Evaluating three proposals for testing independence in non linear spatial processes*

Fernando A. López-Hernández
Departamento de Métodos Cuantitativos e Informáticos
Universidad Politécnica de Cartagena

M. Luz Maté-Sánchez-Val
Departamento de Economía Financiera y Contabilidad
Universidad Politécnica de Cartagena

Andrés Artal-Tur
Departamento de Economía
Universidad Politécnica de Cartagena

Abstract

This paper evaluates the behaviour of different families of tests when checking for spatial independence in the presence of nonlinearities. To reach this goal, we select three representative proposals. The usual parametric tests of I-Moran, the nonparametric proposal of Brett and Pinkse (1997), and the semiparametric Scan test. In order to study how they perform, we simulate different nonlinear spatial structures by Monte Carlo methods, hence conducting empirical tests on the matter. Main results show failures of traditional tests in this framework, and the need to build on new proposals in the presence of nonlinearities. An empirical application to an economic-theory-of-production scenario illustrates the performance of the three tests.

Keywords: Nonlinear spatial processes, spatial dependence tests, Monte Carlo simulations, R&D.

JEL classification: C-14, C-63, O-32, R-12

AMS classification: 62M30, 62H11

* The authors would like to express their thanks to the projects 11897/PHCS/09 of the Fundación Seneca, Comunidad Autónoma de Murcia. Fernando A. López acknowledges financial support from ECEO2009-10534 of the Ministerio de Ciencia e Innovación del Reino de España, Mariluz Maté acknowledges financial support from the Spanish Ministry of Science and Innovation (Project ECO2011-29080) and Andrés Artal-Tur acknowledges financial support from the Spanish Ministry of Science and Innovation (Project MINECO ECO2011-27169) cofinanciado con Fondos FEDER.

Evaluando tres propuestas para contrastar independencia en procesos espaciales no lineales

Resumen

Este artículo evalúa el comportamiento de tres estadísticos utilizados para contrastar la hipótesis de independencia de procesos espaciales cuando subyace una estructura no lineal en los datos: el clásico test paramétrico de Moran, la propuesta no paramétrica de Brett y Pinkse y el test semiparamétrico Scan. Para comparar el comportamiento de estos contrastes se realiza un extenso ejercicio de Montecarlo en el que se proponen diversas estructuras de dependencia espacial no lineal. Los resultados obtenidos señalan la necesidad de aplicar los nuevos contrastes en entornos no lineales, dado que los tradicionales suelen fallar en su detección. Una aplicación a la función de producción empresarial permite ilustrar esta cuestión.

Palabras Clave: Procesos espaciales no lineales, contrastes de independencia espacial, simulaciones de Monte Carlo, I+D.

Clasificación JEL: C-14, C-63, O-32, R-12

Clasificación AMS: 62M30, 62H11

1. Introduction

Literature on spatial econometrics, which builds on the importance of including spatial dimension when explaining real world phenomena, has been largely confined to linear models (Anselin and Bera, 1998). In fact, spatial modelling of the type of linear regressions with spatial interdependence still dominates literature (Anselin, 2010). Nevertheless, as yet noted by Kovach (1960), life can be (so) nonlinear. Therefore, careful modelling requires nonlinear specifications, given that “ignoring the potential nonlinear relationships in spatial dependence models often results in inconsistent estimations of the parameters of interest and misleading conclusions” (Su and Jin, 2010, p. 18). This surely explains the increasing attention to the issue paid by the literature[1]. The availability of new softwares and micro-databases has also led to a boost in empirical applications (de Graaff et al, 2001). In general, the introduction of nonlinearities in spatial analysis opened new ways for modelling, improving our understanding of socio-economic systems.

In certain ways, such a development reminds the process followed by traditional econometrics some years ago, for example in time-series models[2]. Original efforts for introducing nonlinear interdependences in spatial econometrics took the form of nonlinear distance decay functions (Dubin, 1988), as well as flexible functional specifications capturing some types of nonlinear relations between variables (Yang et

[1] See, for example, the 2010 monographic work on that issue in *Journal of Econometrics,* vol 157.

[2] See i.e. *Journal of Econometrics*, vol 157, particularly the contribution of Su and Jin, 2010; see also Pede et al, 2008; Yatchew, 1998.

al., 2006; Pace et al., 2004; Baltagi and Li, 2001; van Gastel and Paelinck, 1995). Most of these papers introduce a parametric transformation (e.g., Box-Cox one) on the response variable and/or on the regressors. Recent approaches include the use of transition processes defined as gradual regime-switching structures in the form of smooth autoregressive models (STAR). The STAR framework allows the parameters of the model to take on different values across regimes in a smooth transitional process, with spatial correlation taken into account through the use of a weight matrix suited to identify the topology of the entire spatial system (Pede et al., 2008). The contributions of Mur et al. (2009) and López et al. (2009) follow the same line of reasoning, but introducing instability in the spatial dependency mechanisms.

From a methodological perspective, researchers are increasingly exploring nonparametric and semiparametric methods in order to identify the presence of spatial structures under nonlinearities (de Graaff et al., 2001; Yatchew, 1998). Their main advantage is that they allow data determining the functional form of the model, with no *a priorism* (Lu, 2009). Following this line of research, the present paper is intended to compare the power of the three usual families of tests for spatial dependence effects: parametric, nonparametric and semiparametric. To reach this goal, we select three representative specifications of every family. Firstly, we analyse one of the most well-known parametric tests, the I-Moran test (Moran, 1950), which is widely employed in the first stages of many exploratory and spatial econometrics studies. Secondly, we select the nonparametric proposal of Brett and Pinkse (1997), the BP test. In comparison with the Moran Index, this test is scarcely present in the literature, probably due to the lack of software. Finally, we analyse the behaviour of a semiparametric test developed in epidemiology literature but not very well-kwon in a spatial econometrics context, namely the Scan test (Kulldorff et al., 2009). This test belongs to the family of scan windows tests and, although it is usually employed with the aim of identifying regional clusters of different behaviour, it can also be used as a test of independence. This test has been incorporated into the freely available SaTScan software (http://www.satscan.org).

To establish a comparison among these proposals for identifying spatial dependence patterns, we employ Monte Carlo methods. After that, we develop an empirical exercise where evaluating the spatial effect in a production function with internal R&D in two Spanish municipalities (Madrid and Barcelona). Anticipating some of our findings, in the Monte Carlo exercise we observe that under nonlinear spatial structures the semiparametric and nonparametric tests show much more power than the classical I-Moran test. Moreover, the I-Moran test fails to capture the presence of spatial dependence for nonlinear models. In the case of our empirical application, we found that the typical linear spatial structure (spatial error model) is not enough to capture all spatial effects present in the data, also recommends employing new proposed tests.

After this introduction, the remainder of the paper is organised as follows. Section 2 presents the three tests to be compared, including some comments about its computation. In Section 3, Monte Carlo simulations are carried out in order to compare the power and main characteristics of the three families of tests. Section 4 includes an

ESTADÍSTICA ESPAÑOLA Vol. 55. Núm. 180 / 2013

application, where testing the power of the three families of tests is completed. Finally, Section 5 gives the conclusion.

2. Three proposals for testing independence in nonlinear spatial processes

This section introduces the previously mentioned spatial dependence tests, indicating the main advantages and disadvantages of their use.

The parametric Moran's I test (Moran, 1950)

Moran's I test is an extension of Pearson product moment correlation coefficient to a univariate series. In this sense, Moran's I tests if the values of a variable x, measured at different locations (x_i and x_j with $i, j=1,2,...,n$ and $i \neq j$), are associated. Formally, Moran´s I test follows the expression [1] which, after standardization, follows asymptotically a normal distribution:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \bar{x}) w_{ij} (x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad [1]$$

where $\bar{x}$ is the sample mean for the variable x, w_{ij} is the (i,j)-element of a known weight matrix (w) which quantifies the different intensities among spatial locations in function of their proximity. Finally, S_0 is the sum of all w elements and n is the number of observations.

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \qquad [2]$$

Moran's I is not, strictly, a test of independence but a no-correlation test between the values of a variable in the different locations and its spatial surroundings. This test has good properties in comparison with other spatial dependence tests (Florax and de Graaff, 2004), which explains its popularity in spatial econometrics literature. Despite its good properties, the Moran's I test has some limitations that should be taken into account when we are testing spatial correlation. Among them, Moran's I test requires that the weight matrix is correctly specified in relation to the alternative hypothesis (Cliff and Ord, 1981). Besides that, the Moran test could fail to detect spatial dependence under nonlinearities (López et al. 2010).

Non-parametric spatial dependence test: τ test of Brett and Pinkse (Brett and Pinkse, 1997)

This is a nonparametric test built taking into consideration the properties of the characteristic functions. Specifically, it is based on the property that if two variables (in our case, x and his spatial lag wx) are independent, the joint characteristic function must factorize into the product of their marginal characteristic functions. To compute the test, an f practitioner-chosen density function with infinite support is considered, with

$h(x) = \int e^{iux} f(u)du$ its Fourier transform. Let $\{y_t\}$ and $\{z_t\}$ be independent copies of the process $\{x_t\}$ and let be the average of proximate observations of x_t. We also need to define, $h_{ts} = h(x_t - x_s)$, $h_{ts}^{NN} = h(x_t^N - x_s^N)$ and $\eta_{n1} = n^{-2}\sum_{s,t} h_{ts}h_{ts}^{NN}$; $\eta_{n2} = n^{-3}\sum_{s,t,u} h_{ts}h_{tu}^{NN}$; $\eta_{n3} = n^{-4}\sum_{s,t,u,v} h_{ts}h_{uv}^{NN}$ with η the number of observations. Let

$$\eta_n = (\eta_{n1} - \eta_{n2})^2 (\eta_{n2} - \eta_{n3})^2 \qquad [3]$$

and

$$\upsilon_n = (\gamma_n - \mu_n^2)^2 n^{-1} \sum_t n_t^{-1} \left(I(n_t > 0) + \sum_s n_s^{-1} \left(I(s \in N_1) I(t \in N_s) \right) \right) \qquad [4]$$

where, $\mu_n = n^{-2}\sum_{t,s} h_{ts}$, $\gamma_n = n^{-3}\sum_{t,s,u} h_{ts}h_{tu}$ N_t the set of proximate observations of point t and n_t, that is, the cardinal of set N_t.

Then, under the null of independence, the Brett and Pinkse statistic

$$\tau = \frac{n\eta_n}{2\upsilon_n} \qquad [5]$$

is asymptotically χ_1^2 distributed.

As in the Moran I test, the BP test considers, under the null hypothesis, some knowledge about the spatial dependence structure. Among the different advantages of the BP test is that it is well-behaved for different sample sizes, even in the case of reduced sample sizes. On the negative side, the BP test could fail when the process is non-stationary and it is sensitive to the scaling of the observations. Additionally, the BP tends to be undersized if the variable does not follow a normal distribution. The Brett and Pinske test needs some information about the spatial dependence structure to be computed. In this sense, at least the neighbourhood connections are required. Once, they are defined, the BP test is quite powerful and its power does not vary even with reduced size samples (Pinkse, 1998). This test is a good candidate to detect non linear spatial dependence structures.

Semiparametric spatial dependence Scan test (Kulldorff et al., 2009)

The Scan Test is a classic technique in the field of epidemiology (Kulldorff and Nagarwalla, 1995). It has been used to identify high incidence clusters of rare sicknesses, mainly cancer, and developed under the assumption that the underlying process follows a discrete distribution. Nevertheless, recent applications of the test are carried out assuming normality. This proposal allows for its use in the field of economics and regional science.

Basically, the test identifies regions of clusters with a different behaviour to the others. To test this hypothesis, central windows of different size and shape are put into each region (circular, elliptical or flexible), comparing the mean value of the observations

inside the window with that which are outside. The window moves across the entire map changing the size and shape to identify the maximum differential between the values inside and outside. Once the window where the maximum difference is identified, it is evaluated by checking if the difference is significant.

The Scan test is a scan statistic in which the null hypothesis is that all the observations are independent and the variable proceeds from a given distribution. The alternative hypothesis states that there is one cluster where the observations have either a larger or a smaller mean than outside that cluster.

In this paper, we present the circular version of the statistic but it can very easily be extended to non-circular windows. The circular spatial scan statistic is defined through a huge number of overlapping circles. For each circle, *z*, a log likelihood ratio *LLR(z)* is calculated, and the test statistic is defined as the maximum *LLR* over all circles. The scanning window will depend on the application, but it is typical to define the window as a circle centred on an observation with a radius varying continuously from zero up to some upper limit. Circles with only one observation are ignored. Let $n_z = \sum_{s \in z} n_s$ be the number of observations in circle *z*, and let $x_z = \sum_{s \in z} x_s$ be the sum of the observed values in circle *z*.

Under the null hypothesis, the maximum likelihood estimates of the mean and variance are $\mu = \frac{1}{n}\sum_i x_i$ and $\sigma^2 = \frac{\sum_i (x_i - \mu)^2}{n}$ respectively. The likelihood under the null hypothesis is then

$$L_0 = \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \qquad [6]$$

and the log likelihood is

$$\ln L_0 = -n\ln\left(\sqrt{2\pi}\right) - n\ln(\sigma) - \sum_i \frac{(x_i-\mu)^2}{2\sigma^2} \qquad [7]$$

Under the alternative hypothesis, we first calculate the maximum likelihood estimators that are specific to each circle *z*, which is $\mu_2 = x_2 / n_2$ for the mean inside the circle and $\lambda_z = (X - x_z)/(n - n_z)$ for the mean outside the circle. The maximum likelihood estimate for the common variance is

$$\sigma_z^2 = \frac{1}{n}\left(\sum_{i \in z} x_i^2 - 2x_z\mu_z + n_z\mu_z^2 + \sum_{i \notin z} x_i^2 - 2(X - x_z)\lambda_z + (n - n_2)\lambda_z^2\right) \qquad [8]$$

The log likelihood for the circle z is

$$\ln L(z) = -n\ln\left(\sqrt{2\pi}\right) - n\ln\left(\sqrt{\sigma_z^2}\right) - n/2 \qquad [9]$$

As the statistic test we use the maximum likelihood ratio

$$\max_z \left(L_z / L_0\right) \qquad [10]$$

or more conveniently the maximum log likelihood ratio

$$Scan = \max_z \left(\ln L_z - \ln L_0\right) = \max_z \left(n\ln(\sigma) + \sum_i -\frac{(x_i-\mu)}{2\sigma^2} - n/2 + n\ln\left(\sqrt{\sigma_z^2}\right) \right) \qquad [11]$$

Only the last term depends on z, so from this formula it can be seen that the most likely cluster selected is the one that minimizes the variance under the alternative hypothesis, which is intuitive. The significance of this statistic is obtained through re-sampling techniques.

The Scan test *scans* the surface, looking for the shape and size that maximises the difference. The other tests of independency introduce a close structure in the null hypothesis that should be specified previously. In this way, the most important difference in comparison with the I-Moran and BP tests is the Scan test does not need the spatial information from the weight matrix. On the negative side, Scan test assumes the null hypothesis of *iid*, following a normal distribution with the same mean value. This is a restrictive assumption. Therefore, the rejection of the null hypothesis could be motivated by the existence of spatial dependence structure or by the existence of outlier observations with different average values. To discriminate between these results further analysis is required, for example, the study of the outliers.

3. The finite sample behaviour of spatial dependence tests under nonlinear processes

In this section, we analyse the finite sample behaviour of three spatial dependence tests under a non linear framework. To attain this purpose: (1) we generate different spatial processes through linear and nonlinear spatial dependence mechanisms and (2) examine the power of the tests in this scenario.

In order to generate non linear spatial processes, we consider that there are i=1,...*n* observations spatially distributed in irregular lattices whose centroids have coordinates (x_i,y_i) generated by a bivariate normal $NMV(0,I_2)$ where I_2 is the identity matrix 2x2. Despite the fact that the use of regular lattice is a very widespread practice, this is not a very realistic situation in many applications with economic data. Due to this, there is an extensive amount of researchers who stresses the need of using non-regular lattices (Farber et al. 2009; Robinson 2010). We also use the 4-nearest neighbours row standardised weight matrix (W_{4nn}) and three sample sizes (n=49,100,225).

MI and BP statistics have an asymptotic distribution which can be used to assess their significance for a particular sample. Nevertheless, in a non linear framework and for a comparative analysis, it is more consistent to evaluate this significance through re-sampling. In some cases, the concepts of non linearity and non normality could be interrelated (Lu, 2009). Therefore, the existence of non linearities may cause the lack of normality. Because of this, in our study, we use the bootstrap technique considering 999 iterations. For the MI and BP tests, we always evaluate the null hypothesis applying the 4-nearest neighbours (W_{4nn}) weight matrix. To carry out the Scan test, it is not necessary to have information about the W weight matrix, only the shape of the window is needed. The p-value for this statistic is obtained through Monte Carlo hypothesis testing (Dwass 1957), by comparing the rank of the maximum likelihood from the real data set with the likelihoods from the random data sets. If the rank is r, then the p-value = r/(1 + number of simulations).

Linear processes

As a previous step in our analysis, Table 1 shows the results corresponding to the size and power of the spatial dependence tests in one of the more frequent linear cases, SAR structures.

$$DGP_0 : Y = (I - \rho W)^{-1} e \text{ with } e \equiv N(0,1) \qquad [12]$$

Table 1

Estimated size and power with DGP_0 (5% significance level)

	n=49			*n=100*			*n=225*		
ρ	*IM*	*BP*	*Scan*	*IM*	*BP*	*Scan*	*IM*	*BP*	*Scan*
0.00	**0.049**	**0.058**	**0.047**	**0.042**	**0.039**	**0.050**	**0.049**	**0.046**	**0.052**
0.10	0.078	0.064	0.100	0.139	0.075	0.105	0.224	0.087	0.121
0.20	0.224	0.136	0.175	0.331	0.178	0.195	0.592	0.286	0.250
0.30	0.379	0.240	0.247	0.604	0.393	0.344	0.883	0.613	0.467
0.40	0.580	0.413	0.365	0.830	0.637	0.534	0.987	0.917	0.679
0.50	0.755	0.612	0.548	0.948	0.881	0.719	1.000	0.994	0.874
0.60	0.928	0.854	0.761	0.991	0.971	0.881	1.000	1.000	0.976
0.70	0.967	0.928	0.842	1.000	0.999	0.961	1.000	1.000	0.997
0.80	0.985	0.967	0.917	1.000	1.000	0.997	1.000	1.000	0.999
0.90	0.998	0.997	0.988	1.000	1.000	1.000	1.000	1.000	1.000
0.98	1.000	0.999	0.990	1.000	1.000	1.000	1.000	1.000	1.000

The size of the three tests is close to the nominal level of 5%. With regard to the power, as was expected, the MI presents the best results. The power of the three tests improves when the value of the parameter (ρ) and/or the size of the lattice n increases.

Nonlinear processes

In this section, we present different experiments which introduce nonlinear spatial dependence structures. We group these experiments attending to three criteria: (i) inadequate specification of the weight matrix (W) (ii) parametric instability of spatial effects and (iii) transformation of the linear processes.

(i) Inadequate specification of the weight matrix

The weight matrix defined to evaluate the existence of spatial dependence in a process plays a fundamental role in the majority of the spatial dependence tests. In this sense, the specification of this matrix (W) should be part of the null hypothesis (Pinkse, 2004). One common specification for the weight matrix is the binary matrix. This matrix does not consider different degrees of intensity in the interrelations among the observations. Only in the case when W is row standardised is there a weighting effect generated by the number of neighbours. Our first approximation to non linearity considers processes where the spatial dependence intensity changes in function of the distance. With this purpose, we consider weight matrices with different number of neighbours and with different weights among them, while the null hypothesis of independence is tested using a classical (W_{4nn}) matrix.

The first DGP that we propose is the following:

$$DGP_1 : Y = (I - \rho W)^{-1} e \text{ with } e \equiv N(0,1) \text{ and } W = \left\{ w_{ij} = 1 / d_{ij} \right\} \qquad [13]$$

Table 2a shows the estimated power of the three tests when the processes are generated according to DGP_1.

A second case, in which the non linearity is related to an incorrect selection of the weight matrix corresponds to a densely connected w matrix. Therefore, our second DGP proposal is:

$$\begin{gathered} DGP_2 : Y = (I - \rho W)^{-1} e \\ \text{with } e \equiv N(0,1) \text{ and } W = \left\{ w_{ij} = 1 \text{ if } d_{ij} < Me\left(d_{ij}\right) \text{ and 0 other wise} \right\} \end{gathered} \qquad [14]$$

Where Me is the median. In this case, the number of neighbours is different for each spatial unit and, therefore, the intensity of each region with the others is also different. The estimated power of the spatial dependence tests when we use (W_{4nn}) to test the null hypothesis appear in Table 2b.

The third case is based on the introduction of spatial dependence only in a reduced number of observations leaving the others without spatial dependence effects. To develop this proposal, we select a observation (k) and interconnect it with its ten nearest neighbours $NN(k,10)$.

$$\begin{gathered} DGP_3 : Y = (I - \rho W)^{-1} e \text{ with } e \equiv N(0,1) \text{ and} \\ W_{ij} = \begin{cases} 1 \text{ if } i, j \in NN\left(k, 10\right) \\ 0 \qquad \text{in other case} \end{cases} \end{gathered} \qquad [15]$$

The estimated power of the tests appears in the Table 2c.

In the three non linear cases (Table 2ab and c), the power of the tests is lower than that obtained in the linear case (Table 1) although the difference is more noticeable for the MI and BP tests. The loss of power is due to the misspecification of the weight matrix

used in the DGP (W) which is different from the weight matrix applied to evaluate the test (W_{4nn}).

Table 2

Estimated size and power with for misspecification in W (5% significance level)

Table 2a: DGP_1

	n=49			n=100			n=225		
ρ	IM	BP	Scan	IM	BP	Scan	IM	BP	Scan
0.20	0.057	0.066	0.070	0.073	0.055	0.073	0.055	0.061	0.088
0.50	0.122	0.087	0.138	0.147	0.091	0.141	0.167	0.079	0.148
0.70	0.222	0.131	0.225	0.183	0.120	0.175	0.189	0.116	0.168
0.98	0.384	0.264	0.368	0.363	0.219	0.307	0.422	0.220	0.404

Table 2b: PGD_2

	n=49			n=100			n=225		
ρ	IM	BP	Scan	IM	BP	Scan	IM	BP	Scan
0.20	0.053	0.073	0.077	0.051	0.046	0.071	0.045	0.055	0.078
0.50	0.135	0.116	0.161	0.103	0.075	0.135	0.070	0.051	0.102
0.70	0.210	0.154	0.265	0.139	0.114	0.178	0.106	0.086	0.157
0.98	0.415	0.343	0.454	0.335	0.277	0.273	0.213	0.103	0.214

Table 2c: PGD_3

	n=49			n=100			n=225		
ρ	IM	BP	Scan	IM	BP	Scan	IM	BP	Scan
0.20	0.064	0.062	0.073	0.043	0.061	0.057	0.050	0.051	0.067
0.50	0.126	0.093	0.145	0.077	0.068	0.135	0.071	0.071	0.130
0.70	0.170	0.109	0.271	0.215	0.203	0.336	0.175	0.191	0.287
0.98	0.922	0.908	0.936	0.925	0.922	0.945	0.937	0.947	0.950

(ii) Parametric instability in the spatial dependence

Another way of introducing nonlinear spatial interaction structures is by assuming that the intensity of the interaction is different among spatial units. Following Mur et al. (2009), we account for the DGP:

$$Y = \left(I - \rho HW\right)^{-1} e \text{ with } e = N\left(0, \sigma^2 I\right) \text{ and } H = diag\left(h_i; i = 1, \dots n\right) \qquad [16]$$

Different specifications for the H matrix generate different spatial dependence structures. In our case, we consider situations in which the non parametric and semi-parametric tests exceed the power of the MI test.

Our proposal consists of introducing spatial dependence in a two-step process,

- In the first step, a process without spatial structure is generated e ≡ N (0,1).
- In the second step, according to the expression [16], a spatial dependence structure is obtained according to the following cases:

case (i) $$h_i = \begin{cases} \rho \text{ if } e_i \geq Q_3 \\ 0 \text{ in other case} \end{cases}$$

case (ii) $$h_i = \begin{cases} \rho \text{ if } e_i \geq Me \\ \text{-}\rho \text{ if } e_i < Me \\ 0 \text{ in other case} \end{cases}$$

case (iii) $$h_i = \begin{cases} \rho \text{ if } e_i \in (\mathrm{Q}_1, \mathrm{Q}_3) \\ -\rho \text{ in other case} \end{cases}$$

case (iv) $$h_i = 2\left|e_i^3\right| \Big/ \max_i \left|e_i^3\right|$$

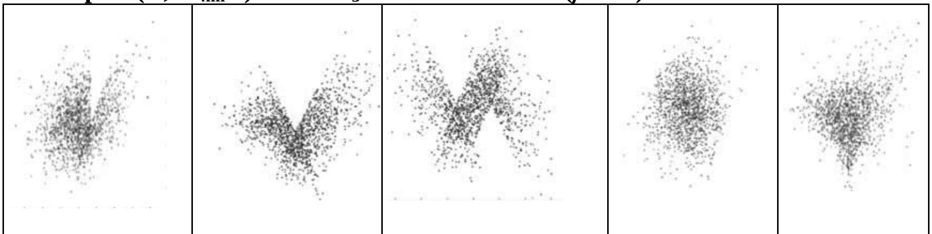
case (v) $$h_i = \sin\left(\frac{\pi}{4} e_i\right)$$

Where Q_k is the k-quartile of the observations obtained from the first step. *Me* is the median of *e*.

Figure 1 shows, by way of example, the scatter-plot of $(Y, W_{4nn}Y)$ for a pair of data generated for each of these cases, for a sample size of n=1600. The spatial interaction structure is clearly nonlinear as we can see when the sample is large and there is a strong spatial dependence structure (ρ=0.8). For the cases of small sample sizes, the symptoms of spatial structure are weaker, in spite of the high or low values of the parameter ρ.

Figure 1

Scatterplot $(Y, W_{4nn}Y)$ in GDP_5 with n=1600 and (ρ=0.8)



$h_i = \begin{cases} \rho \text{ if } e_i \geq Q_3 \\ 0 \text{ in other case} \end{cases}$	$h_i = \begin{cases} \rho \text{ if } e_i \geq Me \\ \text{-}\rho \text{ if } e_i < Me \\ 0 \text{ in other case} \end{cases}$	$h_i = \begin{cases} \rho \text{ if } e_i \in (\mathrm{Q}_1, \mathrm{Q}_3) \\ -\rho \text{ in other case} \end{cases}$	$h_i = 2\left\|e_i^3\right\| \Big/ \max_i \left\|e_i^3\right\|$	$h_i = \sin\left(\frac{\pi}{4} e_i\right)$

The weight matrix used in the DGP and in the tests is the same W_{4nn}. Table 3 presents the results corresponding to the estimated power of the spatial dependence tests for this case.

Table 3

Estimated size and power for nonlinear process due to instability in the spatial dependence. Discrete case. (5% significance level)

Table 3a

$$h_i = \begin{cases} \rho & \text{if } e_i \geq Q_3 \\ 0 & \text{in other case} \end{cases}$$

	n=49			*n=100*			*n=225*		
ρ	*IM*	*BP*	*Scan*	*IM*	*BP*	*Scan*	*IM*	*BP*	*Scan*
0.20	0.050	0.066	0.085	0.088	0.060	0.125	0.093	0.083	0.124
0.50	0.124	0.115	0.177	0.179	0.157	0.240	0.339	0.250	0.372
0.70	0.187	0.146	0.230	0.313	0.242	0.342	0.557	0.465	0.571
0.98	0.280	0.253	0.308	0.523	0.453	0.529	0.814	0.704	0.762

Table 3b

$$h_i = \begin{cases} \rho & \text{if } e_i \geq Me \\ -\rho & \text{if } e_i < Me \\ 0 & \text{in other case} \end{cases}$$

	n=49			*n=100*			*n=225*		
ρ	*IM*	*BP*	*Scan*	*IM*	*BP*	*Scan*	*IM*	*BP*	*Scan*
0.20	0.049	0.090	0.088	0.053	0.110	0.136	0.045	0.149	0.145
0.50	0.089	0.337	0.239	0.115	0.549	0.305	0.154	0.860	0.503
0.70	0.128	0.519	0.319	0.212	0.852	0.501	0.310	0.991	0.754
0.98	0.221	0.752	0.456	0.370	0.973	0.713	0.630	0.999	0.911

Table 3c:

$$h_i = \begin{cases} \rho & \text{if } e_i \in (Q_1, Q_3) \\ -\rho & \text{in other case} \end{cases}$$

	n=49			*n=100*			*n=225*		
ρ	*IM*	*BP*	*Scan*	*IM*	*BP*	*Scan*	*IM*	*BP*	*Scan*
0.20	0.041	0.079	0.051	0.042	0.074	0.057	0.035	0.086	0.036
0.50	0.042	0.181	0.063	0.046	0.340	0.032	0.036	0.638	0.014
0.70	0.062	0.344	0.062	0.050	0.691	0.039	0.030	0.967	0.022
0.98	0.110	0.643	0.100	0.091	0.936	0.093	0.150	1.000	0.149

Table 3

Estimated size and power for nonlinear process due to instability in the spatial dependence. Discrete case. (5% significance level) (Continued)

Table 3d

$$h_i = 2\left|e_i^3\right| \Big/ \max_i \left|e_i^3\right|$$

	n=49			*n=100*			*n=225*		
ρ	*IM*	*BP*	*Scan*	*IM*	*BP*	*Scan*	*IM*	*BP*	*Scan*
0.20	0.078	0.075	0.081	0.063	0.062	0.081	0.060	0.064	0.097
0.50	0.107	0.068	0.124	0.102	0.074	0.142	0.094	0.089	0.171
0.70	0.132	0.067	0.161	0.172	0.099	0.193	0.186	0.118	0.228
0.98	0.233	0.139	0.252	0.231	0.114	0.217	0.294	0.168	0.299

Table 3e

$$h_i = \sin\left(\frac{\pi}{4} e_i\right)$$

	n=49			*n=100*			*n=225*		
ρ	*IM*	*BP*	*Scan*	*IM*	*BP*	*Scan*	*IM*	*BP*	*Scan*
0.20	0.039	0.078	0.078	0.044	0.067	0.108	0.033	0.079	0.109
0.50	0.061	0.150	0.156	0.064	0.181	0.196	0.087	0.306	0.312
0.70	0.079	0.207	0.197	0.102	0.319	0.282	0.141	0.561	0.468
0.98	0.122	0.344	0.262	0.181	0.575	0.450	0.265	0.824	0.636

In all cases, the estimated power of the MI test is lower than the values obtained in Table 1. However, thenon parametric *BP* and the semiparametric *Scan* tests show better results.

(iii) Process transformations.

The transformation of a linear process is another alternative to introduce non linear spatial dependence structures. Following López et al. (2010), we explore two possible transformations of a SAR process:

$$\text{(i)} \quad Y = \exp\left\{-\left(I - \rho W\right)^{-1} e\right\}^3$$

$$\text{(ii)} \quad Y = 1 \Big/ \left\{\left(I - \rho W\right)^{-1} e\right\}^{1/3}$$

Table 4 shows the estimated power for the three spatial dependence tests. MI attains reasonable estimated power only for large sample sizes and with high spatial dependence; the non parametric and semi-parametric tests offer good results for reduced sample sizes and moderate spatial dependence coefficients.

Table 4

Estimated size and power for nonlinear process due to nonlinear transformations

Table 4a

$$Y = \exp\left\{-\left(I-\rho W\right)^{-1} e\right\}^{3}$$

ρ	*n=49* *IM*	*BP*	*Scan*	*n=100* *IM*	*BP*	*Scan*	*n=225* *IM*	*BP*	*Scan*
0.20	0.070	0.140	0.117	0.072	0.140	0.133	0.053	0.175	0.154
0.50	0.095	0.388	0.206	0.099	0.595	0.254	0.095	0.843	0.344
0.70	0.111	0.709	0.356	0.097	0.946	0.488	0.102	0.999	0.726
0.98	0.678	0.919	0.865	0.916	0.993	0.986	0.976	0.999	1.000

Table 4b

$$Y = 1\Big/\left\{\left(I-\rho W\right)^{-1} e\right\}^{1/3}$$

ρ	*n=49* *IM*	*BP*	*Scan*	*n=100* *IM*	*BP*	*Scan*	*n=225* *IM*	*BP*	*Scan*
0.20	0.083	0.086	0.085	0.084	0.096	0.078	0.130	0.157	0.080
0.50	0.262	0.325	0.184	0.309	0.484	0.185	0.531	0.778	0.145
0.70	0.457	0.622	0.364	0.668	0.893	0.388	0.928	0.999	0.337
0.98	0.882	0.957	0.879	0.963	0.999	0.910	0.999	1.000	0.937

4. An application: Production function with internal R&D expenditures

4.1 Empirical model and data issues

This section develops an empirical application of our three spatial tests to an economic-theory-of-production scenario. Specifically, we consider the output of a firm i (Y_i) to be a function of the traditional production factors (capital (K_i) and labour (L_i)), plus an additional input, the internal R&D effort (RD_i) (Tseng, 2008). We define the production function by employing a Cobb-Douglas specification:

$$Y_i = AK_i^{\alpha} L_i^{\beta} RD_i^{\lambda} e^{\varepsilon_i} \qquad [17]$$

Where A is a constant α,β and λ measure output elasticity to capital, labour and internal R&D, respectively; γ is the non-material rate of technical change, and ε is the error term. Applying logs, linealizing, and defining $d = \ln A$, we get:

$$\ln Y_i = d + \alpha 1nK_i + \beta \ln L_i + \lambda \ln RD_i + \varepsilon_i \qquad [18]$$

ESTADÍSTICA ESPAÑOLA Vol. 55. Núm. 180 / 2013

Size of firms is also a relevant factor in our analysis, so we split the sample between big and SMEs companies[3]. This results in the following specification of our empirical model:

$$\ln Y_i = d + \alpha \ln K_i + \alpha' S_i 1n\left(K_i\right) + \beta \ln L_i + \beta' S_i 1n\left(L_i\right) + \lambda \ln RD_i + \lambda' S_i 1n\left(RD_i\right) + \varepsilon_i \quad [19º]$$

S_i is a dummy variable taking value one for the big size (BS) firms and zero otherwise.

Firm level data comes from the Community Innovation Survey (CIS) of Spain, produced by the Spanish Institute of Statistics (INE). The variables correspond to PRODUCTION (*Y*), approached as total sales per company, CAPITAL (*K*), as the stock of net fixed assets of the firm, and LABOUR (*L*) as the number of employees. We use data on the internal R&D (*RD*) expenditure of the firm, as a proxy for technological/knowledge input. We have selected those manufacturing companies answering the CIS questionnaire in years 2004 and 2006 (around 6,100 companies). Equation [19] has been estimated for two groups of firms, those located in the municipality of Madrid and those in Barcelona (NUTS III level, Eurostat nomenclature). Our final data set then comprises a subset of 805 firms (142 big size companies) located in Barcelona, and another subset with 330 firms (112 large companies) in Madrid. Figure 2 shows the spatial distribution of firms for both municipalities in the study.

Figure 2

Spatial distribution of firms in Madrid and Barcelona

Madrid

Municipality of Madrid (detail)

Barcelona

Municipality of Barcelona (detail)

3 According to the European Commission (2002), we consider a firm with more than 250 employees as a big size (BS) company.

4.2 Empirical results

We begin by estimating the model by Ordinary Least Square (OLS) and testing different hypothesis using the residuals. Estimation results are included in Table 5; OLS coefficients are in line with previous literature on the topic. For both Madrid and Barcelona, production output is determined by the capital stock, with greater intensity in the SME establishments, and by employment. The contribution of employment to output, particularly in the case of Barcelona, does not seem to show differences between BS firms and SME, although employment is still an important element of the production function. In the case of Madrid, the capital stock seems to be relatively less important for output in BS firms than in SME, while for labour the correlation seems to be the reverse: greater effects for BS than for SME. Explanations for such differences include the role of Madrid as the "capital" of the country (Turner and Turner, 2011), its greater capacity for attracting BS, and more elaborate arguments linked to the debate about the increasing importance of human capital to the detriment of traditional physical capital in a globalised production process (Leamer, 2007). Finally, the impact of expenditure in R&D activities does not play a significant role for SME, but shows a very interesting role for BS in both municipalities, with a higher and significant coefficient in the case of Madrid.

These results show the important role of R&D activities on the production side of the companies that are able to invest in such activities (as BS companies). The smaller significance of this variable for SME is due to the particularities of our sample, where the segment of small companies just accounts for 16 per cent of total expenditure in R&D in the case of Barcelona, and 10 per cent in that of Madrid. This means that investments in R&D can be attributed to the segment of BS[4].

Regarding the spatial diagnostic measures for both municipalities (Barcelona and Madrid), the three tests (IM, BP and Scan) detect the presence of spatial dependence in the residuals. In order to compute the MI and the BP tests, we built a contiguity weight matrix (W) based on 4-nearest neighbours criteria. According to these results, the next step is to estimate model (19) by including a spatial dependence structure. In order to identify the adequate spatial structure we compute the Lagrange Multiplier (LM) tests. For both estimations, the RS-LE test appears as no significant, therefore, following Hendry's methodology (Florax et al., 2006) we conclude that a Spatial Error Model (SEM) is preferable. Therefore, we re-run the estimation of the model (19) but now including a SEM structure. Results of this estimation are also included in Table 5. As we can see, for both samples, the spatial coefficient (ρ) appears to be positive and significant.

Another important result is related to the selection of the spatial dependence tests to be applied in a nonlinear case, IM, BP or Scan. Once we have modelled the spatial dependence structure, through a SEM specification, the MI test indicates that there are no symptoms of spatial dependence in the residuals of the SEM model (see Table 5), while the nonparametric (BP) and semiparametric (Scan) tests still continue pointing to the presence of such spatial effects. Results are similar for both municipalities, of

[4] We do not extend more on empirical results here, given that our focus is prominently in the performance of the spatial dependence tests. Additional information is as usual available under request to the authors.

Madrid and Barcelona. We conclude that, for this case, the MI test not appears to be as useful as the BP or the Scan test to detect the existence of spatial patterns in the residuals. This application has helped us to highlight the need of applying new spatial dependence test, semiparametric or nonparametric ones, if we have to deal with nonlinear specification of economic models.

Table 5

Models Cobb-Douglas of productivity for Madrid and Barcelona

	Madrid				*Barcelona*			
	OLS		*SEM*		*OLS*		*SEM*	
	Coeff.	*t-ratio*	*Coeff.*	*t-ratio*	*Coeff.*	*t-ratio*	*Coeff.*	*t-ratio*
a	3.273*	6.931	3.303*	7.040	2.941*	11.074	2.892*	10.895
α	0.698*	16.654	0.695*	16.796	0.756*	32.569	0.757*	32.859
α'	-0.135*	-2.991	-0.147*	-3.304	-0.065**	-1.773	-0.067**	-1.848
β	0.421*	6.926	0.426*	7.083	0.315*	10.130	0.315*	10.252
β'	0.103	1.054	0.120	1.239	0.029	0.306	0.042	0.456
λ	-0.005	-0.166	-0.005	-0.177	-0.013**	-0.777	-0.011	-0.652
λ'	0.115*	2.581	0.121*	2.774	0.062*	1.951	0.058*	1.951
ρ			0.160*	2.275				3.438
R^2-adj	0.937				0.933			
Log-Lik	-252.72		-247.39		-450.63			

SPATIAL DIAGNOSTIC MEASURES

	Statistic	*p-value*	*Statistic*	*p-value*	*Statistic*	*p-value*	*Statistic*	*p-value*
IM	3.311	0.012	0.930	0.529	5.483	0.000	1.771	0.516
BP	8.635	0.002	2.968	0.035	33.894	0.000	11.972	0.000
Scan	14.855	0.004	10.701	0.040	15.344	0.040	11.345	0.022
LM-EL	10.217	0.001			29.274	0.000		
RS-EL	7.883	0.005			27.993	0.000		
LM-LE	3.528	0.060			1.457	0.227		
RS-LE	1.195	0.274			0.176	0.675		
LRCOM			6.224	0.514			5.211	0.735

* p-value<0.05; ** p-value< 0.10; LM-EL and LM-LE: Lagrange Multiplier tests for residual autocorrelation and omission of a spatial lag of the endogenous variable in the model. Respectively, robust to local specification errors in the null hypothesis (Anselin et al. 1996) RS-LE and RS-EL: Lagrange Multiplier tests for residual autocorrelation and omission of a spatial lag of the endogenous variable in SEM and SLM models, respectively (Anselin and Bera 1998). LRCOM: Likelihood Ratio test of common factors (Burridge. 1981). Log-lik: estimated log-likelihood.

Finally, it is possible to delve into the analysis of the spatial structure still present in the residuals by means of the Scan test, which allows us to identify the existence of spatial clusters of firms. The Scan test identifies, for the case of Madrid, a set of 27 companies, located in the North, responsible for the spatial dependence pattern detected by Scan and BP tests. This subset of firms also presents an error average value of 0.50, very high in comparison with a global average value of 0.00 in the test results. Figure 3 shows the

location of these companies. According to this result, our hypothesis is that the nonparametric tests were able to detect instability in the spatial dependence structure of the SEM model that were unnoticed after estimating this model.

Figure 3

Spatial Cluster of firms in Madrid

In bold colour, the spatial cluster of firms

In the case of the companies located in Barcelona, results are not as clear. We only detect two contiguous firms with high residuals. Moreover, according to this last result, we would like to point out that the lack of linearity in our spatial model, equation [19], could be caused by the existence of outliers with spatial interactions effects of different intensities among firms (Mur and Lauridsen 2007).

5. Conclusions

Spatial econometrics studies have usually employed the linear regression framework when modelling socio-economic relationships in space, but recently we have assisted to a growing interest in developing new models for dealing with nonlinearities. Improving our understanding of how phenomena occur is the basis of that new research area. As a natural complement, new families of tests also become necessary, particularly those better enabled for a nonlinear scenario. In this paper we have evaluated the behaviour of the three type of tests for checking spatial independence: parametric, nonparametric and semiparametric. To attain this goal we have selected representative proposals of each family of tests: the parametric I-Moran test, the nonparametric proposal of Brett and Pinkse (1997) and a semiparametric test applied on epidemiology studies, the Scan test.

In order to establish a comparison among these proposals, we have begun by generating different nonlinear spatial structures through Monte Carlo simulations, and then conducting empirical testing on the matter. Monte Carlo simulations have shown that the parametric *MI* test tends to fail when the process under study is nonlinear. This result is found in nearly all simulated processes, with *Scan* and *BP* tests clearly showing greater power in nonlinear frameworks.

In addition, we have developed an empirical exercise analysing how R&D expenditures affect the production function in two samples of manufacturing firms from Madrid and

Barcelona. Results show that, in addition to the traditional production factors (capital, labour and R&D), including the spatial dimension in the production function clearly improves the estimation output. We have also noted the higher capacity of new tests in presence of nonlinearities, with nonparametric proposals showing the most robust behaviour.

References

ANSELIN, L. (2010). «Thirty years of spatial econometrics», *Papers in Regional Science,* Vol. 89 pp. 3-25.

ANSELIN, L., BERA, A. (1998). «Spatial dependence in linear regression models with an introduction to spatial econometrics». In A. Ullah and D. Giles (eds.), *Handbook of Applied Economic Statistics*, New York: Marcel Dekker, pp. 237-289.

ANSELIN, L., BERA, A., FLORAX, RJGM. AND YOON, M. (1996). «Simple Diagnostic Tests for Spatial Dependence». *Regional Science and Urban Economics* 26(1), pp. 77-104.

BALTAGI, B. H., AND LI, D. (2001). «LM tests for functional form and spatial correlation», *International Regional Science Review*, Vol. 24, pp. 194-225.

BRETT, C., AND PINKSE, J. (1997). «Those taxes are all over the map! A test for spatial Independence of municipal tax rates in British Columbia». *International Regional Science Review* 20, pp. 131–151.

BURRIDGE, P. (1981). «Testing for a common factor in a spatial autoregression model». *Environment and Planning A*, Vol. 13(7), pp. 795-800.

CLIFF, A.D., AND ORD, J.K. (1981). «Spatial processes: models and applications». *London: Pion*, Vol. 44.

DE GRAAFF ,T., FLORAX, RJGM., AND NIJKAMP, P. (2001). «A general misspecification test for spatial regression models: dependence, heterogeneity and nonlinearity», *Journal of Regional Science*, Vol. 41 (2), pp. 255-276.

DUBIN, R. (1988).«Estimation of regression coefficients in the presence of spatially autocorrelated error terms», *Review of Economic and Statistics*, Vol. 70, pp. 466-474.

DWASS, M. (1957). «Modified randomization tests for nonparametric hypotheses». *Annals of Mathematical Statistics*, Vol. 2, pp. 181-187.

FARBER, S., PÁEZ, A., AND VOLZ, E. (2009). «Topology and dependency tests in spatial and network autoregressive models». *Geographical Analysis* Vol. 41(2), pp. 158-180.

FLORAX, RJGM. AND DE GRAAFF, T. (2004). «The performance of diagnostics tests for spatial dependence in linear regression models a meta-analysis of simulation studies». In: Anselin, L., Florax, R., and Rey, S. (Eds.), *Advances in Spatial Econometrics Methodology, Tools and Applications. Springer, Berlin*, pp. 29–65

FLORAX, RJGM., FOLMER, H., AND REY, S. (2006). «A comment on specification searches in spatial econometrics: the relevance of Hendry's methodology: a reply». *Regional Science and Urban Economics* 36, pp 300–308.

KOVACH, L. D. (1960). «Life can be so non linear», *American Scientist*, Vol. 48.

KULLDORFF, M., AND NAGARWALLA, N. (1995). «Spatial disease clusters: Detection and Inference». *Statistics in Medicine*, 14, pp. 799-810.

KULLDORFF, M., HUANG, L. AND KONTY, K. (2009), «A scan statistic for continuous data based on the normal probability model». *International Journal of Health Geographics*, Vol. 8, pp.58-73.

LEAMER, EE. (2007), «A Flat World, a Level Playing Field, a Small World After All, or None of the Above? A Review of Thomas L. Friedman's The World is Flat». *Journal of Economic Literature*, Vol. XLV, pp.83-126.

LÓPEZ, FA., MATILLA, M., MUR, J. AND RUIZ, M. (2010). «A non-parametric spatial independence test using symbolic entropy», *Regional Science and Urban Economics*40, 106-115

LU, Z. (2009), «Advances in nonlinear spatial times series modelling: A personal view», *mimeo*, School of Mathematical Sciences, University of Adelaide, Australia.

MORAN, P. (1950). «Notes on continuous stochastic phenomena». *Biometrika* Vol. 37, pp. 17–23.

MUR, J, LÓPEZ, FA. AND ANGULO, A. (2009). «Testing the hypothesis of stability in spatial econometric models». *Papers in Regional Science*. Vol. 88, pp. 409-444

MUR, J. AND LAURIDSEN, J. (2007). «Outliers and spatial dependence in cross-sectional regressions». *Environment and Planning A*, Vol. 39, pp. 1752-1769.

PACE, R K, BARRY, R., SLAWSON, V.C. JR., AND SIRMANS, C.F. (2004), «Simultaneous spatial and functional form transformation». In Anselin L, R J G M Floraxand S J Rey (eds.), *Advances in Spatial Econometrics*, Berlin: Springer-Verlag, pp. 197-224.

PEDE, VO., FLORAX, RJGM. AND HOLT, MT. (2008), «Modelling nonlinear spatial dynamics: A family of spatial STAR models and application to US economic growth», selected paper prepared for presentation at the *American Agricultural Economics Association Annual Meeting*, Orlando, Florida, July, pp. 27-29.

PINKSE, J. (1998). «A consistent nonparametric test for serial dependence». *Journal of Econometrics* 84, 205-231.

PINKSE, J. (2004). «Moran-flavoured tests with nuisance parameters: examples». In Anselin L, R Florax and S Rey (eds.), *New Advances in Spatial Econometrics*, pp. 67–77. Berlin: Springer.

ROBINSON, PM. (2010). «Efficient estimation of the semiparametric spatial autoregressive model» *Journal of Econometrics*, Vol. 157, pp. 6-17.

SU, L. AND JIN, S. (2010), «Profile quasi-maximum likelihood estimation of partially linear spatial autoregressive models», *Journal of Econometrics*, Vol. 157, pp. 18-33.

TSENG, CY. (2008). «Internal R&D effort, external imported technology and economic value added: empirical study of Taiwan's electronic industry». *Applied Economics*, 40, pp. 1073–1082

TURNER, SC., AND TURNER, RN. (2011). «Capital cities: A special case in urban development», *Annals of Regional Science*, Vol. 46, pp.19-35.

VAN GASTEL, RAJJ. AND PAELINCK, JHP. (1995). «Computation of Box-Cox transform parameters: A new method and its application to spatial econometrics», in Anselin L, RJGM Florax (eds.), *New Directions in Spatial Econometrics*, Berlin: Springer-Verlag, pp. 136-155.

YANG, Z., LI, C. AND TSE, YG. (2006). «Functional form and spatial dependence in dynamic panels», *Economics Letters*, Vol. 91, pp. 138-145.

YATCHEW, A. (1998). «Nonparametric regression techniques in economics», *Journal of Economics Literature*, Vol. 36, pp. 669-721.