Selecting the Most Adequate Spatial Weighting Matrix:A Study on Criteria

Herrera Gómez, Marcos and Mur Lacambra, Jesús and Ruiz Marín, Manuel

July 2012

Selecting the Most Adequate Spatial Weighting Matrix: A Study on Criteria*

Marcos Herrera (University of Zaragoza); mherreragomez@gmail.com

Jesús Mur (University of Zaragoza); jmur@unizar.es[1]

Manuel Ruiz (University of Murcia); manuel.ruizmarin@um.es

Abstract

In spatial econometrics, it is customary to specify a weighting matrix, the so-called W matrix, by choosing one matrix from a finite set of matrices. The decision is extremely important because, if the W matrix is misspecified, the estimates are likely to be biased and inconsistent. However, the procedure to select W is not well defined and, usually, it reflects the judgments of the user. In this paper, we revise the literature looking for criteria to help with this problem. Also, a new nonparametric procedure is introduced. Our proposal is based on a measure of the information, conditional entropy. We compare these alternatives by means of a Monte Carlo experiment.

1 Introduction

The weighting matrix is a very characteristic element of spatial models and, frequently, is the cause of dispute in relation to what it is and how it should be specified. This matrix is a key element for modeling spatial data and it has received considerable attention (Anselin, 2002). However, from our point of view, we still do not have a complete convincing answer to both questions.

Formally, for any spatial sample, the spatial weights matrix is an $N \times N$ positive matrix, where N is the size of the data set:

*Paper presented at the VIth World Conference - Spatial Econometrics Association. Salvador, Brazil. July 11-13, 2012.

[1]Corresponding author: Department of Economic Analysis, University of Zaragoza. Gran Via 2-4 (50005). Zaragoza (Spain).

$$W = \begin{bmatrix} 0 & w_{1,2} & \cdots & w_{1,j} & \cdots & w_{1,N} \\ w_{2,1} & 0 & \cdots & w_{2,j} & \cdots & w_{2,N} \\ \vdots & \vdots & \ddots & \cdots & \cdots & \cdots \\ w_{i,1} & w_{i,2} & \vdots & 0 & \cdots & w_{i,N} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \cdots \\ w_{N,1} & w_{N,2} & \vdots & w_{N,j} & \vdots & 0 \end{bmatrix}, \tag{1}$$

where the elements w_{ij} are the spatial weights. The spatial weights w_{ij} are non-zero when i and j are hypothesized to be neighbors, and zero otherwise. By convention, the self-neighbor relation is excluded, so the diagonal elements of W, $w_{ii} = 0$.

As a extensively used criterion for choosing the weights w_{ij} is that of geographical proximity or distance. But, but this is not necessarily true for all applications as evidenced by the principle of *allotopy* stated by Ancot et al. (1982): "often what happens in a region is related with other phenomena located in distinct and remote parts of the space". The problem of the identification what observations are neighbors and how to introduce them in the analysis is not a particular problem of spatial econometrics. Similar problems there are in time series although in this case we have some clear indications: you have to look to the past and take into account also the frequency of the data. However, the space is irregular and heterogeneous and the influences may be of any type across space. Nearness, as claimed by Tobler (1970) is just one possibility.

We agree with Haining (2003, p.74) in the sequence of actions: "*The first step of quantifying the structure of spatial dependence in a data set is to define for any set of points or area objects the spatial relationships that exist between them*". This is what Anselin (1988, p. 16) designates "*the need to determine which other units in the spatial system have an influence on the particular unit under consideration (...) expressed in the topological notions of neighborhood and nearest neighbor*". This step is crucial, but how do we do? In some cases, we might have enough information to fully specify a weighting matrix. In other cases, this matrix will be no more than a mere hypothesis. In fact, from our experience, we suspect that the second situation is the most common among practitioners.

The uncertainty that dominates the specification of the weighting matrix results from a problem of underidentification that affects, in general, to the most part of spatial models. Paelinck (1979, p.20) admits that there is an identification problem in the interdependent specifications used to model spatial behaviors. In terms of Lesage and Pace (2009, p.8), an unrestricted spatial autoregressive process like the following:

$$\left.\begin{array}{c} y_i = \alpha_{ij} y_j + \alpha_{ik} y_k + x_i \beta + \varepsilon_i \\ y_j = \alpha_{ji} y_i + \alpha_{jk} y_k + x_j \beta + \varepsilon_j \\ y_k = \alpha_{ki} y_i + \alpha_{kj} y_j + x_k \beta + \varepsilon_k \\ \varepsilon_i; \varepsilon_j; \varepsilon_k \sim N\left(0; \sigma^2\right) \end{array}\right\}, \tag{2}$$

"would be of little practical usefulness since it would result in a system with many more parameters than observations. The solution to the over-parametrization problem that arises when we allow each dependence relation to have relation-specific parameters is to impose structure on the spatial dependence

parameters". This is the reason why we need a spatial weighting matrix. Folmer and Oud (2008) propose what they call a structural approach to the problem of specifying a weighting matrix; Paci and Usai (2009) advocate for the use of proxies for the existence of spillover effects, etc. There are other proposals in the literature trying to amend the paucity of information that involves this problem. However, these methods also have restrictions and limitations.

The question of specifying a matrix seems really complex, although the practitioners prefer simple solutions. By large, the dominant approach involves an exogenous treatment of the problem. Nearby or neighboring units are treated as contiguous in a square binary connectivity matrix as, for example, in the traditional physical adjacent criteria, the k-nearest neighbors or the great circle distance. Afterward, the binary matrix can be normalized in some way (by rows, columns, according to the total sum). Other matrices are constructed using some given function of the geographical distance between the centroids of the spatial units; the inverse of the distance between the two points is a popular decision; then the matrix can be normalized. Geography may be replaced by another domain in order to obtain others measures of distance. Recently various endogenous procedures have appeared like the AMOEBA algorithm (Getis and Aldsdat, 2004, Aldsdat and Getis, 2006), the CCC method of Mur and Paelinck (2010) or the entropy-based approach of Fernandez et al (2009). Although the differences between them, the basic idea of these algorithms appeared in the works of Kooijman (1976) and Openshaw (1977): using the information contained in the raw data, or in the residuals of the model in order to estimate the weighting matrix. This can be done if we have a panel of spatial data like in Conley and Molinari (2007), Bhattacharjee and Jensen-Butler (2006) and Beenstock et al (2010) but it is not easy in the case of a single cross-section. Finally, there are different well-known approaches which combine strong a prioris about the channels of interaction with endogenous inferential algorithms (Bodson and Peters, 1975, Dacey, 1965).

Bavaud (1998, p.153), given this state of affairs, is clearly skeptical, *"there is no such thing as "true", "universal" spatial weights, optimal in all situations*" and continues by stating that the weighting matrix "*must reflect the properties of the particular phenomena, properties which are bound to differ from field to field*". This means that, in the end, the problem of selecting a weighting matrix is a problem of model selection. In fact, different weighting matrices result in different spatial lags of the endogenous or the exogenous variables included in the model. Different equations with different regressors mean a model selection problem, even when the weighting matrix appears in the error term. This is the direction that we want to explore in the present paper as an alternative way to deal with the uncertainty of specifying the spatial weighting matrix.

Section 2 continues with a revision of the techniques of model selection that seem to fit better into our problem. We present our own non-parametric procedure in Section 3. Section 4 discusses a large Monte Carlo experiment in which we compare the small sample behavior of the most promising techniques. Section 5 concludes summarizing the most interesting results of our work.

2 Selecting a Weighting Matrix

The model of equation (2) can be written in matrix form:

$$y = \Gamma y + x\beta + \varepsilon, \tag{3}$$

where y and ε are $(n \times 1)$ vectors, x is a $(n \times k)$ matrix, β is a $(k \times 1)$ vector of parameters and Γ is a $(n \times n)$ matrix of interaction coefficients. The model is underidentified and a common solution to achieve identification consists in introducing some structure in the matrix Γ. This means to impose restrictions on the spatial interaction coefficients as, for example: $\Gamma = \rho W$, where ρ is a parameter and W is a matrix of weights. The term $y_W = Wy$ that appears on the right hand side of the equation is one spatial lag of the endogenous variable. It is worth highlighting a couple of questions:

(i) The weighting matrix can be constructed in different ways using different interaction hypothesis. Each hypothesis results in a different weighting matrix and in a different spatial lag. In conclusion, different weighting matrices means different models containing different variables.

(ii) There are general guidelines about specifying a weighting matrix. For example: nearness, accessibility, influence, etc. However, it will be difficult to say which of these general principles would be better. The problem is clearly dominated by uncertainty.

Corrado and Fingleton (2011) discuss the construction of a weighting matrix from a theoretical perspective (they are worried, for example, about the information that the weights of a weighting matrix should contain). We prefer to focus on the statistical treatment of such uncertainty.

Let us assume that we have a set of N linearly independent weighting matrices, $\Upsilon = \{W_1; W_2; \ldots; W_N\}$. Usually N corresponds to a small number of different competing matrices but in some cases this number may be quite large, reflecting a situation of great uncertainty. For instance, each matrix generates a different spatial lag and a different spatial model. These matrices may be related by different restrictions, resulting in a series of nested models. If the matrices are not related, the sequence of spatial models will be non-nested.

Two weighting matrices may be nested, for example, in the cases of binary rook-type and queen-type movements: all the links of the first matrix are contained in the second matrix which include also some other non-zero links. Discriminating between these two matrices is not difficult using the techniques for selecting between nested models. For example, in a maximum-likelihood approach (we would need the assumption of normality) it may be enough with a likelihood ratio or a Lagrange Multiplier. The last one is very simple as can be seen in Appendix 1.

For the case of non-nested matrices, we might find several proposals in the literature. Anselin (1984) provides the appropriate Cox-statistic for the case of:

$$\left.\begin{array}{l} H_0 : y = \rho_1 W_1 y + x_1\beta_1 + \varepsilon_1 \\ H_A : y = \rho_2 W_2 y + x_2\beta_2 + \varepsilon_2 \end{array}\right\}, \tag{4}$$

that Leenders (2002) converts into the J-test using an augmented regression like the following:

$$y = (1-\alpha)\left[\rho_1 W_1 y + x_1\beta_1\right] + \alpha\left[\hat{\rho}_2 W_2 y + x_2\hat{\beta}_2\right] + \nu, \tag{5}$$

being $\hat{\rho}_2$ and $\hat{\beta}_2$ the corresponding maximum-likelihood estimates (ML from now on) of the respective parameters on a separate estimation of the model of H_A. Leenders shows that the J-test can be

extended to the comparison of a null model against N different models. Kelejian (2008) maintains the approach of Leenders in a $SARAR$ framework, which requires GMM estimators:

$$\begin{aligned} y &= \rho_i W_i y + x_i \beta_i + u_i = Z_i \gamma_i + u_i, \\ u_i &= \lambda_i M_i u_i + v_i, \end{aligned} \tag{6}$$

with $i = 1, 2, ..., N$, $Z_i = (W_i y, x_i)$ and $\gamma_i = (\rho_i, \beta)$. The J-test for selecting a weighting matrix corresponds to the case where $x_i = x$; $W_i = M_i$ but $W_i \neq W_j$. In order to obtain the test, we need the estimation of an augmented regression, similar to that of (5):

$$y(\hat{\lambda}) = S(\hat{\lambda})\eta + \varepsilon, \tag{7}$$

where $S(\hat{\lambda}) = \left[Z(\hat{\lambda}), F\right]$, $Z(\lambda) = (I - \lambda W) Z$ (the same for $y(\hat{\lambda})$), being $\hat{\lambda}$ the estimate of λ for the model of the null. Moreover $F = [Z_1\hat{\gamma}_1, Z_2\hat{\gamma}_2, \ldots, Z_N\hat{\gamma}_N, W_1 Z_1\hat{\gamma}_1, W_2 Z_2\hat{\gamma}_2, \ldots, W_N Z_N\hat{\gamma}_N]$. The equation of (7) can be estimated by 2SLS using a matrix of instruments: $\hat{S} = \left[\hat{Z}(\hat{\lambda}), \hat{F}\right]$, where $\hat{F} = PF$ (similar for $Z(\hat{\lambda})$) with $P = H\left(H'H\right)^{-1} H$ and $H = \left[x, Wx, W^2x\right]$. Under the null that, for example, model 0 is correct and the 2SLS estimate of η is asymptotically normal:

$$\hat{\eta} \sim N\left[\eta_0; \sigma_\epsilon^2 \left(\hat{S}'\hat{S}\right)^{-1}\right], \tag{8}$$

where $\eta_0 = [\gamma'; 0]$. The J-test checks that the last $2N$ parameters of vector η are zero. Define $\hat{\delta} = A\hat{\eta}$ where A is a $2N \times (k + 1 + 2N)$ matrix corresponding to the null hypothesis: $H_0 :\ A\eta = 0$, then the J-test can be formulated as a Wald statistic:

$$\hat{\delta}'\hat{V}^{-1}\hat{\delta} \sim \chi^2(2N), \tag{9}$$

being $\hat{V}$ the estimated sample covariance of $\hat{\delta}$.

Burridge and Fingleton (2010) show that the asymptotic Chi-square distribution can be a poor approximation. They advocate for a bootstrap resampling procedure that appears to improve both the size and the power of the J-test. There, remains implementation problems related to the use of consistent estimates for the parameters of (6) in the corresponding augmented regression. Kelejian (2008) proposes to construct the test using GMM-type estimators and Burridge (2011) suggests a mixture between GMM and likelihood-based moment conditions which controls more effectively the size of the test. Piras and Lozano (2010) present new evidence on the use of the J-test that relates the power of the test to a wise selection of the instruments.

The problem of model selection has been often treated, very successfully, from a Bayesian perspective (Leamer, 1978); this also includes the case of selecting a weight matrix in a spatial model by Hepple (1985a, b). The Bayesian approach, although highly demanding in terms of information, is appealing and powerful (Lesage and Pace, 2009). The same as the J-test, the starting point is a finite set of alternative models, $M = \{M_1; M_2; \ldots; M_N\}$. The specification of each model coincides (regressors, structure of dependence, etc.) but not for the spatial weighting matrix. Denote by θ

the vector of k parameters. Then, the joint probability of the set of N models, k parameters and n observations corresponds to:

$$p\left(M,\theta,y\right)=\pi\left(M\right)\pi\left(\theta|\,M\right)L\left(y|\,\theta,M\right), \tag{10}$$

where $\pi\left(M\right)$ refers to the priors of the models, usually $\pi\left(M\right)=1/N$; $\pi\left(\theta|\,M\right)$ reflects the priors of the vector of conditional parameters to the model and $L\left(y|\,\theta,M\right)$ is the likelihood of the data conditioned on the parameters and models. Using the Bayes' rule:

$$p\left(M,\theta|y\right)=\frac{p\left(M,\theta,y\right)}{p\left(y\right)}=\frac{\pi\left(M\right)\pi\left(\theta|\,M\right)L\left(y|\,\theta,M\right)}{p\left(y\right)}. \tag{11}$$

The posterior probability of the models, conditioned to the data, results from the integration of (11) over the parameter vector θ:

$$p\left(M|y\right)=\int p\left(M,\theta|y\right)d\theta. \tag{12}$$

This is the measure of probability needed in order to compare different weighting matrices. Lesage and Pace (2009) discuss the case of a Gaussian SAR model:

$$\left.\begin{array}{c} y=\rho_i W_i y+X_i\beta_i+\varepsilon_i \\ \varepsilon_i\sim i.i.d.\mathcal{N}(0;\sigma_\epsilon^2) \end{array}\right\}, \tag{13}$$

The log-marginal likelihood of (10) is:

$$p\left(M|y\right)=\int \pi_\beta\left(\beta|\,\sigma^2\right)\pi_\sigma\left(\sigma^2\right)\pi_\rho\left(\rho\right)L\left(y|\,\theta,M\right)d\beta d\sigma^2 d\rho. \tag{14}$$

They assume independence between the priors assigned to β and σ^2, Normal-Inverse-Gamma conjugate priors, and that for ρ, a $Beta(d,d)$ distribution. The calculations are not simple and, finally, "*we must rely on univariate numerical integration over the parameter ρ to convert this* (expression 14) *to the scalar expression necessary to calculate $p\left(M|Y\right)$ needed for model comparison purposes*" (Lesage and Pace, 2009, p 172). The SEM case is solved in Lesage and Parent (2007); to our knowledge, the $SARAR$ model of (6) remains still unsolved.

Model selection techniques may also have a role in this problem, specially if we have any preference for any weighting matrix. In other words, we are not considering the idea of a null hypothesis. There is a huge literature on model selection for nested and non-nested models with different purposes and criteria. In our case, we are looking for the most appropriate weighting matrix for the data. We consider that the Kullback-Leibler information criterion might be a good measure. Apart from Kullback-Leibler criterion, we can use the Akaike information criterion which is simple to obtain. This criterion assures a balance between fit and parsimony (Akaike, 1974). The expression of Akaike criterion is very well-known:

$$AIC_i=-2L\left(\hat{\theta};y\right)+q(k), \tag{15}$$

being $L\left(\hat{\theta};y\right)$ the log-likelihood of the model at the maximum-likelihood estimates, $\hat{\theta}$, and $q(k)$ a penalty function that depends on the number of unknown parameters. Usually, the penalty function is simply equal to $q(k) = 2k$. The decision rule is to select the model, weighting matrix in our case, that produces the lowest AIC.

Recently Hansen (2007) introduced another perspective to the problem of model selection, which tries to reflect the confidence of the practitioner in the different alternatives. In general, the selection criteria that minimize the mean-square estimation error achieve a good balance between bias, due to misspecification errors, and variance due to parameter estimation. The optimal criterion would select the estimator with the lowest risk. This is what happens with the Bayesian concept of posterior probability, which combines prior with sampling information to select the best model; also with the selection criteria as, for example, the AIC or the $SBIC$ statistics. The procedure of the J-test is a classical decision problem solved using only sampling information, with the purpose of minimizing the type II error and assuring a given type I error.

Expressed in another way, given our collection of weighting matrices $W = \{W_1; W_2; \ldots; W_N\}$, all of which are referred to the same spatial model, the purpose is to select the matrix W_n. This matrix combines with the other terms of the model produces a vector of estimates, $\hat{\theta}_n(W_n)$, which minimizes the risk. Hansen (2007) shows that further reductions in the mean-squared error can be attained by averaging across estimators. The averaging estimator for θ is:

$$\hat{\theta}(W) = \sum_{n=1}^{N} \varpi^n \hat{\theta}_n(W_n). \tag{16}$$

As stated by Hansen and Racine (2010), the collection of weights, $\{\varpi^n; n = 1, 2, ..., N\}$ should be non-negative and linked on the unit simplex of $\mathbb{R}^{\mathbb{N}}; \sum_{n=1}^{N} \varpi^n = 1$.

Subsequently, these weights ϖ^n can be used to compare the adjustment of each model (W matrix) with respect to the data.

3 A Non-Parametric Proposal for Selecting a Weighting Matrix

This section presents a new non-parametric procedure for selecting a weighting matrix. The selection criterion is based on the idea that the most adequate matrix should produce more information with respect to the variables that we are trying to relate. The measure of information is a reformulation of the traditional entropy index in terms of what is called *symbolic entropy*, and it does not depend on judgments of the user.

As explained in Matilla and Ruiz (2008), the procedure implies, first, transforming the series into a sequence of symbols which should capture all of the relevant information. Then we translate the inference to the space of symbols using appropriate techniques.

Beginning with the symbolization process, assuming that $\{x_s\}_{s\in S}$ and $\{y_s\}_{s\in S}$ are two spatial processes, where S is a set of locations in space. Denoted by $\Gamma_l = \{\sigma_1, \sigma_2, \ldots, \sigma_l\}$ the set of symbols defined by the practitioner; σ_i, for $i = 1, 2, \ldots, l$, is a symbol. Symbolizing a process is defining a map

$$f : \{x_s\}_{s \in S} \to \Gamma_l, \tag{17}$$

such that each element x_s is associated to a single symbol $f(x_s) = \sigma_{i_s}$ with $i_s \in \{1, 2, \ldots, l\}$. We say that location $s \in S$ is of the $\sigma_i - type$, relative to the series $\{x_s\}_{s \in S}$, if and only if $f(x_s) = \sigma_{i_s}$. We call f the *symbolization map*. The same procedure can be followed for a second series $\{y_s\}_{s \in S}$.

Denoted by $\{Z_s\}_{s \in S}$ a bivariate process as:

$$Z_s = \{x_s, y_s\}. \tag{18}$$

For this case, we define the set of symbols Ω_l as the direct product of the two sets Γ_l, that is, $\Omega_l^2 = \Gamma_l \times \Gamma_l$ whose elements are the form $\eta_{ij} = \left(\sigma_i^x, \sigma_j^y\right)$. The symbolization function of the bivariate process would be

$$g : \{Z_s\}_{s \in S} \to \Omega_l^2 = \Gamma_l \times \Gamma_l, \tag{19}$$

defined by

$$g(Z_s = (x_s, y_s)) = (f(x_s), f(y_s)) = \eta_{ij} = \left(\sigma_i^x, \sigma_j^y\right). \tag{20}$$

We say that s is $\eta_{ij} - type$ for $Z = (x, y)$ if and only if s is $\sigma_i^x - type$ for x and $\sigma_j^y - type$ for y.

In the following, we are going to use a simple symbolization function f . Let M_e^x be the median of the univariate spatial process $\{x_s\}_{s \in S}$ and define an indicator function

$$\tau_s = \begin{cases} 1 & if \quad x_s \geq M_e^x \\ 0 & otherwise \end{cases}. \tag{21}$$

Let $m \geq 2$ be the *embedding dimension*; this is a parameter defined by the practitioner. For each $s \in S$, let N_s be the set formed by the $(m-1)$ neighbours of s. We use the term $m - surrounding$ to denote the set formed by each s and N_s, such that $m - surrounding$ of $x_m(s) = \left(x_s, x_{s_1}, \ldots, x_{s_{m-1}}\right)$. Let us define another indicator function for each $s_i \in N_s$:

$$\iota_{ss_i} = \begin{cases} 0 & if \quad \tau_s \neq \tau_{s_i} \\ 1 & otherwise \end{cases}. \tag{22}$$

Finally, we have a symbolization map for the spatial process $\{x_s\}_{s \in S}$ as $f : \{x_s\}_{s \in S} \to \Gamma_m$:

$$f(x_s) = \sum_{i=1}^{m-1} \iota_{ss_i}, \tag{23}$$

where $\Gamma_m = \{0, 1, \ldots, m-1\}$. The cardinality of Γ_m is equal to m.

Let us introduce some fundamental definitions:

Definition 1: The Shannon entropy, $h(x)$, of a discrete random variable x is: $h(x) = -\sum_{i=1}^{n} p(x_i) \, ln(p(x_i))$.

Definition 2: The entropy $h(x, y)$ of a pair of discrete random variables (x, y) with joint distribution $p(x, y)$ is: $h(x, y) = -\sum_x \sum_y p(x, y) \, ln(p(x, y))$.

Definition 3: Conditional entropy $h(x|y)$ with distribution $p(x,y)$ is defined as: $h(x|y) = -\sum_x \sum_y p(x,y) \, ln\left(p(x|y)\right)$.

The last index, $h(x|y)$, is the entropy of x that remains when y has been observed.

These entropy measures can be easily adapted to the empirical distribution of the symbols. Once the series have been symbolized, for a embedding dimension $m \geq 2$, we can calculate the absolute and relative frequency of the collections of symbols $\sigma_{i_s}^x \in \Gamma_l$ and $\sigma_{j_s}^y \in \Gamma_l$.

The absolute frequency of symbol σ_i^x is:

$$n_{\sigma_i^x} = \#\left\{s \in S | s \quad is \quad \sigma_i^x - type \quad for \quad x\right\}. \tag{24}$$

Similarly, for series $\{y_s\}_{s\in S}$, the absolute frequency of symbol σ_j^y is:

$$n_{\sigma_j^y} = \#\left\{s \in S | s \quad is \quad \sigma_j^y - type \quad for \quad y\right\}. \tag{25}$$

Next, the relative frequencies can also be estimated:

$$p\left(\sigma_i^x\right) \equiv p_{\sigma_i^x} = \frac{\#\left\{s \in S | s \quad is \quad \sigma_i^x - type \quad for \quad x\right\}}{|S|} = \frac{n_{\sigma_i^x}}{|S|}, \tag{26}$$

$$p\left(\sigma_j^y\right) \equiv p_{\sigma_j^y} = \frac{\#\left\{s \in S | s \quad is \quad \sigma_j^y - type \quad for \quad y\right\}}{|S|} = \frac{n_{\sigma_j^y}}{|S|}, \tag{27}$$

where $|S|$ denotes the cardinal of set S; in general $|S| = N$.

Similarly, we calculate the relative frequency for $\eta_{ij} \in \Omega_l^2$:

$$p\left(\eta_{ij}\right) \equiv p_{\eta_{ij}} = \frac{\#\left\{s \in S | s \quad is \quad \eta_{ij} - type\right\}}{|S|} = \frac{n_{\eta_{ij}}}{|S|}. \tag{28}$$

Finally, the *symbolic entropy* for the $two-dimensional$ spatial series $\{Z_s\}_{s\in S}$ is:

$$h_Z(m) = -\sum_{\eta \in \Omega_m^2} p(\eta) \, ln\left(p(\eta)\right). \tag{29}$$

We can obtain the marginal symbolic entropies as

$$h_x(m) = -\sum_{\sigma^x \in \Gamma_m} p\left(\sigma^x\right) ln\left(p\left(\sigma^x\right)\right), \tag{30}$$

$$h_y(m) = -\sum_{\sigma^y \in \Gamma_m} p\left(\sigma^y\right) ln\left(p\left(\sigma^y\right)\right). \tag{31}$$

In turn(tern), we can obtain the symbolic entropy of y, conditioned by the occurrence of symbol σ^x in x as:

$$h_{y|\sigma^x}(m) = -\sum_{\sigma^y \in \Gamma_m} p\left(\sigma^y|\sigma^x\right) ln\left(p\left(\sigma^y|\sigma^x\right)\right). \tag{32}$$

We can also estimate the conditional symbolic entropy of y_s given x_s:

$$h_{y|x}(m) = \sum_{\sigma^x \in \Gamma_m} p(\sigma^x)\, h_{y|\sigma^x}(m). \tag{33}$$

Let us move to the problem of choosing a weighting matrix for the relationship between the variables x and y. This selection will be made from among a finite set of relevant weighting matrices. Denoted by $\mathcal{W}(x,y) = \{W_\jmath | \jmath \in \mathcal{J}\}$ this set of matrices, where $\mathcal{J}$ is a set of index. We refer to $\mathcal{W}(x,y)$ as the spatial-dependence structure set between x and y.

Denoted by $\mathcal{K}$ a subset of Γ_m, the space of symbols, and let $W \in \mathcal{W}(x,y)$ be a member of the set of matrices. We can define

$$\mathcal{K}_W^x = \{\sigma^x \in \mathcal{K} | \sigma^x \text{ is admissible for } Wx\}, \tag{34}$$

where *admissible* indicates that the probability of symbol occurrence is positive.

By Γ_m^x we denote the set of symbols which are admissible for $\{x_s\}_{s\in S}$. Let $W_0 \in \mathcal{W}(x,y)$ be the most informative weighting matrix for the relationship between x and y. Given the spatial process $\{y_s\}_{s\in S}$, there is a subset $\mathcal{K} \subseteq \Gamma_m$ such that $p\left(\mathcal{K}_{W_0}^x | \sigma^y\right) > p\left(\mathcal{K}_W^{*x} | \sigma^y\right)$ for all $\mathcal{K}^* \subseteq \Gamma_m$, $W \in \mathcal{W}(x,y) \setminus \{W_0\}$ and $\sigma^y \in \Gamma_m^y$. Then

$$\begin{aligned} h_{W_0 x|y}(m) &= -\sum_{\sigma^y \in \Gamma^y} p(\sigma^y) \left[\sum_{\sigma^x \in \mathcal{K}_{W_o}^x} p(\sigma^x|\sigma^y) \ln\left(p(\sigma^x|\sigma^y)\right) \right] \\ &\leq -\sum_{\sigma^y \in \Gamma^y} p_{\sigma^y} \left[\sum_{\sigma^x \in K_W^{*x}} p(\sigma^x|\sigma^y) \ln\left(p(\sigma^x|\sigma^y)\right) \right] = h_{Wx|y}(m). \end{aligned} \tag{35}$$

In this way, we have proved the following theorem.

Theorem 1: *Let $\{x_s\}_{s\in S}$ and $\{y_s\}_{s\in S}$ two spatial processes. For a fixed embedding dimension $m \geq 2$, with $m \in \mathbb{N}$, if the most important weighting matrix that reveals the spatial-dependence structure between x and y is $W_0 \in \mathcal{W}(x,y)$ then*

$$h_{W_0 x|y}(m) = \min_{W \in \mathcal{W}(x,y)} \left\{ h_{Wx|y}(m) \right\}. \tag{36}$$

Given the Theorem 1 and using the following property: $h_{Wx|y} \leq h_{Wx}$, we propose the following criterion for selecting between different matrices:

$$pseudo - R^2 = 1 - {}^{h_{Wx|y}(m)}/_{h_{Wx}(m)}.$$

The selection of the matrix is made using the highest value of $pseudo - R^2$.

4 The Monte Carlo Experiment

In this section, we generate a large number of samples from different data generation process (D.G.P.) to study the performance of different proposals: J-test, Bayesian approach, averaging estimator (Racine-Hansen) and conditional symbolic entropy.

Our major interest is to detect the weighting matrix more informative between different alternatives. For this, we have an unique explanatory variable x, the same in all models. But the D.G.P. uses different spatial structures, that is $W = W_i$, where i is the matrix for the $i-th$ alternative model.

Each experiment starts by obtaining a random map in a hypothetical two-dimensional space. This irregular map is reflected on the corresponding normalized W matrix. In the first case, W is based on a matrix of 1s and 0s denoting contiguous and non-contiguous regions, respectively. Afterward, we normalize the W matrix so that the sum of each row is equal to 1.

The following global parameters are involved in the $D.G.P.$:

$$N \in \{100, 400, 700, 1000\}\,,\ k \in \{4, 5, 7\}\,, \tag{37}$$

where N is the sample size and k is the number of neighbors for each observation. The number of replications is equal 1000.

In the cases of nested models, we use the following matrices:

- $W_4 = 4 - nearest - neighbors$
- $W_5 = 5 - nearest - neighbors$
- $W_7 = 7 - nearest - neighbors$

where W_7 contains W_5 matrix and W_5 contains W_4 matrix, before the standardization.

In the cases of non-nested models, all spatial weighting matrices contain 4 neighbors but we modify the criterion of neighborhood. In all cases, we assume the following non-nested matrices:

- $W^{(1)} = 4 - nearest - neighbors$
- $W^{(2)} = 5^{o} - to - 8^{o} - nearest - neighbors$
- $W^{(1-2)} = 1^{o} - 2^{o} - 5^{o} - 6^{o} - nearest - neighbors$

In this experiment, we want to simulate both linear and non-linear relationships between the variables x and y.

In the first case, linearity, we control the relationship between variables using the *expected coefficient of determination* ($R^2_{y/x}$) based on a specification like this:

$$y = \beta x + \theta W x + \varepsilon. \tag{38}$$

Under equation (38), the expected coefficient of determination between the variables is equal to (assuming an unit variance of x and in ε as well as incorrelation between the two variables):

$$R^2_{y/x} = \frac{\beta^2 + \left(\theta^2/_{m-1}\right)}{\beta^2 + \left(\theta^2/_{m-1}\right) + 1}$$

We have considered different values for this coefficient:

$$R^2_{y/x} \in \{0.4; 0.6; 0.8\} \tag{39}$$

For simplicity, in all cases we maintain $\beta = 0.5$. The spatial lag parameter of x, θ, is obtained by deduction: $\theta = \sqrt{\frac{(1-m)(\beta^2(1-R^2)-R^2)}{1-R^2}}$.

Having defined the values of the parameters involved in the simulation, we can present the different processes used in the analysis.

DGP1: Linear

$$y = \beta x + \theta W x + \varepsilon \tag{40}$$

DGP2: Non-linear 1

$$y = \exp\left[(\beta x + \theta W x + \varepsilon)^{1.25}\right] \tag{41}$$

DGP3: Non-linear 2

$$y = {}^{1}\!/_{(\beta x + \theta W x + \varepsilon)^2} \tag{42}$$

In all cases: $x \sim \mathcal{N}(0,1)$, $\varepsilon \sim \mathcal{N}(0,1)$ and $Cov(x, \varepsilon) = 0$.

Results

The performance of Hansen-Racine, Bayesian, J-test and LM for the nested models are presented in Tables 1-9 . When the process is linear, Table 1, the selection made by criteria of Hansen-Racine and Bayesian is near to 100%, in almost all situations. The behavior of J-test and LM is similar, with results that exceed 85% of correct selection in almost all cases (Table 2). The LM is slightly higher for the case of W_7.

Table 1: DGP1: Linear Process. Nested Models

Criterion		Hansen-Racine			Bayesian		
Matrices		W_4	W_5	W_7	W_4	W_5	W_7
N	R^2	% Select	% Select	% Select	% Select	% Select	% Select
$N = 100$	0.4	92.7	82.1	86.0	91.5	83.9	85.6
	0.6	99.7	97.3	98.7	99.6	98.0	98.4
	0.8	100.0	100.0	99.9	100.0	100.0	99.9
$N = 400$	0.4	100.0	98.9	99.6	100.0	99.3	99.4
	0.6	100.0	100.0	100.0	100.0	100.0	100.0
	0.8	100.0	100.0	100.0	100.0	100.0	100.0
$N = 700$	0.4	100.0	100.0	100.0	100.0	100.0	100.0
	0.6	100.0	100.0	100.0	100.0	100.0	100.0
	0.8	100.0	100.0	100.0	100.0	100.0	100.0
$N = 1000$	0.4	100.0	100.0	100.0	100.0	100.0	100.0
	0.6	100.0	100.0	100.0	100.0	100.0	100.0
	0.8	100.0	100.0	100.0	100.0	100.0	100.0

Note: % Select is the number of times that each W is selected correctly. Replications: 1000.

Table 2: DGP1: Linear Process. Nested Models

Criterion		J-test			LM		
Matrices		W_4	W_5	W_7	W_4	W_5	W_7
N	R^2	% Select	% Select	% Select	% Select	% Select	% Select
$N = 100$	0.4	71.2	53.1	55.8	89.7	73.3	60.9
	0.6	89.5	86.5	87.1	90.3	91.0	97.3
	0.8	87.4	85.6	88.1	88.0	90.5	99.9
$N = 400$	0.4	89.5	88.7	88.5	90.4	92.3	99.9
	0.6	87.8	85.3	87.7	88.7	91.5	100.0
	0.8	89.9	86.2	88.0	91.0	90.4	100.0
$N = 700$	0.4	87.4	87.4	89.2	88.8	93.7	100.0
	0.6	89.1	85.2	87.1	89.4	91.7	100.0
	0.8	91.0	86.8	87.4	91.8	92.7	100.0
$N = 1000$	0.4	88.4	86.6	89.5	89.1	92.2	100.0
	0.6	90.6	87.8	90.5	91.7	93.1	100.0
	0.8	87.8	86.2	89.6	89.1	91.9	100.0

Note: % Select is the number of times that each W is selected correctly. Replications: 1000.

When the generating process is non-linear, $DGP2$, the results are significantly altered. The Bayesian approach is the best performance with a maximum value of 89% of correct selection when the sample size is equal to 1000. The LM test tends to select subidentified matrices and due to this we observe high rates of selection for W_4. In this case, R^2's are presented only to identify the value involved in the generation of θ.

Table 3: DGP2: Non-Linear Process 1. Nested Models

Criterion		Hansen-Racine			Bayesian		
Matrices		W_4	W_5	W_7	W_4	W_5	W_7
N	R^2	% Select	% Select	% Select	% Select	% Select	% Select
$N = 100$	0.4	27.4	24.4	37.3	64.5	49.4	60.7
	0.6	26.7	24.2	34.2	75.6	61.8	68.7
	0.8	28.2	20.2	29.9	76.2	59.7	71.2
$N = 400$	0.4	30.0	25.0	41.8	80.5	69.8	75.3
	0.6	28.7	22.0	37.1	85.4	73.9	80.5
	0.8	38.0	17.6	37.3	84.1	68.9	76.9
$N = 700$	0.4	28.5	23.3	42.2	86.9	72.9	80.4
	0.6	28.8	22.0	41.0	88.7	80.0	82.8
	0.8	41.7	18.9	40.8	87.2	75.0	80.7
$N = 1000$	0.4	29.3	23.8	46.9	89.0	79.1	83.2
	0.6	33.0	22.9	40.0	88.8	82.0	85.0
	0.8	39.5	20.7	37.5	88.0	76.3	81.5

Note: % Select is the number of times that each W is selected correctly. Replications: 1000.

Table 4: DGP2: Non-Linear Process 1. Nested Models

Criterion		J-test			LM		
Matrices		W_4	W_5	W_7	W_4	W_5	W_7
N	R^2	% Select	% Select	% Select	% Select	% Select	% Select
$N = 100$	0.4	14.9	4.5	10.3	88.6	21.8	12.7
	0.6	28.4	13.2	18.3	88.1	35.2	23.1
	0.8	28.7	12.7	20.0	91.1	32.1	24.8
$N = 400$	0.4	42.3	26.8	34.4	89.5	50.4	38.7
	0.6	53.8	33.7	42.9	89.1	53.7	46.5
	0.8	48.1	28.0	35.3	91.0	50.6	39.3
$N = 700$	0.4	58.2	35.3	43.4	90.3	55.7	47.7
	0.6	61.7	45.9	53.1	90.5	63.9	58.4
	0.8	54.1	33.8	42.3	88.7	57.7	47.1
$N = 1000$	0.4	66.4	50.1	51.4	91.1	65.4	58.0
	0.6	67.6	52.2	58.8	89.2	68.5	62.2
	0.8	60.5	39.0	47.0	88.6	56.1	52.2

Note: % Select is the number of times that each W is selected correctly. Replications: 1000.

When the non-linearity is incremented, $DGP3$, there is no criterion that provides adequate information about the genuine generating process. In this case, we can observe how the LM test tends to select the matrix with the least neighbors in all cases.

Table 5: DGP3: Non-Linear Process 2. Nested Models

Criterion		Hansen-Racine			Bayesian		
Matrices		W_4	W_5	W_7	W_4	W_5	W_7
N	R^2	% Select	% Select	% Select	% Select	% Select	% Select
$N = 100$	0.4	38.8	25.2	32.7	30.9	23.5	27.2
	0.6	42.7	27.3	36.5	31.1	23.8	31.0
	0.8	42.7	31.7	44.3	34.3	22.2	29.2
$N = 400$	0.4	40.6	23.3	36.2	34.9	24.2	29.3
	0.6	40.7	26.8	38.7	30.1	23.1	28.2
	0.8	44.9	32.5	44.3	29.2	23.1	28.2
$N = 700$	0.4	39.3	25.1	33.1	35.2	22.5	32.3
	0.6	41.4	26.9	38.8	33.4	21.6	29.7
	0.8	46.7	33.1	43.2	29.9	21.1	27.6
$N = 1000$	0.4	37.7	22.6	33.9	31.5	24.4	31.9
	0.6	42.6	26.5	36.4	33.0	21.5	28.3
	0.8	43.2	30.9	42.0	33.0	22.4	29.5

Note: % Select is the number of times that each W is selected correctly. Replications: 1000.

Table 6: DGP3: Non-Linear Process 2. Nested Models

Criterion		J-test			LM		
Matrices		W_4	W_5	W_7	W_4	W_5	W_7
N	R^2	% Select	% Select	% Select	% Select	% Select	% Select
$N = 100$	0.4	0.3	0.4	0.2	89.4	5.5	0.6
	0.6	0.5	0.0	0.6	90.6	5.3	1.6
	0.8	0.4	0.0	0.5	90.0	3.3	0.9
$N = 400$	0.4	0.4	0.0	0.2	89.9	5.1	0.3
	0.6	0.3	0.1	0.3	89.1	4.3	0.9
	0.8	0.2	0.0	0.1	90.8	3.5	0.3
$N = 700$	0.4	0.4	0.2	0.4	90.7	5.1	1.0
	0.6	0.1	0.0	0.0	90.4	4.7	1.0
	0.8	0.1	0.0	0.3	90.3	4.7	0.9
$N = 1000$	0.4	0.7	0.0	0.4	88.6	4.0	1.2
	0.6	0.4	0.0	0.0	89.0	5.0	0.3
	0.8	0.4	0.1	0.3	89.7	3.2	0.7

Note: % Select is the number of times that each W is selected correctly. Replications: 1000.

The behavior of the Conditional Entropy is presented in the Tables 7-9. We apply the following rule to select the embedding dimension m: $m^2 \cdot 5 \approx N$. That is, on average, each symbol should have an expected frequency closed to 5. Therefore, we use for nested models $m = 8$ for all cases because contain W_7, W_5 and W_4. Due to this rule, the minimum sample size is 400.

For the linear process, Table 7, Entropy does not make a good selection in comparison to the other criteria.

Table 7: DGP1: Linear Process. Nested Models

Criterion		Conditional Entropy		
Matrices		W_4	W_5	W_7
N	R^2	% Select	% Select	% Select
$N = 400$	0.4	15.8	23.8	88.9
	0.6	43.5	46.6	92.4
	0.8	86.3	83.1	98.1
$N = 700$	0.4	21.5	31.2	91.4
	0.6	63.7	62.7	96.2
	0.8	96.4	94.1	99.3
$N = 1000$	0.4	27.5	34.5	91.4
	0.6	74.6	70.1	96.5
	0.8	99.4	97.5	99.6

Note: % Select is the number of times that each W is selected correctly.
Replications: 1000.

For the non-linear process 1, $DGP2$, the performance of Conditional Entropy improves and it reachs values over 90% in several cases. The behavior is similar to the Bayesian criterion, except for $R^2 = 0.4$. In the case of $DGP3$, the percentage of correct selection of the matrix is higher than the other criteria in most cases.

Table 8: DGP2: Non-Linear Process 1. Nested Models

Criterion		Conditional Entropy		
Matrices		W_4	W_5	W_7
N	R^2	% Select	% Select	% Select
$N = 400$	0.4	16.8	23.8	88.3
	0.6	45.2	46.6	92.6
	0.8	87.2	83.1	98.2
$N = 700$	0.4	23.4	31.2	88.9
	0.6	59.1	62.7	96.4
	0.8	96.6	94.1	99.8
$N = 1000$	0.4	32.8	34.5	91.4
	0.6	73.3	70.1	96.5
	0.8	98.7	97.5	99.6

Note: % Select is the number of times that each W is selected correctly.
Replications: 1000.

Table 9: DGP3: Non-Linear Process 2. Nested Models

Criterion		Conditional Entropy		
Matrices		W_4	W_5	W_7
N	R^2	% Select	% Select	% Select
$N = 400$	0.4	6.2	14.9	90.8
	0.6	15.5	27.6	92.7
	0.8	37.4	41.0	95.2
$N = 700$	0.4	11.1	20.2	91.9
	0.6	29.2	35.2	93.6
	0.8	57.6	58.2	96.8
$N = 1000$	0.4	12.2	22.9	92.6
	0.6	40.1	42.7	94.2
	0.8	76.5	72.5	98.4

Note: % Select is the number of times that each W is selected correctly.
Replications: 1000.

In the following Tables 10-15, we present the results for non nested models. In similar way as in nested models, when the process is linear, $DGP1$, the percentage of correct selection of Hansen-Racine criterion and Bayesian approach is almost of 100%. The behavior of J-test is stable at around 88% of correct selection. With regard to Conditional Entropy, its performance improves when the R^2 and sample size increases, outperforming in most cases of J-test.

Table 10: DGP1: Linear Process. Non-Nested Models

Criterion		Hansen-Racine			Bayesian		
Matrices		$W^{(1)}$	$W^{(2)}$	$W^{(1-2)}$	$W^{(1)}$	$W^{(2)}$	$W^{(1-2)}$
N	R^2	% Select	% Select	% Select	% Select	% Select	% Select
$N = 100$	0.4	99.6	99.4	99.8	99.6	99.5	99.8
	0.6	100.0	100.0	100.0	100.0	100.0	100.0
	0.8	100.0	100.0	100.0	100.0	100.0	100.0
$N = 400$	0.4	100.0	100.0	100.0	100.0	100.0	100.0
	0.6	100.0	100.0	100.0	100.0	100.0	100.0
	0.8	100.0	100.0	100.0	100.0	100.0	100.0
$N = 700$	0.4	100.0	100.0	100.0	100.0	100.0	100.0
	0.6	100.0	100.0	100.0	100.0	100.0	100.0
	0.8	100.0	100.0	100.0	100.0	100.0	100.0
$N = 1000$	0.4	100.0	100.0	100.0	100.0	100.0	100.0
	0.6	100.0	100.0	100.0	100.0	100.0	100.0
	0.8	100.0	100.0	100.0	100.0	100.0	100.0

Note: % Select is the number of times that each W is selected correctly. Replications: 1000.

Table 11: DGP1:Linear Process. Non-Nested Models

Criterion		J-test			Conditional Entropy		
Matrices		$W^{(1)}$	$W^{(2)}$	$W^{(1-2)}$	$W^{(1)}$	$W^{(2)}$	$W^{(1-2)}$
N	R^2	% Select	% Select	% Select	% Select	% Select	% Select
$N = 100$	0.4	87.1	87.4	86.7	56.2	47.6	43.2
	0.6	88.7	88.1	86.4	81.0	66.6	66.3
	0.8	88.0	89.1	86.8	95.2	91.0	91.8
$N = 400$	0.4	88.7	86.9	88.0	83.7	59.5	67.4
	0.6	86.1	85.7	85.3	99.2	96.3	95.5
	0.8	87.9	88.6	87.8	100.0	100.0	100.0
$N = 700$	0.4	86.5	87.6	87.6	94.8	77.7	80.5
	0.6	86.9	88.7	87.9	100.0	99.5	99.3
	0.8	85.8	89.1	86.6	100.0	100.0	100.0
$N = 1000$	0.4	88.1	87.9	86.8	98.5	85.8	86.0
	0.6	89.0	88.1	87.5	100.0	99.8	99.7
	0.8	87.9	87.7	87.7	100.0	100.0	100.0

Note: % Select is the number of times that each W is selected correctly. Replications: 1000.

In the first case of non-linear process, $DPG2$, the Bayesian criterion has a good performance reaching values over 90% in most cases. The J-test has a good percentage of correct selection with values over 80%, except for N equal to 100. The behavior of Conditional Entropy is correct, reaching values of 100% in many situations.

Table 12: DGP2: Non-Linear Process 1. Non-Nested Models

Criterion		Hansen-Racine			Bayesian		
Matrices		$W^{(1)}$	$W^{(2)}$	$W^{(1-2)}$	$W^{(1)}$	$W^{(2)}$	$W^{(1-2)}$
N	R^2	% Select	% Select	% Select	% Select	% Select	% Select
$N = 100$	0.4	14.6	13.4	26.9	81.8	84.8	79.8
	0.6	13.6	12.8	26.3	93.1	91.6	91.0
	0.8	21.3	22.8	27.8	95.0	94.7	95.0
$N = 400$	0.4	14.6	14.7	26.1	95.0	96.1	95.6
	0.6	19.3	17.4	27.4	97.9	98.5	97.2
	0.8	28.7	27.8	31.2	98.0	97.4	97.5
$N = 700$	0.4	16.8	13.9	26.2	97.5	97.1	97.0
	0.6	20.7	22.7	24.1	98.5	98.3	98.8
	0.8	32.8	31.7	28.8	98.3	98.3	99.9
$N = 1000$	0.4	15.5	15.5	21.9	98.8	98.5	97.9
	0.6	21.2	23.4	28.8	99.1	99.3	99.0
	0.8	35.4	34.3	33.3	99.3	99.0	99.0

Note: % Select is the number of times that each W is selected correctly. Replications: 1000.

Table 13: DGP2:Non-Linear Process 1. Non-Nested Models

Criterion		J-test			Conditional Entropy		
Matrices		$W^{(1)}$	$W^{(2)}$	$W^{(1-2)}$	$W^{(1)}$	$W^{(2)}$	$W^{(1-2)}$
N	R^2	% Select	% Select	% Select	% Select	% Select	% Select
$N = 100$	0.4	51.1	50.2	47.3	56.2	38.6	43.2
	0.6	70.9	68.0	67.5	81.0	66.7	66.3
	0.8	71.6	73.6	73.4	95.2	91.2	91.8
$N = 400$	0.4	80.0	79.9	77.7	85.4	59.5	68.1
	0.6	83.3	83.3	81.6	99.3	96.3	95.5
	0.8	84.0	85.2	83.2	100.0	100.0	100.0
$N = 700$	0.4	80.9	85.0	80.9	96.0	75.8	79.0
	0.6	84.1	86.0	83.3	100.0	99.2	98.7
	0.8	84.6	86.2	86.0	100.0	100.0	100.0
$N = 1000$	0.4	86.3	83.5	83.6	98.5	86.8	86.7
	0.6	85.2	85.9	83.9	100.0	99.8	99.7
	0.8	85.3	86.0	84.7	100.0	100.0	100.0

Note: % Select is the number of times that each W is selected correctly. Replications: 1000.

For the non-linear process 2, as happened in nested cases, the criteria of Hansen-Racine, Bayesian and J-test do not provide useful information about the genuine generating process. The Entropy criterion is clearly the best, except for small sample sizes, reaching to 100% of correct selection when the DGP uses $W^{(1)}$ and $N = 1000$.

Table 14: DGP3: Non-Linear Process 2. Non-Nested Models

Criterion		Hansen-Racine			Bayesian		
Matrices		$W^{(1)}$	$W^{(2)}$	$W^{(1-2)}$	$W^{(1)}$	$W^{(2)}$	$W^{(1-2)}$
N	R^2	% Select	% Select	% Select	% Select	% Select	% Select
$N = 100$	0.4	35.7	34.3	37.0	30.3	30.0	29.9
	0.6	40.0	39.9	41.8	27.9	31.6	27.2
	0.8	46.1	44.9	43.9	21.8	24.2	22.7
$N = 400$	0.4	34.8	32.5	36.1	26.1	27.1	28.2
	0.6	42.2	42.1	41.3	27.5	28.4	29.3
	0.8	48.7	49.0	47.6	24.1	22.8	23.7
$N = 700$	0.4	34.6	34.4	34.8	27.7	28.2	30.8
	0.6	41.5	39.6	40.9	27.4	28.0	30.7
	0.8	48.0	47.1	46.3	22.5	23.1	20.9
$N = 1000$	0.4	35.9	36.2	35.8	28.6	30.8	26.6
	0.6	43.5	41.1	41.6	27.9	27.3	27.9
	0.8	50.3	48.0	47.3	21.3	22.7	23.2

Note: % Select is the number of times that each W is selected correctly. Replications: 1000.

Table 15: DGP3: Non-Linear Process 2. Non-Nested Models

Criterion		J-test			Conditional Entropy		
Matrices		$W^{(1)}$	$W^{(2)}$	$W^{(1-2)}$	$W^{(1)}$	$W^{(2)}$	$W^{(1-2)}$
N	R^2	% Select	% Select	% Select	% Select	% Select	% Select
$N = 100$	0.4	2.9	2.6	2.1	44.2	31.7	31.0
	0.6	2.2	1.7	1.3	52.7	40.3	42.9
	0.8	1.0	1.0	0.5	62.2	51.7	52.4
$N = 400$	0.4	2.2	2.1	2.4	68.0	42.6	44.9
	0.6	1.5	1.4	1.6	84.1	70.4	72.6
	0.8	0.3	0.4	0.3	92.2	84.6	87.2
$N = 700$	0.4	1.8	2.0	2.1	79.4	57.6	63.8
	0.6	0.9	1.3	1.2	95.0	86.2	85.9
	0.8	0.4	0.2	0.3	98.2	96.3	96.2
$N = 1000$	0.4	2.1	2.2	1.5	85.6	61.3	67.2
	0.6	0.9	1.2	1.0	97.7	93.1	93.0
	0.8	0.3	1.0	0.1	100.0	99.2	98.9

Note: % Select is the number of times that each W is selected correctly. Replications: 1000.

5 Conclusions

The paper shows a collection of criteria to select the spatial weighting matrix. Our point of view is that the problem of selecting a weighting matrix is a problem of model selection. In fact, different weighting matrices result in different spatial lags of endogenous or exogenous variables included in the model. This is the direction that we explored in the present paper as an alternative way to deal with the uncertainty of specifying the spatial weighting matrix.

Generally speaking, among the different criteria that we have presented, the Bayesian criterion is the most stable under linear and weak non-linear conditions. The J-test, considered as an important tool to select spatial models, is not adequate in most situations.

Our Conditional Entropy criterion has two advantages: simplicity and good behavior under non-linear processes. In this criterion, it is not necessary any specification. The only assumption is that there is a spatial structure that links the variables under analysis. In previous revised methods we need to assume linearity, correct specification, normality in some cases, and further adequate estimation of parameters.

For future research agenda, we will explore the behavior of these criteria for spatial dynamic models and misspecified models.

References

[1] Akaike, H. (1973): Information Theory and an Extension of the Maximum Likelihood Principle. In Petrow, B. and F. Csaki (eds): *2nd International Symposium on Information Theory* (pp 267-281). Budapest: Akademiai Kiodo.

[2] Aldstadt, J. and A. Getis (2006): Using AMOEBA to Create a Spatial Weights Matrix and Identify Spatial Clusters. *Geographical Analysis* 38 327-343.

[3] Ancot, L, J. Paelinck, L. Klaassen and W Molle (1982): Topics in Regional Development Modelling. In M. Albegov, Å. Andersson and F. Snickars (eds, pp.341-359), *Regional Development Modelling in Theory and Practice*. Amsterdam: North Holland.

[4] Anselin, L. (1984): Specification Tests on the Structure of Interaction in Spatial Econometric Models. *Papers, Regional Science Association* 54 165-182.

[5] Anselin L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer.

[6] Anselin, L. (2002): Under the Hood: Issues in the Specification and Interpretation of Spatial Regression Models. *Agricultural Economics* 17 247–267.

[7] Bavaud, F. (1998): Models for Spatial Weights: a Systematic Look. *Geographical Analysis* 30 153-171.

[8] Beenstock M., Ben Zeev N. and Felsenstein D (2010): Nonparametric Estimation of the Spatial Connectivity Matrix using Spatial Panel Data. *Working Paper*, Department of Geography, Hebrew University of Jerusalem.

[9] Bhattacharjee A, Jensen-Butler C (2006): Estimation of spatial weights matrix, with an application to diffusion in housing demand. *Working Paper*, School of Economics and Finance, University of St.Andrews, UK.

[10] Bodson, P. and D. Peters (1975): Estimation of the Coefficients of a Linear Regression in the Presence of Spatial Autocorrelation: An Application to a Belgium Labor Demand Function. *Environment and Planning A* 7 455-472.

[11] Burridge, P. (2011): Improving the J test in the SARAR model by likelihood-based estimation. *Working Paper*; Department of Economics and Related Studies, University of York .

[12] Burridge, P. and Fingleton, B. (2010): Bootstrap inference in spatial econometrics: the J-test. *Spatial Economic Analysis* 5 93-119.

[13] Conley, T. and F. Molinari (2007): Spatial Correlation Robust Inference with Errors in Location or Distance. *Journal of Econometrics*, 140 76-96.

[14] Corrado, L. and B. Fingleton (2011): Where is Economics in Spatial Econometrics? Working Paper; Department of Economics, University of Strathclyde.

[15] Dacey M. (1965): A Review on Measures of Contiguity for Two and k-Color Maps. In J. Berry and D. Marble (eds.): *A Reader in Statistical Geography*. Englewood Cliffs: Prentice-Hall.

[16] Fernández E., Mayor M. and J. Rodríguez (2009): Estimating spatial autoregressive models by GME-GCE techniques. *International Regional Science Review*, 32 148-172.

[17] Folmer, H. and J. Oud (2008): How to get rid of W? A latent variable approach to modeling spatially lagged variables. *Environment and Planning A* 40 2526-2538

[18] Getis A, and J. Aldstadt (2004): Constructing the Spatial Weights Matrix Using a Local Statistic Spatial. *Geographical Analysis*, 36 90-104.

[19] Haining, R. (2003): *Spatial Data Analysis*. Cambridge: Cambridge University Press.

[20] Hansen, B. (2007): Least Squares Model Averaging. *Econometrica,* 75, 1175-1189.

[21] Hansen, B. and J. Racine (2010): Jackknife Model Averaging. *Working Paper*, Department of Economics, McMaster University

[22] Hepple, L. (1995a): Bayesian Techniques in Spatial and Network Econometrics: 1 Model Comparison and Posterior Odds. *Environment and Planning A,* 27, 447–469.

[23] Hepple, L. (1995b): Bayesian Techniques in Spatial and Network Econometrics: 2 Computational Methods and Algorithms. *Environment and Planning A,* 27, 615–644.

[24] Kelejian, H (2008): A spatial J-test for Model Specification Against a Single or a Set of Non-Nested Alternatives. *Letters in Spatial and Resource Sciences*, 1 3-11.

[25] Kooijman, S. (1976): Some Remarks on the Statistical Analysis of Grids Especially with Respect to Ecology. *Annals of Systems Research* 5.

[26] Leamer, E (1978): *Specification Searches: Ad Hoc Inference with Non Experimental Data*. New York: John Wiley and Sons, Inc.

[27] Leenders, R (2002): Modeling Social Influence through Network Autocorrelation: Constructing the Weight Matrix. *Social Networks*, 24, 21-47.

[28] Lesage, J. and K. Pace (2009): *Introduction to Spatial Econometrics*. Boca Raton: CRC Press.

[29] Lesage, J. and O. Parent (2007): Bayesian Model Averaging for Spatial Econometric Models. *Geographical Analysis,* 39, 241-267.

[30] Matilla, M. and M. Ruiz (2008): A non-parametric independence test using permutation entropy. *Journal of Econometrics*, 144, 139-155.

[31] Moran, P. (1948): The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society B* 10 243-251.

[32] Mur, J. and J Paelinck (2010): Deriving the W-matrix via p-median complete correlation analysis of residuals. *The Annals of Regional Science*, DOI: 10.1007/s00168-010-0379-3.

[33] Openshaw, S. (1977): Optimal Zoning Systems for Spatial Interaction Models. *Environment and Planning A* 9, 169-84.

[34] Ord K. (1975): Estimation Methods for Models of Spatial Interaction. *Journal of the American Statistical Association.* 70 120-126.

[35] Paci, R. and S. Usai (2009): Knowledge flows across European regions. *The Annals of Regional Science*, 43 669-690.

[36] Paelinck, J and L. Klaassen (1979): *Spatial Econometrics*. Farnborough: Saxon House

[37] Piras, G and N Lozano (2010): Spatial J-test: some Monte Carlo evidence. *Statistics and Computing,* DOI: 10.1007/s11222-010-9215-y.

[38] Tobler W. (1970): A computer movie simulating urban growth in the Detroit region. *Economic Geography,* 46 234-240.

[39] Whittle, P. (1954): On Stationary Processes in the Plane. *Biometrika*, 41 434-449.

Appendix 1. A Lagrange Multiplier for discriminating between two weighting matrices

Let us assume a spatial model, of an autoregressive type, with a normally distributed error term:

$$y = \rho W y + x\beta + \varepsilon;\ \varepsilon \sim iidN(0;\sigma^2). \tag{43}$$

We deal with the problem of choosing between two weighting matrices, one of which is nested in the other. For example, we need to decide if the ring formed by the 3 nearest neighbors is enough or we do need also the 4th nearest neighbor. The question is to decide if some weights are zero. In these case, we split the nesting weighting matrix into two matrices: $W = W_1 + W_0$. The null hypothesis is that the weights in W_0 are not relevant in the model of (43), which becomes:

$$y = \rho W_1 y + x\beta + \varepsilon;\ \varepsilon \sim iidN(0;\sigma^2). \tag{44}$$

The model of the alternative can be written as:

$$y = \rho_1 W_1 y + \rho_0 W_0 y + x\beta + \varepsilon;\ \varepsilon \sim iidN(0;\sigma^2). \tag{45}$$

According to (43) parameters ρ_0 and ρ_1 must be the same, although we maintain the unrestricted version of (45) as our testing equation. If ρ_0 is zero in this equation, W_0 is irrelevant and the weighting matrix simplifies into W_1 . We propose the following null and alternative hypothesis:

$$\left.\begin{array}{l} H_0 : \rho_0 = 0 \\ H_A : \rho_0 \neq 0 \end{array}\right\}. \tag{46}$$

Assuming normality in the error terms, the Lagrange Multiplier is the following:

$$LM_{W_0} = \left(\frac{y' W_0^{'} \hat{\varepsilon}_{W_1}}{\hat{\sigma}^2} - tr\left(B_1 W_0\right)\right)^2 \hat{\sigma}^2_{g(\rho_0)} \sim \chi^2(1), \tag{47}$$

where $\hat{\sigma}^2$ is the maximum-likelihood estimation of σ^2 obtained from the model of 45 under the null of 46; $\hat{\varepsilon}_{W_1}$ is the vector of residuals from the model of the null. B_1 is the matrix $B_1 = \left(I - \hat{\rho}_1 W_1\right)^{-1}$ where the maximum likelihood estimation of $\hat{\rho}_1$is used. The second term of the expression, $\hat{\sigma}^2_{g(\rho_0)}$, refers to the inverse of the estimated variance of the element of the score corresponding to the null hypothesis of (46), which expression is:

$\hat{\sigma}^2_{g(\rho_0)} = I^{-1}_{\rho_0\rho_0} + I^{-1}_{\rho_0\rho_0} I^{'}_{\theta\rho_0} I^{-1}_{\theta\theta_0} I_{\theta\rho_0} I^{-1}_{\rho_0\rho_0}$

- $I_{\rho_0\rho_0} = \frac{\hat{y}' W_0^{'} W_0 \hat{y}}{\hat{\sigma}^2} + tr\left(B_1^{'} W_0 + B_1^{'} W_0^{'}\right) B_1 W_0$
- $I^{-1}_{\theta\theta_0} = \left[I^{-1}_{\theta\theta} - I^{'}_{\theta\rho_0} I^{-1}_{\rho_0\rho_0} I_{\theta\rho_0}\right]$
- $I^{'}_{\theta\rho_0} = \frac{1}{\hat{\sigma}^2}\left[\begin{array}{ccc} x' W_0 \hat{y} & \hat{y}' W_0^{'} W_1 \hat{y} + \hat{\sigma}^2 tr\left(W_0 B_1 B_1^{'} W_1^{'} + B_1 W_0 B_1 W_1\right) & tr B_1 W_0 \end{array}\right]$
- $I^{-1}_{\theta\theta} = \frac{1}{\hat{\sigma}^2}\left[\begin{array}{ccc} x'x & x' W_1 \hat{y} & 0 \\ & \frac{\hat{y}' W_1^{'} W_1 \hat{y}}{\hat{\sigma}^2} + tr\left(B_1^{'} W_1 + B_1^{'} W_1^{'}\right) B_1 W_1 & tr B_1 W_1 \\ & & \frac{R}{2\hat{\sigma}^2} \end{array}\right]$

$I^{-1}_{\theta\theta_0}$ is the covariance matrix of the maximum-likelihood estimates of vector $\theta' = \begin{bmatrix} \beta & \rho_1 & \sigma^2 \end{bmatrix}'$, under the restriction of (46), ; $I^{-1}_{\theta\theta}$ is the covariance matrix of the marginal-maximum likelihood estimation of vector θ in the model of (46). $I_{\theta\rho_0}$ is the covariance vector between the maximum-likelihood estimates of the coefficients of the null model, $\theta' = \begin{bmatrix} \beta & \rho_1 & \sigma^2 \end{bmatrix}'$, and the parameter of the null hypothesis, ρ_0. Given a significance level for the test, α, the decision rule for testing the hypothesis of 46 is:

$If \quad 0 \leq LM_{W_0} \leq \chi^2_\alpha(1) \quad Do \quad not \quad reject \quad H_0,$
$If \quad LM_{W_0} > \chi^2_\alpha(1) \quad Reject \quad H_0.$