Estimators of Binary Spatial Autoregressive Models: A Monte Carlo Study

Raffaella Calabrese
University of Milano-Bicocca
raffaella.calabrese1@unimib.it

Johan A. Elkink*
University College Dublin
jos.elkink@ucd.ie

February 15, 2013

Abstract

The goal of this paper is to provide a cohesive description and a critical comparison of the main estimators proposed in the literature for spatial binary choice models. The properties of such estimators are investigated using a theoretical and simulation study. To the authors' knowledge, this is the first paper that provides a comprehensive Monte Carlo study of the estimators' properties. This simulation study shows that the Gibbs estimator (LeSage, 2000) performs best for low spatial autocorrelation, while the Recursive Importance Sampler (Beron and Vijverberg, 2004) performs best for high spatial autocorrelation. The same results are obtained by increasing the sample size. Finally, the linearized General Method of Moments estimator (Klier and McMillen, 2008) is the fastest algorithm that provides accurate estimates for low spatial autocorrelation and large sample size.

1 Introduction

In applied work in economics and political science, there is increased attention to the importance of spatial or network interdependence between observations. Not only does this violate the assumption of independence underlying many econometric methodologies for cross-sectional data, there is also growing interest in estimating the strength of the interdependence itself. While the econometric literature on linear regression models with spatial interdependence is well established, in particular since the publication of Anselin (1988)'s seminal work, the literature on regression models with binary dependent variables and spatial interdependence is still relatively limited.

Many applications with such models can be considered – including the contagion of currency crises (Novo, 2003), firm-level decision-making on locations (Autant-Bernard, 2006), ecological studies of spatial distributions of plants (Collingham et al., 2000), studies in policy diffusions of flat taxes (Baturo and Gray, 2009), anti-smoking laws (Shipan and Volden, 2006) or pension privatization (Weyland, 2007) – across academic disciplines such as economics, political science, sociology, ecology, planning, or even neurology.

*Corresponding author.

Proximity in this context can be interpreted in a broad manner. Whether one defines proximity in a physical or in a cultural or interaction sense, or in a manner that encompasses large distances or the entire space (all units affect all other units), the estimation challenges discussed in this article still hold.[1] The conclusions can thus be directly applied to social network analysis as well as spatial econometric analysis. Anselin (2002, 255) refers to this perspective as the *object view* or this type of data as *lattice data*. The alternative, a *geostatistics* perspective, where we observe only specific monitoring sites and space is seen as a continuous space or a point pattern (Bivand, 1998), leads to an entirely different econometric framework and will not be discussed in his article.

Spatial econometric models raise new difficulties that cannot be dealt with by standard econometric models. Estimation problems arise due to the dependence across observations, in that we must adjust the estimation procedures for the loss of information associated with dependent observations. Indeed, in the presence of spatial dependence, standard logit or probit estimation procedures, which assume independence, result in inconsistent and inefficient estimates (McMillen, 1992). In particular, McMillen (1992) notes that both the spatially dependent error model and the spatial lag model imply heteroskedastic disturbances, which cause the parameter estimates to be inconsistent. For these reasons econometricians began to pay more attention to spatial dependence problems in the last two decades and some important advances have been made in both theoretical and empirical studies (Anselin, Florax and Rey, 2004).

The aim of this article is to compare the main estimators proposed in the literature for estimating the spatial autocorrelation parameter in binary choice models. On the one hand, this goal is achieved by analysing the theoretical characteristics of the main estimators for spatial models for binary response data. This topic has been in part developed by Fleming (2004) but we consider also the recent literature. Moreover, our paper is focused only on binary choice models, instead Fleming (2004) has considered discrete choice models. On the other hand, the most innovative aspect of this work is the comparison of the above-mentioned estimators by Monte Carlo simulations. To our knowledge, this is the first work that performs Monte Carlo simulations on the main estimators of the spatial autocorrelation parameter for binary response data. The importance and the necessity of this analysis is strongly suggested by Fleming (2004).

Currently, the most used methodologies available to estimating spatial regression models are five. McMillen (1992) proposes an EM algorithm based estimation procedure. In particular, McMillen (1992) replaces the latent continuous variable with its expected value and then applies the maximum likelihood method (Ord, 1975). Similarly to McMillen (1992), LeSage (2000) also replaces the latent continuous variable with its expected value, solving thereafter a spatial continuous model using the Gibbs sampling approach. Following the work of Vijverberg (1997) on the simulation from a multivariate normal distribution, Beron and Vijverberg (2004) suggests to apply the recursive importance sampling (RIS) to the maximum likelihood method, since the likelihood function is a multivariate normal distribution. Pinkse and Slade (1998) develop a model based on the generalised method of moments (GMM). Klier and McMillen (2008) linearize Pinkse and Slade (1998)'s model around a convenient starting point.

The present paper is organized as follows. The next section reviews the widely used specifications of the binary choice models with spatial dependence. In section 3 we analyse and compare the main methodologies proposed in the literature to estimate the spatial autocorre-

[1] The complications with the interpretations of the observed effects increase, however, since factors that affect the similarity are also likely to have an effect on the linkages between the units. See Shalizi and Thomas (2010) for a discussion of the inherent confounding of homophily and contagion mechanisms.

lation parameter in binary response models. In section 4 we compare the properties of these estimators by Monte Carlo simulations. The last section concludes.

2 Spatial binary choice models

A widely used representation of a regression model for an observed dichotomous response Y_i is the latent response model (Verbeek, 2008, p.180) with dependent variable the continuous variable Y_i^*, whereby

$$Y_i = \begin{cases} 1, & Y_i^* > 0 \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

with $i = 1, 2, \ldots, n$. A linear model is specified for this latent response, so the model specification is

$$\begin{aligned} \mathbf{Y}^* &= \rho \mathbf{W}\mathbf{Y}^* + \mathbf{X}\boldsymbol{\beta} + \mathbf{d} \\ \mathbf{d} &= \lambda \mathbf{S}\mathbf{d} + \boldsymbol{\epsilon}, \end{aligned} \tag{2}$$

where $\mathbf{Y}^*$ is a continuous random vector, $\mathbf{X}$ represents an $n \times k$ matrix of explanatory variables, the error term $\boldsymbol{\epsilon}$ can follow a multivariate normal distribution in a probit model or a multivariate logistic distribution in a logit model. $\mathbf{W}$ and $\mathbf{S}$ are spatial lag and spatial error weights matrices, respectively, ρ and λ the associated scalar parameters. We highlight that only the latent variable can be used for the spatial lag, since both the models $\mathbf{Y}^* = \rho \mathbf{W}\mathbf{Y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and $\mathbf{Y} = \rho \mathbf{W}\mathbf{Y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ are infeasible (Anselin, 2002; Beron and Vijverberg, 2004; Klier and McMillen, 2008).

Evidence of the absence of a consolidated literature is given by the different denominations of the models – we follow LeSage (2000)'s notation. From the general model (2) two models are derived. Setting $\mathbf{S} = \mathbf{0}$ produces a spatial lag model, which we will refer to as the Binary Spatial AutoRegressive model (BSAR):

$$\mathbf{Y}^* = (I - \rho \mathbf{W})^{-1}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = (\mathbf{I} - \rho \mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{3}$$

where

$$\mathbf{e} = (\mathbf{I} - \rho \mathbf{W})^{-1}\boldsymbol{\epsilon}. \tag{4}$$

Letting $\mathbf{W} = \mathbf{0}$ results in a regression model with spatial autocorrelation in the disturbances, a spatial error model which we label the Binary Spatial Error Model (BSEM):

$$\mathbf{Y}^* = \mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \lambda \mathbf{S})^{-1}\boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

where

$$\mathbf{u} = (\mathbf{I} - \lambda \mathbf{S})^{-1}\boldsymbol{\epsilon}.$$

The two models are based on different assumptions about the causes of the spatial dependence.[2] The spatial lag relates to an explicit spillover effect where one agent copies behavior from neighboring agents. It also relates to a theoretical model where the behavior is dependent on shared resources between different agents. The spatial error model concerns different causal relationships. For example, a typical issue that leads to spatial correlation in the errors is a mismatch between the spatial delineation of the measurement and the empirical presence of the

[2]For a clear interpretation of the spatial lag and spatial error models, see Case (1992).

variable of interest. For example, when studying the presence of a particular natural resource in particular countries, the geographical zones in which this resource is present do not usually match exactly with the country borders. A measurement of the presence of these resources in countries is thus necessarily spatially correlated, but as a nuisance rather than in a theoretically interesting sense. Another common cause of spatial autocorrelation in the errors is an omitted variable that is itself spatially correlated. In terms of estimation, the two types of autocorrelation are often difficult to distinguish (Brueckner, 2003, 184-185). The different theoretical mechanisms are of course not mutually exclusive and a spatial model that incorporates both a spatial lag and spatial residuals is perfectly reasonable.

In this paper we are primarily interested in estimating diffusion effects, and thus our focus is on the estimation of the spatial autocorrelation parameter ρ. For this reason, in this work we only analyse models with spatial lags (BSAR) and leave spatial errors (BSEM) aside.

The contiguity or weight matrix $\mathbf{W}$ is defined by

$$w_{ij} = \begin{cases} 1 & \text{if the } i\text{-th and } j\text{-th observations are contiguous;} \\ 0 & \text{if } i = j \text{ or the } i\text{-th and } j\text{-th observations are not continguous,} \end{cases}$$

so it is a square matrix of order n and its main diagonal elements equal to zero. Contiguity can refer to geographical and alternative vicinity. The use of the weight matrix $\mathbf{W}$ implies that the spatial sites form a countable lattice (Lee, 2004), but part of the literature considers a continuous spatial index (Conley, 1999). Because of the potential of heteroscedasticity due to the variation in the number of neighbors for different observations, $\mathbf{W}$ is commonly normalized as follows $w_{ij}/(\sum_j w_{ij})$ for $i, j = 1, 2, \ldots, n$. This means that the normalized matrix $\mathbf{W}$ is generally asymmetric, while the original weight matrix $\mathbf{W}$ is often symmetric.[3] Although this is the common approach, there are various other ways of defining and normalizing $\mathbf{W}$ (Tiefelsdorf, 2000; Anselin, 2002). Since the aim of this paper is the comparison of the main methodologies to estimate the spatial autocorrelation parameter ρ, we consider for all of them the normalized matrix $\mathbf{W}$.[4]

For binary dependent variables, the most used models are the logistic and the probit models (McCullagh and Nelder, 1989). In the next section we analyse and compare the main estimators of the autocorrelation parameter in both spatial probit (Beron and Vijverberg, 2004; LeSage, 2000; McMillen, 1992) and logit (Klier and McMillen, 2008) models.[5]

[3] Novo (2003) considers an asymmetric non-normalized $\mathbf{W}$.

[4] When the normalized contiguity matrix $\mathbf{W}$ is considered, to ensure the invertibility of the matrix $(I - \rho\mathbf{W})$ in the maximum likelihood method, Anselin (1982) proves that $1/\omega_{min} < \rho < 1$ where ω_{min} is the minimum eigenvalue of the contiguity matrix $\mathbf{W}$.

[5] There are a number of related estimators that, for various reasons, will not be included in the discussion and Monte Carlo analyses. These estimators are related, but make assumptions about the data that are beyond the scope of this paper. For the spatial random effects probit (Case, 1991, 1992), when $\mathbf{W}$ is constrained to be block-diagonal, in other words, when the focus is on membership of a particular geographic region or cluster of units rather than some kind of proximity measure, the spatial model can be substantially simplified (Case, 1991, 1992). The logistic auto-logistic (Gumpertz, Graham and Ristaino, 1997; Bee and Espa, 2008) applies to data on a regular grid, which is not applicable in the type of diffusion studies we have in mind in this paper. Dubin (1995)'s spatial logit model is a straightforward diffusion model that avoids most complications of spatial models by using the temporally lagged, realized dependent variable to create the spatial lag. McMillen (1992, 1995*b*)'s heteroscedastic probit using weighted least squares applies to the spatial error model, but not the spatial autoregressive model we discuss in this paper.

3 Estimators for binary spatial autoregressive models

3.1 Expectation-Maximization algorithm

The Expectation-Maximization (EM) algorithm is designed for cases where the data is incomplete, for example due to missing values (Dempster, Laird and Rubin, 1977). Since the probit model can be viewed as a latent response model, and this latent variable is similarly unobserved, McMillen (1992) proposes to apply the EM algorithm to the probit model with spatially lagged dependent variables and spatial error autocorrelation. In particular, the latent unobserved observations y_i^* are replaced by estimated values. Given estimates of the values y_i^*, the EM algorithm proceeds to estimate the other parameters in the model using the maximum likelihood method.

In the EM algorithm the assumption of homogeneity for the disturbances $\boldsymbol{\epsilon}$ is introduced. This means that the error term $\boldsymbol{\epsilon}$ can follow the n-dimensional multivariate normal distribution $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma_{\boldsymbol{\epsilon}}^2 \mathbf{I})$ in a probit model. The variance of the error term is indeed

$$var(\mathbf{e}) = var\left[(\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}\right] = \sigma_{\boldsymbol{\epsilon}}^2\left[(\mathbf{I} - \rho\mathbf{W})'(\mathbf{I} - \rho\mathbf{W})\right]^{-1}. \tag{5}$$

Let

$$\mathbf{D} = diag(\boldsymbol{\sigma}_{\mathbf{e}}) \tag{6}$$

be the diagonal matrix with diagonal elements $\boldsymbol{\sigma}_{\mathbf{e}}$ that represent the root square of the diagonal elements in the matrix (5) and

$$\mathbf{q} = \mathbf{D}^{-1}(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}. \tag{7}$$

Since $\boldsymbol{\beta}$ and σ_{ε}^2 cannot both be estimated in probit models, McMillen (1992) assumes $\sigma_{\varepsilon}^2 = 1$. In the E-step, the observed dependent variable is replaced by the expectation of the latent variable $\mathbf{Y}^*$ conditional on the observed dependent variable $\mathbf{Y}$, making use of generalized residuals (Cox and Snell, 1968; Chesher and Irish, 1987). To compute this expectation in the first iteration, the starting values of the parameters $\boldsymbol{\beta}$ and ρ are used, in subsequent iterations, the estimated parameters. By computing the conditional expectation of equation (3), in the E-step the following result is used

$$E[\mathbf{Y}^*/\mathbf{Y} = \mathbf{y}] = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{D}\frac{\phi_n(\mathbf{q})\left[\mathbf{y} - \Phi_n(\mathbf{q})\right]}{\Phi_n(\mathbf{q})\left[1 - \Phi_n(\mathbf{q})\right]}, \tag{8}$$

where $\phi_n(\cdot)$ and $\Phi_n(\cdot)$ denote respectively the n-dimensional multivariate probability density and cumulative distribution functions of a standard normal.

Subsequently, setting $\sigma^2 = 1$ in the M-step, new estimates are obtained by maximizing the log-likelihood function

$$k - \frac{1}{2}[(\mathbf{I} - \rho\mathbf{W})\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}]'[(\mathbf{I} - \rho\mathbf{W})\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}] + \sum_{i=1}^{n} ln(1 - \rho\omega_i),$$

where ω_i are the eigenvalues of $\mathbf{W}$. $\prod_{i=1}^{n}(1 - \rho\omega_i)$ is a computationally efficient approximation of the determinant $|\mathbf{I} - \rho\mathbf{W}|$ (Ord, 1975). This process is repeated until convergence.[6]

[6]To obtain a ρ estimate in the interval (-1,1), we apply the one-to-one transformation $\rho = -1 + 2\Phi_1(\rho^*)$, making use of the invariance of maximum likelihood estimators (Davidson and MacKinnon, 1993, p. 253–255).

The main advantage of this methodology is that it avoids to compute an n-dimensional integral. The cost is that the E-step requires the calculation of the inverse of the matrix $(\mathbf{I}-\rho\mathbf{W})$. Although this can be made slightly more efficient by using the eigenvalues of $\mathbf{W}$ to approximate the inverse, it still slows down the algorithm considerably. In addition to the computational burden in the implementation of the algorithm, the main drawback of this proposal is the covariance matrix estimate of dimensions $n \times n$. By considering the spatial probit model as a non-linear weighted least squares model, McMillen (1992) obtains biased but consistent estimates of the covariance matrix. For this reason McMillen (1995*a*) explores computationally simpler alternatives to the methods in McMillen (1992), expressing the belief that the methods proposed in McMillen (1992) are impractical for large sample sizes. Another problem with McMillen (1992)'s approach is the need to specify a functional form for the nonconstant variance over space (LeSage, 2000). In larger models a practitioner would need to devote considerable effort to testing the functional form and variables involved in the model for the variance of the noise elements ε_i. Finally, the EM approach cannot provide an estimate of precision for the spatial autoregressive parameter ρ.

3.2 Gibbs sampling

The Gibbs sampler is a particular Markov Chain Monte Carlo (MCMC) introduced by Geman and Geman (1984) in the context of image restoration. When a direct specification of a joint distribution is not feasible, the Gibbs sampling procedure specifies the complete conditional distributions for all parameters in the model and proceeds to sample from these distributions to collect a large set of parameter draws. During sampling, a conditional distribution for the latent observations y_i^* conditional on all other parameters in the model is considered.[7] This distribution is used to produce a random draw for all y_i^* in the probit model. The conditional distribution for the latent variables takes the form of a normal distribution centered on the predicted value truncated at the left at 0 if $y_i = 1$ and at the right at 1 if $y_i = 0$.

The Bayesian Gibbs sampler approach to estimating spatial discrete choice models (both BSAR and BSEM models) is proposed by LeSage (2000) and is an extension of the Gibbs sampling method suggested by Geman and Geman (1984).[8] This method exhibits a similarity to the EM algorithm, where the latent unobserved observations on the dependent variable y_i^* are replaced by estimated values. The Bayesian approach is different in the way it formulates the likelihood function and the estimates of the unobserved latent variable. The Gibbs estimator remedies the two limitations of McMillen (1992)'s EM estimator, its slow convergence and its bias in the estimation of standard errors.

It is important to underline that LeSage (2000) relaxes the assumption of homogeneity for the disturbances $\boldsymbol{\epsilon}$ used in BSAR and BSEM models. This means that the error term $\boldsymbol{\epsilon}$ can follow a multivariate normal distribution $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma_\epsilon^2\mathbf{V})$ in a probit model, where $\mathbf{V} = diag(v_1, v_2, \ldots, v_n)$ and v_i with $i = 1, 2, \ldots, n$ are the variance parameters to be estimated. Greene (2003) points out that accounting for heteroskedasticity is important for probit models because the estimates are inconsistent in the presence of nonconstant disturbance variances.

In order to assign the priors of a BSAR model, LeSage (2000) assumes that the priors are

[7] Gelfand and Smith (1990) demonstrate that Gibbs sampling from the sequence of complete conditional distributions for all parameters in the model produces a set of estimates that converges in the limit to the true posterior distribution of the parameters.

[8] Bolduc, Fortin and Gordon (1997) take a similar approach for the closely related spatial ordinal probit model.

independent

$$\pi(\rho,\beta,\sigma,\mathbf{V}) = \pi(\rho)\pi(\beta)\pi(\sigma)\pi(\mathbf{V}),$$

where

$$\begin{aligned} \pi(\rho) &\propto constant \\ \pi\left(v_i^{-1}/q\right) &\backsim \frac{\chi^2(q)}{q} \qquad i = 1,2,..,n \\ \pi(\sigma^2) &\propto \frac{1}{\sigma}. \end{aligned}$$

The parameter q controls the amount of dispersion in v_i, with small values of q producing leptokurtic distributions and large values imposing homoskedasticity.[9] We summarize LeSage (2000)'s algorithm by the following steps:[10]

1. Initial values for the parameters ρ_0, β_0, σ_0, $\mathbf{V}_0$ are considered. The residuals ϵ_0 are computed by substituting these values in equation (3). Using a random draw from $\chi^2(n)$ the following value is determined:

$$\sigma_1^2 = \frac{\epsilon_0'\mathbf{V}_0\epsilon_0}{\chi^2(n)}.$$

2. Given the values ρ_0, $\mathbf{V}_0$, σ_1, the parameter β_1 is drawn from the multivariate normal

$$f(\beta_1/\rho_0,\mathbf{V}_0,\sigma_1) \sim N_n\left[(\mathbf{X}'\mathbf{V}_0^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_0^{-1}(\mathbf{I}-\rho\mathbf{W})\mathbf{y}^*, \sigma_1^2(\mathbf{X}'\mathbf{V}_0^{-1}\mathbf{X})^{-1}\right].$$

3. By drawing an n-vector of random $\chi^2(q+1)$ and by using ρ_0, β_1, and σ_1, the values v_i with $i = 1,2,\ldots,n$ are computed with

$$v_i = \frac{\sigma_1^{-2}\varepsilon_{i1}^2 + q}{\chi^2(q+1)}.$$

4. By knowing the values β_1, σ_1, $\mathbf{V}_1$, the *metropolis sampling* algorithm (Hastings, 1970) is applied to determine ρ_1. The conditional posterior for ρ given β_1, σ_1, $\mathbf{V}_1$ is

$$f(\rho_0/\beta_1,\sigma_1,\mathbf{V}_1) \propto |\mathbf{I}-\rho\mathbf{W}| exp\left\{-\frac{1}{2\sigma^2}\epsilon_1'\mathbf{V}_1^{-1}\epsilon_1\right\}. \tag{9}$$

   Let the value $\rho^* = \rho_0 + cZ$ be generated, where Z is a draw from a standard normal distribution and c is a known constant.[11] The acceptance probability $p = min\{1, \frac{f(\rho^*)}{f(\rho_0)}\}$, where $f(\cdot)$ is defined in equation (9). A value m is drawn from a continuous uniform distribution with support $[0,1]$. If $m < p$, the next draw from the density function (9) is given by $\rho_1 = \rho^*$, otherwise the draw is taken to be the current value $\rho_1 = \rho_0$.

[9] $q = 7$ produces estimates similar to logit and use of a large value, e.g. q=100, produces estimates similar to those from probit.

[10] We follow Thomas (2007)'s implementation of LeSage (2000)'s methodology, which follows the suggestion by Fleming (2004, p. 159) to transform the latent variable into one that is distibuted independently by using the Cholesky root of the inverted error covariance matrices (cf. matrix **A** in the RIS estimator below).

[11] For the Monte Carlo simulations in this article we set $c = 0.1$.

5. The values of the latent dependent variable $\mathbf{y}^*$ are sampled from the multivariate truncated normal distribution
$$\mathbf{Y}^* \sim N_n^T((\mathbf{I}-\rho_1\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}_1, \boldsymbol{\Lambda}),$$
where $\boldsymbol{\Lambda}$ is the diagonal matrix whose elements are the elements of the main diagonal of the matrix $(\mathbf{I}-\rho_1\mathbf{W})^{-1}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'[(\mathbf{I}-\rho_1\mathbf{W})^{-1}]'$. The normal distribution is truncated at the left at 0 if $Y=1$ and truncated at the right at 0 if $Y=0$ (Albert and Chib, 1993).

LeSage (2000)'s approach overcomes the problems in estimating the standard error in the EM algorithm since parameter standard errors are derived from the posterior parameter distributions. The first advantage of the Bayesian strategy is to be able to derive the condition distribution of each parameter, and thus compute different moments of the distribution. The second advantage is its flexibility to account for the heteroskedasticity in the error terms.

3.3 Recursive Importance Sampling

Beron and Vijverberg (2004) propose a recursive importance sampling (RIS) estimator to evaluate directly the n-dimensional integral in both the BSAR and the BSEM models. The RIS-normal simulator is identical to what is sometimes called the Geweke-Hajivassiliou-Keane (GHK) simulator (Borsch-Supan and Hajivassiliou, 1993).

Define $\mathbf{Z}$ as an $n \times n$ matrix with

$$z_{ij} = \begin{cases} 1-2y_i & \text{if} \quad i=j \\ 0 & \text{otherwise,} \end{cases}$$

for $i, j = 1, 2, \ldots, n$. This means that $\mathbf{Z}$ is a diagonal matrix that satisfies the equation $\mathbf{Z}\mathbf{Z}' = \mathbf{I}_n$. By defining $\mathbf{t} = \mathbf{Z}\mathbf{e}$, we obtain that

$$Var(\mathbf{t}) = \mathbf{Z}Var(\mathbf{e})\mathbf{Z}' \equiv \boldsymbol{\Sigma}_\rho,$$

where $Var(\mathbf{e})$ is provided by equation (5). By this notation the observed vector $\mathbf{y}$ defined in (1) leads to an upper limit on $\mathbf{t}$:

$$\mathbf{t} < -\mathbf{Z}(\mathbf{I}-\rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta},$$

which means that we can write the log-likelihood function as

$$\ln L = \ln \Phi_n\left[-\mathbf{Z}(\mathbf{I}-\rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}; \mathbf{0}, \boldsymbol{\Sigma}_\rho\right] \equiv \ln \Phi_n\left[\boldsymbol{T}; \mathbf{0}, \boldsymbol{\Sigma}_\rho\right], \tag{10}$$

where $\Phi_n[\boldsymbol{j}; \boldsymbol{\mu}, \boldsymbol{\Omega}]$ is a n-dimensional normal cumulative distribution function with mean vector μ and variance-covariance matrix $\boldsymbol{\Omega}$.

In order to evaluate the probability in equation (10), Beron and Vijverberg (2004) propose to apply the RIS simulator, developed in detail by Vijverberg (1997). Let $\mathbf{A}$ be an upper triangular matrix such that $\mathbf{A}'\mathbf{A} = \Sigma_\rho^{-1}$ and let $\eta = \mathbf{A}\mathbf{t}$. Whether Σ_ρ is standardized or not, the vector η is i.i.d. standard normal. By defining the matrix $\mathbf{B} = \mathbf{A}^{-1}$, $\mathbf{B}$ results an upper triangular matrix with $b_{jj} > 0 \; \forall j$ and $\mathbf{B}\eta = \mathbf{t}$.

Given the upper bound $\{\mathbf{B}\eta = \mathbf{t}\} < \mathbf{T}$, we can apply the following iterative procedure:

$$\begin{aligned} \eta_n &< b_{nn}^{-1} T_n \equiv \eta_{n0} \\ \eta_j &< b_{jj}^{-1}\left[T_j - \sum_{i=j+1}^{n} b_{ji}\eta_i\right] \equiv \eta_{j0}(T_j, \eta_{j+1}, \ldots, \eta_n) \equiv \eta_{j0}. \end{aligned} \tag{11}$$

Let $g(\eta_j)$ be a probability density function with support the whole real axis and let $G(\cdot)$ be the associated cumulative distribution function. By denoting

$$g^c(\eta_j) = \frac{g(\eta_j)}{G(\eta_{j0})}$$

for $\eta_j \leq \eta_{j0}$, we can compute the following probability:

$$p = P\{\mathbf{t} < \mathbf{T}\} = \int_{-\infty}^{\mathbf{T}} \phi(\mathbf{t}; \mathbf{0}, \Omega) d\mathbf{t} = \int_{-\infty}^{\eta_{n0}} \dots \int_{-\infty}^{\eta_{1,0}} \prod_{j=1}^{n} \phi(\eta_j) d\eta_1 \dots d\eta_n$$
$$= \int_{-\infty}^{\eta_{n0}} \frac{\phi(\eta_n)}{g^c(\eta_n)} \left[\int_{-\infty}^{\eta_{n-1,0}} \frac{\phi(\eta_{n-1})}{g^c(\eta_{n-1})} \dots \left(\int_{-\infty}^{\eta_{2,0}} \frac{\phi(\eta_2)}{g^c(\eta_2)} \Phi(\eta_{10}) g^c(\eta_2) d\eta_2 \right) \dots \right] g^c(\eta_n) d\eta_n. \tag{12}$$

The RIS simulator is implemented by drawing a large number R of random vector $\boldsymbol{\eta}$ satisfying the condition $\eta_j \leq \eta_{j0}$ for $j = 1, 2, \dots, n$ from the density function $g(\cdot)$.[12] There are different suitable density functions used to define $g(\cdot)$ (Vijverberg, 1997). Vijverberg (1999) shows that the RIS-normal simulator is often preferred. For this reason we choose the normal density function in the following Monte Carlo simulations and, in particular, we apply the antithetical sampling strategy suggested by Vijverberg (1997) for simulating from a multivariate normal distribution.

The recursive nature of the RIS simulator is due to the fact that the bounds in equation (11) are backwards determined. For every drawing r of the random vector $\boldsymbol{\eta}$, given η_{n0}, the values $\tilde{\eta}_{n,r}$ and $\tilde{\eta}_{n-1,0,r}$ are calculated using equation (11) by using $\tilde{\eta}_{n,r}$ in the place of η_n. This process is repeating until $\tilde{\eta}_{1,0,r}$ is computed. Then for the RIS-normal simulator the simulated value for p, defined in equation (12), is

$$\hat{p} = \frac{1}{R} \sum_{r=1}^{R} \left(\prod_{j=1}^{n} \Phi[\tilde{\eta}_{j,0,r}] \right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the one dimensional standard normal random variable.

Based on the Monte Carlo study that Beron and Vijverberg (2004) performed the RIS simulator can provide accurate estimates for spatial binary choice models. Moreover, this approach is attractive since it is the only one that directly evaluates the n-dimensional probit likelihood function. This means that only this methodology allows for the use of the Likelihood Ratio test. Because of these advantages Beron, Murdoch and Vijverberg (2003) and Novo (2003) apply the RIS simulator.

3.4 Generalized Method of Moments

This section describes a spatially dependent binary choice methodology that considers the problem as a weighted non-linear version of the linear probability (Amemiya, 1985; Greene, 2002; Maddala, 1983) with a variance-covariance matrix that can be estimated with a Generalized Method of Moments (GMM) estimator (Hansen, 1982). Pinkse and Slade (1998) derive the GMM moment equations from the likelihood function. Klier and McMillen (2008) propose a linearized version of the GMM suggested by Pinkse and Slade (1998).

[12] For the Monte Carlo simulations we use $R = 1000$.

3.4.1 Pinkse and Slade's estimator

While Pinkse and Slade (1998) suggest to apply the GMM to a BSEM model, for achieving the aim of this article we present their estimator for a BSAR model. Similar to McMillen (1992), Pinkse and Slade (1998) consider the generalized residuals[13]

$$\tilde{\mathbf{e}}(\boldsymbol{\theta}) = \mathbf{D}^{-1}E[\mathbf{e}/\mathbf{y},\boldsymbol{\theta}] = \frac{\phi_n\left[\mathbf{q}(\boldsymbol{\theta})\right]\left\{\mathbf{y}-\Phi_n[\mathbf{q}(\boldsymbol{\theta})]\right\}}{\Phi_n\left[\mathbf{q}(\boldsymbol{\theta})\right]\left\{1-\Phi_n[\mathbf{q}(\boldsymbol{\theta})]\right\}}, \tag{13}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}', \rho)'$ is the parameter vector and $\mathbf{D}$ and $\mathbf{q}$ are defined in equations (6) and (7), respectively.

By applying the GMM the parameter vector $\boldsymbol{\theta}$ is estimated by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \tilde{\mathbf{e}}'(\boldsymbol{\theta})\mathbf{Z}\mathbf{M}\mathbf{Z}'\tilde{\mathbf{e}}(\boldsymbol{\theta}), \tag{14}$$

where $\tilde{\mathbf{e}}$ is defined in equation (13), $\mathbf{Z}$ is a matrix of instruments,[14] $\mathbf{M}$ is a positive definite matrix[15] and $\boldsymbol{\Theta}$ is the parametric space.

Pinkse and Slade (1998) provide the asymptotic variance of their estimator for a BSEM model and develop also the hypothesis test for spatial error correlation. Their approach overcomes the problems of evaluating a high order integral and the n by n determinants in the Maximum Likelihood method. The main disadvantage of this approach is that it requires the $n \times n$ matrix $(\mathbf{I} - \rho\mathbf{W})^{-1}$ to be inverted in each iteration. Furthermore, since Pinkse and Slade (1998) apply the GMM method, their estimator is less efficient than the ML estimators.

3.4.2 Klier and McMillen's estimator

Klier and McMillen (2008) linearize Pinkse and Slade (1998)'s model around a convenient starting point for a BSAR logit model. In particular, in equation (14), Klier and McMillen (2008) let $\mathbf{M} = (\mathbf{Z}'\mathbf{Z})^{-1}$, so the objective function for the GMM estimator is

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \tilde{\mathbf{e}}'(\boldsymbol{\theta})\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\tilde{\mathbf{e}}(\boldsymbol{\theta}),$$

hence Klier and McMillen (2008) apply a nonlinear two-stage least squares method. In order to analyse Klier and McMillen (2008)'s methodology we define

$$\mathbf{P} = P\{\mathbf{Y} = 1/\boldsymbol{\theta}\} = \frac{\exp[\mathbf{q}(\boldsymbol{\theta})]}{1+\exp[\mathbf{q}(\boldsymbol{\theta})]}. \tag{15}$$

where $\mathbf{q}(\boldsymbol{\theta})$ is defined in equation (7).

Klier and McMillen (2008)'s iterative procedure has the following steps:

1. assume initial values for the parameter vector $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0', \rho_0)'$;

2. compute $\mathbf{e}_0$ defined in equation (4);

[13] Unlike Chesher and Irish (1987) and Cox and Snell (1968) (see also eq. (8)), Pinkse and Slade (1998) define the generalized residuals as $\mathbf{D}^{-1}E[\mathbf{e}/\mathbf{y},\mathbf{e},\rho]$ and not $E[\mathbf{e}/\mathbf{y},\mathbf{e},\rho]$.

[14] In the Monte Carlo simulations we consider $\mathbf{Z} = \mathbf{1} + \mathbf{X} + \mathbf{W}\mathbf{X} + \mathbf{W}^2\mathbf{X} + \mathbf{W}^3\mathbf{X}$.

[15] Pinkse and Slade (1998) consider $\mathbf{M}$ equal to the identity matrix $\mathbf{M} = \mathbf{I}_n$ in their empirical application and we follow this suggestion in the Monte Carlo simulations.

3. compute the gradient terms

$$\begin{aligned} \mathbf{G}_{\beta i} &= \frac{\partial P_i}{\partial \beta} = \hat{P}_i(1-\hat{P}_i)\mathbf{t}_i \\ G_{\rho i} &= \frac{\partial P_i}{\partial \rho} = \hat{P}_i(1-\hat{P}_i)\left[h_i - \frac{q_i}{\sigma_{ei}^2}\Upsilon_{ii}\right], \end{aligned} \tag{16}$$

where $\mathbf{t}_i$ is the i-th row vector of the matrix $\mathbf{T} = \mathbf{D}^{-1}(\mathbf{I}-\rho\mathbf{W})^{-1}\mathbf{X}$, h_i is the i-th element of the vector $\mathbf{h} = (\mathbf{I}-\rho\mathbf{W})^{-1}\mathbf{W}\mathbf{q}$, q_i is the i-th element of the vector $\mathbf{q}$ defined in equation (7) and Υ_{ii} is the i-th element of the diagonal of the matrix $\Upsilon = (\mathbf{I}-\rho\mathbf{W})^{-1}\mathbf{W}(\mathbf{I}-\rho\mathbf{W})^{-1}(\mathbf{I}-\rho\mathbf{W})^{-1}$.[16]

4. regress the gradient terms $\mathbf{G}_\beta$ and G_ρ on $\mathbf{Z}$ and compute the predicted values $\hat{\mathbf{G}}_\beta$ and $\hat{G}_\rho$;

5. regress $\mathbf{e}_0 + \mathbf{G}_\beta\hat{\beta}_0$ on $\hat{\mathbf{G}}_\beta$ and $\hat{G}_\rho$. The coefficients obtained from this regression are the estimated values of β and ρ.

The main advantage of this approach is that it is not iterative and does not require the $n \times n$ matrix $(\mathbf{I}-\rho\mathbf{W})^{-1}$ to be inverted in each iteration, unlike Pinkse and Slade (1998)'s estimator. This characteristic leads to a computationally significantly faster estimator. The main disadvantage of this estimator is that it provides accurate estimates of ρ only as long as ρ is small. Furthermore, linearized approach cannot provide an estimate of precision for the spatial autoregressive parameter ρ. Finally, since Klier and McMillen (2008) propose a linearization around the starting point, a restriction for the parameter ρ to the interval [-1,1] cannot be introduced by using their method.

4 Monte Carlo simulations

In order to make up for the lack of simulation studies for BSAR models (Fleming, 2004), in this section we compare the properties of these five estimators by Monte Carlo simulations. The set up of the simulations is primarily based on the literature on policy and regime diffusion (e.g. Gleditsch and Ward, 2006; Baturo and Gray, 2009) and on broad similarity with simulations as published in accompaniment of the proposals of estimators discussed in this paper (e.g. McMillen, 1995*b*; Beron and Vijverberg, 2004; Klier and McMillen, 2008). In order to understand how the properties of these estimators vary according to the number of observations, we consider two different sample sizes: $n = 50$ and $n = 500$. The first sample size is set because it resembles the number of states in the US, which is a typical application area for studies in policy diffusion (e.g., Mooney, 2001; Volden, 2006; LeSage and Parent, 2007). The larger sample size is added to be able to see the results when the sample size increases.

[16] The derivative in equation (16) assumes that $(\mathbf{I}-\rho\mathbf{W})$ is a symmetric matrix, which is not guaranteed. For example, when $\mathbf{W}$ is standardized, this is not the case. See the appendix for this derivative without assuming a symmetric matrix. The revised derivative in the appendix does not affect the estimator or the Monte Carlo results when the convenient starting point at $\rho = 0$ is used, as in Klier and McMillen (2008).

In our Monte Carlo analysis we generate 1,000 replications.[17] For generating the data sets[18] we consider one covariate X drawn from a normal distribution $N(2,4)$ with expected value 2 and standard deviation 4.[19] Based on equation (3), the residuals vector $\boldsymbol{\epsilon}$ is generated from a multivariate normal distribution $N_n(\vec{0},\mathbf{I})$ and the parameter vector is $\boldsymbol{\beta} = [4,-2]'$. In order to generate $\mathbf{W}$, we apply the method suggested by Beron and Vijverberg (2004, p. 179) by using $d = 0.21$ for $n = 50$ and $d = 0.06$ for $n = 500$. To analyse how the characteristics of the estimators change according to the level of autocorrelation, we consider four different values $(0; 0.1; 0.45; 0.8)$ of the parameter ρ, such that the last three values are equidistant.[20] For the maximization procedure in the EM, RIS, and Pinkse and Slade (1998) estimators, we use the `optim()` function in `R` with a maximum number of iterations of 1,000. Finally, analogously to LeSage (2000) and Beron and Vijverberg (2004) we consider a maximum number of loops equal to 1,000 for the RIS and EM algorithms, and 3,000 for the Gibbs estimator.

In the following tables we report the mean of the bias and the standard deviation of the estimators (in round brackets) computed on 1,000 sets. The data is generated based on a probit model. Since Klier and McMillen (2008) have proposed their estimator for the logit model, we rescale their parameter estimates to allow for comparison. Because the variance of the logistic distribution is $\pi^2/3$, we report the estimates $\hat{\beta}_0\sqrt{3}/\pi$ and $\hat{\beta}_1\sqrt{3}/\pi$ for the linearized GMM estimator.[21]

The primary focus of this paper is on the estimation of the level of spatial autocorrelation in a binary spatial autoregressive model specification. Since we have the study of the diffusion of policies and regimes in mind, the level of diffusion is typically of key interest. The autocorrelation in the residuals is thus not treated as a mere nuisance, but as a structural factor of substantive interest. Figure 1 provides the distribution of the bias ofthe estimators of the spatial autocorrelation parameter ρ in the BSAR model described in equation (3).[22] Table 1 provides the mean and standard deviation of the bias of the above-mentioned estimators..

It is clear from Figure 1 that the performance of the estimators varies depending on the level of autocorrelation in the data, with particularly large differences between estimators under high levels of autocorrelation. As can be seen in Table 1, in the absence of spatial autocorrelation ($\rho = 0$), the Gibbs and the EM estimators are the best estimators of ρ in terms of both the distortion and the dispersion. The linearized GMM estimator also does particularly well, which is unsurprising, since $\rho = 0$ is the value used as the starting point for the linearization. When looking at the distribution as a whole, however, it is clear that while this estimator performs generally well, there are clear outliers among the estimates. The RIS estimator shows the worst performance and it tends to underestimate ρ for both small ($n = 50$) and large ($n = 500$) samples, with in particular some negative outliers. The Pinkse and Slade (1998) estimator, on the other hand, tends to overestimate ρ in this scenario.

When the level of spatial autocorrelation is positive but still limited ($\rho = 0.1$), the EM and Gibbs estimators still show good performance, even if the differences with the other estimators

[17] The same number of replications is considered by Flores-Lagunes and Schnier (2005); Franzese and Hays (2007); Klier and McMillen (2008).

[18] We use the `R` package `rlecuyer` for the parallel generation of random numbers.

[19] Following McMillen (1995*b*), we prefer to consider a standard deviation of X substantially higher than σ_ϵ.

[20] Anselin (1982), Beron and Vijverberg (2004) and Klier and McMillen (2008) consider similar values.

[21] The spatial coefficient ρ is not affected by the scaling.

[22] These plots are generated using the `violin` function in the `lattice` package in `R`, which in turn makes use of the built-in `density` function for the computation of smoothed kernel density estimates. Default settings are used.

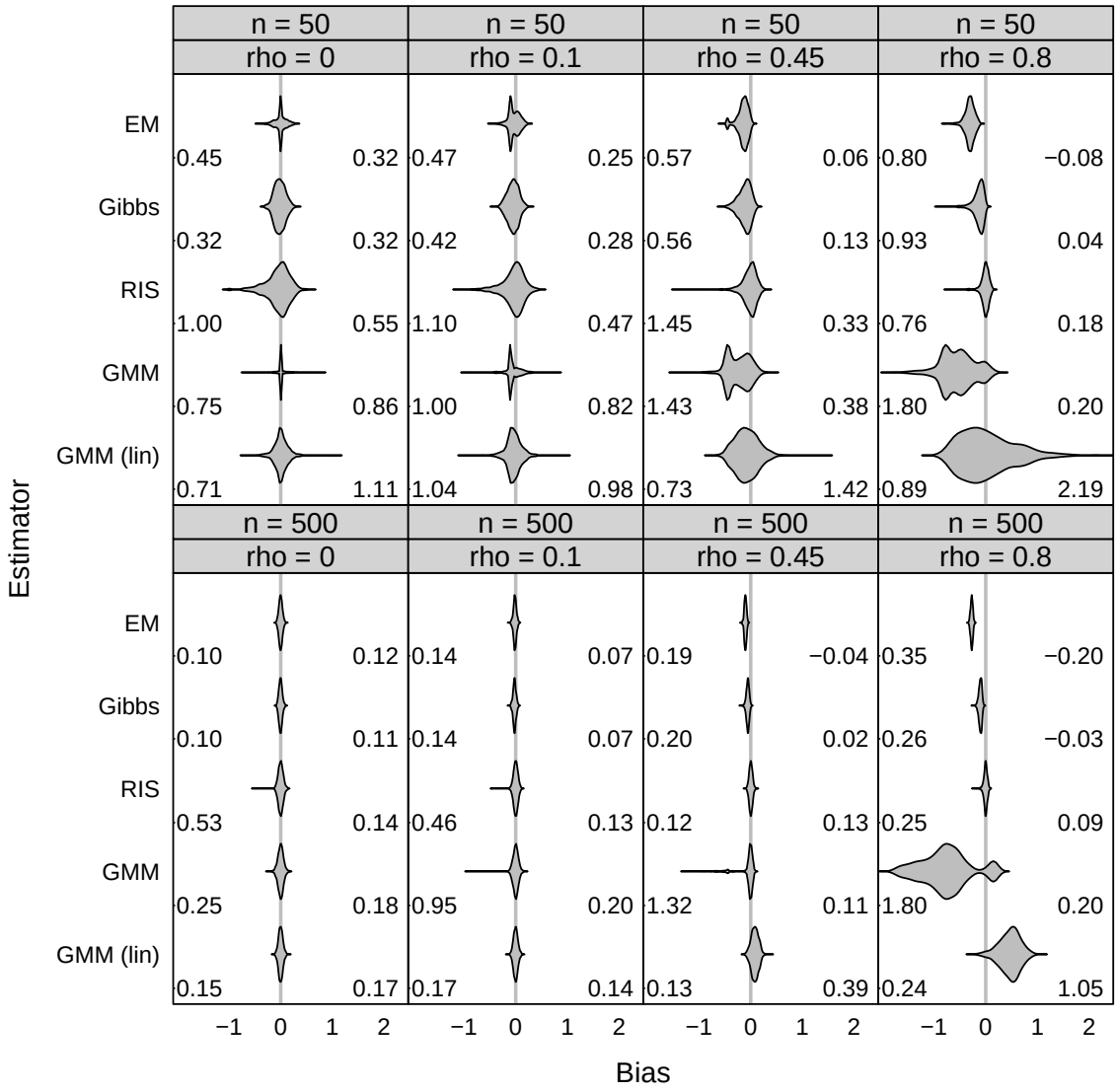

Figure 1: The distribution of the bias of the estimators of the autocorrelation parameter ρ obtained from Monte Carlo simulations on 1,000 samples. Numbers represent minimum and maximum values of the bias.

are less considerable in comparison with the scenario without spatial autocorrelation. Both the EM and the Gibbs estimators tend to underestimate the spatial autocorrelation parameter. As can be expected, the linearized GMM estimator is still performing well this close to the linearization point of $\rho = 0$. The main disadvantage of this estimator is that it is not possible to put a reasonable constraint on $\hat{\rho}$, such that occasional estimates are obtained outside the $[-1,1]$ interval, visible in Figure 1. The RIS estimator is also prone to the occasional outlier in its estimate of the level of autocorrelation. It is striking that all estimators, with the exception of GMM with $n = 50$, tend to underestimate ρ, often by over 50% of the true parameter value, when the sample size is low.

The next scenario contains significant levels of autocorrelation, with $\rho = 0.45$. Under this

	$\rho = 0$		$\rho = 0.1$		$\rho = 0.45$		$\rho = 0.8$	
$\hat{\rho}$ \ n	**50**	**500**	**50**	**500**	**50**	**500**	**50**	**500**
EM	-0.002 (0.108)	-0.002 (0.034)	-0.038 (0.108)	-0.016 (0.030)	-0.148 (0.107)	-0.104 (0.022)	-0.311 (0.098)	-0.272 (0.023)
Gibbs	-0.017 (0.108)	-0.003 (0.032)	-0.050 (0.109)	-0.023 (0.030)	-0.114 (0.120)	-0.058 (0.028)	-0.140 (0.117)	-0.104 (0.030)
RIS	-0.051 (0.235)	-0.002 (0.056)	-0.044 (0.219)	-0.003 (0.053)	-0.027 (0.158)	0.003 (0.034)	-0.015 (0.093)	-0.005 (0.039)
GMM	0.019 (0.165)	-0.001 (0.055)	-0.049 (0.173)	-0.001 (0.060)	-0.243 (0.228)	-0.097 (0.245)	-0.547 (0.336)	-0.775 (0.458)
GMM (lin)	0.011 (0.149)	-0.001 (0.043)	-0.037 (0.166)	-0.001 (0.044)	-0.075 (0.232)	0.080 (0.072)	0.028 (0.528)	0.468 (0.205)

Table 1: The mean bias and the standard deviation (between parentheses) of the estimators of the autocorrelation parameter ρ obtained from Monte Carlo simulations on 1,000 samples.

level of autocorrelation, the RIS estimator starts to perform relatively well compared to the other estimators. While the absolute mean bias is the lowest, the variation is still relatively high, however, and the plot clearly shows the presence of some outliers. For the EM and the Gibbs estimators, the bias is significantly larger than for $\rho = 0.1$ and for the linearized GMM, a clear increase in the dispersion is visible in Figure 1, although this is compensated by a reduction in extreme outliers. The distribution of the bias of the GMM estimator now shows a clear tendency to underestimate the amount of autocorrelation, with a tight distribution under $n = 500$, but with some outliers.

For high spatial autocorrelation ($\rho = 0.8$), the RIS estimator shows the best performance. The linearized GMM now clearly suffers from the large distance from the starting point of the linearization – the extrapolation from $\rho = 0$ to $\rho = 0.8$ leads to significant overestimation of the level of autocorrelation. The plot also shows that the estimator clearly suffers from the lack of a constraint on ρ. The GMM estimator shows rather significant underestimation of ρ, as well as a high level of variance. Furthermore, under this scenario, the simulations suggest asymptotically biased in mean results for the GMM and the linearized GMM estimators, with the mean bias increasing for the larger sample size.[23] Following the trend already visible when moderately increasing ρ, the EM estimator clearly shows greater mean bias under this scenario, while for Gibbs, the results are similar to $\rho = 0.45$.

Even with primary focus on the level of autocorrelation, the estimates of the effects of other independent variables can of course not be ignored. Table 2 provides the mean and the standard deviation of the estimates of β_1, the parameter for independent variable X. The differences in estimate accuracy between the estimators vary more dramatically than for ρ, with the Gibbs estimator clearly outperforming all other estimators under all conditions and the estimator proposed by Pinkse and Slade (1998) clearly providing the worst results for both the

[23] When the observations are "strongly spatially dependent" (Pinkse and Slade, 1998, p. 134, fn. 12), even the consistency is not guaranteed.

	$\rho = 0$		$\rho = 0.1$		$\rho = 0.45$		$\rho = 0.8$	
n / $\hat{\beta}_1$	**50**	**500**	**50**	**500**	**50**	**500**	**50**	**500**
EM	-59.23 (507.24)	-0.12 (0.34)	-43.12 (356.69)	-0.02 (0.33)	-5.52 (56.98)	1.03 (0.11)	1.42 (3.64)	1.69 (0.03)
Gibbs	-0.95 (1.12)	-0.07 (0.28)	-0.86 (1.09)	-0.05 (0.29)	-0.63 (1.02)	0.09 (0.24)	0.30 (0.70)	0.79 (0.19)
RIS	-8.05 (19.89)	0.01 (0.38)	-5.86 (13.06)	0.02 (0.40)	-6.86 (18.59)	0.13 (0.43)	-4.13 (12.62)	1.12 (0.49)
GMM	-5035.95 (28279.94)	-695.99 (6344.00)	-4194.69 (23060.63)	-576.70 (5841.52)	-4777.75 (15011.28)	-17864.07 (250693.49)	-2818.31 (11646.26)	-80663.45 (336721.33)
GMM (lin)	-64.15 (424.80)	-0.04 (1.65)	-41.08 (175.62)	0.07 (1.93)	-8.32 (95.09)	1.13 (0.49)	1.25 (6.82)	1.70 (0.05)

Table 2: The mean bias and the standard deviation (between parentheses) of the estimators of the parameter β_1 obtained from Monte Carlo simulations on 1,000 samples.

bias and the dispersion.

Under absence of spatial autocorrelation ($\rho = 0$) and for $n = 50$, the Gibbs estimator generates the least average bias on $\hat{\beta}_1$, with RIS, linearized GMM, and EM also performing well as long as the sample size is sufficiently large. Under $n = 50$, the Gibbs estimator underestimates β_1 by about 50% on average of the value of β_1 and the other estimators well beyond that. The difference between the smaller and the larger sample sizes is more pronounced than for the estimates of ρ. The GMM estimator is the only estimator that is still significantly biased when the sample size is reasonably large. All estimators tend to underestimate β_1.

Under limited autocorrelation ($\rho = 0.1$), the order of accuracy of the estimators remains more or less the same, with still only Gibbs performing the best under the small sample size and performing the well under the large sample size. The RIS performs similar to EM, Gibbs and GMMlin for large sample size. The GMM estimator shows the worst performance. Increasing the autocorrelation to $\rho = 0.45$, the results for the Gibbs, the RIS and the GMM estimators are very similar to $\rho = 0.1$, but the EM and the linearized GMM estimators show lower mean biases under small sample size.

For the estimation of ρ, we saw significant difference between the moderate and the high levels of autocorrelation – to what extent is this the case for the slope coefficients? Similar to the estimates of ρ, the GMM estimator of Pinkse and Slade (1998) starts to show very significant distortion and dispersion when the autocorrelation is high and, furthermore, shows a significant increase in the mean bias when the sample size increases. The linearized version of this estimator also reflects this increasing mean bias with sample size for high autocorrelation, but nevertheless performs remarkably well, with an average bias of approximately 50 to 75% of β_1, underestimating the magnitude of the negative effect of X. The Gibbs estimator outperforms all other estimators both with small and large sample sizes, followed by the RIS, the EM and the linearized GMM estimators as long as sample size is large. It should be noted that for all estimators, the bias is relatively large, which is an overestimate of the magnitude by slightly over 10%.

Usually of least concern to applied researchers, but relevant for accurate prediction, is the estimate of the intercept of the model, in this case β_0. Table 3 provides the simulation results for this parameter of the model. Not unsurprisingly, the results are closely in line with those for β_1. Indeed, the mean of the bias of β_0 is roughly the bias in Table 2 multiplied by a factor -2. Relative to the size of β, the bias is thus the same on average. Similarly, the standard deviation of the bias is multiplied by a factor 2, with the exception of the GMM estimator under higher levels of autocorrelation, where the dispersion under larger sample size is similar to that for β_1. The same relative results for the different estimators are therefore obtained.

5 Conclusion

In this paper we provide a comprehensive overview of estimators for spatial autoregressive models with binary dependent variables. These models are of particular interest to various applications in economics, political science, and related disciplines, where the outcome might be a policy, a decision, a transition, or otherwise binary outcome. Applications can also be imagined in the field of bioinformatics or neuroscience, although sample sizes tend to be magnitudes larger than those studied here. Many of these outcomes are interdependent through either spatial contiguity or any other form of proximity, including social networks or economic linkages,

	$\rho = 0$		$\rho = 0.1$		$\rho = 0.45$		$\rho = 0.8$	
n / $\hat{\beta}_0$	**50**	**500**	**50**	**500**	**50**	**500**	**50**	**500**
EM	121.85 (1120.25)	0.24 (0.70)	85.58 (689.86)	0.05 (0.69)	12.52 (127.40)	-2.06 (0.23)	-3.02 (11.82)	-3.39 (0.11)
Gibbs	1.92 (2.44)	0.13 (0.59)	1.72 (2.34)	0.09 (0.60)	1.25 (2.21)	-0.19 (0.50)	-0.62 (1.68)	-1.58 (0.39)
RIS	15.59 (38.42)	-0.01 (0.80)	11.62 (26.16)	-0.04 (0.82)	13.41 (36.16)	-0.26 (0.87)	8.14 (25.25)	-2.29 (1.03)
GMM	10553.11 (57014.77)	1497.72 (13763.26)	8927.67 (46734.98)	1215.51 (12202.29)	9761.46 (30576.23)	21968.93 (293214.42)	5390.80 (26072.56)	48233.04 (214363.69)
GMM (lin)	122.91 (725.66)	0.09 (3.30)	81.80 (340.99)	-0.13 (3.88)	17.71 (202.43)	-2.26 (1.01)	-2.72 (22.23)	-3.41 (0.22)

Table 3: The mean bias and the standard deviation (between parentheses) of the estimators of the parameter β_0 obtained from Monte Carlo simulations on 1,000 samples.

and ignoring the inherent spatial structure of the data generates inconsistent and inefficient estimates (McMillen, 1992). Furthermore, in many applications the researcher is explicitly concerned with estimating the level of interdependence – it is indeed this concern that is of primary interest in our discussion and simulation study.

This paper compares five estimators introduced to date for this specific type of model. An extensive simulation study compares the performance of these five estimators under conditions of relatively small sample sizes and varying levels of spatial autocorrelation. When taking both the estimation of the extend of spatial autocorrelation and the coefficients on the other explanatory variables into account, the Gibbs estimator (LeSage, 2000) clearly outperforms the other estimators. When the sample size increases, the difference between the different estimators becomes smaller. When focusing specifically on the spatial autoregressive component alone, the Gibbs estimator (LeSage, 2000) performs best for low spatial autocorrelation, while the Recursive Importance Sampler (Beron and Vijverberg, 2004) performs best for high spatial autocorrelation. The linearized GMM estimator (Klier and McMillen, 2008) is an interesting option when the sample size is large and the autocorrelation relatively low, in particular due to its high computational speed.

References

Albert, J.H. and S. Chib. 1993. "Bayesian analysis of binary and polychotomous response data." *Journal of the American Statistical Association* .

Amemiya, Takeshi. 1985. *Advanced econometrics*. Cambridge: Harvard University Press.

Anselin, Luc. 1982. "A note on small sample properties of estimators in a first-order spatial autoregressive model." *Environment and Planning* 14(1):1023–1030.

Anselin, Luc. 1988. *Spatial econometrics: methods and models*. Dordrecht: Kluwer Academic Publishers.

Anselin, Luc. 2002. "Under the hood. Issues in the specification and interpretation of spatial regression models." *Agricultural Economics* 27:247–267.

Anselin, Luc, Raymond J.G.M. Florax and Sergio J. Rey. 2004. Econometrics for Spatial Models: Recent Advances. In *Advances in Spatial econometrics*, ed. Luc Anselin, Raymond J.G.M. Florax and Sergio J. Rey. Berlin: Springer pp. 1–28.

Autant-Bernard, Corinne. 2006. "Where do firms choose to locate their R&D? A spatial conditional logit analysis on French data." *European Planning Studies* 14(9):1187–1208.

Baturo, Alexander and Julia Gray. 2009. "Flatliners: Ideology and rational learning in the adoption of the flat tax." *European Journal of Political Research* 48:130–159.

Bee, Marco and Giuseppe Espa. 2008. "A Monte Carlo EM algorithm for the estimation of a logistic auto-logistic model with missing data." *Letters in Spatial and Resource Sciences* 1:45–54.

Beron, Kurt J., James C. Murdoch and Wim P.M. Vijverberg. 2003. "Why cooperate? Public goods, economic power, and the Montreal protocol." *Review of Economics and Statistics* 85(2):286–297.

Beron, Kurt J. and Wim P.M. Vijverberg. 2004. Probit in a spatial context: a Monte Carlo analysis. In *Advances in spatial econometrics. Methodology, tools and applications*, ed. Luc Anselin, Raymond J.G.M. Florax and Sergio J. Rey. Berlin: Springer pp. 169–195.

Bivand, Roger. 1998. "A review of spatial statistical techniques for location studies." Paper presented at the CEPR symposium on New Issues in Trade and Location (2277), Lund, Sweden, 28-30 August, 1998.

Bolduc, Denis, B. Fortin and S. Gordon. 1997. "Multinomial probit estimation of spatially interdependent choices: an empirical comparison of two new techniques." *International Regional Science Review* 20(1/2):77–101.

Borsch-Supan, Alex and Vassili A. Hajivassiliou. 1993. "Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models." *Journal of Econometrics* 58:347–368.

Brueckner, Jan K. 2003. "Strategic interaction among governments: an overview of empirical studies." *International Regional Science Review* 26(2):175–188.

Case, Anne C. 1991. "Spatial patterns in household demand." *Econometrica* 59(4):953–965.

Case, Anne C. 1992. "Neighborhood influence and technological change." *Regional Science and Urban Economics* 22:491–508.

Chesher, Andrew and Margaret Irish. 1987. "Residual analysis in the grouped and censored normal linear model." *Journal of Econometrics* 34:33–61.

Collingham, Yvonne C., Richard A. Wadsworth, Brian Huntley and Philip E. Hulme. 2000. "Predicting the spatial distribution of non-indigenous riparian weeds: Issues of spatial scale and extent." *Journal of Applied Ecology* 37:13–27.

Conley, Timothy Gy. 1999. "GMM Estimation with Cross Sectional Dependence." *Journal of econometrics* 92:1–45.

Cox, D.R. and E.J. Snell. 1968. "A general definition of residuals." *Journal of the Royal Statistical Society B* 30(2):248–275.

Davidson, Russell and James G. MacKinnon. 1993. *Estimation and inference in econometrics*. Oxford: Oxford University Press.

Dempster, A.P., N.M. Laird and D.B. Rubin. 1977. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1):1–38.

Dubin, Robin. 1995. Estimating logit models with spatial dependence. In *New directions in spatial econometrics*, ed. Luc Anselin and Raymond J.G.M. Florax. Berlin: Springer Verlag pp. 229–242.

Fleming, Mark M. 2004. Techniques for estimating spatially dependent discrete choice models. In *Advances in spatial econometrics. Methodology, tools and applications*, ed. Luc Anselin, Raymond J.G.M. Florax and Sergio J. Rey. Berlin: Springer pp. 145–167.

Flores-Lagunes, A. and K.E. Schnier. 2005. "Estimation of sample selection models with spatial dependence." Working paper, University of Arizona.

Franzese, Robert J. and Jude C. Hays. 2007. "Spatial econometric models of cross-sectional interdependence in political science panel and time-series-cross-section data." *Political Analysis* 15(2):140–164.

Gelfand, Allan E. and Adrian F. M. Smith. 1990. "Sampling-based Approaches to Calculating Marginal Densities." *Journal of American Statitical Association* 85:398–409.

Geman, S. and D. Geman. 1984. "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12:609–628.

Gleditsch, Kristian S. and Michael D. Ward. 2006. "Diffusion and the international context of democratization." *International Organization* 60(4):911–933.

Greene, William H. 2002. *Econometric Analysis*. London: Prentice Hall.

Greene, William H. 2003. *Econometric Analysis*. 5th ed. Upper Saddle River: Prentice Hall.

Gumpertz, Marcia L., Jonathan M. Graham and Jean B. Ristaino. 1997. "Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: effects of soil variables on disease presence." *Journal of Agricultural, Biological, and Environmental Statistics* 2(2):131–156.

Hansen, L.P. 1982. "Large sample properties of generalized method of moments estimators." *Econometrica* 50:1029–1054.

Hastings, W. K. 1970. "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika* 57:97–109.

Klier, Thomas and Daniel P. McMillen. 2008. "Clustering of auto supplier plants in the United States: generalized method of moments spatial logit for large samples." *Journal of Business & Economic Statistics* 26(4):460–471.

Lee, Lung Fei. 2004. "Asymtotic Distribution of Quasi-maximum Likelihood Estimators for Spatial Autoregressive Models." *Econometrica* 72(6):1899–1925.

LeSage, James P. 2000. "Bayesian estimation of limited dependent variable spatial autoregressive models." *Geographical Analysis* 32(1):19–35.

LeSage, James P. and Olivier Parent. 2007. "Bayesian Model Averaging for Spatial Econometric Models." *Geographical Analysis* 39(3):241–267.

Maddala, G.S. 1983. *Limited-dependent and qualitative variables in econometrics*. New York: Cambridge University Press.

McCullagh, P. and J.A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. New York: Chapman and Hall.

McMillen, Daniel P. 1992. "Probit with spatial autocorrelation." *Journal of Regional Science* 32(3):335–348.

McMillen, Daniel P. 1995*a*. "Selection bias in spatial econometric models." *Journal of Regional Science* 35(3):417–436.

McMillen, Daniel P. 1995*b*. Spatial effects in probit models: a Monte Carlo investigation. In *New directions in spatial econometrics*, ed. Luc Anselin and Raymond J.G.M. Florax. Berlin: Springer Verlag pp. 189–228.

Mooney, Christopher Z. 2001. "Modeling regional effects on state policy diffusion." *Political Research Quarterly* (54):103–124.

Novo, lvaro A. 2003. Contagious Currency Crisis: A Spatial Probit Approach. Working papers Banco de Portugal, Economics and Research Department.
URL: *http://EconPapers.repec.org/RePEc:ptu:wpaper:w200305*

Ord, John. 1975. "Estimation Methods for Models of Spatial Interaction." *Journal of the American Statistical Association* 70:1200–26.

Pinkse, Joris and Margaret E. Slade. 1998. "Contracting in space: an application of spatial statistics to discrete-choice models." *Journal of Econometrics* 85:125–154.

Shalizi, Cosma Rohilla and Andrew C. Thomas. 2010. "Homophily and contagion are generically confounded in observational social network studies." arXiv:1004.4704v3.

Shipan, Charles R. and Craig Volden. 2006. "Bottom-up federalism: The diffusion of antismoking policies from U.S. cities to states." *American Journal of Political Science* 50(4).

Thomas, Timothy S. 2007. "A primer for Bayesian spatial probits, with an application to deforestation in Madagascar." Companion Paper for the World Bank Policy Research Report on Forests, Environment, and Livelihoods.
URL: *http://www.timthomas.net*

Tiefelsdorf, Michael. 2000. *Modeling Spatial Processes: The Identification and Analysis of Spatial Relationships in Regression Residuals by Means of Moran's I*. Vol. 87 of *Lecture notes in earth sciences* Berlin: Springer Verlag Berlin Heidelber.

Verbeek, Marno. 2008. *A guide to modern econometrics*. Chichester: John Wiley & Sons.

Vijverberg, Wim P.M. 1997. "Monte Carlo evaluations of multivariate normal probabilities." *Journal of Econometrics* 76:281–307.

Vijverberg, Wim P.M. 1999. "Rectangular and Wedse-Shape Multivariate Normal Probabilities." Working Paper, School of Social Sceinces, University of Texas at Dallas.

Volden, Craig. 2006. "States as policy laboratories: Emulating success in the children's health insurance program." *American Journal of Political Science* 50(2):294–312.

Weyland, Kurt. 2007. *Bounded rationality and policy diffusion: Social sector reform in Latin America*. Princeton: Princeton University Press.

Appendix: Derivative for the linearization of the GMM BSAR model

The gradient (16) of interest to the linearization proposed in Klier and McMillen (2008) is the derivative of the logistic link function to the spatial autoregressive parameter ρ:

$$G_{\rho i} = \frac{\partial P_i}{\partial \rho} = \frac{\partial}{\partial \rho}\left[1 + exp\left(q_i\right)\right]^{-1} = \frac{\partial}{\partial \rho}\left[1 + exp\left(\frac{-\Psi^{-1}\mathbf{x}_i'\boldsymbol{\beta}}{\sigma_{ei}}\right)\right]^{-1},$$

where $\Psi = (\mathbf{I} - \rho\mathbf{W})$ and $\mathbf{q}$ as in equation (7). We derive the following gradient

$$G_{\rho i} = P_i(1 - P_i)\left(h_i + \frac{q_i}{\sigma_{ei}}\frac{\partial \sigma_{ei}}{\partial \rho}\right),$$

where $\boldsymbol{h}$ is defined in equation (16) and

$$\begin{aligned}\frac{\partial \sigma_{ei}}{\partial \rho} &= \frac{\partial}{\partial \rho}\left[(\Psi'\Psi)^{-1}\right]^{\frac{1}{2}} \\ &= -\frac{1}{2\sigma_{ei}}(\Psi'\Psi)^{-1}(\Psi + \Psi')\mathbf{W}(\Psi'\Psi)^{-1}. \end{aligned} \tag{17}$$

If Ψ is symmetric, equation (17) simplifies to:

$$\frac{\partial \sigma_{ei}}{\partial \rho} = -\frac{1}{\sigma_{ei}}\Psi^{-1}\mathbf{W}\Psi^{-1}\Psi^{-1},$$

which leads to equation (16).