Local weighting matrices or the necessity of flexibility

Jesus Mur, Antonio Paez

University of Zaragoza. email: jmur@unizar.es

McMaster University, email: paezha@mcmaster.ca

Abstract

Local estimation is part of toolbox in current spatial econometric. Geographically Weighted Regressions are very popular algorithms useful to estimate static models in each point of the space, whereas the SALE or the Zoom approaches are solutions in the case of dynamic models. These techniques are well founded and have good properties. However, Farber and Paez (2008) detect some inconsistencies and weaknesses. The point that we want to study in this paper refers to the role of the bandwith. This measure defines how many neighbors are used in the estimation of the local parameters corresponding to each observation. The cross-validation is the most popular technique to define the bandwith, although there are other criteria that merit some consideration. On the other hand, the objective of these algorithms is to relax the restriction of global homogeneity allowing for local peculiarities. However, the definition of local neighborhood is held constant across space. This restriction can be avoided. Specifically, we discuss the procedure of specifying the sequence of local weighting matrices that will be used in the analysis. Our purpose is to develop a procedure for constructing a weighting matrix that reflects also the local surrounding of each observation. We examine two different strategies: the first is a parametric approach which involves the J test, as presented by Kelejian (2008), and the second is a nonparametric approach that uses the guidance of the symbolic entropy measures. The first part of the paper presents the overall problem, including a review of the literature; we discuss the solutions in the second part and the third part consists of a Monte Carlo simulation.

1 Introduction

The difficulties caused by the lack of stability in the parameters of an econometric model are well known: biased and inconsistent estimators, misleading tests and, in general, wrong inference. Their importance explains the attention that the literature has dedicated to the problem. The first formal test of parameter stability is that of Chow (1960), which considers only one break point, known a priori. Dufour (1982) extends the discussion to the case of multiple regimes and Phillips and Ploberger (1994) and Rossi (2005) place it in a context of model selection.

The discussion quickly took on a spatial context with the work of Casetti (1972, 1991), in which a parametric approach predominates. In fact, Casetti proposes explicitly modeling how the break in the parameters is produced through the so-called '*contextual*' variables. In the nineties, there was a great leap forwards when concern about the '*pockets of local nonstationarity*', characteristic of the literature dedicated to the LISA (Getis and Ord 1992; Anselin 1995) coincided with the development of non-parametric procedures for analyzing spatial data (McMillen 1996; McMillen and McDonald 1997). The best known approach in this line is what Brunsdon et al. (1996) call Geographically Weighted Regressions (GWR in what follows), whose immediate precursor are the Locally Weighted Regressions (LWR from now on) proposed in the seminal papers of Cleveland (1979) and Cleveland and Devlin (1988). In all these papers, interest shifts from the general to the local.

The convenience of local approaches is clear when the heterogeneity of the data is high and escapes the control of the model or when the appropriate functional form is doubtful. The GWR algorithm has also been used to correct the problems of spatial correlation that come from an inadequate treatment of the spatial heterogeneity in the data (Páez et al. 2002a, 2002b). In any case, flexible specifications are recommended.

The question that we wish to deal with in this paper continues in the same line but focusing on the need of more flexibility in the sense of what may be called a local spatial weighting matrix. That is, the GWR algorithm allows for a greater heterogeneity but maintaining constant across space the definition of neighborhood. From our point of view, this is an unnecessary restriction that can be relaxed by adjusting the definition of neighborhood to the local characteristics of each point. The problem is introduced in Section 2. Section

3 proposes some solutions which are calibrated in Section 4. Section 5 contains the main conclusions.

2 Why do we need more flexibility

Briefly, the GWR consists in estimating a given, usually linear, equation in each point of the sampling space using only local information. Let us assume that we have specified the following model:

$$y = x\beta + u;\ u \sim iidN(0, \Lambda) \tag{1}$$

where y is the (Rx1) vector of the observations of the endogenous variable, x is an (Rxk) matrix of observations of the k explanatory variables, u is a random vector of error terms not necessarily homoskedastic and, for example, normally distributed. For the moment, we assume that the specification does not include spatial interaction terms. The model of 1 has been specified under the assumption of homogeneity, which may not hold in some circumstances. As indicated by McMillen (2004, p. 232): '*spatial relationships are typically more complicated. Statistical tests based on simple functional forms often reveal that coefficients vary over space*'; in other words, a certain unobserved heterogeneity often persists in the data. The GWR solution is to introduce more flexibility by acting on the systematic part of the equation that now is estimated locally and for each sampling point:

$$\hat{\beta}_i = \left[X'W_iX\right]^{-1}\left[X'W_iy\right];\ i = 1, 2, ..., R \tag{2}$$

where

$$W_i = \begin{bmatrix} \alpha_{i1} & 0 & 0 & \dots & 0 \\ 0 & \alpha_{i2} & 0 & \dots & 0 \\ 0 & 0 & \alpha_{i3} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \alpha_{iR} \end{bmatrix} \tag{3}$$

The terms $\{\alpha_{ir}; r = 1, 2, ..., R\}$ are the local weights corresponding to the local estimation of the model 1 in point i. The local weights usually are constrained to the unit interval: $0 \leqslant \alpha_{ir} \leqslant 1$ as a way of normalizing

the influence that observation r has on the estimation of the coefficients corresponding to observation i. The local estimate of 2 can be expressed as:

$$\hat{\beta}_i = \left(\sum_{r=1}^{R} \alpha_{ir} x_r^{'} x_r \right)^{-1} \left(\sum_{r=1}^{R} \alpha_{ir} x_r^{'} y_r \right) ; \; i = 1, 2, ..., R \tag{4}$$

being x_r the (1xk) vector of observations corresponding to point r. The interesting question with equation 4 is that it clearly reflects that this is a problem of information: the points surrounding observation i do not have the same quantity of information in relation to the behaviour of the equation 1 in point i. Expressed in other terms, there is a problem of heteroskedasticity in the equation pertaining to observation i according to the sequence of weights $\{\alpha_{ir}; r = 1, 2, ..., R\}$. Obviously, the next problem is the quantification of these weights (for every point i!) which amounts to the construction of a global (RxR) weighting matrix:

$$W = \begin{bmatrix} 0 & \alpha_{12} & \alpha_{13} & \dots & \alpha_{1R} \\ \alpha_{21} & 0 & \alpha_{23} & \dots & \alpha_{2R} \\ \alpha_{31} & \alpha_{32} & 0 & \dots & \alpha_{3R} \\ \dots & \dots & \dots & \dots & \dots \\ \alpha_{R1} & \alpha_{R2} & \alpha_{R3} & \dots & 0 \end{bmatrix} = \begin{bmatrix} \alpha_{1.} \\ \alpha_{2.} \\ \alpha_{3.} \\ \dots \\ \alpha_{R.} \end{bmatrix} ; \; W_i = diag\left(\alpha_{i.}\right) \tag{5}$$

It is clear that the GWR estimates will be unbiased only if the assumption of global homogeneity, implicit in 1, is true. In fact, the GWR algorithm will produce a sequence of locally weighted least squares estimates, unbiased in every point of the sampling space and optimal in the sense that the variance of these estimates will be minimal. Obviously, if the instability in the non-random component of equation of 1 is serious, the GWR algorithm (as said, a type of feasible generalized least squares estimate) will of little help.

By convention, it is assumed that the influence of the information of point i in the local estimation of the slopes corresponding to the same point i is zero (the main diagonal of matrix W is zero). The reason, the same as with the construction of the spatial weight matrix used in spatial models, is to assure the identification of the model. The literature on GWR suggest specifying the $\alpha's$ weigths according to some simple function of the distance; for example:

$$\left.\begin{array}{c} \bullet\alpha_{ir} = \left\{ \begin{array}{c} 1\ d_{ir} < d \\ 0\ d_{ir} \geqslant d \end{array} \right. \\ \bullet\alpha_{ir} = exp\left[-\frac{1}{2}\left(\frac{d_{ir}}{d}\right)^2\right] \\ \bullet\alpha_{ir} = \left\{ \begin{array}{c} \left[1-\left(\frac{d_{ir}}{d}\right)^2\right]^2\ d_{ir} < d \\ 0\ d_{ir} \geqslant d \end{array} \right. \end{array}\right\} \tag{6}$$

where d is the *bandwidth* of the algorithm. One of the main problems of this methodology is related to the determination of the optimal value for the bandwidth. There are several alternatives altough the most popular is the so-called *cross-validation approach*, which amounts to choosing the value of d that minimizes the (mean square) prediction error of the GWR estimation.

3 Criteria to define the 'local neighbors'.

$$y = \Gamma y + x\beta + \varepsilon \tag{7}$$

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. A solution, perhaps the most popular, consists of introducing some structure in the matrix Γ, parametrizing the spatial interaction coefficients as, for example: $\Gamma = \rho W$, ρ is a parameter and W a matrix of weights. The term $y_W = Wy$ that, consequently, appears on the right hand side (rhs, form now on) of the equation is called the spatial lag of the endogenous variable. At this point it is worth to highlight a couple of questions:

(i) The weighting matrix can be constructed in different ways following, for example, some interaction hypothesis. Each hypothesis will result in a different weighting matrix leading to a different spatial lag. In sum, different weighting matrices amounts to different models.

(ii) There are some general guidelines about how to specify a weighting matrix using concepts like nearness, accessibility, influence, etc. Different models might require different interaction channels that are not necessarily known. This implies uncertainty and diffuse priors.

Corrado and Fingleton (2011) discuss the construction of a weighting matrix based on theoretical considerations (they wonder 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. As said, 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), a Lagrange Multiplier can be used.

3.1 The J test

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

$$\left.\begin{array}{l} H_0 : y = \rho_1 W_1 y + x_1 \beta_1 + \varepsilon_1 \\ H_A : y = \rho_2 W_2 y + x_2 \beta_2 + \varepsilon_2 \end{array}\right\} \tag{8}$$

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

$$y = (1-\alpha)\left[\rho_1 W_1 y + x_1 \beta_1\right] + \alpha\left[\hat{\rho}_2 W_2 y + x_2 \hat{\beta}_2\right] + \nu \tag{9}$$

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

$$\begin{aligned} y &= \rho_i W_i y + x_i \beta_i + u_i = Z_i \gamma_i + u_i \\ u_i &= \lambda_i M_i u_i + v_i \end{aligned} \tag{10}$$

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 $W_i \neq W_j$. Let us assume that there are two alternatives, of a $SARAR(1,1)$ type. The subindex 0 indicates that it is the model of the null hypothesis:

$$\begin{aligned} y &= X_0 \beta_0 + \lambda_0 W_0 y + u_0 \\ u_0 &= \rho_0 M_0 u_0 + \varepsilon_0 \end{aligned} \tag{11}$$

where y denotes the $R \times 1$ vector of observations of the dependent variable, X_0 denotes the $R \times k$ matrix of regressors (in our case it could contain a single constant term). Both variables, X_0 and y, have been measured without error. W_0 and M_0 are $R \times R$ spatial weighting matrices defined *a priori*, β_0 is a $k \times 1$ vector of unknown parameters, λ_0 and ρ_0 are unknown scalar parameters, u_0 denotes the $R \times 1$ vector of errors terms and ε_0 is an $R \times 1$ vector of innovations, assuming that $\varepsilon_0 \sim i.i.d.\left(0, \sigma^2 I_R\right)$. This is called $Model_0$.

Under the alternative hypothesis, the data-generating process has a similar structure, $Model_1$:

$$\begin{aligned} y &= X_1 \beta_1 + \lambda_1 W_1 y + u_1 \\ u_1 &= \rho_1 M_1 u_1 + \varepsilon_1 \end{aligned} \tag{12}$$

Premultiplying $Model_0$ by $(I_R - \rho_0 M_0)$ yields:

$$y_0(\rho) = Z_0(\rho)\gamma + \varepsilon_0 \tag{13}$$

where $y_0(\rho) = (I_R - \rho_0 M_0) y$, $Z_0(\rho) = (I_R - \rho_0 M_0) Z_0$, with $Z_0 = (X_0 \beta_0, \rho_0 W_0 y)$ and $\gamma' = \left(\beta', \lambda\right)$. The same transformation can be applied to $Model_1$.

In this context, the J-test can been seen as the test of the following augmented equation:

$$y_0(\rho) = Z_0(\rho)\gamma + \phi[Z_1(\rho_1)\hat{\gamma}_1] + \varepsilon_0 \tag{14}$$

where $\hat{\gamma}_1$ represents a consistent estimator of γ_1 and ϕ is a parameter whose value, under the null hypothesis, is $\phi = 0$.

The parameters to be estimated are, for $Model_0$, β_0, λ_0, ρ_0, σ_0^2 and, for $Model_1$, β_1, λ_1, ρ_1 and the variance σ_1^2. These coefficients can be obtained by the generalized method of moments, GMM, suggested by Kelejian and Prucha (1999) or by the recent quasi-maximum likelihood method, QML, proposed by Burridge and Fingleton (2010). Below we present briefly the GMM procedure of Kelejian and Prucha.

As the model (14) contains a spatial lag of the dependent variable, the estimation method proposed is based on instrumental variables. Let the list of instruments be:

$$\begin{aligned}
T_0 &= (X_0, W_0X_0, \ldots, W_0^rX_0, M_0W_0X_0, \ldots, M_0W_0^rX_0)_{LI} \\
T_1 &= (X_1, W_1X_1, \ldots, W_1^rX_1, M_1W_1X_1, \ldots, M_1W_1^rX_1)_{LI} \\
\bar{T} &= \left(\bar{X}, W\bar{X}, \ldots, W^r\bar{X}, MW\bar{X}, \ldots, MW^r\bar{X}\right)_{LI}
\end{aligned}$$

where $\bar{X} = (X_0, X_1)$, subindex LI indicates that the columns of the corresponding matrices are linearly independent; typically, $r \leq 2$. Kelejian suggests the following procedure:

1. Estimate the null hypothesis model of (11) by two-stage least squares, $2SLS$, using the matrix of instruments T_0; we obtain the residual vector $\hat{u}_0$. Repeat this procedure for the alternative model (12) by $2SLS$, using the matrix of instruments T_1.

2. Take $\hat{\gamma}_1$ appearing in (14) as the $2SLS$ estimator based on matrix T_1 for the alternative model.

3. Using the estimated residuals of null model, $\hat{u}_0$, estimate the parameter ρ_0 by the generalized moments procedure, GMM, proposed by Kelejian and Prucha (1998). Replace ρ_0 with $\hat{\rho}_0$ and estimate the resulting model by $2SLS$ using instrument matrix T_0. Obtain the residual vector, $\hat{\varepsilon}$, and use this vector to estimate the corresponding variance: $\hat{\sigma}_\varepsilon^2 = {}^{\hat{\varepsilon}'\hat{\varepsilon}}/_R$. This is the generalized spatial two-stage least squares procedure.

4. Replace ρ in (14) by $\hat{\rho}_0$. Considering $F = (Z_1\hat{\gamma}_1)$ as the empirical counterpart to (14) let

$$y_0(\hat{\rho}) \approx Z_0(\hat{\rho}_0)\gamma + \phi F + \varepsilon_0 \tag{15}$$

5. Estimate (15) by $2SLS$ using $\bar{T}$ as instruments. Specifically, the set of regressors of (15) is denoted by $S = (Z_0(\hat{\rho}), F)$, and the regression parameters as $\eta' = \left(\gamma', \phi\right)$. Note that, under the null hypothesis model, $\eta_0' = \left(\gamma', 0\right)$. Let $\hat{S} = PS \equiv \left(\hat{Z}_0(\hat{\rho}), \hat{F}\right)$ where $P = \bar{T}\left(\bar{T}'\bar{T}\right)^{-1}\bar{T}'$, so the $2SLS$ estimator of η is: $\hat{\eta} = \left(\hat{S}'\hat{S}\right)^{-1}\hat{S}'y_0(\hat{\rho})$.

Kelejian (2008) shows that

$$\begin{aligned} R^{1/2}(\hat{\eta} - \eta) &\xrightarrow{D} \mathcal{N}\left[0, \sigma_\varepsilon^2 \underset{R\to\infty}{plim}\left(\frac{\hat{S}'\hat{S}}{R}\right)^{-1}\right] \\ \hat{\sigma}_\varepsilon^2 &\xrightarrow{P} \sigma_\varepsilon^2 \end{aligned} \tag{16}$$

Clearly, for finite samples the inference can be based on an approximation such as:

$$\hat{\eta} \approx \mathcal{N}\left[\eta, \hat{\sigma}_\varepsilon^2\left(\hat{S}'\hat{S}\right)^{-1}\right] \tag{17}$$

Let $\bar{k} = k + 2$; $\hat{\eta}' = \left(\hat{\gamma}', \hat{\phi}\right)$; $\hat{V}_{\hat{\phi}}$ be the estimated variance corresponding to $\hat{\phi}$, wich appears in the $(k+2)\times(k+2)$ entry of the $\bar{k}\times\bar{k}$ matrix (17), $\hat{\sigma}_\varepsilon^2\left(\hat{S}'\hat{S}\right)^{-1}$. Then, a Wald test of H_0: $\phi = 0$ against H_1: $\phi \neq 0$, at the α% level of significance would be to reject H_0 if

$$\hat{\phi}'\hat{V}_{\hat{\phi}}^{-1}\hat{\phi} > \chi_{1-\alpha}^2(1) \tag{18}$$

As an alternative to the asymptotic distribution, Burridge and Fingleton (2010) suggest a bootstrap procedure with better properties for finite samples. As a generalization of this procedure, Kelejian proposes a limited number, $g \geq 1$, of alternatives of the same type, in which $Model_0$ is not nested.

The J-test works reasonably well for finite samples, although it involves some problems of power, especially when the rival matrices are very close. For further details, see Burridge and Fingleton (2010).

We shall remember that our objective is to select the most informative weighting matrix assuming dependence between x and y. In short, the problem of interest is: $X_0 = X_1 = x$, $\rho_0 = \rho_1 = 0$, but $W_0 \neq W_1$. In other words, there are two models with same explanatory variable, and no spatial autocorrelation in the respective error terms, $Model_0$ and $Model_1$; but the weighting matrices differ. The resulting specification is as follows:

$$y = \beta_j x + \lambda_j W_j x + u_j, \quad j = 0, 1 \tag{19}$$

This expression is a reformulation of (11), considering $W_j y = W_j x$.

In sum, it is worth to highlight that there are few alternative procedures for selecting the correct weighting matrix. This procedure of the J-test aims to determine the spatial setting on which the rest of the analysis is based.

Given the importance of this decision, the non-parametric chapter presents an alternative procedure to compete with the J-test.

3.2 The entropy criterion

The purpose of this section is to present a new non-parametric procedure for selecting a weighting matrix. The selection criterion is based on the information content existing in the Space for the relation we are working with; this relation may be, or not, of a causal type. The measure of information that we use is based on a reformulation of the traditional entropy indices in terms of what is called *symbolic entropy*, and it does not depend on the priors of the practitioner.

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

Beginning with the symbolization process, assume 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. Denote by $\Gamma_n = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ 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

$$f : \{x_s\}_{s \in S} \to \Gamma_l \tag{20}$$

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 process can be followed for the series y_s.

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

$$Z_s = \{x_s, y_s\} \tag{21}$$

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 of the form $\eta_{ij} = \left(\sigma_i^x, \sigma_j^y\right)$. The symbolization function of the bivariate process would be

$$g : \{Z_s\}_{s \in S} \to \Omega_l^2 = \Gamma_l \times \Gamma_l \tag{22}$$

defined by

$$g(Z_s = (x_s, y_s)) = (f(x_s), f(y_s)) = \eta_{ij} = \left(\sigma_i^x, \sigma_j^y\right) \tag{23}$$

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 the following symbolization function f . Let M_e^x be the median of the univariate spatial process $\{x_s\}_{s \in S}$ and define the indicator function

$$\tau_s = \begin{cases} 1 & if \quad x_s \geq M_e^x \\ 0 & otherwise \end{cases} \tag{24}$$

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

$$\iota_{ss_i} = \begin{cases} 0 & if \quad \tau_s \neq \tau_{s_i} \\ 1 & otherwise \end{cases} \tag{25}$$

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

$$f(x_s) = \sum_{i=1}^{m-1} \iota_{ss_i} \tag{26}$$

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

Moreover, we need to 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: $h(x|y) = -\sum_x \sum_y p(x, y)\, ln\,(p(x|y))$.

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

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

The absolute frequency of symbol σ^x_i is:

$$n_{\sigma^x_i} = \#\{s \in S | s \quad is \quad \sigma^x_i - type \quad for \quad x\} \tag{27}$$

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

$$n_{\sigma^y_j} = \#\left\{s \in S | s \quad is \quad \sigma^y_j - type \quad for \quad y\right\} \tag{28}$$

Next, the relative frequencies can also be estimated:

$$p(\sigma^x_i) \equiv p_{\sigma^x_i} = \frac{\#\{s \in S | s \quad is \quad \sigma^x_i - type \quad for \quad x\}}{|S|} = \frac{n_{\sigma^x_i}}{|S|} \tag{29}$$

$$p\left(\sigma^y_j\right) \equiv p_{\sigma^y_j} = \frac{\#\left\{s \in S | s \quad is \quad \sigma^y_j - type \quad for \quad y\right\}}{|S|} = \frac{n_{\sigma^y_j}}{|S|} \tag{30}$$

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

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

$$p(\eta_{ij}) \equiv p_{\eta_{ij}} = \frac{\#\{s \in S | s \quad is \quad \eta_{ij} - type\}}{|S|} = \frac{n_{\eta_{ij}}}{|S|} \tag{31}$$

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

$$h_Z(m) = -\sum_{\eta \in \Omega_m^2} p(\eta) ln(p(\eta)) \tag{32}$$

We can obtain the marginal symbolic entropies as

$$h_x(m) = -\sum_{\sigma^x \in \Gamma_m} p(\sigma^x) ln(p(\sigma^x)) \tag{33}$$

$$h_y(m) = -\sum_{\sigma^y \in \Gamma_m} p(\sigma^y) ln(p(\sigma^y)) \tag{34}$$

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

$$h_{y|\sigma^x}(m) = -\sum_{\sigma^y \in \Gamma_m} p(\sigma^y|\sigma^x) ln(p(\sigma^y|\sigma^x)) \tag{35}$$

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

$$h_{y|x}(m) = \sum_{\sigma^x \in \Gamma_m} p(\sigma^x) h_{y|\sigma^x}(m) \tag{36}$$

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

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

$$\mathcal{K}_W^x = \{\sigma^x \in \mathcal{K} | \sigma^x \text{ is admissible for } Wx\}. \tag{37}$$

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

By Γ_m^x we denote the set of symbols that 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

$$\begin{aligned} h_{W_0 x|y}(m) &= -\sum_{\sigma^y \in \Gamma^y} p(\sigma^y) \left[\sum_{\sigma^x \in \mathcal{K}_{W_o}^x} p(\sigma^x|\sigma^y) \ln\left(p(\sigma^x|\sigma^y)\right) \right] \leq \qquad (38) \\ &\leq -\sum_{\sigma^y \in \Gamma^y} p_{\sigma^y} \left[\sum_{\sigma^x \in K_W^{*x}} p(\sigma^x|\sigma^y) \ln\left(p(\sigma^x|\sigma^y)\right) \right] = h_{Wx|y}(m) \end{aligned}$$

We have thus 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*

$$h_{W_0 x|y}(m) = \min_{W \in \mathcal{W}(x,y)} \left\{ h_{Wx|y}(m) \right\}. \qquad (39)$$

4 Monte Carlo Experiment

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

Our principal interest is to detect the weighting matrix more informative between different alternatives. For this, we having the explanatory variable, x, the same in the all models, but the spatial structures differ, so that $W_0 = W_i$, where i is the matrix for the $i-th$ alternative model.

A great variety of alternative of weighting matrices are possible for our study, however we restrict our attention to k-nearest neighbors and weights distance-based. Also, we can work with different models: Spatial autoregressive process (SAR) or spatial error model (SEM) or SARAR(p,q).

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, subsequently normalized so that rows sum to 1. For the second case, distance-based weigth, W is constructed using $w_{ij} = d_{ij}^{-2}$ for $d_{ij} < D$, where D is a cut-distance, and $d_{ij} = 0$ otherwise, denoting d_{ij}as the straigh-line (Euclidean) distance between regions i ans j.

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

$$N \in \{100, 300, 600, 1000\}\,,\ \rho \in \{0.1; 0.3; 0.5; 0.7; 0.9\}\,,\ m \in \{4, 5, 6, 7, 8\} \quad (40)$$

where N is the sample size, ρ is the spatial autocorrelation parameter and m is usually known as the *embedding dimension*. Briefly, the latter corresponds to the set made by each observation and its $m - 1$ neighbours.

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

In the first case, linearity, we control the relation by, for instance, the coefficient of determination expected from the equation. Based on a specification like this:

$$y = \beta x + \theta W x + \varepsilon, \quad (41)$$

the strength of the relation can be deduced by the expected $R^2_{y/x}$ coefficient.

Under equation (41), 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:

$$R^2_{y/x} \in \{0.3; 0.5; 0.7; 0.9\} \quad (42)$$

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

[TO BE COMPLETED]

5 Conclusions

[TO BE COMPLETED]

References

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

[2] 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.

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

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

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

[5] 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.

[6] 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.

[7] Bodson, P. and D. Peters (1975): Estimation of the Coeffcients 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.

[29] 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 .

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

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

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

[9] 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.

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

[11] 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

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

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

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

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

[27] 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.

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

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

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

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

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

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

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

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

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

[18] 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.

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

[20] 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.

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

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

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

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