

Wachowiak, Mark; Wachowiak-Smolikova, Renata; Zimmerling, Jonathan

Conference Paper

The Viability of Global Optimization for Parameter Estimation in Spatial Econometrics Models

52nd Congress of the European Regional Science Association: "Regions in Motion - Breaking the Path", 21-25 August 2012, Bratislava, Slovakia

Provided in Cooperation with:

European Regional Science Association (ERSA)

Suggested Citation: Wachowiak, Mark; Wachowiak-Smolikova, Renata; Zimmerling, Jonathan (2012) : The Viability of Global Optimization for Parameter Estimation in Spatial Econometrics Models, 52nd Congress of the European Regional Science Association: "Regions in Motion - Breaking the Path", 21-25 August 2012, Bratislava, Slovakia, European Regional Science Association (ERSA), Louvain-la-Neuve

This Version is available at:

<https://hdl.handle.net/10419/120602>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

The Viability of Global Optimization for Parameter Estimation in Spatial Econometrics Models

Mark P. Wachowiak, Renata Wachowiak-Smolíková, Jonathan Zimmerling
Department of Computer Science and Mathematics
Nipissing University, North Bay, ON P1B 8L7 Canada
{markw, renatas}@nipissingu.ca, jzimmerling794@community.nipissingu.ca

Theme: O. Spatial econometrics and analysis

Keywords: Spatial econometrics, global optimization, parameter estimation, particle swarm optimization

JEL: C02, C13, C61

Abstract

This paper addresses parameter estimation of spatial regression models incorporating spatial lag. These models are very important in spatial econometrics, where spatial interaction and structure are introduced into regression analysis. However, parameters of spatial lag models are difficult to estimate due to simultaneity bias. These parameter estimation problems are generally intractable by standard numerical methods, and, consequently, robust and efficient optimization techniques are needed. In this paper, global optimization (specifically, particle swarm optimization, or PSO) is used to estimate parameters of spatial autoregressive models. PSO was tested with an autoregressive spatial model for which no analytic initial guess can be computed, and for which no analytic parameter estimation method is known. The results indicate that global optimization is a viable approach to estimating the parameters of spatial autoregressive models, and suggest that future directions should focus on more advanced global techniques, such as branch-and-bound, dividing rectangles, and differential evolution, which may further improve parameter estimation in spatial econometrics applications.

1. Introduction

Econometrics combines economics and statistics to analyze and model economic data. These data are heavily based on observations, and are used to create models to either support or contradict specific economic theories. Since economic units and processes have varying levels of interaction, quantitative methods often estimate the equilibrium of the observed system.

As one of the newer econometrics specializations, spatial econometrics focuses on incorporating spatial relationships into economic models (LeSage, Kelley Pace, 2009). Spatial factors, which are most often geographic distances, may also include cultural or political relationships. It is important to accurately estimate the parameters in spatial models to under-

stand the effects of spatial relationships and to develop more robust predictive models. In many cases, a best estimator exists for parameters for many of these models. However, in other cases, no exact estimator exists.

For dynamic panel data models extended with a spatially lagged dependent variable, two approximations are used, as there is no exact estimator (Elhorst, 2003). Even with these approximations, the parameters are very difficult to estimate, as the derivative of the likelihood function is, in most practical cases, unattainable. Consequently, iterative numerical methods are required. Little work has been done towards determining suitable methods for maximizing these likelihoods, and the literature currently focuses on simpler models.

This paper focuses on using global optimization to estimate the parameters of the dynamic panel data model extended with a spatially lagged dependent variable (Elhorst, 2003, 2005). Given the complexity of the likelihood functions involved, a global approach is immediately preferred over a local one.

2. Spatial Regression Models

In regional science, relationships may be geographic, cultural, political, or economic, and multiple relationships can be considered in the same model. These relationships are denoted by $N \times N$ matrix (where N is the number of units) of spatial weights, denoted as \mathbf{W} . The i th row of \mathbf{W} represents the relationships of the i th unit with all other units, where the row is normalized to sum to unity, and the relationship between a unit and itself is assumed to be 0. Spatial dependence or autocorrelation can be described in different ways in the context of the standard linear regression model. The spatial lag model, or spatial autoregressive model, is given by (Anselin, 1988):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \kappa\mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon} \quad (1)$$

where \mathbf{y} is a vector of observations on the dependent variable, $\boldsymbol{\beta}$ is a vector of coefficients (usually to be determined), κ is a spatial autoregressive coefficient, \mathbf{W} is the spatial weights matrix with characteristic roots ω_i , \mathbf{X} is a matrix of observations on independent variables, and $\boldsymbol{\varepsilon}$ is the vector of error terms. The dependent variable for a unit relies on both the observed data associated with it and on that of all other units as well – the *spatial lag*. This system is difficult to estimate because of the simultaneity bias (the dependent variable \mathbf{y} is in a feedback relationship with the independent variable \mathbf{X}) encountered when utilizing spatial

analogues of ordinary least squares (Franzese and Hays, 2007), but can be computed with maximum likelihood estimation (Ord, 1975).

A system with some spatial interdependence that does not directly affect the explanatory observations is the spatial error model, and is given by (Anselin, 1988):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varphi} \quad (2)$$

where

$$\boldsymbol{\varphi} = \lambda \mathbf{W}\boldsymbol{\varphi} + \boldsymbol{\varepsilon}, \quad (3)$$

and λ is a spatial error coefficient. It is assumed that the errors are normally distributed with finite variance. The parameters of this model can be estimated with maximum likelihood estimation (Ord, 1975).

Spatial lag and spatial error dependence can be introduced into the cross-sectional dimension of panel data models. The spatial lag specification of such model is given by (Anselin, 1988):

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + \kappa \mathbf{W}\mathbf{y}_t + \boldsymbol{\varepsilon}_t \quad (4)$$

where observations are indexed by spatial unit ($i = 1, \dots, N$) and time period ($t = 1, \dots, T$). Although the relationship matrices may vary over time, this does not affect the model as long as the relationships are known. It can be assumed that the relationship matrix is invariant over time, as it will not affect computational complexity, and the error terms are not spatially or temporally interdependent. Spatial lag can be modeled entirely by interactions between units. The system can be estimated via maximum likelihood estimation (Beck et al., 2006).

Systems may lag temporally as well as spatially. That is, a unit is dependent not only upon the current state of other units, but upon its own previous states (and thus upon the previous states of other units). This model is given as (Beck et al., 2006):

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + \kappa \mathbf{W}\mathbf{y}_{t-1} + \phi \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \quad (5)$$

where ϕ is the temporal lag coefficient. The spatial lag occurs at the previous time, which is realistic in many situations (i.e. a unit does not react instantaneously to other units, but rather after some delay), assuming that the error terms are temporally independent (verified with a Lagrange multiplier test). This model can be estimated using ordinary least squares, as the time delay on the spatial lag removes the simultaneity bias (Beck et al., 2006).

In a system where the spatial lag occurs instantaneously, the model is given by:

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + \kappa \mathbf{W}\mathbf{y}_t + \phi \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \quad (6)$$

which is significantly more difficult to estimate than the previous models. Again, it is assumed that the errors are temporally independent, and that $x_{i,0}$ and $y_{i,0}$ are observable. OLS is ineffective due to simultaneity bias. Maximum likelihood becomes unattainable due to the fact that the covariance matrix of the error term relies on the expected values of pre-sample observations, of which nothing is known. In fact, no satisfactory estimation has been found (Beck et al., 2006). To solve this problem, two approximations, the first based on the Bhargava and Sargan method (BS) and the second on the Nerlove and Balestra method (NB), have been proposed (Elhorst, 2003). The brief derivation that follows [see (Elhorst, 2003) for additional details] is intended to clarify the cost functions that are optimized in this paper.

Taking first differences of Equation (6), the model changes to:

$$\Delta \mathbf{y}_t = \kappa \mathbf{W} \Delta \mathbf{y}_t + \phi \Delta \mathbf{y}_{t-1} + \Delta \boldsymbol{\varepsilon}_t \quad (7)$$

$\Delta \mathbf{y}_t$ is well defined for all $2 \leq t \leq T$, but not for $\Delta \mathbf{y}_1$ because $\Delta \mathbf{y}_0$ is not observed. Thus, the probability function of $\Delta \mathbf{y}_1$ must first be derived. Let $\mathbf{B} = \mathbf{I}_N - \kappa \mathbf{W}$, $\mathbf{A} = \phi \mathbf{B}^{-1} - \mathbf{I}_N$ and let \mathbf{V}_b be the $N \times N$ matrix defined as

$$\mathbf{V}_b = \mathbf{I}_N + \mathbf{A}(\mathbf{I}_N - \phi^2 (\mathbf{B}'\mathbf{B})^{-1})^{-1} \mathbf{A}' - \mathbf{A} \phi^{m-1} \mathbf{B}^{-(m-1)} (\mathbf{I}_N - \phi^2 (\mathbf{B}'\mathbf{B})^{-1})^{-1} \phi^{m-1} \mathbf{B}^{-(m-1)} \mathbf{A}' + \phi^{m-1} \mathbf{B}^{-(m-1)} \phi^{m-1} \mathbf{B}^{-(m-1)}. \quad (8)$$

Let \mathbf{H}_V be an $NT \times NT$ matrix defined as

$$\mathbf{H}_V \equiv \begin{bmatrix} V & -\mathbf{I}_N & 0 & \cdot & 0 & 0 \\ -\mathbf{I}_N & 2 \times \mathbf{I}_N & -\mathbf{I}_N & \cdot & 0 & 0 \\ 0 & -\mathbf{I}_N & 2 \times \mathbf{I}_N & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & 2 \times \mathbf{I}_N & -\mathbf{I}_N \\ 0 & 0 & 0 & \cdot & -\mathbf{I}_N & 2 \times \mathbf{I}_N \end{bmatrix} \quad (9)$$

where \mathbf{V} is an $N \times N$ matrix, and 0 represents an $N \times N$ matrix of zeroes. For fixed $m > 1$,

$$\begin{aligned} \mathbf{B} \Delta \mathbf{y}_t &= \phi^m \mathbf{B}^{-(m-1)} \Delta \mathbf{y}_{t-m} + \Delta \boldsymbol{\varepsilon}_t + \phi \mathbf{B}^{-1} \Delta \boldsymbol{\varepsilon}_{t-1} + \dots \\ &+ \phi^{m-1} \mathbf{B}^{-(m-1)} \Delta \boldsymbol{\varepsilon}_{t-(m-1)} + \sum_{j=0}^{m-1} \phi^j \mathbf{B}^{-j} \Delta \mathbf{X}_{t-j} \boldsymbol{\beta} \\ &= \phi^m \mathbf{B}^{-(m-1)} \Delta \mathbf{y}_{t-m} + \Delta \boldsymbol{\varepsilon}_t + \mathbf{X}^* \end{aligned} \quad (10)$$

Since \mathbf{X}_t is stationary, it follows that $E[\Delta \mathbf{X}_t] = 0$. Thus $E[\mathbf{B}\Delta \mathbf{y}_1] = \phi^m \mathbf{B}^{-(m-1)} \Delta \mathbf{y}_{t-m}$. Because \mathbf{X}^* is not observed, $\text{Var}[\mathbf{B}\Delta \mathbf{y}_1]$ cannot be determined. The BS approximation suggests that the optimal predictor of \mathbf{X}^* when $t = 1$ is $\mathbf{X}^* = \pi_0 \mathbf{1}_N + \Delta \mathbf{X}_1 \pi_1 + \dots + \Delta \mathbf{X}_T \pi_T + \xi$ where $\xi_i \sim N(0, \sigma_\xi^2 \mathbf{I}_N)$, π_0 is a scalar and π_t , ($t = 1, \dots, T$) are $K \times 1$ vectors of parameters. It follows that:

$$\mathbf{B}\Delta \mathbf{y}_1 = \pi_0 \mathbf{1}_N + \Delta \mathbf{X}_1 \pi_1 + \dots + \Delta \mathbf{X}_T \pi_T + \Delta \mathbf{e}_1 \text{ where } \Delta \mathbf{e}_1 = \xi + \sum_{j=0}^{m-1} \phi^j \mathbf{B}^{-j} \Delta \varepsilon_{1-j}. \quad (11)$$

$$E[\Delta \mathbf{e}_1] = 0, \quad E[\Delta \mathbf{e}_1 \Delta \mathbf{e}_2'] = -\sigma^2 \mathbf{I}_N, \quad E[\Delta \mathbf{e}_1 \Delta \mathbf{e}_t'] = 0, (t = 3, \dots, T)$$

$$E[\Delta \mathbf{e}_1 \Delta \mathbf{e}_1'] = \sigma_\xi^2 \mathbf{I}_N + \sigma^2 \mathbf{V}_b \quad (12)$$

Now, let $\theta^2 = \sigma_\xi^2 / \sigma^2$, yielding $E[\Delta \mathbf{e}_1 \Delta \mathbf{e}_1'] = \sigma^2 (\theta^2 \mathbf{I}_N + \mathbf{V}_b)$. Define $\mathbf{V}_{BS} = \theta^2 \mathbf{I}_N + \mathbf{V}_b$. The covariance matrix of $\Delta \mathbf{e}$ can be written as $\text{Var}(\Delta \mathbf{e}) = \sigma^2 [(\mathbf{I}_T \otimes \mathbf{B}^{-1}) \mathbf{H}_{V_{BS}} (\mathbf{I}_T \otimes \mathbf{B}^{-1})]$ where

$\mathbf{H}_V |_{V=V_{BS}}$ is given in (9). Then the log-likelihood function becomes:

$$\begin{aligned} \ln \mathcal{L} = & \frac{-NT}{2} \ln(2\pi\sigma^2) + T \sum_{i=1}^N \ln(1 - \kappa \omega_i) \\ & - \frac{1}{2} \sum_{i=1}^N \ln \left[1 - T + \frac{2T(1 - \kappa \omega_i)}{(1 - \kappa \omega_i + \phi)} \left(1 + \left(\frac{\phi}{1 - \kappa \omega_i} \right)^{2m-1} + T\phi^2 \right) \right] \\ & - \frac{1}{2\sigma^2} \Delta \mathbf{e}' \mathbf{H}_{V_{BS}}^{-1} \Delta \mathbf{e} \end{aligned} \quad (13)$$

with

$$\Delta \mathbf{e} = \begin{bmatrix} \mathbf{B}\Delta \mathbf{y}_1 - (\pi_0 \mathbf{1}_N + \Delta \mathbf{X}_1 \pi_1 + \dots + \Delta \mathbf{X}_T \pi_T) \\ \mathbf{B}\Delta \mathbf{y}_2 - \phi \Delta \mathbf{y}_1 - \Delta \mathbf{X}_2 \beta \\ \dots \\ \mathbf{B}\Delta \mathbf{y}_T - \phi \Delta \mathbf{y}_{T-1} - \Delta \mathbf{X}_T \beta \end{bmatrix}, \quad E[\Delta \mathbf{e} \Delta \mathbf{e}'] = \sigma^2 \mathbf{H}_{V_{BS}}. \quad (14)$$

The BS approximation gives a log-likelihood function containing $KT+K+5$ parameters to be estimated: $\pi_1, \dots, \pi_T, \beta, \pi_0, \theta^2, \phi, \kappa$, and σ^2 . The parameters σ^2, π, β can be solved from their first-order maximizing conditions (Elhorst, 2003).

With the Nerlove and Balestra approximation, $\text{Var}[\mathbf{B}\Delta \mathbf{y}_1]$ can be approached by:

$\text{Var}[\mathbf{B}\Delta y_1] = \text{var}[\Delta \mathbf{e}_1] + \text{var}[\mathbf{X}^*] = \sigma^2 \mathbf{V}_b + \sum_{\mathbf{X}^*}$, where
 $\sum_{\mathbf{X}^*} = (\mathbf{I}_N - \phi \mathbf{B}^{-1})^{-1} (\mathbf{I}_N - \phi^m \mathbf{B}^{-m}) \beta' \sum_{\Delta \mathbf{X}} \beta (\mathbf{I}_N - \phi^m \mathbf{B}^{-m}) (\mathbf{I}_N - \phi \mathbf{B}^{-1})^{-1}$ and $\sum_{\Delta \mathbf{X}}$ is the covariance matrix of $\Delta \mathbf{X}$. Let $\mathbf{v}_{NB} = \mathbf{v}_b + \frac{1}{\sigma^2} \sum_{\mathbf{X}^*}$. Then the covariance matrix of $\Delta \mathbf{e}$ can be written as $\text{Var}(\Delta \mathbf{e}) = \sigma^2 [(\mathbf{I}_T \otimes \mathbf{B}^{-1}) \mathbf{H}_{V_{NB}} (\mathbf{I}_T \otimes \mathbf{B}^{-1})]$. Using the matrix properties of $\mathbf{H}_{V_{NB}}$ and \mathbf{W} , the log likelihood function becomes:

$$\begin{aligned}
\ln \mathcal{L} = & \frac{-NT}{2} \ln(2\pi\sigma^2) + T \sum_{i=1}^N \ln(1 - \kappa \omega_i) \\
& - \frac{1}{2\sigma^2} \Delta \mathbf{e}' \mathbf{H}_{V_{NB}}^{-1} \Delta \mathbf{e} - \frac{1}{2} \sum_{i=1}^N \ln \left[1 - T + \frac{2T(1 - \kappa \omega_i)}{(1 - \kappa \omega_i + \phi)} \left(1 + \left(\frac{\phi}{1 - \kappa \omega_i} \right)^{2m-1} \right) \right. \\
& \left. + T \frac{\beta' \sum_{\Delta \mathbf{X}} \beta (1 - \kappa \omega_i + \phi)^2}{\sigma^2 (1 - \kappa \omega_i)^2} \left(1 - \left(\frac{\phi}{1 - \kappa \omega_i} \right)^m \right)^2 \right]
\end{aligned} \tag{15}$$

where

$$\Delta \mathbf{e} = \begin{bmatrix} \mathbf{B}\Delta \mathbf{Y}_1 - \pi_0 \mathbf{1}_N \\ \dots \\ \mathbf{B}\Delta \mathbf{Y}_T - \phi \Delta \mathbf{Y}_{T-1} - \Delta \mathbf{X}_T \beta \end{bmatrix}, \quad E[\Delta \mathbf{e} \Delta \mathbf{e}'] = \sigma^2 \mathbf{H}_{V_{NB}}. \tag{16}$$

This approximation results in a log-likelihood function containing $K+4$ parameters to be estimated: β , π_0 , ϕ , κ , and σ^2 . None of these parameters can be solved analytically from the first-order maximizing conditions.

To the authors' knowledge, these approximations have not received much attention in the literature. The complexity of the covariance matrices makes derivative calculations of the likelihood very difficult. For instance, consider the NB approach. $\Delta \mathbf{e}' \mathbf{H}_{V_{NB}}^{-1} \Delta \mathbf{e}$ will resolve to a term relying on $N^4 T^4$ matrix entries, each of which contains at least one model parameter. The complex structure of $\mathbf{H}_{V_{NB}}^{-1}$ precludes developing a closed form solution, so a derivative must be developed for all combinations of values of N and T . For any meaningful sample sizes, computing these derivatives becomes prohibitively difficult, and even if derivatives can be found, none of them will be isolated, and would have to be solved via a multi-stage iterative approach, a subject to simultaneity bias. Thus, a numerical solution must be employed. Using generated test data, the PSO global optimization method is employed to find the maximum likelihood estimators of the model parameters.

3. Particle Swarm Optimization

Particle swarm optimization (PSO) (Kennedy et al., 2001) is an effective, iterative global optimization method for continuous, complicated functions with complex search spaces, and has received much attention in the literature. PSO begins with a population of “agents” or “particles”, each with a random starting location and velocity in n dimensions (n is the number of parameters of the cost function). Each particle has a **pbest** vector containing the location of that particle’s best cost function value of the cost function. The vector of the best location achieved by all the agents is called **gbest**. Two uniformly distributed random numbers u_1 and u_2 are generated as weights for the local (personal) and global component of the particle’s velocity. Both these values are multiplied by a constant, generally 2.0, so that the particles overshoot their desired location about half the time. The velocity update for the d^{th} dimension of the i^{th} agent (whose position will be denoted by \mathbf{x}_i) is updated as:

$$\mathbf{v}_{i,d} \leftarrow K[\mathbf{v}_{i,d} + c_1 u_1(\mathbf{pbest}_{i,d} - \mathbf{x}_{i,d}) + c_2 u_2(\mathbf{gbest}_{i,d} - \mathbf{x}_{i,d})] \quad (17)$$

where $K = \frac{2}{|2 - \varphi - \sqrt{\varphi^2 - 4\varphi}|}$, $\varphi = c_1 + c_2$, $c_1 = c_2$, and $\varphi > 4$. φ is generally chosen as 4.1,

yielding $K = 0.729$. The particle’s position in the search space in the next iteration is then given as: $\mathbf{x}_{i,d} \leftarrow \mathbf{x}_{i,d} + \mathbf{v}_{i,d}$.

If a maximum velocity V_{\max} is set, the velocity of each particle is clipped to lie within $[-V_{\max}, V_{\max}]$. The initial search is very global, but the constriction coefficient will decrease the effect of each particle’s previous velocity over time, resulting in a more local search.

4. Methods

4.1 Data Generation

The data for model (6) were generated as follows:

$$\mathbf{X}_i \sim \mathbf{N}(0,1), \quad \varepsilon_i \sim \mathbf{N}(0,0.25), \quad \mathbf{w}_{i,j} = 1, \quad i = j+1, i = j-1, \quad \mathbf{w}_{i,j} = 0, \text{ otherwise}$$

$$\kappa \sim \mathbf{U}(0,1), \quad \phi \sim \mathbf{N}(0,4), \quad \beta \sim \mathbf{N}(0,4), \quad \mathbf{Y}_t = (\mathbf{I}_N - \kappa \mathbf{W})^{-1}(\mathbf{X}_t \beta + \phi \mathbf{Y}_{t-1} + \varepsilon_t).$$

The condition of $|\phi(\mathbf{I}_N - \kappa \mathbf{W})| < 1$ was imposed to ensure that as $T \rightarrow \infty$, y_t does not diverge more quickly than T . The first value of \mathbf{y} was generated by $\mathbf{y}_{\text{initial}} = (\mathbf{I}_N - \kappa \mathbf{W})^{-1}(\mathbf{X}_{\text{initial}} \beta + \varepsilon_{\text{initial}})$.

For any desired sample size (N, T) , $3T + 1$ sets of N data were generated. The last T data sets are chosen to be passed to the estimators, and $\mathbf{y}_0 = \mathbf{y}_{2T}$, $\mathbf{X}_0 = \mathbf{X}_{2T}$. This is done to

mitigate the effects of the initial seed value, and to more accurately simulate a system where the process has started at some point before observation begins.

4.2 Estimation of Parameters Using Particle Swarm Optimization

The particle swarm procedure, also implemented in Matlab, was thoroughly tested with ground truth parameters, and optimized for efficiency. The maximum velocity of $\hat{\kappa}$ was set as 0.005. The procedure ran for 3000 iterations, or until the average distance between the particles and the centroid of all the particles was less than 0.001. The R^2 values of each model were computed to determine the goodness-of-fit for each estimation, which was deemed successful if an R^2 value greater than 0 was achieved ($R^2 < 0$ indicates that the model was misspecified).

The values for \mathbf{X} , \mathbf{Y} , and \mathbf{W} were passed to a PSO procedure for each approximation of the model, with 100 agents with starting parameters for each agent being set as:

$$\hat{\kappa} \sim \mathbf{N}(0.5, 0.0225), \hat{\phi} \sim \mathbf{N}(0, 4), \hat{\beta} \sim \mathbf{N}(0, 4), \hat{\sigma}^2 \sim \mathbf{N}(1, 0.25), \\ \hat{\pi}_0 \sim \mathbf{N}(0, 4), \hat{\sigma}_\epsilon^2 \sim \mathbf{N}(1, 1).$$

The maximum velocity was set to 0.5 for $\hat{\beta}$, and to 0.1 for all other parameters. If the particles did not converge in 1000 iterations, the procedure was restart, up to a maximum of 3000 iterations. 200 trials for each sample size were executed.

5. Results

For the first set of trials, a sample size of $N = 6$, $T = 4$ was generated. The Nerlove and Bales-tra approximation yielded 175 successful trials (87.5% success rate), the Bhargava and Sar-gan approximation yielded 176 (88% success), and there were 159 (79.5%) sets of data where both approximations were successful. The data on which both approximations were success-ful were used to compute the following averages:

Table 1. Results of parameter estimation, $N = 6$, $T = 4$.

	Bias Kappa	Absolute Bias Kappa	Bias Phi	Absolute Bias Phi	Bias Beta	Absolute Bias Beta	R^2	Iterations
NB	0.033	0.073	0.030	0.179	0.017	0.439	0.867	1454.260
BS	0.032	0.065	-0.040	0.178	0.731	1.536	0.919	1146.481

The PSO procedure converged in 103 (51.5%) of trials when used with the NB method, taking an average of 474 iterations, and converging in 128 (64%) of trials on the BS approximation taking an average of 1198 iterations.

The second sample size was chosen as $N = 10$ and $T = 6$. PSO returned $R^2 > 0$ for BS on 170 (85%) of trials, and for NB 192 (96%) of trials. There were 168 (84%) sets of data in which both approximations were successful. For those data, the procedure converged in 95 (47.5%) of NB trials with a mean of 434 iterations, and converged in 114 (57%) of BS trials (mean of 443 iterations).

Table 2. Results of parameter estimation, $N = 10$, $T = 6$.

	Bias Kappa	Absolute Bias Kappa	Bias Phi	Absolute Bias Phi	Bias Beta	Absolute Bias Beta	R^2	Iterations
NB	0.024	0.048	0.003	0.141	-0.262	0.471	0.913	1581.339
BS	0.031	0.050	-0.001	0.146	-0.364	0.676	0.909	1281.601

The third sample size was chosen as $N = 20$, $T = 10$. The NB approximation produced 180 (90%) successes, and the Bhargava and Sargan approximation produced 196 (98%) successes. There were 178 (89%) trials where both approximations were successfully estimated.

Table 3. Results of parameter estimation, $N = 20$, $T = 10$.

	Bias Kappa	Absolute Bias Kappa	Bias Phi	Absolute Bias Phi	Bias Beta	Absolute Bias Beta	R^2	Iterations
NB	0.005	0.025	-0.023	0.076	0.089	0.674	0.908	1902.185
BS	0.004	0.028	-0.046	0.084	0.148	0.609	0.912	1706.023

The PSO procedure converged for 91 (40.5%) of NB trials, with an average of 853 iterations. There was convergence for 100 (50%) of BS trials, with an average of 689 iterations.

The final trial size was $N = 40$, $T = 20$. Due to the computational time required by the likelihood functions on such a large sample size, only 134 trials were run. The NB approximation produced 64 (47.8%) successes, while the BS approximation produced 81 (60.4%) successes. There were 40 (29.8%) trials where both approximations were successful. The following averages are taken from the 40 jointly successful trials.

Table 4. Results of parameter estimation, $N = 40$, $T = 20$.

	Bias Kappa	Absolute Bias Kappa	Bias Phi	Absolute Bias Phi	Bias Beta	Absolute Bias Beta	R^2	Iterations
NB	-0.002	0.013	-0.032	0.043	0.049	0.999	0.924	3000
BS	-0.013	0.019	-0.025	0.037	-0.337	1.639	0.876	3000

The PSO procedure did not converge for any of the trials, and therefore the maximum number of iterations (3000) was expended.

5. Discussion

It is apparent that even for a simple function such as the likelihood function of the general spatial model, a global optimization method rather than a local one should be used when no good initial guess of the parameters is available. The relatively low computational cost of the function makes PSO an attractive option even for large sample sizes. Moreover, PSO exhibited good convergence on the model used, further reducing computational cost.

PSO is also a viable tool for optimizing the likelihood functions of the approximations of the time series cross-sectional spatiotemporal autoregressive model, although there is room for improvement. For $\hat{\kappa}$ and $\hat{\phi}$, a decrease in the average of the absolute bias as N and T increase was observed, indicating consistency. The estimation of these parameters also appears to be unbiased, as the average bias is quite close to zero, neither overestimating nor underestimating the parameters. The approximations of the model are expected to improve as the sample size grows.

However, the estimation of $\hat{\beta}$ is not consistent. The mean of the absolute values of the bias do not decrease as the sample size increases. In fact, it is larger when $N = 40$, $T = 20$ than when $N = 6$, $T = 4$. Furthermore, the BS approximation seems to overestimate this parameter, which may be due to the choice of maximum velocity for particles traveling along the $\hat{\beta}$ dimension, chosen as 0.5 rather than 0.1, as this value led to better and more frequent convergence during preliminary trials. There are currently no well-defined heuristics for choosing maximum velocities, and further study is necessary to determine whether this is due to the choice in velocity, and what would be a more suitable choice.

The convergence rate is also problematic, as there is a clear upward trend. As the sample size increases, a small change in the value of a parameter has a larger impact. The search space becomes more complex as the sample size increases, and, consequently, more iterations are required to achieve better results.

A problem with increasing the number of iterations is the computational cost of PSO. As the sample size increases so does the time required to compute the likelihood. For example, when the sample size is $N = 40$, $T = 20$ the likelihood functions begin to take over to a tenth of a second for each computation; with 100 particles, the computation quickly time consuming. Even if the likelihood function can be computed in a tenth of a second, for a data set that does not converge, 30,000 seconds, or over 8 hours are required, which is too long for estimating a large number of data sets. Parallelization is one way to address the efficiency issue. In 1995, Schnabel (Schnabel, 1995) outlined three key ways in which optimization procedures can be effectively parallelized. The first option is to parallelize the cost function itself. In the case presented, this would not be greatly beneficial, as the routines for distributed matrix computing do not offer a significant advantage. The second strategy is to parallelize the linear algebra routines, which may offer some benefit, as many matrix operations and inverse computations are involved. The third approach is to parallelize the procedure itself. That is, the particles can be distributed across processors, allowing multiple likelihood functions to be computed simultaneously. Since most of the computation time is consumed with cost functions, large efficiency gains can result.

Although there is room for improvement in optimizing the two approximations can be improved with PSO, other methods may be considered as well. Although the usual iterative method for solving generalized method of moments leads to overestimation, generalized method of moments in concert with PSO may lead to unbiased results. In addition, although the literature currently suggests that BS and NB are the most accurate, proven unbiased and consistent estimators may yet be developed.

6. Conclusions

This paper explored the estimation of the parameters of the time-series cross-sectional spatio-temporal autoregressive model using two different approximations of the maximum likelihood combined with particle swarm optimization. The results suggest that unbiased estimation of two of the three parameters of interest can be performed, with suggestions on improving the estimation of the third. For specific sets of data these methods are sufficient, although it is important to investigate in future work whether the failure of the estimation technique in some cases is due to the approximation, to the choice of PSO parameters, or to other factors. The model presented in this paper remains difficult, but, with PSO, a step has been taken towards making this model more practical.

References

1. Anselin, L. (1988). *Spatial econometrics: methods and models*. Springer.
2. Beck, N., Gleditsch, K.S., Beardsley, K. (2006). Space is more than geography: using spatial econometrics in the study of political economy. *International Studies Quarterly* 50: 27-44.
3. Bhargava, A., Sargan, J.D. (1983). Estimating dynamic random effects models from panel data covering short time periods. *Econometrica* 51(6): 1635-1659.
4. Elhorst, J.P. (2003). Unconditional maximum likelihood estimation of dynamic models for spatial panels. Research School SOM, Groningen (<http://som.eldoc.ub.rug.nl/03C27>).
5. Elhorst, J.P. (2005). Unconditional maximum likelihood estimation of linear and log-linear dynamic models for spatial panels. *Geographical Analysis* 37: 85-106.
6. Franzese, R.J. Jr., Hays, J.C. (2007). Spatial econometric models of cross-sectional interdependence in political science panel and time-series-cross-section data. *Political Analysis* 15: 140-164.
7. Kennedy, J., Eberhart, R.C., Shi, Y. (2001). *Swarm Intelligence*, Morgan Kaufman.
8. LeSage, J., Kelley Pace, R. (2009) *Introduction to spatial econometrics*. CRC Press.
9. Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association* 70(349): 120-126.
10. Schnabel, R.B. (1995). A view of the limitations, opportunities, and challenges in parallel nonlinear optimization. *Parallel Computing* 21(6): 875-905.