accepted: 6 December 2010 / Published online: 23 December 2010

Abstract The paper evaluates by means of Monte Carlo simulations the estimators of regression coefficients in the presence of spillover effects from one or more hotspots by the classical W-based spatial autoregressive model and the structural equation model with latent variables (SEM). The estimators are evaluated in terms of bias and root mean squared error (RMSE) for different values of the spatial autoregressive coefficient, different sample sizes and different specifications of weight matrices. The simulation results show that both approaches perform better for smaller values of the spatial autoregressive coefficient and larger sample sizes. SEM tends to outperform the classical approach in term of bias but the classical model based on first-order contiguity matrix has lowest RMSE in most cases. Furthermore, SEM provides a more stable performance in terms of variations of bias and RMSE with respect to changes in the value of autoregressive coefficient, sample size and number of hotspots. It follows that compared to the classical approach, SEM does not only have favorable behavioral properties in that it straightforwardly allows inclusion of different types

A. Liu (✉) · H. Folmer
Department of Spatial Sciences, University of Groningen, P.O. Box 800, 9700AV Groningen, The Netherlands
e-mail: an.liu@rug.nl

H. Folmer
e-mail: h.folmer@rug.nl

H. Folmer
College of Economics and Management, Northwest A&F University, Yangling, Shaanxi 712100, China
e-mail: henk.folmer@wur.nl

J.H.L. Oud
Behavioural Science Institute, Radboud University Nijmegen, P.O. Box 9104, 6500HE Nijmegen, The Netherlands
e-mail: j.oud@pwo.ru.nl

of spatial dependence in one model framework and of testing distance decay, but also favorable econometric properties.

Keywords Spatial autoregressive model · Monte Carlo simulation · Bias · RMSE

JEL Classification C01 · C13 · C51

1 Introduction

The main purpose of W-based spatial econometrics is to control for spatial dependence in the dependent variable or in the error term to obtain consistent, unbiased or efficient estimators. Spatial dependence or spill-overs among the spatial units of observation are typically modeled by means of a spatial weights matrix, often denoted W. Most common in practice is a priori selection and specification of W (usually, a first-order, double rook contiguity matrix) on the basis of intuition or ad hoc considerations, followed by specification searches to decide upon the error or lag model (Florax et al. 2003), with little attention being paid to behavioral aspects underlying spatial dependence.

Folmer and Oud (2008) introduced structural equation models with latent variables (denoted SEM below) as an alternative to the W-based approach. In contrast to the W-based approach, SEM allows explicit modeling and testing of theoretical considerations underlying the spatial structure.

A (general) structural equation model is made up of a structural model that represents the relationships among the latent variables and a measurement model that contains the relationships between the latent variables and their observed indicators. Observe that some or all of the variables in the structural model may be observed in which case the corresponding relationships in the measurement model reduce to identity relationships.

SEM replaces the spatially lagged variables in the W-based regression model by one or more latent variables in the structural model and specifies the relationship between the latent spatially lagged variables and their observed indicators in the measurement model. The selection of the indicators of the latent spatially lagged variables is based on theoretical considerations. In addition, it allows handling of various kinds of spatial dependence in one framework, for instance, spatial dependence due to spillover from neighbors, from hotspots and from spatial units that share certain properties but are no neighbors. Moreover, both the coefficient of the latent spatially lagged variable and the coefficients of the indicators can be tested. While the former is the analog of testing the autoregressive coefficient in W-based modeling, the latter is additional and allows testing of specific behavioral relationships underlying spatial dependence.

Folmer and Oud (2008) show for one specific example, i.e. Anselin's (1988) Columbus, Ohio, crime data set, that SEM produces estimates of the regression coefficients of the explanatory variables that are virtually identical to those obtained by the W-based approach, while the autoregression coefficients slightly differ. To gain further insight into the properties of the W-based approach and SEM, Liu et al. (2010)

carried out a series of Monte Carlo simulations on the basis of the spatial structure in Anselin (1988). The latent spatially lagged variable in the SEM model was measured by a number of nearest neighbors. Data was generated on the basis of the first-order contiguity or inverse distance matrix. The main result was that the W-based approach (with weight matrix consistent with the data generation matrix) had lowest bias and RMSE in the majority of cases. SEM outperformed the W-based approach for some types of W matrices. Particularly, it had the smallest bias in several cases. Generally speaking, however, the results of both approaches did not differ much.

In this paper we further evaluate the performances of the W-based approach and SEM in a different setting, viz. in the presence of spillovers from hotspots. A hotspot is a geographic area that exhibits a high volume or intensity of a certain activity and impacts on the activity in other cells of the geographic system. For instance, a crime hotspot is an area containing higher concentration of criminal incidents than its surrounding areas and impacts on the crime rate in the other neighborhood. Adequate modeling requires that the activity in the hotspot and its impact are taken into account. For details on crime hotspots see amongst others Short et al. (2010).

The remainder of the paper is organized as follows. Section 2 briefly specifies the model structures of the W-based approach and SEM. A detailed description of the simulation study design is given in Sect. 3. In Sect. 4 we report the simulation results and Sect. 5 concludes the paper.

2 Model specifications

The W-based spatial autoregressive model reads:[1]

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

$$\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right), \tag{2}$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations on the dependent variable, $\mathbf{X}$ is an $n \times q$ data matrix of explanatory variables with associated coefficient vector $\boldsymbol{\beta}$, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of error terms. $\mathbf{W}$ is the $n \times n$ spatial weight matrix, with spatial autoregressive coefficient ρ. (For further details see amongst others, LeSage and Pace 2009.)

A structural equation model in general form consists of three basic equations:[2]

$$\mathbf{y} = \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon} \quad \text{with} \quad \mathbf{cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta}_\varepsilon, \tag{3}$$

$$\mathbf{x} = \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta} \quad \text{with} \quad \mathbf{cov}(\boldsymbol{\delta}) = \boldsymbol{\Theta}_\delta, \tag{4}$$

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \quad \text{with} \quad \mathbf{cov}(\boldsymbol{\xi}) = \boldsymbol{\Phi}, \quad \mathbf{cov}(\boldsymbol{\zeta}) = \boldsymbol{\Psi}. \tag{5}$$

[1] Below matrices and vectors are bold-face, scalars in italics.

[2] Observe that models (3)–(5) and (6)–(12) are in terms of variables, while model (1) is in terms of observations. Estimation of a SEM is by minimizing the distance between the theoretical and sample covariance or moment matrix of the y and x variables. The theoretical covariance matrix is in terms of the eight parameter matrices in (3)–(5). For details see Folmer and Oud (2008).

Equations (3) and (4) are the measurement models with $\mathbf{y}$ and $\mathbf{x}$ the p- and q-variate vectors of observable endogenous and exogenous variables or indicators, $\mathbf{\Lambda}_y$ and $\mathbf{\Lambda}_x$ the $p \times k$ and $q \times l$ matrices of loadings of the observable variables (indicators) on the k- and l-vectors of latent variables $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$, and $\mathbf{\Theta}_\varepsilon$ and $\mathbf{\Theta}_\delta$ are the $p \times p$ and $q \times q$ measurement error covariance matrices. In the structural model (5), the $k \times k$ matrix $\mathbf{B}$ specifies the structural relationships among the latent endogenous variables and the $k \times l$ matrix $\mathbf{\Gamma}$ contains the impacts of the exogenous latent variables on the endogenous latent variables. $\mathbf{\Phi}$ is the $l \times l$ covariance matrix of the latent exogenous variables and $\mathbf{\Psi}$ is the $k \times k$ covariance matrix of the errors in the structural model. The measurement errors $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$ are assumed to be uncorrelated with the latent variables $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ as well as with the structural errors $\boldsymbol{\zeta}$. For details on identification, estimation, testing and specification of structural equation models see Jöreskog and Sörbom (1996).

The SEM approach to models with spatial dependence replaces the spatially lagged variable $\mathbf{W}y$ in (1) by a latent variable η in the structural model.[3] In the measurement model η is measured by the observed variables that capture spatial dependence.

As an illustration and in preparation for the simulations below, we present the SEM specification of the simulation models for two hotspots. We consider spillovers from hotspots that decrease with distance from the hotspots.

First, the structural equation model corresponding to (1) is the structural model (5) which in this case is a one-equation relationship. It is presented in (6). There is an identity relationship between y in (1) and the latent dependent variable η_1 in (6). In addition, the spillover effects from the two hotspots to all the other spatial units are represented by $\rho\eta_W$ in (5). Hence, the structural model (5) reads:

$$\eta_1 = \rho\eta_W + \boldsymbol{\gamma}'\boldsymbol{\xi} + \varepsilon. \tag{6}$$

The measurement model for the endogenous vector $\boldsymbol{\eta}$ (made up of η_1 and η_W) takes the form:

$$y = \eta_1, \tag{7}$$

$$y_{h,1} = \eta_W + \varepsilon_1, \tag{8}$$

$$y_{h,2} = \lambda_2\eta_W + \varepsilon_2, \tag{9}$$

where $y_{h,j}$ is the distance weighted spillover from hotspot j. Finally, for the exogenous variables we have identity relationships:

$$\mathbf{x} = \boldsymbol{\xi}, \tag{10}$$

or in terms of individual variables:

$$x_1 = \xi_1, \tag{11}$$

$$x_2 = \xi_2. \tag{12}$$

[3] Observe that if several kinds of spatial dependence need to be distinguished, several latent spatially lagged variables can be applied, each with its own indicators (Folmer and Oud 2008).

Observe that a SEM is not 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. See (8) where $\lambda_1 = 1$. For (10)–(12), $\mathbf{\Lambda}_x$ is an identity matrix and error terms are zeros. See Folmer and Oud (2008), (16)–(27) for further details on SEM model specifications.

3 Simulation study design

In the hotspot model considered below the dependent variable in each spatial unit is affected by the dependent variable in one or more hotspots. We assume that the spillover decreases with distance as the impact from the hotspots weakens over distance. For data generation this setting implies that the hotspots need to be known in advance. However, it is not until the samples are generated that we know which regions are hotspots. To solve this problem we take a step backward and choose the 'potential' hotspots on the basis of an independent variable instead. That is, we take an independent variable, say x_1, which is considered as the key variable that leads to variations in the dependent variable y in each region and designate the hotspot according to the values of x_1.

We consider regular lattice structures of dimensions 7×7 ($n = 49$), 10×10 ($n = 100$) and 15×15 ($n = 225$). The spatial weights matrices are defined on these lattice maps.

For sample generation we specify (1) as:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \boldsymbol{\varepsilon}, \tag{13}$$

or

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}(\mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \boldsymbol{\varepsilon}). \tag{14}$$

Next, $\mathbf{y}$ is generated as follows:

1. Generate the exogenous variable matrix by drawing two variables (x_1, x_2) from a uniform (0, 10) distribution. The exogenous variables are fixed over all simulation runs.
2. Fix $\beta_1 = 1$, $\beta_2 = 0.3$ for all simulation runs.
3. The spatial autoregressive parameter ρ takes the values 0 (the benchmark model), 0.5 and 0.9.
4. Generate values for the error term $\boldsymbol{\varepsilon}$ by randomly drawing from a normal distribution with mean zero and variance 2.0.
5. Choose 1–5 hotspots according to the largest values of x_1 generated in step 1 and compute $\mathbf{y}$ according to (14) adopting the inverse distance matrix with elements equal to $1/d_{ij}$ for cell i and hotspot j, and zero elsewhere. In the case $i = j$, d_{ij} takes the value 1, which equals the minimum possible distance between any cell i and hotspot j on the map. Hence matrix $\mathbf{W}$ consists of nonzero columns j corresponding to hotspot j, and zero columns elsewhere.

The W-based models are estimated on the basis of a first-order queen contiguity matrix and an inverse distance matrix. The contiguity weights matrix is row-standardized. SEM takes the scores per hotspot weighted by inverse distance as indicators of the single latent spatially lagged variable in the structural model. (For the two-hotspot case see (8) and (9).) Given these specifications as well as (6), (7) and (10), estimation of SEM is standard and can be done by means of the software package Mx (Neale et al. 2003).[4]

As in Liu et al. (2010), the performances of the approaches will be compared in terms of bias and RMSE for various sample sizes, specifications of weights matrices and values of the spatial autoregressive coefficient. The bias of an estimator $\hat{\theta}$ with respect to the true value of the parameter θ is defined to be:

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta = E\big[(\hat{\theta} - \theta)\big]; \tag{15}$$

and the RMSE of this estimator is defined as:

$$\text{RMSE}(\hat{\theta}) = \sqrt{E\big[(\hat{\theta} - \theta)^2\big]}. \tag{16}$$

We restrict the comparison to the main coefficients of interest, i.e. the coefficients of the regressors x_1 and x_2. The reason for leaving out the spatial autoregressive coefficient is that in SEM spatial dependence is captured by parameters in the structural and measurement models rather than only the single spatial autoregressive coefficient, as in the W-based approach. Comparison of spatial dependence in both types of approaches would lead to overrunning the size of a letters paper. The number of replications is 1000.

4 Simulation results

Table 1 reports the bias of the estimators of β_1. It shows that for $n = 49$, $\rho = 0$ and $\rho = 0.9$, SEM has lowest bias and outperforms both the first order contiguity matrix (denoted CONT) and the distance inverse (denoted DINV) based W-based model for one to four hotspots. Moreover, CONT always outperform DINV. In the case of five hotspots, CONT has the lowest bias. The bias of β_1 is again lowest for SEM in most cases when $\rho = 0.5$. In the case of two and five hotspots, CONT performs slightly better than SEM.

For 100 observations SEM still outperforms both types of W-based models in most cases. However, CONT is best for one, three and five hotspots when $\rho = 0$ while DINV is best in the case of two and three hotspots when $\rho = 0.5$. When ρ goes up to 0.9, SEM remains a winner for one, four and five hotspots and CONT has the smallest bias in the rest of the cases.

[4] There exist various other software packages to estimate SEMs, e.g. LISREL (Jöreskog and Sörbom 1996). However, these packages cannot be used to estimate models with spatially lagged dependent variables because they do not allow expansion of the likelihood function with the Jacobian correction (Folmer and Oud 2008). Mx can also be applied to estimate the standard lag model (Folmer and Oud 2008).

Table 1 Bias of estimators of β_1

	1 hotspot			2 hotspots			3 hotspots			4 hotspots			5 hotspots		
ρ	CONT	DINV	SEM	CONT	DINV	SEM	CONT	DINV	SEM	CONT	DINV	SEM	CONT	DINV	SEM
	49 observations														
0	−0.003	−0.007	−0.001	−0.003	−0.007	−0.002	−0.003	−0.007	0.000	−0.003	−0.007	−0.003	−0.003	−0.007	−0.004
0.5	−0.013	0.145	−0.001	0.001	0.157	−0.005	−0.017	0.055	−0.012	−0.019	0.022	−0.009	0.017	0.048	0.030
0.9	0.053	1.144	−0.001	0.317	0.935	−0.031	−0.014	0.192	−0.014	−0.025	0.090	−0.005	0.030	0.122	0.090
	100 observations														
0	0.000	0.002	0.002	0.000	0.002	0.000	0.000	0.002	0.003	0.000	0.002	0.000	0.000	0.002	0.005
0.5	0.006	−0.015	0.002	0.027	0.016	−0.034	0.027	0.023	−0.027	0.024	0.020	0.006	0.027	0.024	0.015
0.9	−0.255	−0.505	0.002	0.026	0.036	0.036	0.035	0.041	−0.066	0.039	0.037	−0.023	0.042	0.042	−0.030
	225 observations														
0	−0.003	−0.002	−0.002	−0.003	−0.002	−0.006	−0.003	−0.002	−0.004	−0.003	−0.002	−0.004	−0.003	−0.002	−0.005
0.5	−0.002	−0.011	−0.002	−0.004	−0.007	0.006	0.003	0.003	0.008	0.000	0.000	0.003	0.004	0.004	0.006
0.9	0.043	0.052	−0.002	−0.002	−0.012	0.009	0.007	0.008	0.059	0.003	0.003	0.017	0.010	0.011	0.081

As sample size increases to 225, the biases of the estimator of β_1 in the three models continue to decrease and so do the differences among them. Observe that DINV has the lowest bias in most cases. For $\rho = 0$, DINV outperforms the other two models in the cases of two to five hotspots. Moreover, for $\rho = 0.5$ CONT and DINV have virtually equal bias in the cases of three and five hotspots. For $\rho = 0.9$ the winners are the same as when $\rho = 0$.

The results for β_2 are very much in line with the results for β_1 and are therefore not discussed.

For each model the RMSEs of β_1 and β_2 are very close. Therefore, we add them up and consider the sum, denoted as total RMSE. The results are presented in Table 2. Generally speaking, the RMSEs of both W-based models are lower than those of SEM. However, there are a few exceptions. For instance, in the case of 49 observations, SEM has lowest total RMSE almost everywhere for one and two hotspots and all values of ρ. In another five cases with more than one hotspot, SEM is only slightly trailing the best model, CONT. For 100 and 225 observations, SEM outperforms CONT and DINV in all the cases with one hotspot. In most other cases CONT has lowest RMSE, followed by DINV whose RMSE is only slightly higher for each value of ρ. Notice that as sample size goes up, the total RMSE of β_1 plus β_2 for all three models decreases and becomes more stable. Also observe that the total RMSEs of the three models tend to converge when the number of hotspots increases from one to five for $\rho \geq 0.5$. When $\rho = 0$, the RMSEs do not differ much but there is no convergence by number of hotspots. Another interesting finding is that the total RMSE of SEM is the most stable for changes in sample size, number of hotspots and autoregressive coefficient.

From the above it follows that in line with expectations all three models perform better as the sample size increases. Moreover, for all three models the bias and RMSE get larger when ρ increases; but there is no clear trend when the number of hotspots increases. In terms of bias SEM outperforms both W-based models quite often but not always. DINV is worst in most cases. Although SEM slightly trails the W-based models in terms of RMSE, it does outperform one or the other W-based model in some cases when it is not a winner. For 49 observations SEM has a better chance to outperform the W-based models than for 100 and 225 observations. The differences in total RMSE between SEM and the W-based models are actually very small for large sample sizes. Also note that SEM probably has larger RMSE, because the model has more parameters to be estimated and therefore the parameter estimates have a larger variance. Finally, SEM's bias and RMSE are more stable than its alternatives.

5 Conclusions

In this paper we further evaluate by means of Monte Carlo simulations the estimators of regression coefficients in the presence of spillover effects from hotspots by the W-based and SEM approaches. The former accounts for spatial dependence and spillover effects by means of a spatial weight matrix W and the latter by means of a latent variable in the structural model, measured by means of observed spatially lagged variables in the measurement model. The estimators are evaluated in terms of

Table 2 Total RMSE of estimators of β_1 and β_2

	1 hotspot			2 hotspots			3 hotspots			4 hotspots			5 hotspots		
ρ	CONT	DINV	SEM	CONT	DINV	SEM	CONT	DINV	SEM	CONT	DINV	SEM	CONT	DINV	SEM
	49 observations														
0	0.112	0.115	0.112	0.112	0.115	0.114	0.112	0.115	0.125	0.112	0.115	0.152	0.112	0.115	0.151
0.5	0.121	0.321	0.112	0.122	0.308	0.115	0.119	0.177	0.121	0.120	0.146	0.136	0.119	0.142	0.149
0.9	0.538	2.362	0.112	0.720	1.912	0.138	0.123	0.462	0.327	0.123	0.291	0.174	0.126	0.247	0.244
	100 observations														
0	0.077	0.077	0.077	0.077	0.077	0.086	0.077	0.077	0.087	0.077	0.077	0.091	0.077	0.077	0.097
0.5	0.079	0.079	0.077	0.097	0.107	0.111	0.091	0.093	0.111	0.086	0.085	0.095	0.084	0.082	0.094
0.9	0.332	0.578	0.077	0.098	0.195	0.274	0.106	0.137	0.158	0.094	0.100	0.134	0.092	0.094	0.115
	225 observations														
0	0.052	0.052	0.052	0.052	0.052	0.076	0.052	0.052	0.063	0.052	0.052	0.065	0.052	0.052	0.065
0.5	0.056	0.109	0.052	0.057	0.069	0.088	0.051	0.055	0.073	0.051	0.053	0.065	0.051	0.053	0.068
0.9	0.302	0.385	0.052	0.055	0.126	0.102	0.052	0.059	0.162	0.052	0.057	0.081	0.053	0.054	0.166

bias and RMSE for different values of the spatial autoregressive coefficient, sample sizes and numbers of hotspots.

The simulation results show that both approaches perform better for smaller values of the spatial autoregressive coefficient and larger sample sizes. There is no uniform tendency for increasing number of hotspots. SEM tends to outperform the W-based models in term of bias but the W-based model based on first-order contiguity matrix has lowest RMSE in most cases. In all, combining the evaluation of both bias and RMSE for the estimators we find that the W-based model based on the first-order contiguity weights matrix performs slightly better than SEM while the W-based model based on the inverse distance weights matrix comes last. But there is no uniform winner over all dimensions. However, SEM is the most stable over variations in sample size, number of hotspots and spatial autoregressive coefficient.

References

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

Florax, R.J.G.M., Folmer, H., Rey, S.J.: Specification searches in spatial econometrics: the relevance of Hendry's methodology. Reg. Sci. Urban Econ. **33**(5), 557–579 (2003)

Folmer, H., Oud, J.: How to get rid of W: a latent variables approach to modeling spatially lagged variables. Environ. Plan. A **40**(10), 2526–2538 (2008)

Jöreskog, K.G., Sörbom, D.: Lisrel 8. User's reference guide. Scientific Software International, Chicago, IL (1996)

LeSage, J., Pace, R.K.: Introduction to Spatial Econometrics. Chapman & Hall, London (2009)

Liu, A., Folmer, H., Oud, J.H.L.: W-based versus latent variables spatial autoregressive models: evidence from Monte Carlo simulations. The Annuals of Regional Science. Online first. http://www.springerlink.com/content/e7782311u175713x/fulltext.pdf (2010)

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

Short, M.B., Brantingham, P.J., Bertozzi, A.L., Tita, G.E.: Dissipation and displacement of hotspots in reaction-diffusion models of crime. Proc. Natl. Acad. Sci. USA **107**, 3961–3965 (2010)