Ann Reg Sci (2011) 47:619–639
DOI 10.1007/s00168-010-0398-0

SPECIAL ISSUE PAPER

W-based versus latent variables spatial autoregressive models: evidence from Monte Carlo simulations

An Liu · Henk Folmer · Johan H. L. Oud

Received: 7 July 2009 / Accepted: 1 April 2010 / Published online: 23 June 2010

Abstract In this paper, we compare by means of Monte Carlo simulations two approaches to take spatial autocorrelation into account: the classical spatial autoregressive model and the structural equations model with latent variables. The former accounts for spatial dependence and spillover effects in georeferenced data by means of a spatial weights matrix W. The latter represents spatial dependence and spillover effects by means of a latent variable in the structural (regression) model while the observed spatially lagged variables are related to the latent spatial dependence variable in the measurement model. The simulation results based on Anselin's Columbus, Ohio, crime data set show that the misspecified latent variables approach slightly trails the correctly specified classical approach in terms of bias and root mean squared error of the coefficient estimators.

JEL Classification C13 · C15 · C52 · R15

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

H. Folmer
Department of Social Sciences, Wageningen University,
PO Box 8130, 6700 EW Wageningen, The Netherlands
e-mail: henk.folmer@wur.nl

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

1 Introduction

When it comes to applying econometric models to analyze georeferenced data, researchers are well aware of the fact that ignoring spatial dependencies leads to inefficient and biased estimators. A substantial theoretical and simulation literature on efficient and consistent estimators for spatial dependence models has developed. A major issue concerns the specification of the structure of spatial dependence including the type of spatial weights to be incorporated into the model. This paper focuses on the evaluation of two approaches that explicitly model spatial dependence: the classical W-based regression model and the recently proposed latent variables approach.[1]

In the spatial econometrics literature the W-based spatial regression approach has been dominant and is most commonly used. The approach is based on a spatial weights matrix, usually denoted W, that accounts for spatial dependence or spill-over effects among the spatial units of observation. The selection of a spatial weights matrix is a crucial step in spatial modelling because it a priori imposes a model structure which affects estimates (Bhattacharjee and Jensen-Butler 2006; Anselin 2002; Fingleton 2003) and the substantive interpretation of the research findings (Hepple 1995).

Several types of spatial weights matrices can be used to represent spatial dependence. Most common are the contiguity-based matrices. Two regions are said to be first-order contiguous if they share a common border (rook) or vertex (bishop) or both (queen). The concept of spatial contiguity can be extended to higher orders. Another common type of weights matrix is distance based, such as inverse distance or inverse distance squared, or a fixed distance band.

Much progress has been made with respect to the the construction and comparison of spatial weights matrices including estimation of spatial weights matrices that are consistent with an observed pattern of spatial dependence rather than assuming a priori the nature of spatial interaction dependence (Hepple 1995; Aldstadt and Getis 2004, 2006). In spite of all these developments the most common procedure in applied research is still to assume a priori first-order contiguity, as expressed by a spatial weights 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 under consideration here, the latent variables approach was introduced by Folmer and Oud (2008). It proceeds on the basis of a structural equations 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. An example of a latent variable is socio-economic status. This concept refers to an individual's standing in society which cannot be observed or measured directly. However, it can be measured via observable indicators like educational attainment, income, occupational status, etc. In a similar vein, the latent variable regional welfare can be measured by observable indicators

[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; Tiefelsdorf and Griffith 2007). Spatial filtering is not considered in this paper.

like GDP per capita, features of the income distribution, labour market opportunities, indicators of the healthcare system, environment quality and so on.

A SEM is made up of two submodels: (1) the structural model representing the causal relationships among the latent variables and (2) the measurement model representing the relationships between the latent variables and their observable indicators (Oud and Folmer 2008). The latent variables approach accounts for spatial dependence with the spatially lagged variables represented by latent variables in the structural model and models the relationship between latent spatially lagged variables and their observed indicators in the measurement model. Since one latent variable can be measured by several indicators, this approach allows for the straightforward inclusion of various kinds of spatial dependence in the model, e.g. spatial dependence due to the impact of the nearest neighbours and distance-weighted impact from hotspots. It is also capable of including different types of contiguity, including nonspatial contiguity, for instance, a dependent relationship between regions due to economic, social or demographic similarity (see Case et al. 1993) rather than e.g. conventional spatial dependence such as first-order contiguity.[2] Moreover, several types of spatial dependences can be included in a given model by using different latent variables in the structural model with corresponding sets of indicators in the measurement model, e.g. one latent variable for conventional spatial contiguity and another for socio-economic dependency. Another important feature of the latent variable approach is that it allows testing of the relationship of a latent variable and its indicators, e.g. whether or not the next nearest neighbour contributes to the latent dependence variable.

Folmer and Oud (2008) show that the latent variables approach can produce estimates that are virtually identical to those obtained by the W-based approach but also that it is more general than the latter. They argue, however, that further comparison is needed to draw up the pros and cons of each approach. To gain more insight into the quality of the estimators produced by both approaches, we carry out Monte Carlo simulations. The simulations are performed 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 performance of the approaches will be analyzed in terms of bias and root mean squared error (RMSE) for various values of spatial autocorrelation and kinds of weight matrices.

The remainder of the paper is organized as follows. Section 2 summarizes the W-based spatial regression approach and the latent variables approach and specifies their model structures. A description of the experimental design is given in Sect. 3. In Sect. 4 we present the simulation results while Sect. 5 concludes.

[2] Non-spatial contiguity can also be handled by conventional W methods. However, if both spatial and non-spatial contiguity are taken into account in a conventional model, more than one W matrix is required, particularly one corresponding to the former and one to the latter type of dependence.

2 Model specifications

2.1 The W-based autoregressive model

There are two types of W-based spatial models that are most commonly used in applied research: the spatial lag model and the spatial error model. Here, we restrict ourselves to the former. The lag model assumes that the dependent variable in a given region, say *i*, is a function of exogenous variables in *i* and of the dependent variable in other regions, different from *i*. Particularly, the classical W-based spatial lag or autoregressive model reads:

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

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

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 associated coefficient vector β, ε is an $n \times 1$ vector of error terms. $W$is the $n \times n$ spatial weights matrix with diagonal elements equal to zero and off-diagonal elements unequal to zero, if the regions corresponding to that element meet the adopted spatial dependence definition. The parameter ρ is the spatial autoregressive or spatial lag parameter.

The log-likelihood function for model (1) reads Anselin's (1988):

$$L = -\frac{N}{2}\ln\pi - \frac{N}{2}\ln\sigma^2 + \ln|A| - \frac{1}{2\sigma^2}(Ay - X\beta)'(Ay - X\beta) \quad \text{with } A = I - \rho W. \tag{3}$$

Since the likelihood function contains a Jacobian term $\ln|A|$ and 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. Ord (1975) has derived a simplification of determinants such as $|A|$ in terms of its eigenvalues. Specifically:

$$\ln|I - \rho W| = \ln \Pi_i (1 - \rho w_i) = \Sigma_i \ln(1 - \rho w_i) \tag{4}$$

where the w_i are the eigenvalues of W.

2.2 The latent variables spatial autoregressive model

Before going into detail, we observe that a SEM is usually specified in terms of variables in contrast to spatial regression models like model (1) which are specified in terms of units of observation. Below we adopt the SEM convention. When we refer to a SEM in terms of units of observation we use a tilde (∼).

A SEM in general form consists of two basic equations:

$$\boldsymbol{y} = \Lambda\eta + \varepsilon \quad \text{with } \operatorname{cov}(\varepsilon) = \Theta \tag{5}$$

$$\eta = \mathrm{B}\eta + \zeta \quad \text{with } \operatorname{cov}(\zeta) = \Psi \tag{6}$$

Equation (5) is the measurement model where the vector $\boldsymbol{y}$ contains m observed variables. Λ contains the loadings of the observed variables (indicators) on the vector of k latent variables η[3], and Θ is the $m \times m$ measurement error covariance matrix. In the structural model (6), 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 in ε are assumed to be uncorrelated with the latent variables in η. Observe that directly observed variables can be included in the structural model by specifying in the measurement model an identity relationship between a given observed variable and the corresponding latent variable.

A SEM model is estimated by minimizing the distance between the model implied covariance matrix (on the basis of hypotheses relating to the model structure as specified in the parameter matrices B,Ψ, Λ, and Θ) 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 is the most commonly used estimator and the default in the statistical packages Mx and LISREL. Below we discuss the ML estimator. It maximizes the log-likelihood function:

$$l(\theta|Y) = -\frac{N}{2}\ln|\Sigma| - \frac{N}{2}\mathrm{tr}(S\Sigma^{-1}) - \frac{pN}{2}\ln 2\pi \tag{7}$$

where Σ is the theoretical covariance matrix in terms of the free and constrained elements in the four parameter matrices. That is:

$$\Sigma = \Lambda(I-B)^{-1}\Psi(I-B')^{-1}\Lambda' + \Theta \tag{8}$$

and S is the observed covariance matrix for given data set Y.

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

$$F_{\mathrm{ML}} = \ln|\Sigma| + \mathrm{tr}(S\Sigma^{-1}) - \ln|S| - p \tag{9}$$

gives the same results as maximizing the above likelihood function (Oud and Folmer 2008). The LISREL software package also contains a variety of tests and model evaluation statistics and gives reliable information about the identification status of the model (see Jöreskog and Sörbom 1996 for details).

The SEM analog to model (1) is specified as follows. First, the spatially lagged dependent variable Wy is taken as a latent variable in the structural model (i.e. Wy in (1) is replaced by a latent variable η). Specifically:

$$y = \rho\eta + \gamma' x + \zeta \tag{10}$$

[3] 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 $\lambda_1 = 1$.

The structural model (10) is completed by a measurement equation:

$$\boldsymbol{y} = \Lambda \eta + \varepsilon \tag{11}$$[4]

with

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \end{bmatrix}, \quad \Theta = \begin{bmatrix} \sigma^2_{\varepsilon_1} & 0 & 0 & 0 \\ 0 & \sigma^2_{\varepsilon_2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma^2_{\varepsilon_m} \end{bmatrix} \tag{12}$$

Observe that in (12) the latent spatial lag η is measured by more than one observed variable. For instance, for Anselin's crime model Folmer and Oud (2008) consider the three, respectively, six, nearest neighbours as well as the distance from the crime center.

The measurement model is constructed by means of selection functions or selection matrices $\boldsymbol{S_i}$ which select relevant observations from the vector of observations as follows [as indicated above, a tilde ($\sim$) denotes (a vector of) observations]:

$$\begin{aligned} \tilde{\boldsymbol{y}}_1 &= \boldsymbol{S}_1\tilde{\boldsymbol{y}}, \\ \tilde{\boldsymbol{y}}_2 &= \boldsymbol{S}_2\tilde{\boldsymbol{y}}, \\ &\vdots \\ \tilde{\boldsymbol{y}}_{\boldsymbol{m}} &= \boldsymbol{S_m}\tilde{\boldsymbol{y}}. \end{aligned} \tag{13}$$

That is, $\boldsymbol{S}_1$ selects the values for the first indicator vector $\tilde{\boldsymbol{y}}_1$, $\boldsymbol{S}_2$ for the second indicator vector $\tilde{\mathbf{y}}_2$, etc. For example, for the simulations based on Anselin's crime model, $\boldsymbol{S}_1$ could select the value of crime in the nearest contiguous neighbour, $\boldsymbol{S}_2$ the crime value in the next nearest contiguous neighbour and so on. Thus, for the measurement model we obtain:

$$\begin{aligned} \tilde{\boldsymbol{y}}_1 &= \boldsymbol{S}_1\tilde{\boldsymbol{y}} = \lambda_1\tilde{\eta} + \tilde{\varepsilon}_1, \\ \tilde{\boldsymbol{y}}_2 &= \boldsymbol{S}_2\tilde{\boldsymbol{y}} = \lambda_2\tilde{\eta} + \tilde{\varepsilon}_2, \\ &\vdots \\ \tilde{\boldsymbol{y}}_{\boldsymbol{m}} &= \boldsymbol{S_m}\tilde{\boldsymbol{y}} = \lambda_{\boldsymbol{m}}\tilde{\eta} + \tilde{\varepsilon}_{\boldsymbol{m}}. \end{aligned} \tag{14}$$

From the above it follows that spatial dependence is captured by two kinds of parameters, ρ and λ_i, whereas in the standard lag model only the "average" effect $\rho W y$ shows up. This means that the latent variable approach offers a much richer representation of the spatial structure than the standard spatial econometric approaches. Moreover, it allows testing of the relationship between indicators and the corresponding latent variables, for instance, whether the sixth nearest neighbour is a significant indicator.

The standard SEM log-likelihood function needs correction so as to account for the presence of the latent spatial lag variable among the explanatory variables.

[4] Observe that in (10) y is a scalar while in (11) $\boldsymbol{y}$ is a vector.

Folmer and Oud (2008) show that $\ln |A|$ needs to be added to the log-likelihood function (7) with[5]

$$A = I - \frac{\rho}{m\lambda_1} S_1 - \frac{\rho}{m\lambda_2} S_2 - \cdots - \frac{\rho}{m\lambda_m} S_m \tag{15}$$

3 Experimental design

To investigate the performances of the two modelling approaches specified in Sect. 2, we conduct a number of Monte Carlo simulations. For that purpose we make use of Anselin's (1988) Columbus, Ohio, crime data set. Particularly, we use its spatial structure, the observed values of the exogenous variables and the parameter estimates.

The first step is sample generation. For that purpose we specify equation (1) in terms of Anselin's crime model:

$$y = \rho W y + \beta_0 + x_1\beta_1 + x_2\beta_2 + \varepsilon \tag{16}$$

or:

$$y = (I - \rho W)^{-1}(\beta_0 + x_1\beta_1 + x_2\beta_2 + \varepsilon) \tag{17}$$

Next, y is generated as follows (cf. Florax and Folmer 1992; Florax et al. 2003):

1. Fix the exogenous variables x_1 (income) and x_2 (housing value) at the values in Anselin's (1988) crime data set.
2. Choose Anselin's (1988) estimate of the vector β[6]:

$$\beta_0 = 45.079, \quad \beta_1 = -1.032, \quad \beta_2 = -0.266.$$

3. Generate values for the error term ε by randomly drawing from a normal (0,10) distribution. The variance of ε is approximately equal to Anselin's (1988) estimate.
4. Vary ρ over the interval [0.1, 0.9] using increments of 0.2. Observe that $\rho = 0$ is the non-spatial benchmark model.
5. Compute y according to (17) on the basis of Anselin's (1988) first-order queen contiguity matrix, inverse distance matrix and inverse distance squared matrix, respectively.

[5] First-order and higher order contiguity models with equal number of neighbours and inverse distance models are nested within the class of latent variables models. For the first case each λ_i in the measurement model is fixed at $1/n$ where n is the number of neighbours. For the inverse distance model each λ_i is fixed at $1/n$ if the observed value for each neighbour is weighted by the inverse of distance. Contiguity models with unequal numbers of neighbours are not nested within the latent variables model.

[6] The regression coefficients are chosen in line with Anselin's estimates, since these values are known from the literature. Other values could have been chosen, however.

For the data set thus generated we estimate both the classical spatial lag model and the latent model. The classical models are estimated on the basis of a first-order contiguity matrix, inverse distance matrix and inverse distance squared matrix for each set of samples. For the latent model the first six nearest neighbors are taken as indicators of the latent dependence variable.

The number of replications is 1,000. The estimates are evaluated in terms of bias and RMSE of the spatial lag parameter ρ and the main regression coefficients of interest, β_1 and β_2.

4 Simulation results

Table 1 and Fig. 1 report the biases of the estimators of ρ, β_1 and β_2 of the three classical models and the latent variables model based on samples generated by first-order contiguity weights matrix. The comparison involves three types of classical models—Ccont (estimated on the basis of the first-order contiguity matrix), Cdinv (estimated on the basis of the inverse distance matrix), Cdinv2 (estimated on the basis of the inverse distance squared matrix)—and one latent model—Latent6n (estimated on the basis of the first six nearest neighbors as indicators of the latent spatial dependence variable).

- From Table 1 and Fig. 1a it follows that when $\rho = 0$, the 'true' model—Ccont—has the smallest bias for ρ. For $0 < \rho < 0.5$ this holds for Cdinv and for $\rho \geq 0.5$ for Cdinv2. Latent6n is outperformed by its three alternatives for$\rho < 0.5$. For $\rho \geq 0.5$, however, it outperforms Ccont but not the other two alternatives. For Latent6n we observe a constantly descending trend of the bias.
- Table 1 and Fig. 1b also show that for β_1 Cdinv2 has the smallest bias except for $\rho = 0$ and $\rho = 0.1$ when Ccont and Cdinv have the smallest bias, respectively. For $\rho = 0.3$ Latent6n has smaller bias than Ccont but is outperformed by the two alternative classical models. The bias of Latent6n (in absolute value) follows a U curve with minimum at $\rho = 0.3$.
- Table 1 and Fig. 1c show that for β_2 Ccont has the smallest bias everywhere except for $\rho = 0$ when Cdinv has the smallest bias and for $\rho = 0.1$ when this holds for Latent6n. Moreover, Latent6n outperforms Cdinv and Cdinv2 for $\rho = 0.3$ and 0.9.

It follows from the above that there are several cases where the latent model has smaller bias than one or several of the classical models. Particularly, Latent6n outperforms the "true" or correctly specified Ccont more frequently than the other two misspecified classical models.

The RMSEs of the estimators are shown in Table 2. It shows that there are only very small differences between the four models for each parameter. Figure 2 presents total RMSE, i.e., the sum of the RMSEs of all three parameters. The figure shows that the classical models outperform the latent model. One reason for this is probably that there are more parameters and thus bigger standard errors in the latent model than in its alternatives. Among the classical models the "true" Ccont outperforms the other classical models for $\rho \leq 0.5$. However, Cdinv2 has the smallest RMSE when $\rho > 0.5$, closely followed by Cdinv. Notice that when ρ increases, the RMSE of Latent6n decreases somewhat more than the RMSE of Ccont.

Table 1 Bias of the estimators for samples generated on the basis of the first-order contiguity matrix

	ρ				β_1				β_2			
ρ	Ccont	Cdinv	Cdinv2	Latent6n	Ccont	Cdinv	Cdinv2	Latent6n	Ccont	Cdinv	Cdinv2	Latent6n
0.0	−0.0593	−0.0806	−0.0765	−0.0871	−0.0291	−0.0367	−0.0347	−0.0659	0.0016	0.0011	0.0013	0.0031
0.1	−0.0606	−0.0521	−0.0528	−0.0807	−0.0329	−0.0243	−0.0253	−0.0373	0.0013	0.0019	0.0019	0.0009
0.3	−0.0603	−0.0039	−0.0117	−0.0644	−0.0414	0.0032	−0.0029	0.0173	0.0006	0.0037	0.0033	−0.0028
0.5	−0.0561	0.0254	0.0158	−0.0466	−0.0518	0.0322	0.0224	0.0631	0.0001	0.0052	0.0046	−0.0061
0.7	−0.0473	0.0313	0.0242	−0.0330	−0.0659	0.0554	0.0447	0.0998	0.0001	0.0055	0.0050	−0.0084
0.9	−0.0329	0.0104	0.0081	−0.0128	−0.0943	0.0409	0.0339	0.1459	0.0008	0.0029	0.0027	−0.0022

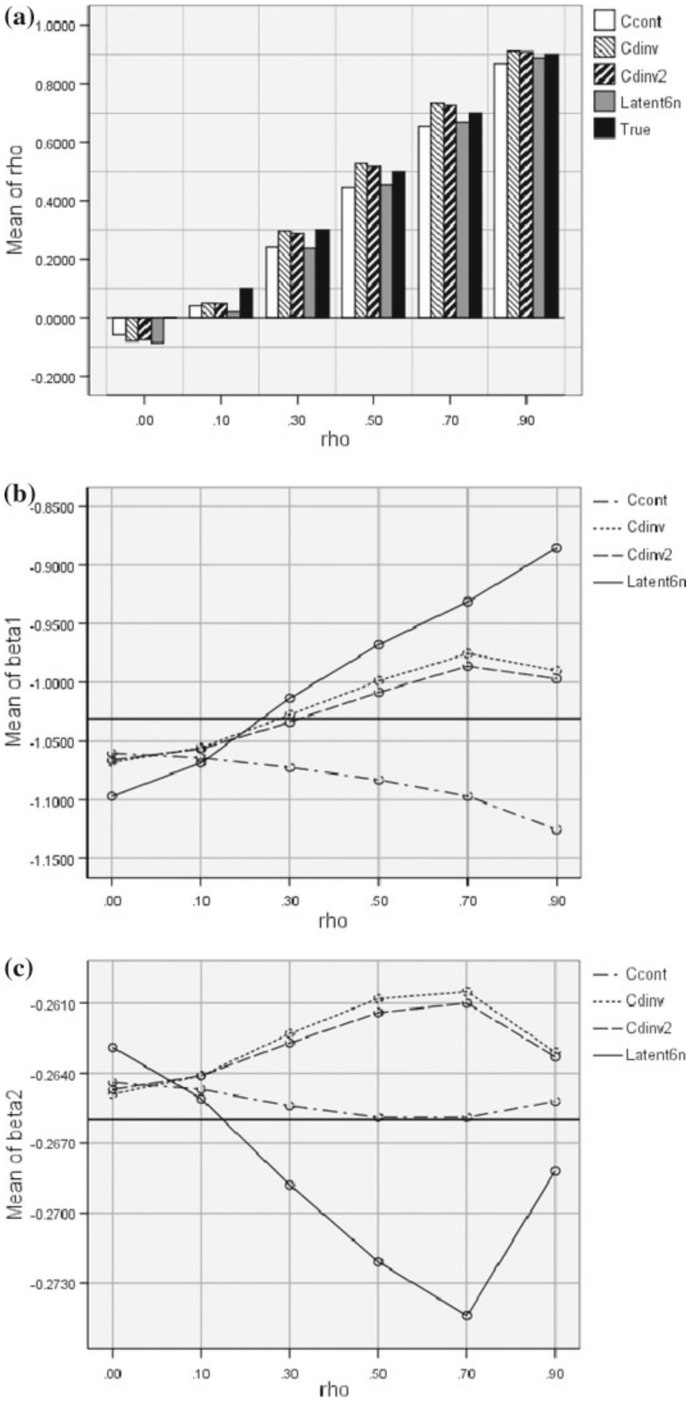

Fig. 1 Bias of the estimators of ρ (**a**), β_1 (**b**) and β_2 (**c**) for the classical models (Ccont, Cdinv, Cdinv2) and the latent model (Latent6n) for samples generated by first-order contiguity matrix

Table 2 RMSE of the estimators for samples generated on the basis of the first-order contiguity matrix

	ρ				β_1				β_2			
ρ	Ccont	Cdinv	Cdinv2	Latent6n	Ccont	Cdinv	Cdinv2	Latent6n	Ccont	Cdinv	Cdinv2	Latent6n
0.0	0.1397	0.1849	0.1772	0.1888	0.2496	0.2577	0.2554	0.2797	0.0721	0.0725	0.0723	0.0733
0.1	0.1352	0.1695	0.1636	0.1788	0.2515	0.2568	0.2558	0.2738	0.0722	0.0723	0.0723	0.0735
0.3	0.1220	0.1402	0.1368	0.1524	0.2556	0.2578	0.2571	0.2700	0.0725	0.0721	0.0722	0.0742
0.5	0.1029	0.1119	0.1092	0.1204	0.2605	0.2611	0.2601	0.2761	0.0728	0.0718	0.0719	0.0759
0.7	0.0774	0.0790	0.0773	0.0849	0.2660	0.2654	0.2641	0.2816	0.0731	0.0714	0.0716	0.0786
0.9	0.0437	0.0345	0.0345	0.0441	0.2736	0.2596	0.2595	0.2934	0.0735	0.0715	0.0716	0.0743

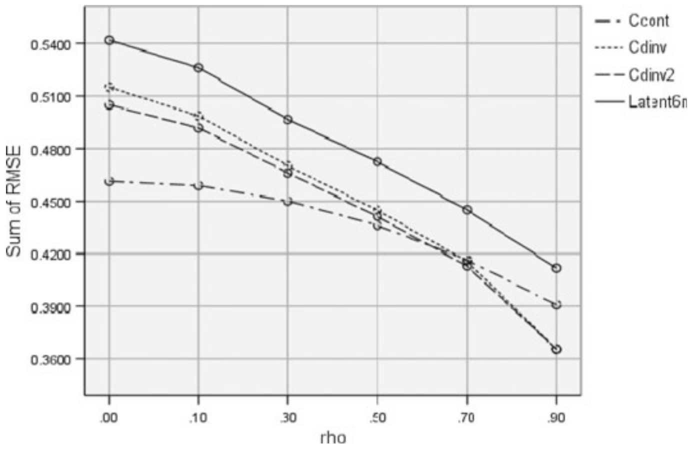

Fig. 2 Total RMSE for the four models for samples generated by first-order contiguity matrix

Taking bias and RMSE together, it follows that Ccont and Cdinv2 perform better than Cdinv and Latent6n.

The results for samples generated by inverse distance weights matrix are summarized in Tables 3 and 4 and shown in Figs. 3 and 4. From Table 3 it follows that all the biases are negative for all four models and all parameters. Moreover, Cdinv, as the 'correct' model, outperforms the other classical models and the latent model most of the time in terms of bias of all three parameters.

Figure 3a shows that for ρ Ccont has the smallest bias when $\rho = 0$ and Cdinv2 is the best when $\rho = 0.1$. Other than that, Cdinv outperforms the other three models. Latent6n is better than Ccont when $\rho = 0.9$. With regard to the bias of β_1, it can be concluded from Fig. 3b that the 'winners' are the same as in the case of ρ except for $\rho = 0.9$ where Latent6n has the lowest bias. Moreover, Latent6n also outperforms Ccont when $\rho \geq 0.5$. From Fig. 3c it follows that Cdinv is best again and Latent6n is outperformed by Ccont except for $\rho = 0.1$ when their biases are equal.

Table 4 gives the RMSEs for ρ, β_1 and β_2 for the four models. As in the case of samples based on the first-order contiguity matrix, the differences between the four models are very small. The latent model sometimes outperforms the classical models. For instance, for $\rho = 0.1$, Latent6n has smaller RMSE for ρ than Cdinv and it outperforms both Cdinv and Cdinv2 when $\rho = 0$. Also for values of $\rho \geq 0.7$ Latent6n outperforms Ccont for β_1.

Figure 4 presents the total RMSE of all three parameters. It follows that total RMSE is smallest for Ccont for the first three values of ρ while Cdinv leads for the next three values of ρ, closely followed by Cdinv2. However, unlike the previous case where samples are generated by first-order contiguity matrix, the classical models do not uniformly outperform the latent approach. Particularly, Latent6n beats Cdinv when $\rho = 0$ and outperforms Ccont when $\rho = 0.9$.

Considering both bias and RMSE, the "true" model Cdinv outperforms its three alternatives most of time. Furthermore, Cdinv2 and Ccont perform about equally well.

Table 3 Bias of the estimators for samples generated by inverse distance matrix

	ρ				β_1				β_2			
ρ	Ccont	Cdinv	Cdinv2	Latent6n	Ccont	Cdinv	Cdinv2	Latent6n	Ccont	Cdinv	Cdinv2	Latent6n
0.0	−0.0554	−0.0955	−0.0847	−0.0871	−0.0387	−0.0597	−0.0539	−0.0659	0.0001	−0.0018	−0.0013	0.0031
0.1	−0.0869	−0.0850	−0.0838	−0.1268	−0.0638	−0.0597	−0.0594	−0.0867	−0.0021	−0.0018	−0.0018	−0.0021
0.3	−0.1216	−0.0639	−0.0724	−0.1685	−0.1117	−0.0591	−0.0669	−0.1154	−0.0062	−0.0017	−0.0024	−0.0166
0.5	−0.1180	−0.0434	−0.0527	−0.1569	−0.1516	−0.0580	−0.0695	−0.1138	−0.0095	−0.0016	−0.0025	−0.0145
0.7	−0.0805	−0.0243	−0.0295	−0.1073	−0.1743	−0.0561	−0.0670	−0.0695	−0.0111	−0.0013	−0.0022	−0.0137
0.9	−0.0228	−0.0074	−0.0082	−0.0186	−0.1518	−0.0535	−0.0587	−0.0412	−0.0088	−0.0009	−0.0013	−0.0123

Table 4 RMSE of the estimators for samples generated by inverse distance matrix

	ρ				β_1				β_2			
ρ	Ccont	Cdinv	Cdinv2	Latent6n	Ccont	Cdinv	Cdinv2	Latent6n	Ccont	Cdinv	Cdinv2	Latent6n
0.0	0.1345	0.2186	0.1998	0.1888	0.2530	0.2728	0.2677	0.2797	0.0718	0.0720	0.0719	0.0733
0.1	0.1403	0.1971	0.1854	0.1955	0.2580	0.2736	0.2700	0.2820	0.0718	0.0720	0.0719	0.0729
0.3	0.1468	0.1531	0.1504	0.1964	0.2717	0.2749	0.2737	0.2868	0.0720	0.0720	0.0720	0.0745
0.5	0.1306	0.1085	0.1094	0.1702	0.2881	0.2760	0.2764	0.2909	0.0723	0.0720	0.0720	0.0741
0.7	0.0876	0.0642	0.0653	0.1129	0.3006	0.2768	0.2778	0.2867	0.0724	0.0719	0.0719	0.0731
0.9	0.0271	0.0210	0.0212	0.0343	0.2969	0.2773	0.2779	0.2817	0.0722	0.0719	0.0719	0.0724

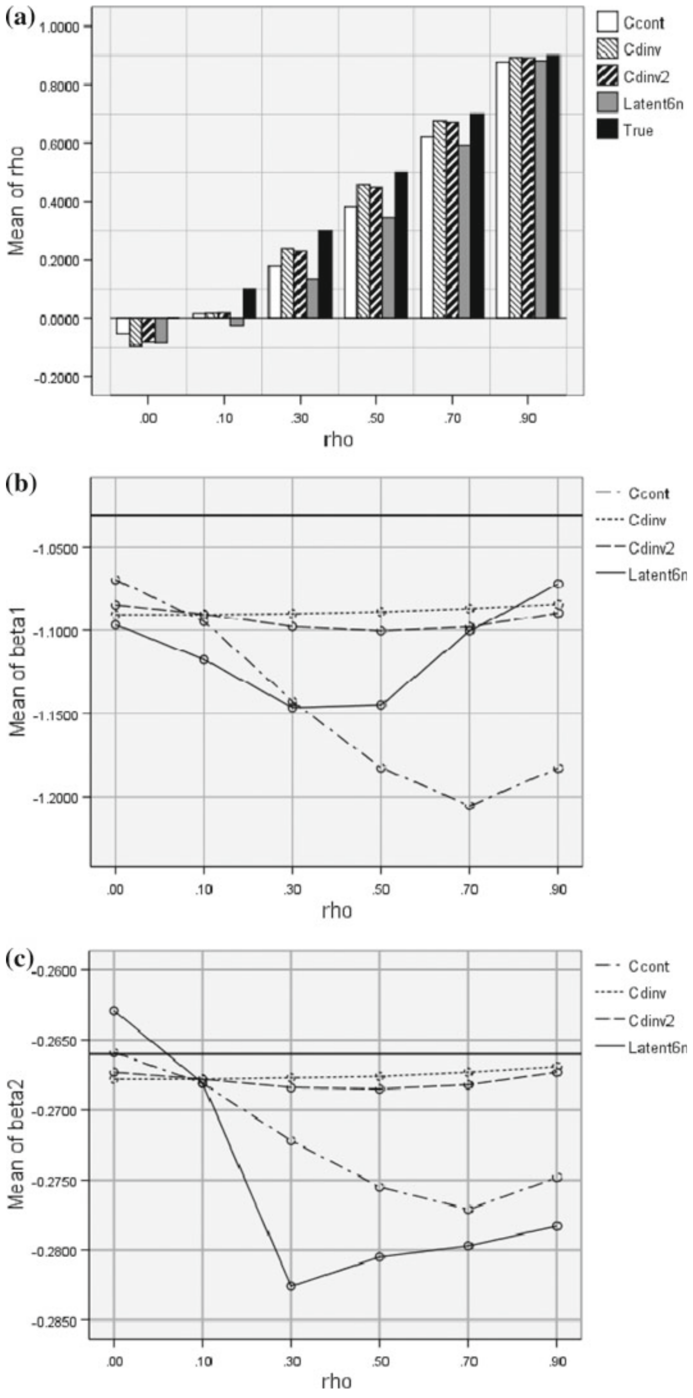

Fig. 3 Bias of the estimators of ρ (**a**), β_1 (**b**) and β_2 (**c**) for the classical models (Ccont, Cdinv, Cdinv2) and the latent model (Latent6n) for samples generated by inverse distance matrix

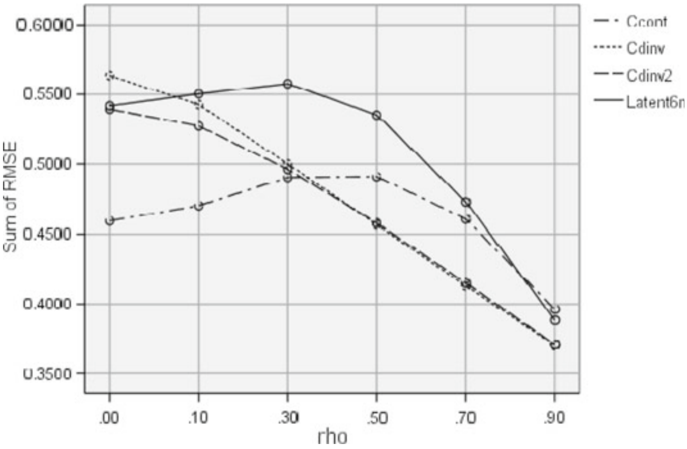

Fig. 4 Total RMSE for the four models for samples generated by inverse distance matrix

The latent model trails the “true” model, and, to a lesser extent, the misspecified classical models. The differences are very small, however.

Table 5 gives the bias for samples generated by the inverse distance squared matrix. With a few exceptions Cdinv performs best, closely followed by the ‘correct’ model Cdinv2. Figure 5a–c show that Ccont has the lowest bias for ρ and β_1 when $\rho = 0$ and for β_2 when $\rho = 0.5$ and 0.7. There are several cases where Latent6n outperforms Ccont: for ρ when $\rho \geq 0.7$, for β_1 when $\rho \geq 0.5$ and for β_2 when $\rho = 0.9$.

The RMSE results are shown in Table 6. As in the previous cases, for each parameter there are only slight differences between the models. Figure 6 which represents the total RMSE summed over all estimators, shows that Ccont has the smallest total RMSE for the first two values of ρ while Cdinv and Cdinv2 produce better and almost identical results for $\rho \geq 0.3$. For $\rho \geq 0.7$ the total RMSE of four models strongly converge.

Considering bias and RMSE together we conclude that Cdinv and Cdinv2 both perform better than Ccont and Latent6n in the case samples are generated on the basis of the inverse distance squared matrix. Latent6n trails the classical approaches, though the differences are small.

The above results show that the “true” model tends to outperform the misspecified alternatives. This applies especially to Latent6n. However, the differences between Latent6n and the classical models tend to be very small. Moreover, in several cases Latent6n performs best or is the best among the misspecified models.

5 Conclusions

This paper evaluates the performance of estimators of two types of spatial autoregressive models: the classical W-based spatial autoregressive model and the structural equations model with latent variables. The former accounts for spatial dependence and spillover effects by means of a spatial weights matrix W and the latter by means

Table 5 Bias of the estimators for samples generated by inverse distance squared matrix

	ρ				β_1				β_2			
ρ	Ccont	Cdinv	Cdinv2	Latent6n	Ccont	Cdinv	Cdinv2	Latent6n	Ccont	Cdinv	Cdinv2	Latent6n
0.0	−0.0353	−0.0494	−0.0463	−0.0871	−0.0307	−0.0397	−0.0377	−0.0659	0.0007	−0.0001	0.0001	0.0031
0.1	−0.0489	−0.0410	−0.0419	−0.1029	−0.0457	−0.0368	−0.0379	−0.0935	−0.0002	0.0002	0.0001	−0.0018
0.3	−0.0595	−0.0275	−0.0317	−0.0939	−0.0705	−0.0322	−0.0372	−0.0951	−0.0011	0.0007	0.0005	−0.0144
0.5	−0.0512	−0.0173	−0.0211	−0.0573	−0.0854	−0.0289	−0.0352	−0.0429	−0.0005	0.0013	0.0011	−0.0104
0.7	−0.0304	−0.0092	−0.0110	−0.0230	−0.0860	−0.0265	−0.0315	0.0311	0.0014	0.0019	0.0019	−0.0099
0.9	−0.0072	−0.0028	−0.0028	−0.0029	−0.0630	−0.0239	−0.0258	0.0283	0.0036	0.0028	0.0028	−0.0031

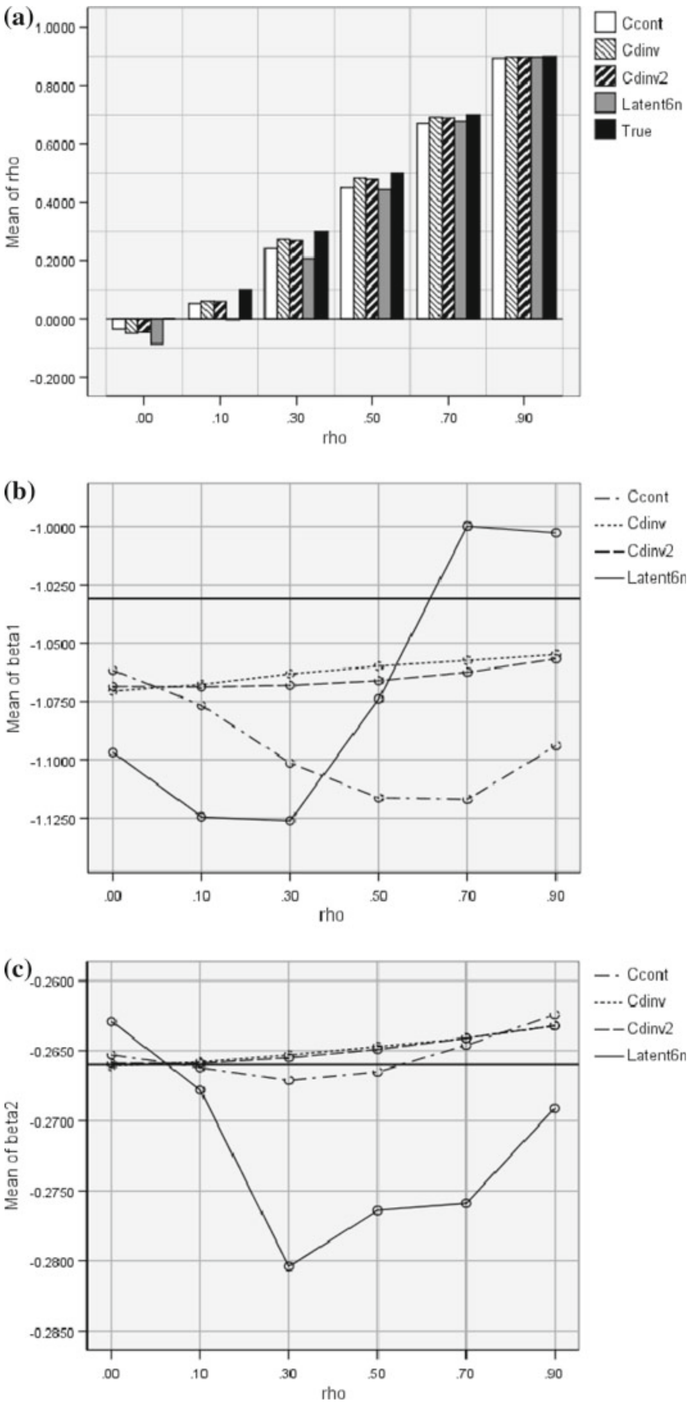

Fig. 5 Bias of the estimators of ρ (**a**), β_1 (**b**) and β_2 (**c**) for the classical models (Ccont, Cdinv, Cdinv2) and the latent model (Latent6n) based on samples generated by inverse distance squared matrix

Table 6 RMSE of the estimators for samples generated by inverse distance squared matrix

	ρ				β_1				β_2			
ρ	Ccont	Cdinv	Cdinv2	Latent6n	Ccont	Cdinv	Cdinv2	Latent6n	Ccont	Cdinv	Cdinv2	Latent6n
0.0	0.1144	0.1469	0.1410	0.1888	0.2587	0.2686	0.2666	0.2797	0.0717	0.0717	0.0717	0.0733
0.1	0.1101	0.1299	0.1266	0.1778	0.2614	0.2683	0.2671	0.2839	0.0717	0.0717	0.0717	0.0729
0.3	0.0967	0.0969	0.0966	0.1449	0.2665	0.2675	0.2672	0.2843	0.0718	0.0717	0.0717	0.0733
0.5	0.0741	0.0658	0.0663	0.1044	0.2701	0.2661	0.2664	0.2773	0.0719	0.0717	0.0717	0.0729
0.7	0.0439	0.0369	0.0373	0.0479	0.2701	0.2640	0.2644	0.2738	0.0719	0.0718	0.0718	0.0725
0.9	0.0126	0.0112	0.0112	0.0131	0.2646	0.2607	0.2609	0.2717	0.0721	0.0721	0.0721	0.0723

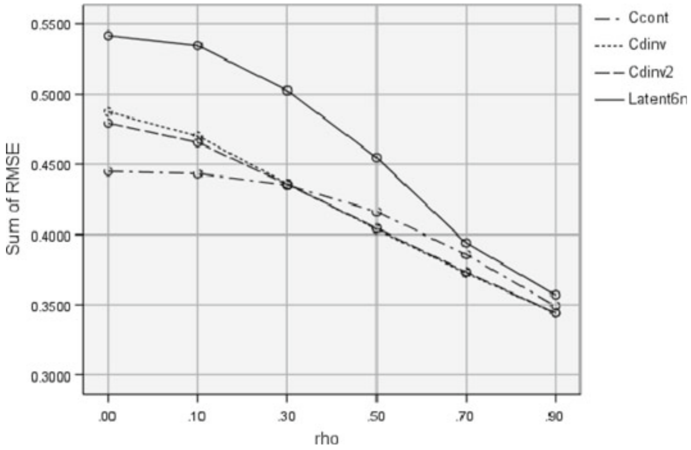

Fig. 6 Total RMSE for the four models for samples generated by inverse distance squared matrix

of a latent variable in the structural model while the relationships between observed spatially lagged variables and the latent spatial dependence variable(s) are given in the measurement model. The two classes of approaches are compared by means of Monte Carlo simulations based on Anselin's Columbus, Ohio, crime data set in terms of bias and RMSE of the coefficient estimators for various values of the spatial lag parameter and different types of weight matrices. Data are generated on the basis of exogenous variable values, parameter values and the first-order contiguity matrix, inverse distance matrix and inverse distance squared matrix, as given by Anselin's (1988).

The simulation results show that the "true" model, i.e. the model whose W matrix is the same as used for data generation, tends to perform best in terms of bias and RMSE. However, the differences between the "true" and the misspecified models including the structural equation model, are very small. Moreover, for several combinations of the value of spatial autoregressive parameter and the weights matrix used in data generation the latter produces results with the smallest bias or RMSE. In other words, the "true" models do not uniformly outperform the latent variables approach.

The fact that only the W-based model structure has been used to generate data is strongly in favor of the W-based estimators, since it implies that in all cases the latent model is misspecified. The reason for analyzing the performance of the latent model on the basis of data generated by means of a W-based model is that we wanted to get insight into the performance of the latent variables approach in the case of misspecification. Subsequent simulations will explore the behavior of both approaches under level playing field conditions.

It should also be noted that in contrast to the classical models, not all model specification options of the latent variable approach have been fully expolited in the simulations. Particularly, in order to keep the simulations simple, the number of observed spatially lagged variables was a priori fixed at six, since in Folmer and Oud (2008) this number gave parameter estimates virtually identical to those obtained by Anselin's (1988). In the practice of structural equation modelling the optimal number of indicators, i.e. observed spatial lags, which may differ by e.g. type of W matrix, can be identified by means of tests. Another unexploited specification search is the possibility of covariances among the error terms in the measurement model (Θ).

To sum up, given the small differences between the latent variables approach and the "true" classical alternatives with respect to bias and RMSE, its unexploited specification search options, its flexibility and information content, it is worthwhile further considering its suitability to model spatial dependence.

References

Aldstadt J, Getis A (2004) Constructing the spatial weights matrix using a local statistic. Geogr Anal 36(2): 90–104

Aldstadt J, Getis A (2006) Using AMOEBA to create a spatial weights matrix and identify spatial clusters. Geogr Anal 38(4):327–343

Anselin L (1988) Spatial econometrics: methods and models. Kluwer, Dordrecht

Anselin L (2002) Under the hood: issues in the specification and interpretation of spatial regression models. Agric Econ 27(3):247–267

Bhattacharjee A, 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

Case AC, Rosen S, Hines JR Jr (1993) Budget spillovers and fiscal policy interdependence: evidence from the states. J Public Econ 52(3):285–307

Fingleton B (2003) Externalities, economic geography and spatial econometrics: conceptual and modeling developments. Int Reg Sci Rev 26(2):197–207

Florax R, Folmer H (1992) Specification and estimation of spatial linear regression models: Monte Carlo evaluation of pre-test estimators. Reg Sci Urban Econ 22:405–432

Florax RJGM, Folmer H, Rey SJ (2003) Specification searches in spatial econometrics: the relevance of Hendry's methodology. Reg Sci Urban Econ 33(5):557–579

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

Getis A, Griffith DA (2002) Comparative spatial filtering in regression analysis. Geogr Anal 34(2):130–140

Hepple LW (1995) Bayesian techniques in spatial and network econometrics: 1. Model comparison and posterior odds. Environ Plan A 27:447–469

Jöreskog KG, Sörbom D (1996) Lisrel 8: user's reference guide. Scientific Software International, Chicago

Neale MC, Boker SM, Xie G, Maes HH (2003) Mx: Statistical Modeling, 6th edn. VCU Box 900126, Richmond, VA 23298: Department of Psychiatry

Ord JK (1975) Estimation methods for models of spatial interaction. J Am Stat Assoc 70:120–126

Oud JHL, Folmer H (2008) A structural equation approach to models with spatial dependence. Geogr Anal 40:152–166

Tiefelsdorf M, Griffith DA (2007) Semiparametric filtering of spatial autocorrelation: the eigenvector approach. Environ Plan A 39(5):1193–1221