

## SOFTWARE FOR BAYESIAN SPATIAL MODEL COMPARISON

James P. LESAGE\*

**Abstract** - Taking a Bayesian perspective on model comparison for cross-sectional and static panel data models considerably simplifies the task of selecting an appropriate model. A wide variety of alternative specifications that include various combinations spatial dependence in lagged values of the dependent variable, spatial lags of the explanatory variables, as well as dependence in the model disturbances have been the focus of a literature on various statistical tests used by practitioners to distinguishing between alternative specifications. LeSage and Pace (2009) make a theoretical argument that implies the task of model selection can be simplified by focusing on only two model specifications, one reflecting theoretical situations involving global spillovers (the spatial Durbin model, SDM) and the other theoretical scenarios involving local spillovers (the spatial Durbin error model, SDEM). LeSage (2014) extends this theoretical argument to the case of static panel data models. MATLAB software functions for carrying out Bayesian cross-sectional and static spatial panel data model comparisons is described here along with a number of illustrative applications.

**Keywords** - STATIC SPACE-TIME PANEL DATA MODELS, BAYES FACTORS, LOCAL VERSUS GLOBAL SPATIAL SPILLOVERS

**Classification JEL** - C11, C23, O47, O52

---

\* Fields Endowed Chair ; Texas State University-San Marcos, Department of Finance & Economics, San Marcos, TX 78666; E-mail: jlesage@spatial-econometrics.com

## 1. INTRODUCTION

LeSage and Pace (2009) define spatial spillovers for a causal relationship between characteristics/actions of entity/agent ( $X_i$ ) located at position  $i$  in space and the outcomes/decisions/actions ( $Y_j$ ) of an agent/entity located at position  $j$  as a situation where:  $\partial Y_j / \partial X_i \neq 0$ .

LeSage and Pace (2014) argue that practitioners should focus on a theoretical motivation for either *local* or *global* spatial spillovers, which would facilitate choice of an appropriate spatial regression specification. If locations  $j$  are neighbors to location  $i$ , we have a *local* spatial spillover, whereas situations involving locations  $j$  that include not only neighbors to  $i$ , but neighbors to neighbors of  $i$ , neighbors to neighbors to neighbors, and so on, are defined as *global* spillover.

Two of the most popular spatial regression specifications are: 1) global spillover models that include a spatial lag of the dependent variable, and 2) local spillover models that include spatial lags of the explanatory variables but not the dependent variable. LeSage (2014) shows that practitioners need logically only consider the single SDM global spillover model versus the SDEM local spillover specification. The logic behind his argument is that other specifications such as the spatial autoregressive (SAR) and spatial error (SEM) are subsumed by these two, as is the spatially lagged explanatory variables (SLX) specification from LeSage and Pace (2009).

## 2. TESTING FOR SLX, SDM VERSUS SDEM IN CROSS-SECTIONAL MODEL SPECIFICATIONS

There is a large literature on various statistical tests that can be used to distinguish between alternative spatial model specifications. For example, the (robust) LM tests developed by Anselin et al. (1996) for cross-sectional data to test whether spatial interaction effects ( $\rho(I_T \otimes W)y$ ) should be included in the model. Elhorst (2010) extends these tests to the case of static panel data models. Another alternative has been to estimate a model that includes spatial interaction effects and test whether the model can be simplified to one that excludes these effects using Wald or Likelihood ratio tests. These tests and the numerical outcomes for a widely used panel data set on cigarette demand (Baltagi, 2001) are illustrated in Elhorst (2012).

LeSage and Pace (2014) point out that before engaging in statistical tests to determine an appropriate model specification, the theoretical context might point to either a local or global specification. For example, if the substantive (or theoretical) aspects of our applied modeling situation point to *local* spillovers as the only reasonable possibility, there should be no need for statistical tests. LeSage (2014) points out that there are cases where theory is ambiguous regarding the question of local versus global spillovers. As a concrete illustration, he points to the case of spillovers from state-level cigarette taxes, noting that substantial commercial smuggling activity could produce global spillovers, where-

as mere cross-border shopping by consumers suggests local spillovers (see LaFaive and Nesbit, 2013).

LeSage and Pace (2009) describe calculation of Bayesian posterior model probabilities for cross-sectional spatial regression models based on work of LeSage and Parent (2007). LeSage (2014) sets forth the theoretical underpinnings for calculation of Bayesian posterior model probabilities for the SLX, SDM and SDEM spatial panel data model specifications. Our focus here is on describing a set of MATLAB functions that implement these approaches to Bayesian model comparison that produce model probabilities. Applied examples are provided that should assist practitioners with intelligent use of these functions.

### 2.1. The `lmarginal_cross_section()` function

A single MATLAB function named `lmarginal_cross_section` was constructed to carry out Bayesian posterior model probabilities that compare the three relevant cross-sectional spatial regression models, SLX, SDM and SDEM. In a cross-sectional spatial regression, an  $N \times 1$  vector  $y$  contains the dependent variable for  $N$  regions, an  $N \times k$  matrix  $X$  contains explanatory variable vectors, and an  $N \times N$  matrix  $W$  holds the spatial weight matrix.

A call of the function using these model data with no options produces a *results structure variable* with a  $3 \times 1$  vector of log marginal magnitudes (these form the basis for Bayesian model comparison) and a  $3 \times 1$  vector of posterior model probabilities for the three models, SLX, SDM and SDEM. Since no input options were specified, these results are based on default values used by the function. We can print out the results using the `mprint` function from the toolbox.<sup>1</sup>

1.0

```
result = lmarginal_cross_section(y,X,W);
in.cnames = strvcat('log-marginal','model probs');
in.rnames = strvcat('model','slx','sdm','sdem');
in.fmt = '%10.4f';
out = [result.lmarginal result.probs];
mprint(out,in);
```

The results printed are shown below, where we see that all three models exhibit similar log-marginal likelihoods and posterior model probabilities. This type of result will occur when spatial dependence in the dependent variable for the SDM model and the disturbances of the SDEM model is low. Note that in the presence of no spatial dependence, all three models collapse to the same specification (the SLX).

---

<sup>1</sup>All references to the toolbox are to the Spatial Econometrics Toolbox for MATLAB available for free at [www.spatial-econometrics.com](http://www.spatial-econometrics.com)

model	log-marginal	model probs
slx	-1088.6072	0.3686
sdm	-1088.8197	0.2980
sdem	-1088.7077	0.3334

This should be clear from examining equations for the three model specifications, where  $\rho$  is the scalar parameter reflecting the strength of spatial dependence. As  $\rho \rightarrow 0$ , all specifications collapse to the SLX. The sample data vector  $y$  in this example was in fact generated using the SLX specification, so the posterior probabilities produced by the model comparison method are correct.

$$\text{SDEM: } y = X\beta + WX\gamma + u, \quad u = \rho Wu + \varepsilon \quad (1)$$

$$\text{SDM: } y = \rho Wy + X\beta + WX\gamma + \varepsilon \quad (2)$$

$$\text{SLX: } y = X\beta + WX\gamma + \varepsilon \quad (3)$$

$$\varepsilon \sim N(0, \sigma_\varepsilon^2 I_N)$$

As with all functions in the *Spatial Econometrics Toolbox*, typing ‘help `lmarginal_cross_section`’ in the MATLAB Command Window will show documentation for the function.

```
>> help lmarginal_cross_section
PURPOSE: Bayesian log-marginal posterior for cross-sectional spatial
regression models
        no priors on beta, sige
        uniform prior on rho, lambda over eigenvalue bounds
-----
USAGE: results = lmarginal_cross_section(y,x,W,info)
where:
  y = dependent variable vector (N x 1)
  x = independent variables matrix, WITHOUT INTERCEPT or W*x
      variables
  W = N by N spatial weight matrix (for W*y and W*e)
  info.lflag = 0 for full lndet computation (default = 1,fastest)
              = 1 for MC lndet approximation (fast for large
              problems)
  info.order = order to use with info.
              lflag = 1 option (default = 50)
  info.iter = iterations to use with info.
              lflag = 1 (default = 30)
  info.rmin = (optional) minimum value of rho in search
              (default = -1)
  info.rmax = (optional) maximum value of rho in search
              (default = +1)
  info.eig  = 0 for default rmin = -1,rmax = +1,
              = 1 for eigenvalue calculation of these
-----
RETURNS: a structure:
  results.meth = 'lmarginal_cross_section'
  results.nobs = # of cross-sectional observations
  results.y    = N x 1 vector of y from input
  results.nvar = # of variables in x-matrix
  results.rmin = minimum value of rho used (default +1)
  results.rmax = maximum value of rho used (default -1)
  results.lflag = lflag value from input (or default value used)
```

```

results.eig = 0,1 from input (or default value used)
results.lmarginal = a 3 x 1 column-vector with [log-marginal]
[ logm_slx logm_sdm logm_sdem ]'
results.probs = a 3 x 1 column-vector with model probs
results.logm_slx
results.logm_sdm
results.logm_sdem
-----

```

This documentation shows that when calling the function an  $N \times k$  explanatory variables matrix without an intercept is used, since the function adds the intercept term as well as the spatially lagged explanatory variables  $WX$ . These spatial lags of  $X$  are part of all three model specifications being compared. The user-controlled options are input using a structure variable (named 'info' here) that allow for various choices that influence the speed and accuracy of the computations involved.<sup>2</sup> Generally, faster computational speed is accompanied by slightly lower numerical accuracy, for example one can use the default (info.eig = 0) which uses the range of  $-1 < \rho < 1$  for the spatial dependence parameter rather than the true range of  $1/\omega_{\min} < \rho < 1/\omega_{\max}$ , where  $\omega_{\min}$  and  $\omega_{\max}$  denote the smallest (i.e., most negative) and largest characteristics roots of the matrix  $W$ . Calculating a minimum eigenvalue can take time for problems involving a large number of observations  $N$ .

An example of the trade-offs between speed and accuracy is shown below, where an SLX model was generated using 605 observations (reflecting Ohio school districts). Despite the difference in accuracy, one should draw the same inference that an SLX model is the specification most consistent with the sample data from both sets of results.

```

results for defaults (no info input arguments)
uses -1, 1 interval and lndetmc
true model is SLX
time taken is:      0.1720 seconds
model log-marginal model probs
slx  -1088.6072    0.4447
sdm  -1089.1336    0.2627
sdem -1089.0255    0.2927
results for info.lflag = 0, info.eig = 1
true model is SLX
time taken is:      8.2840 seconds
model log-marginal model probs
slx  -1088.6072    0.3686
sdm  -1088.8197    0.2980
sdem -1088.7077    0.3334

```

## 2.2. Comparing models and spatial weight matrices

The log-marginal values returned in the function *lmarginal\_cross\_section* result structure variable can be viewed as a common denominator for comparing a number of alternative SLX, SDM, SDEM models based on varying spatial

<sup>2</sup> Users can rely on any structure variable name, for example options.lflag = 0 would work, with the name 'options' entered in place of 'info'.

weight matrices. Note that these cannot be used to compare models with differing sets of explanatory variables (see LeSage and Parent, 2007 regarding this type of comparison). To illustrate this, we use the per capita income growth rates over the 1987 to 1993 period for  $N = 3,111$  US counties (in the lower 48 states plus the district of Columbia) as the dependent variable  $y$  and the initial period (1987) logged income level and logged population level as explanatory variables to form the matrix  $X$ .

A series of nearest neighbor spatial weight matrices are compared along with the three model specifications using a loop, with the code shown below.

```
latt = uscounties(:,2); % extract latt-long coordinates
long = uscounties(:,3); % from the dataset
lmarginals = []; % empty matrix for storing results
neighbors = [];
for ii=5:16; % loop over varying nearest neighbors W-matrices
W = make_neighborsw(latt,long,ii);
neighbors = [neighbors ii];
result = lmarginal_cross_section(pci_growth,xmatrix,W);
% extract log-marginals and put them into a matrix
lmarginals = [lmarginals
              result.lmarginal']; % we transpose the 3 x 1 vector
                                % each column is a different model (column 1 = slx,
                                %                                column 2 = sdm column 3 = sdem)
end;
% calculate model probabilities for each column (model)
nmodels = length(lmarginals);
adj = max(lmarginals);
madj = matsub(lmarginals,adj);
xx = exp(madj);
% compute posterior probabilities
psum = sum(xx);
probs = [neighbors matdiv(xx,psum)];
in.fmt = strvcat('%10d', '%16.4f', '%16.4f', '%16.4f');
in.cnames = strvcat('#neighbors', 'slx', 'sdm', 'sdem');
mprint(probs,in);
```

Since the log-marginals require exponentiation, we use a normalization that subtracts the  $1 \times 3$  vector of maximum values from each column of the  $8 \times 3$  matrix (returned by the MATLAB *max* function) using the *matsub* function from the toolbox. This function performs non-conformable matrix subtraction allowing us to subtract the  $1 \times 3$  vector of column maximum values from each row of the matrix. This normalization avoids numerical over- and under-flows that may arise from exponentiation, but does not alter the posterior probability results.

The results which took 9.8 seconds to calculate are shown below, where we see that for all three models a 14 nearest neighbors weight matrix  $W$  has the highest probability.<sup>3</sup> The probabilities sum to unity down the columns, since the 8 different weight matrices are being compared for each model here.

---

<sup>3</sup> When using the default MC determinant option, `info.lflag = 1`, every run of the program will produce slightly different results because a Monte Carlo approximation to the log-determinant calculation is being used, (see Barry and Pace, 1999).

Changing the input options to the function to produce the most accurate but slowest results produced the second set of results shown, and took 2,786 seconds. It should be apparent that in cases where the sample size is large, the default input options produce reasonably accurate results.<sup>4</sup>

#neig	default options results (no info structure input)			most accurate options results (info.lflag = 0, info.eig = 1)		
	slx	sdm	sdem	slx	sdm	sdem
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	0.0126	0.0000	0.0000	0.0126	0.0000	0.0000
8	0.0168	0.0000	0.0000	0.0168	0.0000	0.0000
9	0.0109	0.0000	0.0000	0.0109	0.0000	0.0000
10	0.0050	0.0000	0.0000	0.0050	0.0000	0.0000
11	0.0289	0.0000	0.0000	0.0289	0.0000	0.0000
12	0.0501	0.0044	0.0000	0.0501	0.0098	0.0001
13	0.0814	0.0232	0.0028	0.0814	0.0310	0.0037
14	0.7646	0.8901	0.9683	0.7646	0.8710	0.9815
15	0.0210	0.0112	0.0021	0.0210	0.0056	0.0008
16	0.0086	0.0711	0.0268	0.0086	0.0826	0.0139

There is still the question of which specification (SLX, SDM, SDEM) is most consistent with this sample data. As noted, log-marginal likelihoods can be viewed as a common denominator for comparing models, so the following code re-normalizes the log-marginals using the maximum element of the entire  $8 \times 3$  matrix, not just the maximum of each column. This produces posterior model probabilities that compare all 24 model specifications such that the 24 probabilities sum to unity.

Interestingly, the results point to the SDEM specification (and a weight matrix based on 14 nearest neighbors) as most consistent with the sample data for this simple county-level spatial growth regression. Most empirical studies in the spatial growth regression literature use the SAR or SDM specification which are global spillovers models, in contrast to the local spillovers SDEM pointed to here. These results indicate that spatial spillovers at the county level do not extend (on average over the sample of counties) beyond first-order (contiguous) counties, and that the spillovers are the result of *contextual effects*, not *endogenous interaction* (see LeSage and Pace, 2014 for detailed discussion of these issues).

```
% find best model, slx, sdm, sdem
adj2 = max(max(lmarginals)); % maximum element of the matrix
madj2 = lmarginals-adj2;
% calculate probabilities across all models and weight matrices
xx2 = exp(madj2);
psum2 = sum(vec(xx2));
probs2 = [neighbors xx2./psum2];
mprint(probs2,in);
```

<sup>4</sup> The SLX model results are identical because no numerical integration is required to calculate the log-marginal likelihood in this case. Varying degrees of accuracy in numerical integration are what explains the difference in results (see LeSage, 2014 for details regarding this.)

#neig	default options results (no info structure input)			most accurate options results (info.lflag = 0, info.eig = 1)		
	slx	sdm	sdem	slx	sdm	sdem
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
12	0.0000	0.0000	0.0001	0.0000	0.0000	0.0001
13	0.0000	0.0000	0.0064	0.0000	0.0000	0.0037
14	0.0000	0.0004	0.9468	0.0000	0.0002	0.9812
15	0.0000	0.0000	0.0024	0.0000	0.0000	0.0008
16	0.0000	0.0002	0.0437	0.0000	0.0000	0.0139

A discussion of the way in which the levels of spatial dependence, and number of observations which impact signal-to-noise in spatial regression models impacts the accuracy of Bayesian model comparison methods can be found in LeSage and Pace (2009, Chapter 6).

### 3. PANEL DATA MODELS

Taking the same approach used to integration of the model parameters LeSage (2014) extended the cross-sectional approach to the case of a panel data model setting. The function *lmarginal\_static\_panel* can be used to compare static panel data specifications for SLX, SDM and SDEM.<sup>5</sup> A static panel model does not contain variables and parameters that allow for time dependence (see Parent and LeSage, 2012 for an example of a dynamic spatial panel data model specification).

A vector of panel data observations for  $N$  regions over  $T$  time periods should be organized such that the first  $N$  elements of the vector  $y$  represent the first time period, the next  $N$  elements the second time period, and so on. The  $NT \times k$  explanatory variables matrix  $X$  is organized in the same fashion, with the intercept excluded. For the moment, we ignore the issue of fixed or random effects as part of the model, with this issue discussed later. In addition to the  $NT \times 1$  vector  $y$  and  $NT \times k$  explanatory variables matrix  $X$ , we assume a single spatial configuration between the  $N$  regions for all time periods, and let  $W$  be an  $N \times N$  row-standardized spatial weight matrix.

The local spatial spillover specification for the case of static panel data models is the *spatial Durbin error model* (SDEM) shown in (4), where  $v$  represent fixed effects and  $\psi_t$  time period specific effects.

---

<sup>5</sup> LeSage (2014) shows that these three models subsume other models such as SAR and SEM as special cases, following the same logic as set forth in LeSage and Pace (2009).

$$\begin{aligned}
y_t &= X_t\beta_1 + WX_t\beta_2 + v + \psi_t + u_t \\
u_t &= \rho Wu_t + \varepsilon_t \quad t = 1, \dots, T \\
\varepsilon_t &\sim N(0, \sigma_\varepsilon^2 I_N), \quad t = 1, \dots, T
\end{aligned} \tag{4}$$

The model allows for local spillovers to immediately neighboring observations through spatial lag terms for the explanatory variables:  $WX_t$ , which have been labeled *contextual effects* in the social networking literature. To differentiate these from the *endogenous interaction* arising from spatial lags of the dependent variable, they have also been labeled *exogenous interaction effects*.

As in the case of cross-sectional models, as the spatial dependence parameter  $\rho \rightarrow 0$ , both the SDM and SDEM collapse to the panel data SLX specification. All three specifications are shown below using matrix notation, where the vector  $y$  is  $NT \times 1$ , the matrix  $X$  has dimension  $NT \times k$ , and we use  $F_S, F_T$  to denote spatial and time period fixed effects, and  $\otimes$  is the Kronecker product.

$$SLX: y = X\beta_1 + (I_T \otimes W)X\beta_2 + F_S + F_T + \varepsilon \tag{5}$$

$$SDM: y = \rho(I_T \otimes W) + X\beta_1 + (I_T \otimes W)X\beta_2 + F_S + F_T + \varepsilon \tag{6}$$

$$SDEM: y = X\beta_1 + (I_T \otimes W)X\beta_2 + F_S + F_T + u \tag{7}$$

$$u = \rho(I_T \otimes W)u + \varepsilon$$

Both SDM and SDEM can be viewed as *encompassing models* for the SLX, since they nest this model. Because of this, when  $\rho$  takes a value of zero, the sample data provide no basis for selecting between these three models. In this circumstance, the log-marginal likelihoods (and associated model probabilities) should approach equality. In fact, despite our interest in comparing the SDM versus SDEM models, practitioners would be interested in knowing when the level of spatial dependence in the data is low enough to allow use of the simpler SLX specification.

Lee and Yu (2010) have proposed a transformation procedure to eliminate both space and time fixed effects from the model. A MATLAB function *demeanF* written by Elhorst can be used to accomplish this task, and the documentation (shown below) for the function *lmarginal\_static\_panel* indicates that it is the user's responsibility to carry out this transformation. There is also a new input option, 'info.iflag' to indicate that this type of transformation has been used prior to inputting the (transformed) data vector  $\tilde{y}$ , (transformed) data matrix  $\tilde{X}$  and (transformed) weight matrix  $\tilde{W}$ .

```
>> help lmarginal_static_panel
PURPOSE: Bayesian log-
marginal posterior for static spatial panel models
user should eliminate fixed effects using differencing transfor-
mations
      no priors on beta, sige
      uniform prior on rho, lambda over eigenvalue bounds
-----
```

```

USAGE: results = lmarginal_static_panel(y,x,W,N,T,info)
where:
  y = dependent variable vector (N*T x 1)
  x = independent variables matrix, WITHOUT INTERCEPT TERM
  W = N by N spatial weight matrix (for W*y and W*e)
  N = # of cross-sectional units
  T = # of time periods
  info.lflag = 0 for full lndet computation (default = 1,fastest)
             = 1 for MC lndet approximation (fast for large problems)
  info.order = order to use with info.lflag = 1 option
             (default = 50)
  info.iter = iterations to use with info.lflag = 1(default = 30)
  info.rmin = (optional) minimum value of rho in search
             (default = -1)
  info.rmax = (optional) maximum value of rho in search
             (default = +1)
  info.iflag = 0 for conventional W-matrix
  info.iflag = 1 for transformed W-matrix (using dmeanF())

```

```

-----
RETURNS: a structure:
  results.meth = 'lmarginal_static_panel'
  results.nobs = # of cross-sectional observations
  results.ntime = # of time periods
  results.y = N*T x 1 vector of y from input
  results.nvar = # of variables in x-matrix
  results.rmin = minimum value of rho used (default +1)
  results.rmax = maximum value of rho used (default -1)
  results.lflag = lflag value from input (or default value used)
  results.iflag = iflag value from input (or default value used)
  results.lmarginal = a 3 x 1 column-vector with [log-marginal]
                    [ logm_slx logm_sdm logm_sdem ]'
  results.probs = a 3 x 1 column-vector with model probs
  results.logm_slx
  results.logm_sdm
  results.logm_sdem
-----

```

An example of using the transformation function *demeanF* is shown below, where ‘ted=0’ is an input option to request only spatial fixed effects, and ‘ted=1’ both space and time-period fixed effects. The example represents a data-generated  $y$ -vector based on  $T = 30$  years and  $N = 49$  states plus District of Columbia.

```

ted = 0;
[yf,Xf,No,To,Wf]=demeanF(y,X,N,T,ted,W);
info.iflag = 1;
result1 = lmarginal_static_panel(yf,Xf,Wf,No,To,info);
in.cnames = strvcats('log-marginal','model probs');
in.rnames = strvcats('model','slx','sdm','sdem');
in.width = 10000;
in.fmt = '%10.4f';
out = [result1.lmarginal result1.probs];
mprint(out,in);

```

```

true model is SLX
model log-marginal model probs
slx    -2896.3385    0.3182
sdm    -2895.9531    0.4679
sdem   -2896.7358    0.2139

```

The time taken to produce these results was round 0.2 seconds on an Intel i7-4510U CPU, 64-bit operating system with MATLAB version 8.3.0.532 (R2014a). We interpret these results to indicate that the SLX model (not the SDM) is the correct specification, and the data-generated  $y$ -vector was indeed created to follow the SLX specification. A lack of high probability in favor of the SDM or SDEM specifications should always be interpreted as evidence in favor of the SLX specification. In the face of low levels of spatial dependence, this is the correct inference regarding the alternative model specifications. LeSage (2014) reports results of a Monte Carlo experiment assessing performance of this Bayesian approach to static panel data model comparison with regard to changes in levels of spatial dependence and sample size. LeSage (2014) summarizes the Monte Carlo results by indicating that for the purpose of drawing inferences regarding the SDM versus SDEM specifications low-to-moderate levels of spatial dependence ( $\rho \geq 0.3$ ) will produce results that identify the correct specification.

As an illustration, we compare SLX, SDM and SDEM specifications for the case  $N = 49, T = 30$  and two matrices  $W$ , one based on first-order contiguity of the states and the other based on lengths of border miles in common between each state and its neighbors. A data-generated  $y$ -vector was constructed based on  $\rho = 0.3$ , with the true weight matrix being that based on border miles in common. Results for the case where the true specification was SDM when comparing three specifications (SLX, SDM and SDEM) based on each weight matrix as follows:

In this comparison the sum of the three model probabilities in each column is one, and the higher log-marginal values for the  $W_m$  miles weight matrix points to these three models as superior to those based on the  $W_c$  contiguity matrix. However, use of the  $W_c$  matrix produces the wrong inference regarding SLX, SDM, SDEM, erroneously pointing to the SDEM as the correct specification. Results are also shown for the case where  $\rho = 0.6$ , and here we see that even in the case of the wrong  $W_c$  matrix, the correct SDM specification is identified. Intuitively, it is difficult to draw inferences regarding the correct spatial weight matrix when there is low or moderate spatial dependence.

```

true model is SDM Wm (rho = 0.3)
model      log-marginal  model probs      log-marginal  model probs
slx Wm     -2952.5556         0.0000          slx Wc        -3043.9501     0.0000
sdm Wm     -2916.6129         0.8251          sdm Wc        -3014.6607     0.0372
sdem Wm    -2918.1642         0.1749          sdem Wc       -3011.4081     0.9628

true model is SDM Wm (rho = 0.6)
model      log-marginal  model probs      log-marginal  model probs
slx Wm     -3263.9848         0.0000          slx Wc        -3344.8898     0.0000
sdm Wm     -2992.2356         1.0000          sdm Wc        -3138.2277     1.0000
sdem Wm    -3038.2632         0.0000          sdem Wc       -3155.8827     0.0000

```

For a comparison of all six models such that the model probabilities sum to one, we use the two  $3 \times 1$  vectors of log-marginals based on the two different  $W$  matrices to construct new probabilities as follows:

```
% find best W-matrix using a 6 x 1 vector of log-marginals
lmarginals = [results1.lmarginal % 3 x 1 vector based on Wm
              results2.lmarginal]; % 3 x 1 vector based on Wc

nmodels = length(lmarginals);
adj = max(lmarginals(:,1));
madj = matsub(lmarginals,adj);
xx = exp(madj);
% compute posterior probabilities
psum = sum(xx);
probs = [matdiv(xx,psum)];
in2.fmt = strvcat('%16.4f');
in2.rnames = strvcat('models','slx Wm','sdm Wm','sdem Wm','slx Wc',
                    'sdm Wc','sdem Wc');
mprint(probs,in2);

models (rho = 0.3)
slx Wm      0.0000
sdm Wm      0.8251 % true model
sdem Wm     0.1749
slx Wc      0.0000
sdm Wc      0.0000
sdem Wc     0.0000
```

Here the posterior model probabilities identify the true SDM specification based on  $W_m$ , even in the case of low levels of spatial dependence ( $\rho = 0.3$ ).

#### 4. CONCLUSIONS

A set of MATLAB functions for calculating log-marginal likelihoods for comparing SLX, SDM and SDEM model specifications for both cross-sectional and static panel data scenarios was described here. The calculations are fast and Monte Carlo results reported in LeSage (2014) indicate that they should work well provided that moderate levels of spatial dependence exist (levels of the scalar dependence parameter  $\rho \geq 0.3$ .)

The functions can be used to draw inferences regarding local versus global spatial spillover specifications in cases where theoretical aspects of the modeling exercise do not clearly indicate an appropriate specification. Applied examples also show how to use these functions to compare alternative spatial weight matrices in an effort to find a weight matrix most consistent with the sample data.

An important caveat is that the model comparison approach illustrated here should not be used to compare model specifications based on differing sets of explanatory variables. For this type of model comparison, Parent and LeSage (2007) set forth an approach for spatial regression models similar to that of Fernandez, Ley, and Steel (2001) for non-spatial regression that have become popular in the Markov Chain Monte Carlo Model Comparison (*MCMC*<sup>3</sup>) literature. Comparing specifications based on different sets of explanatory variables requires assigning a Bayesian prior distribution (for example, the  $g$ -prior from Fernandez, Ley, and Steel, 2001) to parameters in the model, and future work should examine the effectiveness this and other related priors in the case of spatial regression models (see Han and Lee, 2013).

## REFERENCES

- Anselin L., A.K. Bera, R. Florax and M.J. Yoon (1996) Simple Diagnostic Tests for Spatial Dependence, *Regional Science & Urban Economics*, 26, 77-104.
- Baltagi B.H. (2001) *Econometric Analysis of Panel Data*, 2nd edition. Wiley, Chichester.
- Barry, R.P. and R.K. Pace (1999) A Monte Carlo Estimator of the Log Determinant of Large Sparse Matrices, *Linear Algebra and its Applications*, 289, 41-54.
- Elhorst, J.P. (2010) Applied spatial econometrics: raising the bar, *Spatial Economic Analysis*, 5, 928.
- Elhorst J.P. (2012) Matlab Software for Spatial Panels, *International Regional Science Review*, DOI: 10.1177/0160017612452429.
- Fernandez, C., E. Ley, and M.F.J. Steel (2001) Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100, 381-427.
- Han, X. and L. Lee (2013) Bayesian estimation and model selection for Spatial Durbin Error Models with Finite distributed lags, *Regional Science & Urban Economics*, 43, 816-837.
- LaFaive, M.D. and T. Nesbit (2013) Higher Cigarette Taxes Create Lucrative, Dangerous Black Market, *Mackinac Center For Public Policy*, [www.mackinac.org/18128](http://www.mackinac.org/18128);
- Lee, L.F. and J. Yu (2010) Some recent developments in spatial panel data models, *Regional Science & Urban Economics*, 40, 255-271.
- LeSage J.P. (2014) Spatial econometric panel data model specification: A Bayesian approach, *Spatial Statistics*, 9, 122-145.
- LeSage, J.P. and O. Parent (2007) Bayesian Model Averaging for Spatial Econometric Models, *Geographical Analysis*, 39, 3, 241-267.
- LeSage, J.P. and R.K. Pace (2009) *Introduction to Spatial Econometrics*, CRC Press Taylor & Francis Group, Boca Raton.
- LeSage, J.P. and R.K. Pace (2014) Interpreting Spatial Econometric Models, in *Handbook of Regional Science*, M. M. Fischer and Peter Nijkamp (Eds.) Springer, Berlin, 1535-1552.
- Parent, O. and J.P. LeSage (2012) "Spatial dynamic panel data models with random effects," *Regional Science & Urban Economics*, 42: 4, 727-738.

## SOFTWARE POUR LA COMPARAISON DES MODÈLES SPATIAUX BAYÉSIENS

**Résumé** - *L'analyse bayésienne a permis d'améliorer sensiblement la comparaison et la sélection des modèles spatiaux statiques sur des données de panel. Durant, ces dernières années, les statisticiens et les économètres se sont focalisés, plus particulièrement, sur l'étude et la comparaison d'un ensemble de spécifications alternatives des modèles spatiaux existants, permettant de combiner des effets de dépendance spatiale sur la variable expliquée, sur les variables explicatives, voire même sur le terme d'erreur. Sur un plan théorique, LeSage et Pace (2009) montrent que les procédures de sélection du modèle spatial le plus adéquat peuvent être simplifiées, si l'on considère une hypothèse binaire dans laquelle soit on observe des effets d'autocorrélation spatiale sur un plan global (appréhendés par le modèle Durbin spatial) soit on observe des effets d'autocorrélation spatiale dont le caractère est purement local (appréhendés par le modèle Durbin spatial à erreurs). Lesage (2014) introduit cette hypothèse également dans la sélection des modèles spatiaux sur des données de panel. L'objectif de cet article est de présenter la programmation de la comparaison des modèles spatiaux sur des données de panel, en s'appuyant sur les méthodes bayésiennes. Cette programmation est faite sous MATLAB, agrémentée de nombreuses applications empiriques.*

**Mots-clés** - MODÈLES SPATIAUX STATIQUES SUR DES DONNÉES DE PANEL, FACTEURS BAYÉSIENS, EFFETS GLOBAUX OU LOCAUX D'AUTOCORRÉLATION SPATIALE