Geographical Analysis ISSN 0016-7363

Estimation Bias in Spatial Models with Strongly Connected Weight Matrices

Tony E. Smith

Department of Systems and Electrical Engineering, University of Pennsylvania, Philadelphia, PA

This article shows that, for both spatial lag and spatial error models with strongly connected weight matrices, maximum likelihood estimates of the spatial dependence parameter are necessarily biased downward. In addition, this bias is shown to be present in general Moran tests of spatial dependency. Thus, positive dependencies may often fail to be detected when weight matrices are strongly connected. The analysis begins with a detailed examination of downward bias for the extreme case of maximally connected weight matrices. Results for this case are then extended by continuity to a broader range of (appropriately defined) strongly connected matrices. Finally, a simulated numerical example is presented to illustrate some of the practical consequences of these biases.

Introduction

In a recent simulation study, Mizruchi and Neuman (2008) showed that, for spatial lag (SL) models with strongly connected (high-density) weight matrices, a severe downward bias is often present in maximum likelihood estimates of the spatial dependency parameter.[1] A similar finding is reported by Farber, Páez, and Volz (2009) in their simulation analysis of the influence of network topology on tests of spatial dependencies. Hence, the central purpose of this article is to clarify the nature of this bias from an analytical perspective. In addition, the same type of bias is present in both spatial error (SE) models and in the more general Moran test of spatial dependency. In all cases, this bias implies that significantly positive spatial dependencies may *not be detected* when weight matrices are strongly connected.

To establish these results, the analytical strategy employed considers the extreme case of *maximally connected* weight matrices and obtains exact results for this case. The rest follows from simple continuity considerations. To avoid repetition, the analytical development of spatial regression models here focuses on SL models. Parallel results for SE models are simply sketched.

Correspondence: Tony E. Smith, Department of Systems and Electrical Engineering, University of Pennsylvania, 220 South 33rd St., Philadelphia, PA 19104
e-mail: tesmith@seas.upenn.edu

Submitted: December 18, 2007. Revised version accepted: September 26, 2008.

We begin with the following a standard SL model for n spatial units:

$$y = \rho W y + X\beta + \varepsilon\,,\ \varepsilon \sim N(0, \sigma^2 I_n) \tag{1}$$

where $y \in R^n$ is some variable of interest, and $X = [1_n, x_1, \ldots, x_k] \in R^{n\times(k+1)}$ represents a relevant set of k explanatory variables, with $1_n = (1, \ldots, 1)'$ denoting the unit n-vector (corresponding to the intercept term in this linear model). (Throughout the following analysis X is always assumed to have *full* column rank, $k+1$, so that $(X'X)^{-1}$ exists.) The unknown parameters of the model include the vector of beta coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_k)'$, the variance σ^2 of each residual in ε, and the spatial dependence parameter ρ, which is of primary interest in the present analysis.

Also of major interest is the structure of the *spatial weight matrix W*. For our analysis, it is convenient to begin by characterizing these matrices in the following way. First, we choose a fixed positive scalar, b, to serve as an *upper bound* on weight values. With respect to this bound, an n-square matrix, $W = (w_{ij} : i, j = 1, \ldots, n)$, is designated as a weight matrix if and only if (i) $w_{ii} = 0$ and (ii) $0 \leq w_{ij} \leq b$ for all $i, j = 1, \ldots, n$. As usual, condition (i) specifies that dependencies are defined only between distinct spatial units. Condition (ii) can be thought of as a normalization condition that allows each weight, w_{ij}, to be interpreted as the "degree of connectivity" between i and j, where $w_{ij} = b$ implies a maximal degree of connectivity. This is particularly appropriate for applications of model (1) to *social networks* among n agents. For the present, the bound b only serves as a convenient conceptual device and can be set equal to one without loss of generality.[2] However, the question of appropriate matrix normalizations for the estimation of ρ is of some importance and is addressed later.

If the class of all n-square weight matrices is denoted by $\mathbf{W}_n \subset R^{n\times n}$ (where the fixed scale parameter b is taken to be implicit), then the relevant geometry of this set can be depicted for the $n = 2$ case as follows. Observe that each matrix $W \in \mathbf{W}_n$ is of the form

$$W = \begin{pmatrix} 0 & w_{12} \\ w_{21} & 0 \end{pmatrix} \tag{2}$$

and thus is fully characterized by the 2-vector (w_{12}, w_{21}). Hence, the entire class $\mathbf{W}_2$ is seen to be equivalent to the points in the square $[0, b]^2$ (Fig. 1). Here, the lower left-hand corner corresponds to the *minimally connected weight matrix* W_* with all zero components, and the upper right-hand corner corresponds to the *maximally connected weight matrix* W^*,[3] with all off-diagonal elements equal to b. This depiction for the 2-by-2 case clarifies that W_* and W^* are the two natural extreme weight matrices in $\mathbf{W}_n$ for all n.[4] Because W_* corresponds to complete *statistical independence* in model (1), attention naturally focuses on those weight matrices, $W \in \mathbf{W}_n$, that are "sufficiently close" to W_* to inherit all of its desirable large-sample properties (such as consistency and asymptotic normality of parameter estimates). Thus, most of the literature focuses on matrices in the lower left-hand neighborhood of Fig. 1.[5]

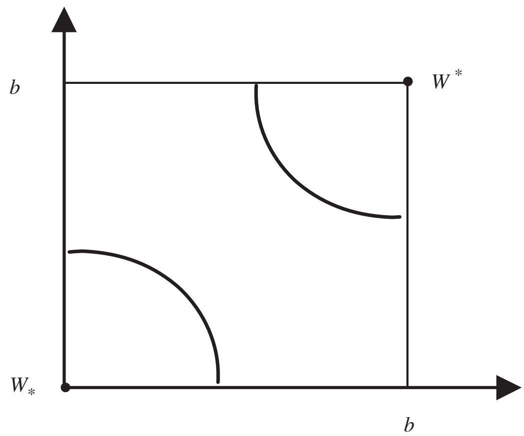

Figure 1. 2 × 2 weight matrices.

In this context, the distinguishing feature of this analysis is its focus on the upper right-hand neighborhood of Fig. 1, which for the moment we loosely designate as ''strongly connected'' weight matrices.[6] The central objective is not only to show that such weight matrices fail to share the desirable properties of the independence case, but also to determine the exact nature of this failure. Of particular interest is the severe *downward bias* in maximum likelihood estimates of the spatial dependency parameter ρ.

To establish this result in a self-contained manner, we begin with a detailed development of the maximum likelihood estimation problem for the SL model in the next section.[7] This is followed in the next section by an analysis of the maximally connected case, W^*, in the upper right-hand corner. In the section, the results for this case are extended by continuity to all matrices ''sufficiently close'' to W^* in an appropriate sense and are illustrated by numerical examples. In the next section, these results are shown to be essentially the same for SE models. Finally, the next section shows that strong connectivity also has consequences for Moran diagnostic tests of spatial independence.

Maximum likelihood estimation for SL models

Model (1) implies that y is multinormally distributed and, in particular, that, for any given data (y, X), the *log-likelihood function* for parameters (β, σ^2, ρ) takes the following form:

$$
\begin{aligned}
L(\beta, \sigma^2, \rho | y, X) = \text{const} - \frac{n}{2}\ln(\sigma^2) + \ln|\det(I_n - \rho W)| \\
- \frac{1}{2\sigma^2}((I_n - \rho W)y - X\beta)'((I_n - \rho W)y - X\beta)
\end{aligned}
\quad (3)
$$

where I_n is the n-square identity matrix, and where all terms not involving the parameters are subsumed in the constant term, const. As with all generalized linear models, one proceeds by first fixing the covariance parameters (in this case ρ) and then maximizing the likelihood function in β and σ^2 to produce the well-known closed-form *conditional estimates*:

$$\hat{\beta}_{sl}(\rho) = (X'X)^{-1}X'(I_n - \rho W)y, \text{ and} \tag{4}$$

$$\hat{\sigma}^2_{sl}(\rho) = (1/n)((I_n - \rho W)y - X\hat{\beta}_{sl}(\rho))'((I_n - \rho W)y - X\hat{\beta}_{sl}(\rho)) \tag{5}$$

where the subscript ''*sl*'' denotes the SL model. These are then substituted into (3) to yield the *concentrated likelihood function*, L_{sl}, for ρ.[8] After some simple canceling of terms, this function takes the form:

$$L_{sl}(\rho|y, X) = const + \ln|\det(I_n - \rho W)| - (n/2)\ln[\hat{\sigma}^2(\rho)] \tag{6}$$

One then maximizes this function to obtain the maximum likelihood estimate $\hat{\rho}_n$ of ρ and then substitutes this value into equations (4) and (5) to obtain corresponding maximum likelihood estimates $\hat{\beta}_n = \hat{\beta}_{sl}(\hat{\rho}_n)$ and $\hat{\sigma}^2_n = \hat{\sigma}^2_{sl}(\hat{\rho}_n)$ of β and σ^2, respectively. However, our primary interest here is in $\hat{\rho}_n$ itself.

To analyze the function L_{sl}, one can make further reductions as follows (see also Anselin 1988, section 12.1.1). First let

$$M = I_n - X(X'X)^{-1}X' \tag{7}$$

denote the *orthogonal projection* onto the complement of the span of X, which by construction satisfies $M = M'$,

$$MX = X - X(X'X)^{-1}X'X = X - X = 0 \tag{8}$$

and

$$MM = (I_n - X(X'X)^{-1}X')(I_n - X(X'X)^{-1}X') = I_n - X(X'X)^{-1}X' = M \tag{9}$$

Substitution of (4) and (7) into (5) then yields the more compact form of the conditional variance estimate:

$$\begin{aligned}\hat{\sigma}^2_{sl}(\rho) &= (1/n)([I_n - X(X'X)^{-1}X'](I_n - \rho W)y)'([I_n - X(X'X)^{-1}X'](I_n - \rho W)y)\\ &= (1/n)(y'(I_n - \rho W)'M(I_n - \rho W)y)\end{aligned} \tag{10}$$

This in turn allows the concentrated likelihood in (6) to be written as

$$L_{sl}(\rho|y, X) = const + \ln|\det(I_n - \rho W)| - (n/2)\ln[y'(I_n - \rho W)'M(I_n - \rho W)y] \tag{11}$$

where the term $-(n/2)\ln(1/n)$ has now been absorbed into the constant.

Further reduction is possible by observing that, if the eigenvalues of W are denoted by $\lambda(W) = \{\lambda_i : i = 1, \ldots, n\}$, then the corresponding eigenvalues of $(I_n - \rho W)$ are well known to be given by $\lambda(I_n - \rho W) = \{1 - \rho\lambda_i : i = 1, \ldots, n\}$. To avoid complications in the following analysis, we restrict our attention to weight matrices, W, with *real* eigenvalues (which, most importantly, includes all W that either are *symmetric* or are row normalizations of symmetric matrices). In addition, the maximum eigenvalue, $\lambda_{\max}(W)$, of W is also assumed to be positive.[9] (In particular, this includes all *nonzero symmetric* weight matrices.) Hence, we now focus on the following subset of weight matrices:

$$\mathbf{W}_n^+ = \{W \in \mathbf{W}_n : \lambda(W) \textit{ is real}, \textit{ and } \lambda_{\max}(W) > 0\} \tag{12}$$

Because the determinant of any matrix is the product of its eigenvalues (Horn and Johnson 1985, theorem 1.2.12), it follows that, for every $W \in \mathbf{W}_n^+$,

$$\det(I_n - \rho W) = \prod\nolimits_i (1 - \rho\lambda_i) \Rightarrow \ln|\det(I_n - \rho W)| = \sum\nolimits_i \ln|1 - \rho\lambda_i| \tag{13}$$

as long as each term $1 - \rho\lambda_i$ on the right-hand side is nonzero. This of course requires further restrictions on ρ. To specify these restrictions, we first note that, because the trace of every matrix is the sum of its eigenvalues, it follows that

$$\sum\nolimits_i \lambda_i = tr(W) = \sum\nolimits_i w_{ii} = 0 \tag{14}$$

for all $W \in \mathbf{W}_n$. But because $\lambda_{\max}(W) > 0$ for all $W \in \mathbf{W}_n^+$, this implies that $\lambda_{\min}(W)$ must be *negative*. These observations together imply that, for any $W \in \mathbf{W}_n^+$, all terms $1 - \rho\lambda_i$ in (13) will be *positive* if the admissible values of ρ are restricted to the *open interval*

$$[W] = \left(\frac{1}{\lambda_{\min}(W)}, \frac{1}{\lambda_{\max}(W)}\right) \tag{15}$$

Hence, we now restrict ρ to the interval $[W]$. Under this restriction, (13) allows (11) to be reduced to the explicit form of

$$L_{sl}(\rho|y, X) = const + \sum\nolimits_i \ln|1 - \rho\lambda_i| - (n/2)\ln[y'(I_n - \rho W)'M(I_n - \rho W)y] \tag{16}$$

which is more readily analyzed (and computed).

At this point one typically proceeds by observing that, because $\ln|\det(I_n - \rho W)| = -\infty$ on the boundaries of $[W]$, it is reasonable to assume that L_{sl} has a well-defined differentiable maximum in the open interval $[W]$. This will be true as long as the second term in (16) is *bounded above*. The following assumption ensures this condition:

$$M(I_n - \rho W)y \neq 0 \textit{ for all } \rho \in [W] \tag{17}$$

To interpret this assumption, observe that model (1) can be equivalently written as $(I_n - \rho W)y = X\beta + \varepsilon$, where $(I_n - \rho W)y$ represents the value of y after SL effects have been accounted for. If this variable is designated as the *effective value* of y,[10]

$$y_W(\rho) = (I_n - \rho W)y \tag{18}$$

in model (1), then as a parallel to classical regression, it is here assumed that none of the effective values $\{y_W(\rho) : \rho \in [W]\}$ of the given data vector y is perfectly fitted by X (i.e., lies in the span of X). We designate data sets (y, X) satisfying (17) as *W-regular*. Notice that for $\rho = 0$ this implies the usual regularity condition that $My \neq 0$. Data (y, X) satisfying only this (classical regression) condition is simply said to be *regular*.

Biased estimation for the maximally connected case in SL models

Given the simple form of the concentrated likelihood function L_{sl} in (16), one can search for a maximum, $\hat{\rho}_n$, in the interval $[W]$ (typically by standard line search procedures). However, for the maximally connected case, $W^* \in \mathbf{W}_n^+$, this maximization procedure is doomed to fail. Indeed, the main outcome of this section is that, even for regular data sets, *L_{sl} is always unbounded on* $[W^*]$. To establish this, we begin by analyzing the properties of W^*. First, observe that since the n-square unit matrix is constructible as the outer product of $1_n 1_n'$, the maximally connected weight matrix $W^* \in \mathbf{W}_n^+$ can be written as

$$W^* = b \cdot (1_n 1_n' - I_n) = \begin{pmatrix} 0 & b & \cdots & b \\ b & 0 & & \vdots \\ \vdots & & \ddots & b \\ b & \cdots & b & 0 \end{pmatrix} \tag{19}$$

With this explicit form, the following shows that the eigenvalues of W^* are computable in closed form (all proofs are in Appendix A):

Lemma 1. *For all $b > 0$, the eigenvalues of W^* in* (19) *are given by*

$$\lambda(W^*) = \{-b, \ldots, -b, b(n-1)\} \tag{20}$$

where the eigenvalue $-b$ has multiplicity $n-1$.

The second (and most important) property of maximally connected weight matrices is the following identity:

Lemma 2. *If M is the orthogonal projection matrix in* (7) *associated with any data matrix $X = [1_n, x_1, \ldots, x_k]$ for model* (1), *then*

$$M \cdot W^* = -b \cdot M = W^* \cdot M \tag{21}$$

An additional consequence of this result is

Lemma 3. *Every regular data set, (y, X), is W^*-regular.*

With these properties, we are now ready to establish our main result, namely that L_{sl} is unbounded on $[W^*]$. In particular, we show that L_{sl} increases without bound as ρ approaches the lower boundary of $[W^*]$. Observe from (15) and (20) that this *lower boundary point*, ρ_*, is given by

$$\rho_* = 1/\lambda_{\min}(W^*) = -1/b \tag{22}$$

With this definition we now have

Proposition 1. *If* $W = W^*$ *in model* (1), *then for all regular data sets (y, X) and all decreasing sequences* (ρ_m) *in* $[W^*]$, *with* $\lim_{m\to\infty} \rho_m = \rho_*$,

$$\lim_{m\to\infty} L_{sl}(\rho_m|y, X) = \infty \tag{23}$$

From a formal viewpoint, Proposition 1 implies that *no maximum likelihood estimator of* ρ *exists* for model (1) when $W = W^*$.[11] This finding is somewhat surprising, given that the existence of maximum likelihood estimators for model (1) is generally assumed to hold as long as $W \in \mathbf{W}_n$ and $\rho \in [W]$. Moreover, it is interesting that, from a practical viewpoint, such a failure would probably not even be detected by standard software. Indeed, one would typically observe that the line-search algorithm has "converged" to some value of ρ very close to ρ_*.

To gain further insight, this finding can be illustrated by the concentrated likelihood function shown in Fig. 2 (corresponding to the numerical simulation example presented below in the "Consequences for strongly connected weight matrices in SL models" section [for a sample size of $n = 50$]). The "First term" and "Second term" shown in Fig. 2 correspond, respectively, to the log-determinant expression and the log-quadratic expression in the concentrated likelihood function (16) above (see also expressions [A8] and [A9] in Appendix A). Notice that the log-determinant term is always well behaved, because it is a sum of simple concave functions, $\ln(1 - \rho\lambda_i)$, on $[W^*]$. Hence, the culprit here is the log-quadratic term, which in the present case not only diverges to $+\infty$ at ρ_* but does so at a faster rate than the corresponding divergence of the log-determinant to $-\infty$.

Before examining the practical consequences of this result for strongly connected weight matrices, we give an alternative statement of Proposition 1 that also proves useful for applications (as discussed further in the next section). Recall that our basic regularity assumption on data (y, X) was designed to avoid cases where some effective y-value, $y_W(\rho)$, was perfectly fitted by the data, X, in model (1). We now show that Proposition 1 arises from the fact that for *every* data set (y, X) in model (1), X must yield a *perfect fit* to the "effective" y-value $y_{W^*}(\rho_*)$ on the lower boundary of $[W^*]$. This depends critically on the presence of an intercept term in model (1) (as is also apparent from the proof of Lemma 2 in Appendix A). This intercept term can now be made explicit by rewriting model (1) as

$$y = \rho W y + \beta_0 1_n + \tilde{X}\tilde{\beta} + \varepsilon, \varepsilon \sim N(0, \sigma^2) \tag{24}$$

where $\tilde{X} = [x_1, \ldots, x_k]$ and $\tilde{\beta} = (\beta_1, \ldots, \beta_k)'$. For the particular case of W^*, it then follows that, for any choice of $\tilde{X}$,

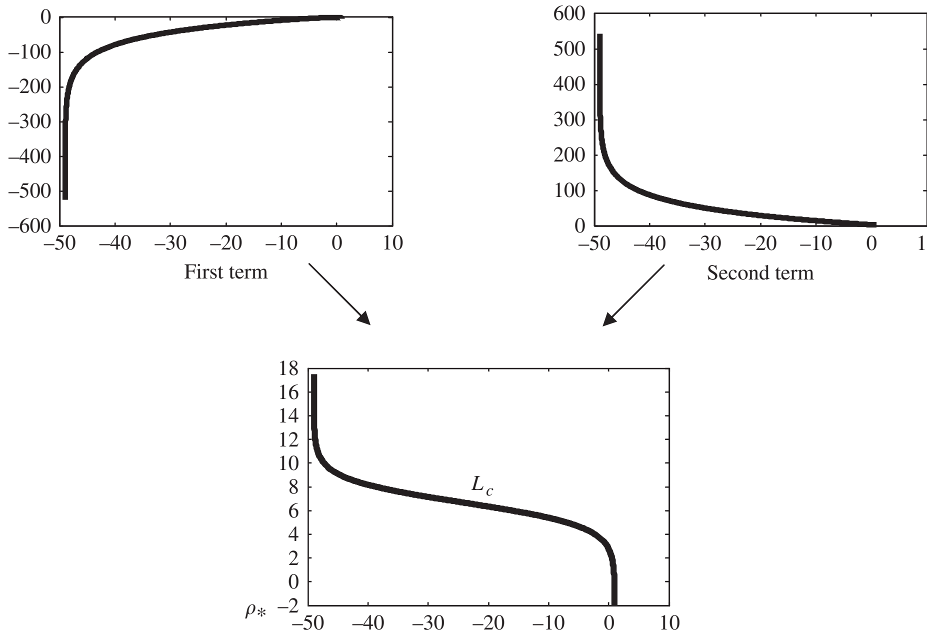

Figure 2. Concentrated likelihood function for W^*.

$$\begin{aligned} y &= \rho W^* y + \beta_0 1_n + \tilde{X}\tilde{\beta} + \varepsilon \\ &\Rightarrow (I_n - \rho W^*)y = \beta_0 1_n + \tilde{X}\tilde{\beta} + \varepsilon \\ &\Rightarrow y_{W^*}(\rho) = \beta_0 1_n + \tilde{X}\tilde{\beta} + \varepsilon \end{aligned} \tag{25}$$

With this notation, recall from Lemma 3 and expression (17) that for any regular data set $(y, X) = (y, [1_n, \tilde{X}])$ there exists no $\rho \in [W^*]$ such that the effective value $y_{W^*}(\rho)$ is a perfect fit to X; that is,

$$y_{W^*}(\rho) = \beta_0(\rho)1_n + \tilde{X}\tilde{\beta}(\rho) \tag{26}$$

for some choice of $[\beta_0(\rho), \tilde{\beta}(\rho)]$. In spite of this, we now show that condition (26) must *always hold* at the lower boundary value ρ_*, of $[W^*]$[12]:

Proposition 2. *If* $W = W^*$ *in model* (1), *then, for any data set (y, X),*

$$y_{W^*}(\rho_*) = \beta_0(\rho_*) \cdot 1_n + \tilde{X}\tilde{\beta}(\rho_*) \tag{27}$$

where $\beta_0(\rho_*) = 1_n' y$ *and* $\tilde{\beta}(\rho_*) = 0$.

Examining this result in terms of perfect fits provides information about the bias of other parameter estimates. For when $\hat{\rho}_n \approx \rho_*$, expression (27) also suggests that $\hat{\beta}_0 \approx 1_n' y$ and $\hat{\beta}_j \approx 0$ for all $j = 1, \ldots, k$. Moreover, because a perfect fit necessarily implies zero variance for residuals, this suggests that $\hat{\sigma}^2 \approx 0$.[13] The practical consequences of these findings are explored in the next section.

Consequences for strongly connected weight matrices in SL models

The preceding results show that, for the extreme case of maximally connected matrices, we can obtain an exact analytical formulation of the bias inherent in maximum likelihood estimates for SL models. This in turn suggests that such bias should be inherited by matrices $W \in \mathbf{W}_n^+$ that are "close" to W^* in some appropriate sense. To do so, we begin by endowing $\mathbf{W}_n^+$ with a matrix norm that will allow an explicit measure of "closeness." Here there are many choices. For example, the ℓ_1-*norm* of any matrix $A = (a_{ij}) \in R^{n\times n}$ is $\|A\|_1 = \sum_{ij} |a_{ij}|$, and the ℓ_2-*norm* (*Euclidean norm*) of A is $\|A\|_2 = \left(\sum_{ij} a_{ij}^2\right)^{1/2}$.[14] But for weight matrices $W \in \mathbf{W}_n^+$, a more natural choice is the following scaled version of the ℓ_1-norm, which we now designate as the *relative connectivity norm*[15]:

$$\|W\|_{rc} = \frac{\|W\|_1}{\|W^*\|_1} = \frac{1}{b \cdot n(n-1)} \sum_{ij} w_{ij} = \frac{1}{n(n-1)} \sum_{ij} (w_{ij}/b) \tag{28}$$

If (w_{ij}/b) denotes the relative connectivity between units (agents) i and j, then this is simply the average of these relative connectivities over all distinct (i, j) pairs. In the case of binary matrices $W \in \mathbf{W}_n^+$, this easily reduces to the graph-theoretical notion of *average link density*. Given this norm (or any other matrix norm), the induced *distance* between W and W^* is given by

$$\|W - W^*\|_{rc} = \frac{1}{b \cdot n(n-1)} \sum_{ij} |w_{ij} - b| = \frac{1}{n(n-1)} \sum_{ij} [1 - (w_{ij}/b)] \tag{29}$$

The discussion thus far has not considered the actual *magnitude* of ρ. All that has been required for a given weight matrix, W, is that these values lie in the open interval $[W]$ of expression (15) and that this interval contains zero (so that both positive and negative values of ρ are always possible). Further insight can be gained by evaluating this interval in specific cases. The numerical illustration below uses a sample of size $n = 50$. By setting the bound at $b = 1$, it follows from Lemma 1 that for the maximally connected matrix $W^* \in \mathbf{W}_{50}^+$ we obtain the values $\lambda_{\min}(W^*) = -1$ and $\lambda_{\max}(W^*) = 49$. Thus, the corresponding bounds on ρ for this case are

$$\rho \in (-1/b, 1/b(n-1)) = \left(-1, \frac{1}{49}\right) \approx (-1, 0.02) \tag{30}$$

which, from a practical viewpoint, offer only a narrow range for *positive* spatial dependencies. However, because positive dependencies are by far the most common in practice, it seems most prudent to choose b to yield a normalized range of positive values. A natural choice here is to set $b = 1/(n-1)$, so that

$$\lambda_{\max}(W^*) = b(n-1) = (n-1)/(n-1) = 1 \tag{31}$$

This will ensure that the interval $[0,1)$ of nonnegative ρ values used for most applications always lies in $[W^*]$. For $n=50$ we then have

$$[W^*] = (-1/b, 1) = (-(n-1), 1) = (-49, 1) \tag{32}$$

More generally, this normalization implies that[16]

$$[0,1) \subset [W] \textit{for all } W \in \mathbf{W}_n^+ \tag{33}$$

So this same interval of ρ values is available for every choice of $W \in \mathbf{W}_n^+$.[17] Given this normalization, the objective of this section is to extend the bias results for maximally connected weight matrices W^* in Proposition 1 to all weight matrices, $W \in \mathbf{W}_n^+$, that are *strongly connected* in the sense that they are ''sufficiently close'' to W^* in the relative connectivity norm. To do so, we employ the following additional conventions. First, for any given $W \in \mathbf{W}_n^+$ and data set (y, X) for model (1), let $\hat{\rho}_W(y, X)$ denote the maximum likelihood estimator of ρ. As pointed out above, this estimator can fail to exist even when (y, X) is W-regular. But for weight matrices W close to W^* (in relative connectivity), if (y, X) is W-regular, then a differentiable maximum, $\hat{\rho}_W(y, X)$, fails to exist only when L_{sl} is unbounded at the lower boundary of $[W]$. In such cases, we set $\hat{\rho}_W(y, X)$ equal to this lower boundary, so that $\hat{\rho}_W(y, X)$ can be treated as a well-defined value for each W. Next, to quantify the possible bias of these estimates, we focus only on the most important case of *positive* dependencies in model (1), that is, $\rho > 0$,[18] and quantify various degrees of underestimation by inequalities of the form,

$$\hat{\rho}_W(y, X) < \rho/(1+\alpha) \tag{34}$$

where parameter $\alpha > 0$ can be interpreted as a *bias factor*. For example, a bias factor of $\alpha = 1$ implies that $\hat{\rho}_W(y, X)$ is less than half the true value of ρ. More generally, higher bias factors correspond to a more severe underestimation of ρ. With these conventions, we now have the following consequence of Proposition 1:

Proposition 3. *For any regular data set* (y, X) *with* $n \geq 3$ *and any given value* $\rho_0 \in (0,1)$ *of the spatial dependency parameter for model* (1), *there exists for each choice of bias factor* $\alpha \in (0,1)$ *a sufficiently small* $\varepsilon = \varepsilon(\alpha, \rho_0, y, X) > 0$ *such that for all* $W \in \mathbf{W}_n^+$,

$$\|W - W^*\|_{rc} < \varepsilon \Rightarrow \hat{\rho}_W(y, X) < \rho_0/(1+\alpha) \tag{35}$$

In other words, for any degree of bias $\alpha > 0$, there is some threshold level of ''strong connectivity,'' $\|W - W^*\| < \varepsilon$, which is sufficient to ensure this degree of bias. As with all such continuity results, however, Proposition 3 still leaves open the question of how strong this connectivity must be in order to see a substantial effect. While such a question can only be answered definitively by extensive simulations, it is nonetheless possible to illustrate the potential significance of these results by means of a typical example.[19]

Here we set $n = 50$, $k = 2$, and construct x-data (x_1, x_2) by simulating two uniformly distributed random vectors, so that $X = [1_{50}, x_1, x_2]$. Model (1) is then parameterized with $\beta = (\beta_0, \beta_1, \beta_2)' = (1, 2, 3)$ and standard deviation $\sigma = 1$. Again for the sake of illustration, the single value $\rho = 0.5$ was chosen to represent (substantial) positive spatial dependency in model (1). To allow an *average-link-density* interpretation of the matrix norm in (40), only symmetric binary weight matrices were used. A number of such matrices W with different average link densities, $d = \|W\|_{rc} \in (0, 1)$, were randomly sampled. In particular, the values $d \in \{0.30, 0.50, 0.80, 0.90, 0.95, 0.99\}$ were chosen for study, and for each such d a single matrix, $W_{(d)} \in \mathbf{W}_n^+$, was randomly sampled by independently assigning $w_{ij} = 1$ with probability d and $w_{ij} = 0$ otherwise.[20]

To make the results at different density levels more comparable, each matrix W_d was normalized in the same manner as W^*, by dividing W_d by its maximum eigenvalue. This rescaling ensures that the positive values of ρ in each simulated model are *exactly the same*, namely, $\rho \in (0, 1)$.[21] For each of these matrices, 1,000 y-vectors were then simulated from model (1), and the corresponding maximum likelihood estimates $\{\hat{\rho}_d(s) : s = 1, .., 1000\}$ were computed.[22] Perhaps the simplest way to summarize these results is to compare the sample mean values of $\hat{\rho}_d$ for each of these average link densities with the true value, $\rho = 0.50$, as in column 2 of Table 1.

As expected, one sees underestimation in all cases, with steadily increasing severity for higher densities. For comparison, the maximally connected case, $d = 1$, has been added to show that this extreme case is vastly worse than all others. The continuity properties in Proposition 3, however, are still evident. Underestimation becomes quite severe as average link density increases. Note also that in Table 1 the corresponding ρ-intervals, $[W_d]$, in (15) above are given in column 4 (column 3 is discussed below).

To provide a fuller comparison, selected histograms of $\{\hat{\rho}_d(s) : s = 1, .., 1000\}$ are shown for the cases $d = 0.50$, 0.80, 0.90, 0.99 in Fig. 3.[23] Here the true value, $\rho = 0.50$, is indicated by a bold arrow in each case to facilitate the visual comparison of these estimates. So at average-link-density levels of at least 80% ($d \geq 0.80$), there is a substantial downward bias in ρ estimates. Another way to see

Table 1 Mean Estimates of ρ

Average link density	Mean $\hat{\rho}$ for SL models	Mean $\hat{\rho}$ for SE models	ρ-interval
0.30	0.481	0.195	(−2.49, 1)
0.50	0.454	−0.038	(−3.51, 1)
0.80	0.369	−0.801	(−6.31, 1)
0.90	0.168	−1.880	(−9.02, 1)
0.95	0.033	−2.281	(−10.7, 1)
0.99	−0.830	−6.363	(−18.9, 1)
1.00	−8.999	−48.999	(−49.0, 1)

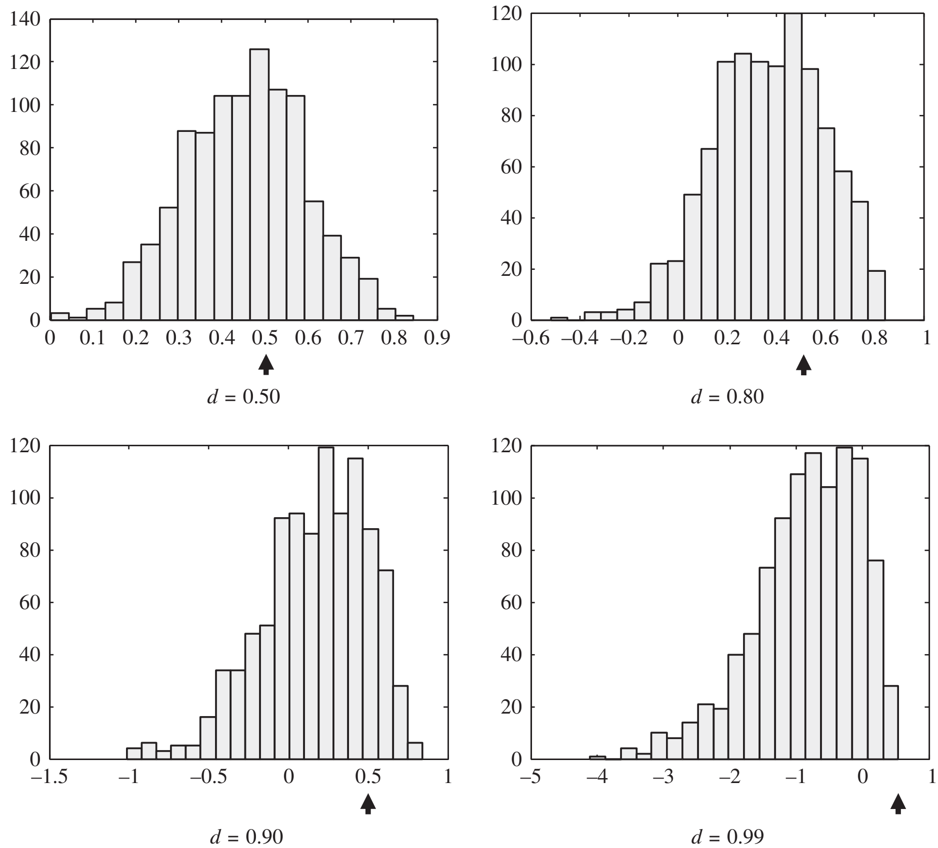

Figure 3. Histograms of ρ estimates.

this is to consider the fraction of estimated values that are significantly different from zero in the standard two-sided test (using asymptotic z values).[24] For a true value of $\rho = 0.50$, only the upper 42% of sample estimates at $d = 0.80$ are significantly different from zero. When the average link density is increased to $d = 0.90$, this drops to $<15\%$. Such significance differences will be investigated further in the next section on ''Extension to SE models.''

Finally, recall from the discussion following Proposition 2 that this underestimation of ρ has consequences for the bias of other parameter estimates. While it is difficult to place definitive magnitudes on the degree of these biases, they can at least be illustrated for the simulations of model (1) above. Here the mean estimates for all parameters are shown in Table 2 (where the means for $\hat{\rho}$ have been repeated from Table 1).

Recall from Proposition 2 that, for the ''perfect fit'' case in the last row of Table 2, one would predict the intercept coefficient $\hat{\beta}_0 \approx 1'_n y$. In the present case, the mean value of $1'_n y$ is about 351, which is in clear agreement with Table 2. Hence,

Table 2 Mean Values of Parameter Estimates for the Spatial Lag Model

Average link density	Mean $\hat{\rho}(\rho = 0.5)$	Mean $\hat{\beta}_0(\beta_0 = 1)$	Mean $\hat{\beta}_1(\beta_1 = 2)$	Mean $\hat{\beta}_2(\beta_2 = 3)$	Mean $\hat{\sigma}^2(\sigma^2 = 1)$
0.30	0.481	1.138	1.978	3.003	0.92887
0.50	0.454	1.336	1.946	2.9973	0.92644
0.80	0.369	1.942	1.922	3.0143	0.91802
0.90	0.168	3.384	1.944	2.9201	0.91908
0.95	0.033	4.302	1.985	2.9209	0.90547
0.99	− 0.830	10.330	1.994	3.012	0.89564
1.00	− 48.999	351.020	0.00004	0.00006	3.8e–010

for strongly connected weight matrices this results in extreme overestimation of β_0 in the present case. Note that, while the limiting estimates of $\beta = (\beta_1, \beta_2)'$ and σ^2 in Table 2 also agree with the zero values predicted by Proposition 2, these biases seem to disappear much more rapidly as link density decreases. However, even a slight downward bias in $\hat{\sigma}^2$ (and hence $\hat{\sigma}$) can have potentially serious consequences for testing and, in particular, can lead to erroneous significance of beta parameters.

Extension to SE models

These results demonstrate that strong connectivity of weight matrices can lead to severe bias in the estimation of spatial dependencies in SL models. It is thus natural to ask whether spatial error models exhibit similar behavior. Our main result shows that for spatial dependence parameters the bias properties of these two models are essentially identical. To establish this, we begin by formulating the SE model and sketching the parallel maximum likelihood estimation problem for this model. As a parallel to model (1), the SE model[25] for n spatial units is defined as

$$y = X\beta + u, u = \rho W u + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n) \tag{36}$$

where now the spatial dependence parameter ρ and the spatial weight matrix W characterize possible spatial dependencies among the residuals rather than in the dependent variable y.[26] If one solves for u and writes this model in reduced form as

$$y = X\beta + (I_n - \rho W)^{-1}\varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n) \tag{37}$$

it becomes clear that ρ and W directly influence the covariance structure of the residuals ε. Again y is multinormally distributed, where the log-likelihood function for parameters (β, σ^2, ρ) in (3) now takes the form

$$\begin{aligned} L(\beta, \sigma^2, \rho | y, X) = {} & const - \frac{n}{2}\ln(\sigma^2) + \ln|\det(I_n - \rho W)| \\ & - \frac{1}{2\sigma^2}(y - X\beta)'(I_n - \rho W)'(I_n - \rho W)(y - X\beta) \end{aligned} \tag{38}$$

The parallel between (3) and (38) is even clearer when one solves for the conditional estimates of β and σ^2 given ρ (with "*se*" denoting the SE model):

$$\hat{\beta}_{se}(\rho) = [X'(I_n - \rho W)'(I_n - \rho W)X]^{-1} X'(I_n - \rho W)'(I_n - \rho W)y \qquad (39)$$

$$\hat{\sigma}^2_{se}(\rho) = (1/n)[y - X\hat{\beta}_{se}(\rho)]'(I_n - \rho W)'(I_n - \rho W)[y - X\hat{\beta}_{sar}(\rho)] \qquad (40)$$

and substitutes these into (38) to obtain the *concentrated likelihood function*, L_{se}, for ρ. Again, after canceling terms, this function reduces to

$$L_{se}(\rho|y, X) = const + \ln|\det(I_n - \rho W)| - (n/2)\ln[\hat{\sigma}^2(\rho)] \qquad (41)$$

which is identical in form to (6).[27] Hence, these *concentrated likelihood functions* differ only with respect to their corresponding conditional variance estimates in (5) and (40). However, for the special case of maximally connected weight matrices, W^*, these conditional variance estimates are identical, as we now show.

We begin with the following preliminary result on a certain class of orthogonal projections, which are exemplified by the key projection $X(X'X)^{-1}X'$, embodied in expression (7) for *M*. If for any matrix $A \in R^{n\times k}$ of full column rank $k \leq n$, we let $P_A = A(A'A)^{-1}A'$ denote the orthogonal projection of $R^{n\times k}$ into the span of *A* (so that, by definition, $P_A A = A$), then we have the following useful condition for equality between such projections:

Lemma 4. *For any matrices* $A, B \in R^{n\times k}$ *of full column rank,*

$$P_A = P_B \Leftrightarrow P_A B = B \qquad (42)$$

As shown in Appendix A, this result yields the following key identity between SL models (1) and SE models (36) for the case of maximally connected weight matrices.

Proposition 4. *If* $W = W^*$ *in models* (1) *and* (36), *then the concentrated likelihood functions* L_{sl} *and* L_{se} *are identical for all* $\rho \in [W^*]$.

In particular, Proposition 4 shows that for maximally connected weight matrices, W^*, the maximum likelihood estimates of ρ in corresponding SL and SE models must always be identical. This in turn implies that Proposition 1 must hold in tact if the SL model in (1) is replaced by the SE model in (36). Hence, the same type of continuity argument in Proposition 2 can be used to show that the spatial dependency parameter ρ in SE models will be underestimated for strongly connected weight matrices.

Rather than repeat such arguments here, we simply report the corresponding estimation results for the SE model based on the same data *X*, parameters (β, σ^2, ρ), and weight matrices W_d, $d \in \{0.30, 0.50, 0.80, 0.90, 0.95, 0.99, 1.00\}$ used above in the section on "Biased estimation for the maximally connected case in SL models." The results for ρ displayed in column 3 of Table 1 show that, as predicted by

Proposition 4, these estimates converge to the same extreme value as d approaches unity. However, it is also clear that (at least in this particular example) the underestimation of ρ is even more severe than for the SL model above.

The results for other parameter estimates are exhibited in Table 3. Notice first that all mean beta estimates appear to be remarkably accurate—even in the maximally connected case. This is explained by the well-known fact that, for the SE model, $\hat{\beta}$ is *always* an unbiased estimator of β for a correctly specified model, because

$$\begin{aligned} E(\hat{\beta}|X) &= [X'(I_n-\rho W)'(I_n-\rho W)X]^{-1}X'(I_n-\rho W)'(I_n-\rho W)E(y|X) \\ &= [X'(I_n-\rho W)'(I_n-\rho W)X]^{-1}X'(I_n-\rho W)'(I_n-\rho W)X\beta = \beta \end{aligned} \tag{43}$$

However, in the extreme case of maximal connectivity these estimates are, in fact, completely unstable (as can be seen by the dependency of $\hat{\beta}$ on $\hat{\rho}$ in the conditional beta estimator of (39)). In particular, if we set

$$\hat{\rho}_n = 1/\lambda_{\min}(W^*) = -1/b \tag{44}$$

in this extreme case, and let $B_{\hat{\rho}_n} = I_n - \hat{\rho}_n W^*$, then

$$B_{\hat{\rho}_n} = I_n - (-1/b)[b(1_n 1_n' - I_n)] = 1_n 1_n' \tag{45}$$

together with $1_n' 1_n = n$, implies that

$$\hat{\beta}_{se}(\hat{\rho}_n) = (X'B_{\hat{\rho}_n}' B_{\hat{\rho}_n} X)^{-1} X' B_{\hat{\rho}_n}' B_{\hat{\rho}_n} y = (X' 1_n 1_n' X)^{-1} X' 1_n 1_n' y \tag{46}$$

Hence, if there is at least one explanatory variable other than the intercept (i.e., if $k \geq 1$), then the matrix $X'1_n1_n'X$ is *singular*, and the inverse in (46) does not exist. In practice, however, what typically happens is that estimation algorithms converge to values close to $-1/b$, which will yield well-defined answers. In the case illustrated above, where $-1/(1/49) = -49$, even values of -48.999 continue to produce reasonable-looking estimates on average.

Consequences for Moran tests of spatial autocorrelation

Aside from the above consequences for spatial regression models, such as SL and SE, strong connectivity of weight matrices has broader implications for diagnostic

Table 3 Mean Values of Parameter Estimates for the Spatial Error Model

Average link density	Mean $\hat{\rho}(\rho = 0.5)$	Mean $\hat{\beta}_0(\beta_0 = 1)$	Mean $\hat{\beta}_1(\beta_1 = 2)$	Mean $\hat{\beta}_2(\beta_2 = 3)$	Mean $\hat{\sigma}^2(\sigma^2 = 1)$
0.30	0.195	1.064	2.010	2.939	0.937
0.50	– 0.038	1.040	1.960	2.958	0.933
0.80	– 0.801	0.956	2.039	2.039	0.904
0.90	– 1.880	0.997	2.011	3.006	0.864
0.95	– 2.281	0.998	1.994	3.037	0.823
0.99	– 6.363	1.047	1.945	3.032	0.706
1.00	– 48.999	1.025	1.985	2.999	0.159

analyses of spatial autocorrelation. This is most evident in the single most widely used test for spatial autocorrelation, namely the *Moran test*. In particular, consider the null hypothesis of independence ($\rho = 0$), under which both SL and SE models reduce to the standard linear model:

$$y = X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n) \tag{47}$$

If one constructs the standard maximum likelihood (ordinary least squares) estimate,

$$\hat{\beta} = (X'X)^{-1}X'y \tag{48}$$

of β under this hypothesis and forms the corresponding vector of *residual estimates*,

$$\hat{\varepsilon} = y - \hat{y} = y - X\hat{\beta} \tag{49}$$

then for any given candidate choice of a spatial weight matrix W the associated Moran statistic I_W is defined by (see, e.g., Anselin 1988, section 8.1.1)

$$I_W = \alpha_W \frac{\hat{\varepsilon}' W \hat{\varepsilon}}{\hat{\varepsilon}'\hat{\varepsilon}} \tag{50}$$

where the positive constant $\alpha_W = n/\|W\|_1 = n/\sum_{ij} w_{ij}$ plays no substantive role in the analysis to follow. This can be expressed in a more convenient form (again following Anselin) by noting from (48) and (49) that

$$\hat{\varepsilon} = y - X(X'X)^{-1}X'y = [I_n - X(X'X)^{-1}X']y = My \tag{51}$$

and hence from (9) that I_W can be equivalently written as

$$I_W = \alpha_W \frac{y'MWMy}{y'My} \tag{52}$$

Under the hypothesis of independence in (47), the mean and variance of I_W are well known to be given by Cliff and Ord (1981, section 8.3) and Anselin (1988, section 8.1.1).

$$E(I_W) = \frac{\alpha_W tr(MW)}{n - (k+1)} \tag{53}$$

and

$$\text{var}(I_W) = \frac{(\alpha_W)^2\{tr(MWMW') + tr(MWMW) + [tr(MW)]^2\}}{[n-(k+1)]\cdot[n-(k-1)]} - [E(I_W)]^2 \tag{54}$$

In this setting, our main result shows that, for maximally connected weight matrices, W^*, this Moran statistic is *degenerate*.[28] In particular it is *completely concentrated at the mean*, $E(I_{W^*})$, and hence can never detect spatial autocorrelation. To establish this result, we note from (52) that this statistic is only meaningful for data sets (y, X) with $y'My \neq 0$. However, because $y'My = 0 \Leftrightarrow My = 0$ (as shown in Lemma 3), this is equivalent to the condition that $My \neq 0$. Hence, for purposes of this section we again assume *regularity* of (y, X). In addition, we employ the normalization convention $b = 1/(n-1)$ for W^* so that $\lambda_{\max}(W^*) = 1$. Finally, for each regular data set (y, X), let $I_W(y, X)$ denote the corresponding sample value of I_W in (52). With these conventions, we have the following result:

Proposition 5. *For all regular data sets (y, X),*

$$I_{W^*}(y, X) = E(I_{W^*}) \tag{55}$$

This result implies that (with probability one)[29] the realized value of I_W must be precisely its *expected value* under independence. Hence, no evidence for spatial dependence can ever be detected in this extreme case. More generally, the same type of continuity argument used in Proposition 3 shows that, for weight matrices W that are sufficiently close to W^* (say in terms of the relative connectivity norm), it must be true that the possible values of I_W are concentrated close to the mean $E(I_W)$. So again this statistic should have little ability to detect spatial dependence.

These ideas can be made more concrete in terms of the standard z test for Moran statistics found in most software. If the standard deviation of I_W under independence is denoted by $\sigma(I_W) = \text{var}(I_W)^{1/2}$, then it is well known that the standardized z value

$$Z_W = \frac{I_W - E(I_W)}{\sigma(I_W)} \tag{56}$$

is approximately distributed as $N(0,1)$ for large n (Cliff and Ord 1981, section 8.5.1). Hence, one can use this distribution theory to test the hypothesis of spatial independence with respect to weight matrix W.[30]

To study the behavior of this test for strongly connected weight matrices, we shall focus only on the simulation results provided above in the section on ''Biased estimation for the maximally connected case in SL models.'' Here, it was assumed that $\rho = 0.5$ and hence that a substantial degree of positive spatial dependence is present. To determine whether this dependence can be detected by the Moran statistic for a given weight matrix, W, it suffices to compute $I_W(y, X)$ for simulated data sets from model (1) and then to examine the frequency distribution of z values, $Z_W(y, X)$, generated by this data. For a one-sided test of $\rho > 0$ at the $\alpha = 0.05$ level, we need only count the fraction of z values above $z_\alpha = 1.65$ to determine the power of this test to detect positive spatial dependence, given the true value $\rho = 0.5$. For the 1,000 simulated values at each link density level in the section above on biased estimation in SL models, the resulting estimated power levels are shown in Table 4.

It is clear that at link densities above 0.80 the distribution is so concentrated around the null mean $E(I_W)$ that even a dependency level of $\rho = 0.5$ is detectable less than 10% of the time.[31] Also, even though the distribution of I_W concentrates at the null mean as link density approaches 1, the power levels do not appear to fall to zero in Table 4. The concentration of I_W values drives the *variance* in (53) to zero (easily verifiable by the same calculations used for the mean in the proof of Proposition 5[32]). Hence, when I_W is highly concentrated, the standardized value Z_W becomes unstable (as it approaches the limiting indeterminate values 0/0 for W^*).

Table 4 Power of Moran for a test at $\rho = 0.5$

Average link density	Sample mean	Null mean	Power ($\rho = 0.5$)
0.30	0.0416	−0.0182	0.383
0.50	−0.0046	−0.0183	0.137
0.80	−0.0152	−0.0184	0.091
0.90	−0.0190	−0.0187	0.059
0.95	−0.0187	−0.0189	0.055
0.99	−0.0201	−0.0190	0.054

Concluding remarks

We have shown that the presence of strongly connected spatial weight matrices can introduce serious biases into both the estimating and testing of spatial autocorrelation. Hence, one may ask whether there is any simple intuitive explanation. One possibility relates to the notion of "effective sample size." It has long been observed that the presence of statistical dependencies essentially reduce the amount of information gained from each individual observation. For example, the observation of a sequence of perfectly correlated coin tosses will offer no more information than the observation of only the first toss, no matter how long the sequence is. Hence, insofar as strong spatial connectivity reflects strong dependencies among units (or agents), there should be less statistical information available for estimations or tests of hypotheses.

While this argument has intuitive appeal and is no doubt true to some extent, it fails to explain, for example, why maximum likelihood methods should systematically *underestimate* the ρ parameter in SL and SE models. The findings presented here suggest that much can be learned by studying the extreme case of maximally connected weight matrices, W^*. In particular, both concentrated likelihood functions and Moran statistics reduce to particularly simple forms in this case and can thus be studied in detail. Even in this extreme case, however, the subtlety of the underestimation question is underscored by the quite different arguments used to bound the values for each term in the concentrated log-likelihood function. In particular, both the eigenvalue structure of W^* and the relation of W^* to the regression projection operator, $I_n - X(X'X)^{-1}X'$, were involved. Thus, in some respects, these results raise as many theoretical questions as they answer.

Even more important are questions relating to the practical consequences of these results. Although the single simulation example presented here is very suggestive, it can provide no definitive guidelines for applications. Hence, the actual severity of these biases can be determined only by more extensive and systematic simulation studies, such as those already begun by Mizruchi and Neuman (2008) and Farber, Páez, and Volz (2009).

Acknowledgements

I am grateful to Mark Mizruchi, Eric Neuman, Oleg Smirnov, Harry Kelejian, and Federico Martellosio for many helpful comments on an earlier draft of this article.

Notes

1 I am indebted to a referee for pointing out that similar observations were made by Bao and Ullah (2007) about the second order bias of these estimates in the context of a pure SL model with circular weight matrices of varying degrees of connectivity.

2 A specific choice for b is considered below in the section on "Consequences for strongly connected weight matrices in SL models" below.

3 This terminology is not to be confused with the graph-theoretical notion of "totally connected," which refers only to the presence of nonzero links between all distinct node pairs.

4 This can also be expressed in terms of the (cell-wise) matrix inequalities $W_* \leq W \leq W^*$ for all $W \in \mathrm{W}_n$.

5 One important example, pioneered by Kelejian and Prucha (1998) in the context of increasing domain asymptotics, is to require uniform boundedness of the row and column sums of n-square weight matrices, W_n. This implies that W_n must "approach" the corresponding n-square matrix W_* (in an appropriate sense) as n becomes large. A more explicit graph-theoretical condition of the same type, designated as "uniformly bounded maximum local degree," is employed by Griffith and Lagona (1998).

6 Maximally connected spatial weight matrices have been studied in a somewhat different context by Kelejian and Prucha (2002), who described them simply as models with "equal spatial weights" (see also Kelejian, Prucha, and Yuzefovich 2006 and Baltagi 2006). These matrices are also closely related to *equicorrelation* (or *intraclass correlation*) matrices, as discussed, for example, in the review paper by Donner (1986).

7 While this development is quite standard (as, e.g., in Anselin 1988 and Anselin and Bera 1998, section III.B), the present results depend critically on the details of this maximum likelihood formulation.

8 Concentrated likelihood functions are also designated as *profile* likelihood functions (as, e.g., in Pace and Salvan 1997, section 4.6).

9 This maximum eigenvalue is always nonnegative (Horn and Johnson 1985, theorem 8.1.3) but need not be positive even when W has positive elements. Even for $n = 2$, the matrix $W = [0\ 1\ ;\ 0\ 0]$ has $\lambda(W) = \{0, 0\}$.

10 This is also referred to as a "spatially filtered" version of y (e.g., in Anselin and Bera 1998).

11 This failure of existence is an instance of the more general result of Arnold (1979, theorem 3) regarding the nonexistence of maximum likelihood estimators for covariance parameters in linear models with exchangeably distributed errors. I am indebted to Federico Martellosio for pointing this out to me.

12 The following result is essentially contained in theorem 1 of Kelejian, Prucha, and Yuzefovich (2006), where it is employed to analyze the consistency properties of 2SLS estimation in the case of equal spatial weights.

13 These results illustrate the more general finding of Arnold (1979, p. 196) regarding the inconsistency of standard parameter estimates for linear models with exchangeably distributed errors.

14 Many other choices are illustrated in Horn and Johnson (1985, section 5.6).

15 Because every positive scaling of a norm is also a norm, the first equality shows that this is indeed a matrix norm.

16 Expression (33) follows from the fact that $0 \leq W \leq W^* \Rightarrow \lambda_{\max}(W) \leq \lambda_{\max}(W^*) = 1 \Rightarrow 1/\lambda_{\max}(W) \geq 1$ (see Horn and Johnson 1985, corollary 8.1.19).
17 An alternative normalization that also shares this property is to set b equal to the reciprocal of the smallest row or column sum, as proposed by Kelejian and Prucha (in press, lemma 2). Although less standard than the present convention, this normalization has the advantage of being much easier to compute for large weight matrices. A number of additional normalizations, or ''coding schemes,'' are discussed in Tiefelsdorf, Griffith, and Boots (1999).
18 Cases involving nonpositive dependencies ($\rho \leq 0$) are discussed briefly in the Remark following the proof sketch for Proposition 3 in Appendix A.
19 This example is meant only to illustrate the practical consequences of the analytical results above. As mentioned in the introduction, more extensive and systematic simulations can be found in Mizruchi and Neuman (2008) and Farber, Páez, and Volz (2009).
20 Note that density values d can only be approximated by this sampling procedure. However, repeated samples at each density level yielded variations that were too small to warrant reporting. In all cases, the matrix $W_{(d)}$ had an average link density well within 0.01 of d.
21 The normalization $b = 1/(n-1) = 1/49$ used has the theoretical advantage of preserving all relative connectivity relationships. But the present scaling to unit maximum eigenvalues is a more typical normalization in practice. For comparison, calculations were also done for the 1/49 scaling, which produced even more dramatic underestimation results than those presented here.
22 The estimation was done in Matlab using a modified version of the LeSage (1999) suite of programs.
23 Cases $d = 0.30$ and $d = 0.95$ are, respectively, very similar to $d = 0.50$ and $d = 0.90$ and are omitted.
24 Note that for tests of positive ρ, it is theoretically more appropriate to consider a one-sided test ($\rho > 0$). However, such results are not reported in standard spatial regression software.
25 This terminology follows Anselin and Bera (1998).
26 Although the spatial dependence parameter in this model acts on residuals rather than y, we choose to keep the same notation, ρ, to emphasize the parallels between these two models.
27 In particular, the constant terms (const) are also easily shown to be identical.
28 This degeneracy is also an instance of the more general result in Arnold (1979, theorem 5) for the class of invariant test statistics for linear models with exchangeably distributed errors. A more explicit version relating to the present case is given in Martellosio (2008, props. 3.4 and 3.6).
29 It is a simple matter to show that, for any X, the set of y with $My = 0$ has probability measure zero.
30 The *exact* distribution of I_W under independence has been obtained by Tiefelsdorf and Boots (1995). However, most statistical packages rely on the asymptotic approximation above.
31 The 0.054 value for density 0.99 is consistent with a limiting value of $\alpha = 0.05$ for the maximally connected case, as implied by the results of Martellosio (2008, prop. 3.5).

32 Note, in particular, from Lemma 2 that for W^*, $tr(MW^*MW^*) = tr[(-bM)(-bM)] = b^2 tr(MM) = b^2 tr(M) = [n-(k+1)]/(n-1)^2$.

33 For the case of $b = 1/(n-1)$, this result appears in section 2.5 of Kelejian and Prucha (2002).

34 Note that L_{sl} is also unbounded at the upper boundary of $[W^*]$, namely, $\rho* = 1/\lambda_{\max}(W^*) = 1/[b(n-1)]$, but because $L_{sl}(\rho * |y, X) = -\infty$, this is of little interest for maximum likelihood estimation.

35 A full proof of Proposition 3 is available in Appendix A of the online version of this article (under ''Recent Papers'') at http://www.seas.upenn.edu/~tesmith.

Appendix A. Proofs of results in the text

Proof of Lemma 1: It follows from Searle (1982, section 12.3.d) that the eigenvalues of any matrix of the form $A = aI + c11'$ are given by

$$\lambda(A) = \{a, \ldots, a, (a+nc)\} \tag{A1}$$

where a has multiplicity $n-1$. Hence the eigenvalues of

$$W^* = b \cdot (1_n 1_n' - I_n) = (-b)I_n + (b)1_n 1_n' \tag{A2}$$

are immediately seen to be those in (20) of Lemma 1. □

Proof of Lemma 2: Observe from (19) and (7) that

$$\begin{aligned} M \cdot W^* &= (I_n - X(X'X)^{-1}X') \cdot b \cdot (1_n 1_n' - I_n) \\ &= b \cdot [1_n 1_n' - I_n - X(X'X)^{-1}X'1_n 1_n' + X(X'X)^{-1}X'] \end{aligned} \tag{A3}$$

But because $[X(X'X)^{-1}X']X = X$ and 1_n is the first column of X, it follows in particular that $[X(X'X)^{-1}X']\,1_n = 1_n$. Hence, we see that

$$\begin{aligned} M \cdot W^* &= b \cdot [1_n 1_n' - I_n - 1_n 1_n' + X(X'X)^{-1}X'] \\ &= b \cdot [X(X'X)^{-1}X' - I_n] = -b \cdot M \end{aligned} \tag{A4}$$

Next observe that, because $[X(X'X)^{-1}X']1_n = 1_n \Rightarrow 1_n' = 1_n'[X(X'X)^{-1}X']$, it also follows that

$$\begin{aligned} W^* \cdot M &= b \cdot (1_n 1_n' - I_n) \cdot (I_n - X(X'X)^{-1}X') \\ &= b \cdot [1_n 1_n' - 1_n 1_n' X(X'X)^{-1}X' - I_n + X(X'X)^{-1}X'] \\ &= b \cdot [1_n 1_n' - 1_n 1_n' - I_n + X(X'X)^{-1}X'] \\ &= b \cdot [-I_n + X(X'X)^{-1}X'] = -b \cdot M \quad \square \end{aligned} \tag{A5}$$

Proof of Lemma 3: Observe from Lemma 2, together with the symmetry of W^* and M, that for any $\rho \in [W^*]$ and dataset (y, X),

$$\begin{aligned} y'(I_n - \rho W^*)'M(I_n - \rho W^*)y &= y'(I_n - \rho W^*)'(M - \rho MW^*)y \\ &= y'(M - \rho MW^* - \rho W^*M + \rho^2 W^*MW^*)y \\ &= y'(M + \rho bM + \rho bM + \rho^2 b^2 M)y \\ &= (1 + 2\rho b + \rho^2 b^2) \cdot y'My \\ &= (1 + \rho b)^2 \cdot y'My \end{aligned} \tag{A6}$$

But $\rho \in [W^*]$ then implies that $\rho > -1/b$ and hence that $1 + \rho b > 0$. Thus, W^*-regularity of (y, X) will follow if it can be shown that $y'My > 0$. But because M is an orthogonal project matrix and hence is *positive semidefinite*, $y'My \geq 0$ for all y and, moreover, that $y'My = 0 \Leftrightarrow My = 0$ (Horn and Johnson 1985, p. 400). Finally, because the regularity of (y, X) implies that $My \neq 0$, it must then be true that $y'My > 0$ and thus that W^*-regularity holds. □

Proof of Proposition 1: The strategy is to use Lemmas 1 and 2 to show that the concentrated likelihood function (expression [16] in the text),

$$\begin{aligned} L_{sl}(\rho|y, X) = const. + \sum\nolimits_i \ln|1 - \rho\lambda_i| - (n/2)\ln[y'(I_n - \rho W)'M(I_n \\ - \rho W)y] \end{aligned} \tag{A7}$$

is reducible to a simple analytical form for which the result is obvious. To do so, we first observe from Lemma 1 and the positivity of $\min_i\{1 - \rho\lambda_i(W)\}$ on $[W]$ that for any $\rho \in [W^*]$ we must have [33]

$$\begin{aligned} \sum\nolimits_i \ln|1 - \rho\lambda_i| &= \sum\nolimits_i \ln(1 - \rho\lambda_i) = (n-1)\ln[1 - \rho(-b)] + \ln[1 - \rho b(n-1)] \\ &= (n-1)\ln(1 + \rho b) + \ln[1 - \rho b(n-1)] \end{aligned} \tag{A8}$$

Moreover, we see from (A6) that

$$\begin{aligned} &-(n/2)\ln[y'(I_n - \rho W^*)'M(I_n - \rho W^*)y] = -(n/2)\ln[(1 + \rho b)^2 y'My] \\ &= -\{n\ln(1 + \rho b) + (n/2)\ln(y'My)\} \end{aligned} \tag{A9}$$

Notice also from Lemma 3 that this log expression is well defined for all $\rho \in [W^*]$. Hence, by substituting (A8) and (A9) into (A7) we obtain the following simple expression for the concentrated likelihood function:

$$\begin{aligned} L_{sl}(\rho|y, X) &= const + \{(n-1)\ln(1 + \rho b) + \ln[1 - \rho b(n-1)]\} \\ &\quad - \{n\ln(1 + \rho b) + (n/2)\ln(y'My)\} \\ &= const - \ln(1 + \rho b) + \ln[1 - \rho b(n-1)] \end{aligned} \tag{A10}$$

where the term $(n/2)\ln(y'My)$, not containing ρ, has again been absorbed in *const*. From here we need only observe that because $\rho_* = -1/b$ it follows that for any decreasing sequence (ρ_m) in $[W^*]$, with $\lim_{m\to\infty} \rho_m = \rho_*$, we must have

$$\begin{aligned}\lim_{m\to\infty} L_{sl}(\rho_m|y,X) &= const - \lim_{m\to\infty}\ln(1+\rho_m b) + \lim_{m\to\infty}\ln[1-\rho_m b(n-1)] \\ &= const - \ln(1+\rho_* b) + \ln[1-\rho_* b(n-1)] \\ &= const - \ln(0) + \ln(n) = \infty\end{aligned} \tag{A11}$$

and the result is established.[34] □

Proof of Proposition 2: To establish (27) in Proposition 2, recall from (22) that for any positive bound b,

$$\rho_* = 1/\lambda_{\min}(W^*) = -1/b < 0 \tag{A12}$$

Hence, it follows that

$$\begin{aligned}y_{W^*}(\rho_*) &= (I_n - \rho_* W^*)y = \{I_n - (-1/b)[b\cdot(1_n 1_n' - I_n)]\}y \\ &= [I_n + (1_n 1_n' - I_n)]y = 1_n 1_n' y = (1_n' y)\cdot 1_n + X\cdot 0 \\ &= \beta_0(\rho_*)\cdot 1_n + X\beta(\rho_*)\end{aligned} \tag{A13}$$

and the result is established. □

Proof Sketch for Proposition 3: While the proof of this result is rather technical,[35] the basic idea is simple. Observe from Fig. 2 that not only does the concentrated likelihood function L_{sl} diverge to $+\infty$ at ρ_*, but its derivative is *negative everywhere* in $[W^*]$. Hence, if we now write the concentrated likelihood function as $L_{sl}(\rho|y,X,W)$ to emphasize its dependence on W (as well as data (y, X)), then the strategy of the proof is to show that the corresponding derivative, $L_{sl}'(\rho|y,X,W)$, with respect to ρ is *continuous* in W at the point W^*. Using this continuity property, it is possible to show that for any choice of bias factor α when W is sufficiently close to W^* (i.e., when ε in expression [35] of Proposition 3 is sufficiently small), one can guarantee that $L_{sl}'(\rho|y,X,W)$ will be negative for all $\rho \in [W]$ with $\rho \geq \rho_0/(1+\alpha)$ and thus that $L_{sl}(\rho|y,X,W)$ can only achieve a maximum on $[\rho_*, \rho/(1+\alpha))$. □

Remark:

The proof sketched above also shows (from the persistence of negative slopes) that for strongly connected weight matrices under conditions of *no* spatial dependence, the null hypothesis, $\rho = 0$, will tend to be falsely rejected in favor of *negative* dependencies ($\rho < 0$). Moreover, in cases where such dependencies are actually negative, it is equally clear the strength of these dependencies will tend to be overestimated.

Proof of Lemma 4: **Proof:** Because $P_B B = B$, it follows at once that $P_A = P_B \Rightarrow P_A B = P_B B = B$. Thus, we need only establish the converse. To do so, observe that

$$P_A B = B \;\Rightarrow\; P_A B\,[(B'B)^{-1}B'] = B\,[(B'B)^{-1}B'] \;\Rightarrow\; P_A P_B = P_B \tag{A14}$$

Moreover, it also follows that

$$\begin{aligned} B &= P_A B = A(A'A)^{-1}A'B = A(A'A)^{-1}(A'B) \\ &\Rightarrow B'B = B'A(A'A)^{-1}(A'B) \Rightarrow |B'B| = |B'A| \cdot |A'A|^{-1} \cdot |A'B| \\ &\Rightarrow |A'B|^2 = |B'B| \cdot |A'A| > 0 \end{aligned} \tag{A15}$$

and hence that $A'B$ is nonsingular. Thus, by the first line of (A15) we have

$$P_B B = B \Rightarrow P_B A(A'A)^{-1}(A'B) = A(A'A)^{-1}(A'B) \Rightarrow P_B A = A \tag{A16}$$

where the last implication follows by post-multiplication of both sides by the inverse of the nonsingular matrix $(A'A)^{-1}(A'B)$. Hence, by the argument in (A14)

$$P_B A = A \Rightarrow P_B P_A = P_A \Rightarrow P_A P_B = P_A \tag{A17}$$

where the last implication follows by taking transposes of both sides and using the symmetry of P_A and P_B. It can be concluded from (A14) and (A17) that

$$P_A = P_A P_B = P_B \tag{A18}$$

and the result is established. □

Proof of Proposition 4: To establish this result, it is clear from (6) and (41) in the text that it suffices to show that the conditional variance estimates in (10) and (40) are identical on $[W^*]$. If for notational convenience we now let

$$B_\rho = I_n - \rho W^* = I_n - \rho b(1_n 1_n' - I_n) \tag{A19}$$

(where $b = 1/(n-1)$ is one possibility) then by the first line of (10), it follows that for the SL model (1),

$$\begin{aligned} \hat{\sigma}^2_{sl}(\rho) &= (1/n)([I_n - X(X'X)^{-1}X']B_\rho y)'([I_n - X(X'X)^{-1}X']B_\rho y) \\ &= (1/n)\left\|[I_n - X(X'X)^{-1}X']B_\rho y\right\|^2 \\ &= (1/n)\left\|(I_n - P_X)B_\rho y\right\|^2 \end{aligned} \tag{A20}$$

To compare this with the SE model (37), observe from (39) that

$$\hat{\beta}_{se}(\rho) = (X'B_\rho' B_\rho X)^{-1}X'B_\rho' B_\rho y = [(B_\rho X_\rho)'(B_\rho X)]^{-1}(B_\rho X)'B_\rho y \tag{A21}$$

and hence from (40) that

$$\begin{aligned} \hat{\sigma}^2_{se}(\rho) &= (1/n)(y - X\hat{\beta}_{se}(\rho))'B_\rho' B_\rho(y - X\hat{\beta}_{se}(\rho)) \\ &= (1/n)\left\|B_\rho(y - X\hat{\beta}_{se}(\rho))\right\|^2 \\ &= (1/n)\left\|B_\rho(y - X[(B_\rho X_\rho)'(B_\rho X)]^{-1}(B_\rho X)'B_\rho y)\right\|^2 \\ &= (1/n)\left\|\{I_n - (B_\rho X)[(B_\rho X)'(B_\rho X)]^{-1}(B_\rho X)'\}B_\rho y)\right\|^2 \\ &= (1/n)\left\|(I_n - P_{B_\rho X})B_\rho y\right\|^2 \end{aligned} \tag{A22}$$

In this form, the result will follow if it can be shown that

$$P_X = P_{B_\rho X} \text{ for all } \rho \in [W^*]. \tag{A23}$$

But because $X = [1_n, \tilde{X}]$ and $P_X X = X$ together imply that $P_X 1_n = 1_n$, we must have

$$\begin{aligned} P_X(B_\rho X) &= bP_X(1_n 1_n' - I_n)X = b(P_X 1_n)1_n' X - bP_X X \\ &= b(1_n 1_n')X - bX = b(1_n 1_n' - I_n)X = B_\rho X \end{aligned} \tag{A24}$$

and may conclude from Lemma 4 that (A23) holds for all $\rho \in [W^*]$. □

Proof of Proposition 5: Because $b = 1/(n-1) \Rightarrow \|W^*\|_1 = n(n-1)[1/(n-1)] = n$, it follows that

$$\alpha_{W^*} = n/\|W^*\|_1 n/n = 1 \tag{A25}$$

and hence that the Moran statistic for this case reduces to

$$I_{W^*}(y, X) = \frac{y' M W^* y}{y' M y} \tag{A26}$$

Thus, we see from Lemma 3 that

$$I_{W^*}(y, X) = \frac{y'(-bM)y}{y'My} = -\frac{1}{n-1} \cdot \frac{y'My}{y'My} = -\frac{1}{n-1} \tag{A27}$$

and may conclude that I_{W^*} is indeed concentrated at a single value. To show that this value is precisely the mean $E(I_{W^*})$, under independence, we note that, because the trace of orthogonal projection (symmetric idempotent) matrix M is equal to the dimension of its image space (Searle 1982, section 12.2), and that, because the dimension of the complement of the span of X is $n-(k+1)$, it follows that

$$tr(M) = n - (k+1) \tag{A28}$$

This in turn implies from Lemma 3 that

$$tr(MW^*) = tr(-bM) = \left(-\frac{1}{n-1}\right) tr(M) = -\frac{n-(k+1)}{n-1} \tag{A29}$$

and hence from (A25), together with (53) in the text, that

$$E(I_{W^*}) = \frac{tr(MW^*)}{n-(k+1)} = -\frac{1}{n-1} \tag{A30}$$

Thus, the result follows from (A27) and (A30). □

References

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht, The Netherlands: Kluwer.

Anselin, L., and A. Bera. (1998). ''Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics.'' In *Handbook of Applied Economic Statistics*, 237–89, edited by A. Ullah and D. Giles. New York: Marcel Dekker.

Arnold, S. F. (1979). ''Linear Models with Exchangeably Distributed Errors.'' *Journal of the American Statistical Association* 74, 194–99.

Baltagi, B. H. (2006). ''Random Effects and Spatial Autocorrelation with Equal Weights.'' *Econometric Theory* 22, 973–84.

Bao, Y., and A. Ullah. (2007). ''Finite Sample Properties of Maximum Likelihood Estimator in Spatial Models.'' *Journal of Econometrics* 137, 396–413.

Cliff, A. D., and J. K. Ord. (1981). *Spatial Processes: Models and Applications*. London: Pion.

Donner, A. (1986). ''A Review of Inference Procedures for the Intraclass Correlation Coefficient in the One-Way Random Effects Model.'' *International Statistical Review* 54, 67–82.

Farber, S., A. Páez, and E. Volz. (2009). ''Topology and Dependency Tests in Spatial and Network Autoregressive Models.'' *Geographical Analysis* 41, 158–80.

Griffith, D. A., and F. Lagona. (1998). ''On the Quality of Likelihood-Based Estimators in Spatial Autoregressive Models When the Data Dependence Structure is Misspecified.'' *Journal of Statistical Planning and Inference* 69, 153–74.

Horn, R. A., and C. R. Johnson. (1985). *Matrix Analysis*. Cambridge, UK: Cambridge University Press.

Kelejian, H. H., and I. R. Prucha. (2002). ''2SLS and OLS in a Spatial Autoregressive Model with Equal Spatial Weights.'' *Regional Science and Urban Economics* 32, 601–707.

Kelejian, H. H., and I. R. Prucha. (in press). ''Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances.'' *Journal of Econometrics*.

Kelejian, H. H., I. R. Prucha, and Y. Yuzefovich. (2006). ''Estimation Problems in Models with Spatial Weighting Matrices Which Have Blocks of Equal Elements.'' *Journal of Regional Science* 46, 507–15.

Kelejian, H. H., and I. R. Prucha. (1998). ''Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances.'' *The Journal of Real Estate Finance and Economics* 17, 99–121.

LeSage, J. (1999). *Spatial Econometrics Toolbox*. Available at http://www.spatial-econometrics.com (accessed Dec. 2007).

Martellosio, F. (2008). ''Testing for Spatial Autocorrelation: The Regressors That Make the Power Disappear.'' Working Paper, Department of Economics, University of Reading, Reading RG6 6AW, UK.

Mizruchi, M. S., and E. J. Neuman. (2008). ''The Effect of Density on the Levels of Bias in the Network Autocorrelation Model.'' *Social Networks* 30, 190–200.

Pace, L., and A. Salvan. (1997). *Principles of Statistical Inference from a Non-Fisherian Perspective*. Singapore: World Scientific.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. New York: Wiley.

Tiefelsdorf, M., and B. N. Boots. (1995). ''The Exact Distribution of Moran's I.'' *Environment and Planning A* 27, 985–99.

Tiefelsdorf, M., D. A. Griffith, and B. Boots. (1999). ''A Variance-Stabilizing Coding Scheme for Spatial Link Matrices.'' *Environment and Planning A* 31, 165–80.