Detecting dependence between spatial processes

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

2013

Detecting Dependence Between Spatial Processes

Marcos Herrera [(1)], Manuel Ruiz [(2),] Jesús Mur [(3)],

(1). CONICET - IELDE, Salta National University (Argentina);

email: mherreragomez@gmail.com

(2). Department of Computing and Quantitative Methods. Technical University of Cartagena, Spain

email: manuel.ruiz@upct.es

(3). Department of Economic Analysis. University of Zaragoza (Spain)

email: jmur@unizar.es

Detecting Dependence Between Spatial Processes

Abstract

Testing the assumption of independence between variables is a crucial aspect of spatial data analysis. However, the literature is limited and somewhat confusing. To our knowledge, we can mention only the bivariate generalization of Moran's statistic. This test suffers from several restrictions: it is applicable only to pairs of variables, a weighting matrix and the assumption of linearity are needed; the null hypothesis of the test is not totally clear. Given these limitations, we develop a new non-parametric test, $\Upsilon(m)$, based on symbolic dynamics with better properties. We show that the $\Upsilon(m)$ test can be extended to a multivariate framework, it is robust to departures from linearity, it does not need a weighting matrix and can be adapted to different specifications of the null. The test is consistent, computationally simple and with good size and power, as shown by a Monte Carlo experiment. An application to the case of the productivity of the manufacturing sector in the Ebro Valley illustrates our approach.

Keywords: Non-parametric methods; Spatial bootstrapping; Spatial independence; Symbolic dynamics.

JEL Classification: C21; C50; R15

1 Introduction

Dependence is a distinguishing feature of spatial data. The notion that near locations will exhibit similar values in a given variable is natural. The literature uses the term *spatial correlation* to refer to this feature, which has attracted considerable attention (Lesage and Pace, 2009, and references therein). We may think in terms of one or several variables but the situation is the same: data may exhibit positive or negative spatial autocorrelation depending on the interaction of the agents.

The detection of these dependencies is crucial in order to (i) model complex spatial relationships in which the spatial arrangement plays a fundamental role, and (ii) predict the spatial layout of a variable from known information about others variables, possibly, in other locations. This has been widely recognized in different fields such as epidemiology (Lawson, 2006), biology (Schmitz, 2010), economics (Anselin, 1988) and demography (Goodchild and Janelle, 2003).

The bivariate generalization of Moran's I (Wartenberg, 1985) is the most popular technique for testing the assumption of independence between spatial variables. The expression of the statistic is as follows:

$$I_{xy} = \frac{\sum_{i=1}^{L}\sum_{\substack{j=1 \\ i\neq j}}^{L} y_i w_{ij} x_j}{S_0\sqrt{\hat{V}ar(y)\hat{V}ar(x)}}, \tag{1}$$

where w_{ij} is the $(i,j) - th$ element of the weighting matrix W and S_0 is the sum of its elements; $\hat{V}ar(y)$ and $\hat{V}ar(x)$ refer to the (estimated) variance of the two series, y and x. Wartenberg's objective is to account for the spatial dependence of the data and their multivariate covariance, '*developing a strategy for the explanatory analysis of the spatial pattern in the multivariate domain*' (p. 264). Assuming that k different variables are observed in L observation points in space, he combines the relevant information of this dataset in a $(k \times k)$ M matrix of Mantel coefficients. The diagonal values of this matrix are the Moran's I of

each variable and the off-diagonal elements are the bivariate cross-correlation coefficients of (1).

The work of Wartenberg focuses on the statistical modelling of the covariance structure of the data, using generalized spatial principal component analysis. Consequently, he does not address the problem of assessing significance of the bivariate coefficients (their expected value and variance were obtained by Mantel, 1967, using a permutation approach). Anselin *et al.* (2002) take up the question adopting the so-called randomization approach which means the assignation of values to locations using random permutations (Cliff and Ord, 1981, Hall, 1985, or Upton and Fingleton, 1985). Then the observed value of the statistic is evaluated against the empirical randomized distribution. This is the procedure implemented in, for example, OpenGeoda (Geoda, 2012) in the Multivariate Moran's I Menu. The null hypothesis is that the two series are *i.i.d.* and independent.

The approach of Czaplewski and Reich (1993) to the use of the bivariate Moran's I is different. Their null hypothesis is that there is no spatial autocorrelation in the bivariate process $\{y_s; x_s\}_{s \in S}$ where S is a set of locations in space; in their words: '*assume the bivariate observation* $(y_s; x_s)$ *for location* s *is a random, spatially independent drawing form one (or separate identical) population(s), and the joint distribution function(s) for* Y *and* X *are unknown*' (p.2). The null of no spatial correlation is conditional on the observed linear dependence of the pairs $(y_s; x_s)_{s \in S}$. The expected value of the bivariate Moran's I_{xy} of (1), considering the $L!$ random permutations of the pairs $(y_s; x_s)_{s \in S}$ is:

$$E\left[I_{xy}\right] = -\frac{\rho_{yx}}{L-1}, \tag{2}$$

ρ_{yx} being a measure of linear correlation between y and x. The expression of the variance is rather awkward, as shown in equation (60) of Czaplewski and Reich (1993). Moreover, they show that, for moderate to large sample sizes (L should be greater than 40), the bivariate Moran's I_{yx} tends to normality under the null (Reich *et al.*, 1994, for a nice application).

Moreover, we should not forget the risk of spurious correlation when using spatial data. Fingleton (2001) showed that it is very likely to find correlation between two unrelated series, in the case they have a spatial unit root or are near to nonstationarity; this result is, obviously, spurious. Mur and Trivez (2003) extend this problem to series that are spatially autocorrelated, not necessarily near nonstationarity, but with a strong deterministic component. If this is the case, the bivariate Moran's I, both in the Geoda and in the Czaplewski-Reich version, will reject the null hypothesis due to the strong autocorrelation existing in the series.

Our impression is that the bivariate Moran's I is a good statistic to explore spatial relationships but it is not enough. We would need some new techniques capable of working with three different null hypotheses: (i) the series are $i.i.d.$ and independent (ii) the bivariate process is spatially non-autocorrelated and (iii) the series are independent. This is the purpose of the new nonparametric statistic that we call $\Upsilon(m)$. The$\Upsilon(m)$ test is free from distributional assumptions, it does need the specification of a W weighting matrix, it is robust to strong departures from linearity, and it is flexible to the specification of the null hypothesis. We show that this test can be generalized to the case of more than two variables, is consistent and with good size and power in the small sample case.

In Section 2 we present some definitions and basic concepts. Section 3 obtains the independence test under different versions of the null hypothesis. Section 4 presents the results of a Monte Carlo experiment in which we also include the bivariate Moran's I. In Section 5, we discuss an application to the case of the productivity in the manufacturing sector in the Ebro valley. Conclusions appear in Section 6.

2 Tools for Symbolic Analysis

Symbolic dynamics is based on the transformation of a series into a sequence of symbols that captures useful information that cannot be directly observed. The idea is to consider a

space where all the possible states of a system can be represented. This space is partitioned into a finite number of regions and each region is represented by a symbol. In other words, symbolic dynamics is a simplified description of a dynamical system. For further details, see Hao and Zheng (1998).

2.1 Symbolization Process

This section presents a symbolization procedure for spatial series. The symbolization can be improved if additional information for the processes under study is available.

Let $\{x_s\}_{s\in S}$ and $\{y_s\}_{s\in S}$ be two spatial processes of real data, where S is a set of locations in space. Let us define a non-empty finite set of symbols, denoted by $\Gamma_n = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$.

Symbolizing a process means defining a map

$$f_x : \{x_s\}_{s\in S} \to \Gamma_n, \tag{3}$$

such that each element of the series, x_s, is associated with a unique symbol $f_x(x_s) = \sigma_{i_s}$ with $i_s \in \{1, 2, \ldots, n\}$. 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 $f_x(x_s) = \sigma_{i_s}$. We call f_x the symbolization map; the same can be done for y.

Now we introduce a bivariate process $\{Z_s\}_{s\in S}$:

$$Z_s = \{x_s, y_s\}, \tag{4}$$

where x_s and y_s are two univariate spatial processes. We define the set of symbols Ω_n^2 as the Cartesian product of the two individual sets Γ_n, that is, $\Omega_n^2 = \Gamma_n \times \Gamma_n$. The new symbols are $\eta_{ij} = \left(\sigma_i^x, \sigma_j^y\right)$. The symbolization function of the bivariate process can be expressed as:

$$g : \{Z_s\}_{s\in S} \to \Omega_n^2 = \Gamma_n \times \Gamma_n, \tag{5}$$

where

$$g\left(Z_s = (x_s, y_s)\right) = \left(f_x\left(x_s\right), f_y\left(y_s\right)\right) = \eta_{ij} = \left(\sigma_i^x, \sigma_j^y\right). \tag{6}$$

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

The analysis can be extended to a multivariate framework by considering the following $k - dimensional$ process, $\{Z_s\}_{s\in S} = \{x_{1s}, x_{2s}, \ldots, x_{ks}\}$. Let

$$\Omega_n^k = \Gamma_n \times \Gamma_n \cdots \times \Gamma_n$$

be the Cartesian product of k copies of Γ_n and $\eta_{i_1,i_2,\ldots i_k} = (\sigma_{i_1}, \sigma_{i_2}, \ldots, \sigma_{i_k}) \in \Omega_n^k$. We say that s is $\eta_{i_1,i_2,\ldots i_k} - type$ if and only if s is $\sigma_{i_j} - type$ for x_{js} for all $j = 1, 2, \ldots, k$.

Different symbolization schemes can be used depending on the characteristics of the series. For the case of spatial data, the following procedure is simple and offers good results (other alternatives can be found in Matilla and Ruiz, 2008, 2009, López *et al.*, 2010, and Ruiz *et al.*, 2010). Let M_e^x be the median of the spatial series $\{x_s\}_{s\in S}$. Define the indicator function

$$\tau_s = \begin{cases} 1 & if \quad x_s \geq M_e^x. \\ 0 & otherwise. \end{cases} \tag{7}$$

For each $s \in S$, let N_s be the set formed by the $(m-1)$ neighbours of s, $N_s = \{s_1, \ldots, s_{m-1}\}$, where $m \geq 2$ is the embedding dimension of the symbolization process. We use the term $m - surrounding$ to denote the group formed by location s and its corresponding set of indices N_s. Then for each location we build the $(m \times 1)$ vector:

$$x_m\left(s\right) = \left(x_s, x_{s_1}, \ldots, x_{s_{m-1}}\right).$$

Next, we define the indicator function for each s_i with $i = 1, 2, \ldots, m-1$:

$$\iota_{ss_i} = \begin{cases} 0 & if \quad \tau_s \neq \tau_{s_i}. \\ 1 & otherwise. \end{cases} \tag{8}$$

Finally, we can symbolize the spatial process $\{x_s\}_{s \in S}$ as $f_x : \{x_s\}_{s \in S} \rightarrow \Gamma_m$, where function f_x is:

$$f_x(x_s) = \sum_{i=1}^{m-1} \iota_{ss_i}, \tag{9}$$

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

In sum, this symbolization process consists of comparing, for each location s, the value τ_s with τ_{s_i} for all $s_i \in N_s$. We repeat the discussion for y. Once the two univariate processes have been symbolized, we obtain the symbol corresponding to the bivariate process in each location.

Next, we calculate the absolute and relative frequency of the symbols. The absolute frequency of symbol σ_i^x is simply:

$$n_{\sigma_i^x} = \#\left\{s \in S | s \quad is \quad \sigma_i^x - type \quad for \quad x\right\}, \tag{10}$$

whose relative frequencies can be estimated:

$$p(\sigma_i^x) \equiv p_{\sigma_i^x} = \frac{\#\left\{s \in S | s \quad is \quad \sigma_i^x - type \quad for \quad x\right\}}{|S|} = \frac{n_{\sigma_i^x}}{|S|}. \tag{11}$$

$|S|$ denotes the cardinality of set S. The same applies for series $\{y_s\}_{s \in S}$. The relative frequency for $\eta_{ij} \in \Omega_n^2$ is:

$$p(\eta_{ij}) \equiv p_{\eta_{ij}} = \frac{\#\left\{s \in S | s \quad is \quad \eta_{ij} - type\right\}}{|S|} = \frac{n_{\eta_{ij}}}{|S|}, \tag{12}$$

We introduce the concept of symbolic entropy for a $two - dimensional$ spatial series

$\{Z_s\}_{s\in S}$ through the Shannon statistic:

$$h_Z(m) = -\sum_{\eta\in\Omega_m^2} p(\eta)\, ln\,(p(\eta))\,. \tag{13}$$

As is well-known, $h_Z(m)$ is a measure of information contained, in this case, in the bivariate process.

Similarly, we can define the marginal symbolic entropies as

$$h_x(m) = -\sum_{\sigma^x\in\Gamma_m} p(\sigma^x)\, ln\,(p(\sigma^x))\,, \tag{14}$$

$$h_y(m) = -\sum_{\sigma^y\in\Gamma_m} p(\sigma^y)\, ln\,(p(\sigma^y))\,, \tag{15}$$

The joint and marginal entropies satisfy that $0 \leq h(m) \leq ln(n)$. The lower limit is reached when only a single symbol appears and the upper limit is reached when all the symbols have the same probability (they are uniformly distributed).

Note that under the assumption that each univariate process is *i.i.d.,* and using the symbolization procedure of (7)-(8)-(9), the probability of a given symbol is given by

$$p(\sigma) = {}^{C_\sigma^{m-1}}\!/_{2^{(m-1)}}, \tag{16}$$

where $C_\sigma^{m-1} = {}^{(m-1)!}\!/_{[(m-1-\sigma)!\sigma!]}$ denotes the number of σ-combinations of symbols, $\sigma \in \{0,\ldots,m-1\}$, for the set of $m-1$ elements.

2.2 A Simple Example

The following example presents a simple symbolization process. Let us assume that we have two spatial processes x and y, observed in a 3×3 regular lattice as shown in Figure 1.

[FIGURE 1 ABOUT HERE]

In the first place, we must define the value of m, the embedding dimension, for example, $m = 4$. Then we identify the 3 nearest neigbours for each location:

$(N_{s_1} = \{s_2, s_4, s_5\})$,$(N_{s_2} = \{s_3, s_1, s_5\})$, $(N_{s_3} = \{s_2, s_6, s_5\})$, $(N_{s_4} = \{s_5, s_1, s_7\})$,

$(N_{s_5} = \{s_6, s_2, s_4\})$, $(N_{s_6} = \{s_3, s_5, s_9\})$, $(N_{s_7} = \{s_8, s_4, s_5\})$, $(N_{s_8} = \{s_9, s_5, s_7\})$,

$(N_{s_9} = \{s_6, s_8, s_5\})$.

Now we can form the $4 - surroundings$ for each observation. In the case of variable x in location s_1:

$x_4(s_1) = (x_{s_1} = 4, x_{s_2} = 1, x_{s_4} = 6, x_{s_5} = 2)$.

Similarly with the other locations and processes. The next step is to obtain of the symbols. In accordance with (7)-(8)-(9), and for variable x in location s_1:

$f_x(x_{s_1}) = (\iota_{s_1 s_2} = 0) + (\iota_{s_1 s_4} = 1) + (\iota_{s_1 s_5} = 0) = 1$.

Likewise, we can obtain the symbols associated with the other locations:

$f_x(x_{s_2}) = 1$; $f_x(x_{s_3}) = 1$; $f_x(x_{s_4}) = 1$; $f_x(x_{s_5}) = 1$; $f_x(x_{s_6}) = 2$; $f_x(x_{s_7}) = 2$; $f_x(x_{s_8}) = 2$; $f_x(x_{s_9}) = 1$.

The symbols associated with series y are: $f_y(y_{s_1}) = 0$; $f_y(y_{s_2}) = 1$; $f_y(y_{s_3}) = 1$; $f_y(y_{s_4}) = 1$; $f_y(y_{s_5}) = 2$; $f_y(y_{s_6}) = 2$; $f_y(y_{s_7}) = 1$; $f_y(y_{s_8}) = 2$; $f_y(y_{s_9}) = 2$.

Finally, the symbols for the bivariate process $Z = (x, y)$ are just the Cartesian product of the two previous sets of symbols:

$f_z(z_{s_1}) = (1;0)$; $f_z(z_{s_2}) = (1,1)$; $f_z(z_{s_3}) = (1,1)$; $f_z(z_{s_4}) = (1,1)$; $f_z(z_{s_5}) = (1,2)$; $f_z(z_{s_6}) = (2,2)$; $f_z(z_{s_7}) = (2,1)$; $f_z(z_{s_8}) = (2,2)$; $f_z(z_{s_9}) = (2,1)$.

3 A Test for Independence Between Spatial Processes

In this section, we develop two non-parametric tests for spatial dependence using the results of the previous section.

3.1 A composite null hypothesis of independence and *i.i.d.*

Let us consider a two-dimensional spatial series $\{Z_s = (x_s, y_s)\}_{s \in S}$ with a fixed embedding dimension, $m \geq 2$. Our interest focuses on the following null and alternative hypotheses:

$$\left.\begin{array}{c} H_0 : \{x_s\}_{s \in S} \textit{ and } \{y_s\}_{s \in S} \textit{ are i.i.d. and independent.} \\ H_1 : \{x_s\}_{s \in S} \textit{ and } \{y_s\}_{s \in S} \textit{ are not i.i.d. and independent.} \end{array}\right\} \tag{17}$$

The null hypothesis is tested using the symbolization procedure of Section 2.1.

For each symbol $\eta \in \Omega_n^2$ of the bivariate process, we define the random variable J_{η_s} as follows:

$$J_{\eta_s} = \begin{cases} 1 & if \textit{ location } s \textit{ is } \eta - type. \\ 0 & otherwise. \end{cases} \tag{18}$$

J_{η_s} is a Bernoulli variable with probability of "success" p_η, where "success" means that s is $\eta - type$.

This restriction is obvious:

$$\sum_{\eta \in \Omega_n^2} p_\eta = 1. \tag{19}$$

We assume that the set of locations S is finite, and that $|S| = L$. We are interested in the number of locations that are of $\eta - type$, for all symbols $\eta \in \Omega_n^2$, this is measured by:

$$Q_\eta = \sum_{s \in S} J_{\eta_s}. \tag{20}$$

Q_η is a random variable whose values are integers in the range $\{0, 1, 2, \ldots, L\}$.

Notice that not all J_η variables are independent (due to the possible overlapping of the $m - surroundings$) and, therefore, Q_η may not be a Binomial random variable. Nevertheless, as shown by Soon (1996), the sum of dependent indicators can be approximated to a Binomial if the following two conditions are satisfied (the details appear in Appendix A):

1. Dependency among the indicators is *weak*.
2. The probability of occurrence of the indicators is *small*.

The first condition can be achived by controling for overlapping. The second condition depends on the number of symbols, n. The symbolization procedure presented in Section 2.1 entails, in general, a low probability for each indicator.

Ruiz *et al.* (2010) discuss the case of controlling for the degree of overlapping of the $m-surroundings$ to assure weak dependence among the J_{η_s} indicators. They show that it is possible to attain a good approximation to the Binomial by considering only a subset of locations, $\tilde{S} \subseteq S$, with a fixed degree of overlapping. Less overlapping means weaker dependence among the indicators, but at the cost of not symbolizing all the observations.

Under the null hypothesis H_0, if the dependence among the J_η indicators is weak, the variable Q_η can be approximated to a Binomial random variable:

$$Q_\eta \approx B\left(L, p_\eta\right). \tag{21}$$

The joint probability function of the n^2 variables $(Q_{\eta_{11}}, Q_{\eta_{12}}, \ldots, Q_{\eta_{nn}})$ is a multinomial such that:

$$P\left(Q_{\eta_{11}} = a_{11}, \ldots, Q_{\eta_{nn}} = a_{nn}\right) = \frac{(a_{11} + a_{12} + \cdots + a_{nn})!}{a_{11}!a_{12}! \cdots a_{nn}!} p_{\eta_{11}}^{a_{11}} p_{\eta_{12}}^{a_{12}} \cdots p_{\eta_{nn}}^{a_{nn}}, \tag{22}$$

where $a_{11} + a_{12} + \cdots + a_{nn} = L$.

As a result of our symbolization procedure, the following Theorem can be proved (see web Appendix B for proofs):

Theorem 3.1. *Let $\{x_s\}_{s\in S}$ and $\{y_s\}_{s\in S}$ be two spatial processes with $|S| = L$. Assume that both processes have been symbolized such that the dependence among indicators J_η is weak. Denote by $h_Z(m)$ the entropy defined in (13) for a fixed embedding dimension $m \geq 2$, with $m \in \mathbb{N}$. If the spatial series $\{x_s\}_{s\in S}$ and $\{y_s\}_{s\in S}$ are i.i.d. and independent, then the statistic*

$$\Upsilon(m) = 2L\left[2(m-1)\,ln(2) - \left[\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{n_{\eta_{ij}}}{L} ln\left(C_{\sigma_i^x}^{m-1} C_{\sigma_j^y}^{m-1}\right)\right] - h_Z(m)\right] \tag{23}$$

is asymptotically distributed as a $\chi^2_{m^2+1}$.

Let α be a real number with $0 \leq \alpha \leq 1$ and χ^2_α is such that $Pr\left(\chi^2_k > \chi^2_\alpha\right) = \alpha$. To test

$$H_0 : \{x_s\}_{s\in S} \text{ and } \{y_s\}_{s\in S} \quad are \quad i.i.d. \quad and \quad independent, \tag{24}$$

the decision rule, with a $100\,(1-\alpha)\,\%$ confidence level, is:

$$If \quad 0 \leq \Upsilon\,(m) \leq \chi^2_\alpha, \qquad Do \quad not \quad reject \quad H_0.$$

$$Otherwise, \quad Reject \quad H_0.$$

This test can be generalized to a $k - dimensional$ spatial processes Z. The expression of the multivariate test is:

$$\Upsilon\,(m) = 2L\left[k\,(m-1)\,ln\,(2) - \left[\sum_{i_1=1}^{n}\cdots\sum_{i_k=1}^{n}\frac{n_{\eta_{ij}}}{L}ln\left(\prod_{j=1}^{k}C^{m-1}_{\sigma^{x_i}_{ij}}\right)\right] - h_Z\,(m)\right] \tag{25}$$

which is asymptotically distributed as a $\chi^2_{(m^k+k-1)}$.

The composite null hypothesis of (24) can be rejected, in the first place, due to the lack of independence of the two spatial processes and, in second place, because they are not *i.i.d.* (one or both of them are spatially autocorrelated). We reserve the term *interdependence* for the first case and the term *intradependence* for the second.

3.2 Consistency of the test $\Upsilon\,(m)$

In the previous section, we have obtained the $\Upsilon\,(m)$ test and its asymptotic distribution under a composite null hypothesis. We now add the property of consistency for a wide range of spatial processes (see web Appendix C for proofs).

Theorem 3.2. *Let* $\{x_s\}_{s\in S}$ *and* $\{y_s\}_{s\in S}$ *be two stationary processes, and* $m > 2$ *with* $m \in \mathbb{N}$. *Under dependence of order lower than* m,

$$\lim_{L\to\infty} Pr\left(\widehat{\Upsilon}\,(m) > C\right) = 1,$$

for all $0 < C < \infty$, $C \in \mathbb{R}$.

Parameter m is crucial to ensure the property of consistency. In fact, the $\Upsilon(m)$ test will reject the composite null hypothesis of (24) provided that the dependence (inter or intra) is of a smaller order than m; that is, it occurs inside the m-surroundings. The researcher must define this parameter considering the following conditions:

1. The minimum value of m is 2.

2. The maximum value of m depends on the sample size L. Note that L must be larger than the number of symbols ($n^2 \leq L$). Moreover, in order to have a good approximation to the χ^2 distribution, the expected value of the frequencies for each symbol should be, at least, 5 (Rohatgi, 1976). This means that the embedding dimension should be fixed so that $5 \times n^2 \leq L$.

In the case of the univariate symbolization procedure of (7)-(8)-(9), $m = n$ so the restriction reads $m \leq \sqrt{\frac{L}{5}}$. For example, if we establish $m = 5$, the bivariate distribution would have 25 symbols and we need a sample of, at least, 125 observations.

3.3 A permutation alternative to the asymptotic approximation

We may come across situations where the conditions described above cannot be fulfilled because there is a high overlapping or an insufficient number of observations. In these cases, the random variable Q_η of (20) will be a poor approximation to the Binomial distribution and, consequently, the convergence of $\Upsilon(m)$ to a χ^2 distribution is not guaranteed. Below we present an alternative strategy for testing independence using a traditional permutation procedure that does not depend on the Binomial approximation.

Let us define $\Psi_1 = \frac{1}{2L}\Upsilon$. The procedure, with a number B of permutations, is as follows:

1. Compute the value of the statistic $\hat{\Psi}_1$ from the original sample $\{x_s\}_{s\in S}$ and $\{y_s\}_{s\in S}$.

2. Randomly permuting $\{x_s\}_{s\in S}$ and $\{y_s\}_{s\in S}$, we obtain two permuted series $\{x_s(b)\}_{s\in S}$ and $\{y_s(b)\}_{s\in S}$, where b is the number of the permuted sample.

3. For series $\{x_s(b)\}_{s\in S}$ and $\{y_s(b)\}_{s\in S}$, estimate the statistic $\hat{\Psi}_1^{(b)}$.

4. Repeat steps 2 and 3 $B-1$ times to obtain B permuted values of the statistic $\left\{\hat{\Psi}_1^{(b)}\right\}_{b=1}^{B}$.

5. Compute the estimated $p_{permutation} - value$:

$$p_{permutation} - value\left(\hat{\Psi}_1\right) = \frac{1}{B}\sum_{b=1}^{B} 1\left(\hat{\Psi}_1^{(b)} > \hat{\Psi}_1\right), \tag{26}$$

where $1(\cdot)$ is an indicator function that assigns 1 if the inequality is true and 0 otherwise.

6. Reject the null hypothesis H_0 of (24) if

$$p_{permutation} - value\left(\hat{\Psi}_1\right) < \alpha, \tag{27}$$

for a nominal size α.

We can proceed in the same way for the case of the bivariate Moran's I_{yx}, either in what we have called the Geoda version or in the Czaplewski-Reich version. In the first version, the randomization algorithm is as follows:

1. Compute the value of the statistic $\hat{I}_{yx}$ from the original sample $\{x_s\}_{s\in S}$ and $\{y_s\}_{s\in S}$.

2. Randomly permuting $\{x_s\}_{s\in S}$ and $\{y_s\}_{s\in S}$, obtain two permuted series $\{x_s(b)\}_{s\in S}$ and $\{y_s(b)\}_{s\in S}$, where b is the number of the permuted sample.

3. For series $\{x_s(b)\}_{s\in S}$ and $\{y_s(b)\}_{s\in S}$, estimate the statistic $\hat{I}_{yx}^{(b)}$.

4. Repeat $B-1$ times steps 2 and 3 to obtain B permuted realizations of the statistic $\left\{\hat{I}_{yx}^{(b)}\right\}_{b=1}^{B}$.

5. Compute the estimated $p_{permutation} - value$:

$$p_{permutation} - value\left(\hat{I}_{yx}\right) = \frac{1}{B}\sum_{b=1}^{B} 1\left(\hat{I}_{yx}^{(b)} > \hat{I}_{yx}\right). \tag{28}$$

6. Reject the null hypothesis H_0 of (24) if

$$p_{permutation} - value\left(\hat{I}_{yx}\right) > 1 - (\alpha/2)\ or\ p_{permutation} - value\left(\hat{I}_{yx}\right) < \alpha/2, \tag{29}$$

for a nominal size α.

Note that, in step 2, we randomly permute the two series separately. This procedure breaks the possible dependence between the series (both are tests of independence) and also destroys the spatial structure which may exist in either series (they are also test of $i.i.d$).

The only difference in the Czaplewski-Reich version is that, in step 2, we randomly permute the pairs $\{x_s, y_s\}_{s\in S}$ of the $\{Z_s\}_{s\in S}$ bivariate process, because we want to preserve the dependence between the two series and destroy their spatial structure.

3.4 A simple null hypothesis of independence

The framework of Section 2 also allows us to test for the hypothesis of independence, allowing the series not to be $i.i.d$. The new null hypothesis is that the series, $\{x_s\}_{s\in S}$ and $\{y_s\}_{s\in S}$ are independent, regardless of their spatial structure. In this section, we obtain a bootstrapped test for this hypothesis that preserves the spatial structure of each series.

Let us recall that, under the assumption of inter-independence, we have:

$$p_{\eta_{ij}} = p_{\sigma_i} p_{\sigma_j}. \tag{30}$$

So that, for the bivariate process of (4), the measure of entropy can be decomposed as $h_Z(m) = h_x(m) + h_y(m)$, which leads us to the statistic:

$$\Psi_2 = h_x(m) + h_y(m) - h_Z(m), \tag{31}$$

which is the basis of the testing procedure of the null hypothesis:

$$H_0^{'} : \{x_s\}_{s\in S} \; and \; \{y_s\}_{s\in S} \quad are \quad independent. \tag{32}$$

The distribution function of the Ψ_2 statistics of (31), under the null of (32), is unknown even in the case of large samples. We resort to bootstrap techniques. Few results have been

established for the case of bootstrapping spatial models (Pinkse and Slade, 1998, Fingleton 2005 and 2008, Fingleton and LeGallo, 2008, Monchuk *et al.*, 2011) and many of them focus on the case of the J test (Burridge and Fingleton, 2010, Burridge, 2012, Han and Lee, 2012)

Our approach follows the general guidelines of the non-overlapping time block bootstrap of Carlstein (1986). We had to develop a *spatial block bootstrap* (SBB) to break the dependence of the series but preserving most part of their spatial structure. To our knowledge, this is the first time that a block bootstrap scheme has been applied to spatial data. The SBB is as follows:

1. Divide the spatial system into $b = L/l$ contiguous observational blocks of length l. By contiguous observational block we mean that the observations of each block are contiguous according to the W weighting matrix. The b blocks cannot overlap and must exhaust the whole spatial system.

2. Divide the two spatial series according to (1).

3. We have b different sub-samples of length l for each series. The next step is to randomly resample the blocks, with replacement, in order to build the new bootstrapped series of length L.

4. The resampling has been completed.

Note that changes in the spatial structure, in terms of the network of connections between the locations, should be small because blocks of observations are resampled, not individual observations. The blocks of step (1) are formed according to a distance criterion to b previously defined fixed points. Let us call them the *buoys* of the SBB. Once the *buoys* have been defined, each observation is assigned to the nearest *buoy* to form a block of length l (other procedures can be used depending on the purpose of the bootstrap).

The second part of the SSB has to do with the assignment of the data, taken from their original location, to the new observation points. The criterion in this stage of the algorithm

is the distance to the *buoy*: (i) the observations of each block were ordered according to the distance to the *buoy* of the block; (ii) the blocks are resampled; (iii) data are assigned to the observation points that, in the destination block, occupy the same relative position as in the block of origin.

The example of Figure 2 illustrates the procedure. The original sample space appears in the left-hand panel, Figure (a), with 40 data points. We are going to use four blocks, $l = 4$. The *buoys* are the four vertices of the rectangle: $\{b(0,0), b(1,0), b(1,1), b(0,1)\}$. The resulting blocks appear in the right-hand panel, Figure (b), where each data point has been assigned to the nearest *buoy*. Then we resample the four blocks, with replacement. Assume that the result of a given bootstrap is $\{b(0,0), b(1,0), b(1,0), b(1,1)\}$. This means that, in the bootstrapped sample, the block $b(0,0)$ will remain in its original position, the same as the second block $b(1,0)$. The data of the second block will be copied in the spatial layout of the third block, $b(1,1)$, and distributed according to the distance criterion. Finally, the 10 observations of the third block will also be copied and distributed in the spatial layout of the fourth block, $b(0,1)$.

[Figure 2 about here]

The SSB adapts well to problems where the W matrix has been built according to some measure of distance, as in the k-nearest neighbours criterion. Obviously, changes in the contiguity criterion should also involve changes in the spatial bootstrapping scheme. The SSB will, inevitably, produce some missmatches in relation to the original spatial ordering of the data but, according to our results, it works efficiently, preserving much of the spatial structure of the original series.

4 Finite Sample Performance

This section focuses on the behaviour of the the new tests in a finite sampling context. We use a Monte Carlo approach.

4.1 Global parameters of the Monte Carlo

For the null hypothesis of *i.i.d.* and independence, we compare the $\hat{\Psi}_1$ test with the bivariate Moran's I_{yx}. The two tests coincide in the definition of the null but their structure is very different. The $\hat{\Psi}_2$ test is not directly comparable with the previous two. Its null hypothesis is that the series are independent, whatever their univariate structure. To our knowledge, this is the first test proposed in the literature to deal with such a problem.

In each simulation, we use a randomly generated spatial map of the L locations. The spatial support is non-regular and overlapping can be a problem, which makes it advisable an empirical approximation to the independence tests.

The global parameters we use are the following:

$$L \in \{100, 400, 1000\}\,,\ m \in \{4, 6, 8\}\,. \tag{33}$$

L is the sample size and m is the embedding dimension.

In the experiment, we want to simulate linear and non-linear models. In the first case, linearity, we control the relation through the expected coefficient of determination, R^2. Based on a specification like this:

$$y = \beta x + \theta W x + \varepsilon,$$

where the variance of x and ε are 1, $Var[x] = Var[\varepsilon] = 1$, assuming independence, $Cov[x, \varepsilon] = 0$, the R^2 statistic is:

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

To measure the empirical size, we use the Data Generation Process (DGP from now on):

$$\begin{aligned} y &\sim \mathcal{N}(0,1), \\ x &\sim \mathcal{N}(0,1), \end{aligned} \tag{34}$$

For the case of the power, we consider bivariate processes of the following type:

$$\begin{aligned} y = F[x, Wy, Wx, \varepsilon]\,;\, \varepsilon \backsim \mathcal{N}(0,1). \\ x \backsim \mathcal{N}(0,1). \end{aligned} \tag{35}$$

where W is specified using the $(m-1)$ nearest neighbours criterion, and F is a functional form to be defined. We have simulated three linear processes and three non-linear processes.

DGP1: Intra-dependence and inter-independence. Linear relation.

$$y = \rho W y + \varepsilon. \tag{36}$$

DGP2 Intra-independence and inter-dependence. Linear relation.

$$y = \beta x + \theta W x + \varepsilon. \tag{37}$$

DGP3 Intra-dependence and inter-dependence. Linear relation.

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

DGP4 Intra-dependence and inter-independence. Non-linear relation.

$$y = {}^{1}\!/\!\left[(I-\rho W)^{-1}\varepsilon\right]. \tag{39}$$

DGP5 Intra-independence and inter-dependence. Non-linear relation.

$$y = {}^{1}/_{(\beta x + \theta W x + \varepsilon)}. \tag{40}$$

DGP6 Intra-dependence and inter-dependence. Non-linear relation.

$$y = {}^{1}/_{\left[(I - \rho W)^{-1}(\theta W x + \varepsilon)\right]}. \tag{41}$$

Three values for the R^2 coefficient have been used: $R^2 \in \{0.4; 0.6; 0.8\}$. The values of ρ, the spatial correlation coefficient, are $\rho \in \{0.4; 0.7; 0.9\}$. Parameter β is fixed at 0.5. The values of θ are obtained as:

$$\theta = \sqrt{\frac{(m-1)\left(\beta^2\left(1 - R^2\right) - R^2\right)}{R^2 - 1}}. \tag{42}$$

In the case of **DGP3**, coefficient ρ has been fixed to 0.5 and θ has been obtained as in (42), where $\beta = 0$.

Note that **DGP4-DGP6** are just the inverses of the corresponding linear models **DGP1-DGP3**. The reason for this (apparently) strange specification is that we would like to use strongly non-linear models to better appreciate the differences between the statistics.

The most important parameters of the simulation are m, which defines the embedding dimension, ρ, which defines the intensity of the spatial intra-dependence and the R^2 which defines the intensity of the relations. Each experiment has been repeated 400 times.

4.2 Results of the Monte Carlo

First, we present the results corresponding to the most restrictive null hypothesis, independence and *i.i.d.*. Then we discuss the results for the null of independence between the series.

4.2.1 Null hypothesis: independence and *i.i.d.*

Table 1 shows the estimated size for the two tests, $\hat{I}_{xy}$ and $\hat{\Psi}_1$. The results are satisfactory although there is a slight tendency to underestimate the size in small samples, especially for the $\hat{I}_{xy}$ statistic. The restriction that, for each symbol, we must have at least 5 observations has consequences: for a sample size $L = 100$, we can only consider the case of $m = 4$.

[Table 1 about here]

$DGP1$ presents intra-dependence and the results appear in Table 2. This case falls under the alternative hypothesis of the bivariate Moran's $\hat{I}_{yx}$ test of not to be *i.i.d.* (the clause of independence is true). The behaviour of $\hat{I}_{yx}$ is very poor, with a maximum estimated power of 44.75% . Results are satisfactory for the $\hat{\Psi}_1$ test: power increases rapidly with the sample size and reacts very quickly to the symptoms of spatial dependence.

[Table 2 about here]

Table 3 shows the estimated power functions for $DGP2$ (inter-dependence only). The estimated power of the two tests is quite good, attaining the maximum value, 100, for samples of medium size. The estimated power of the $\hat{I}_{yx}$ test tends to be higher, especially for small sample sizes. The distance that separates the two tests diminishes as the R^2 coefficient increases.

[Table 3 about here]

For $DGP3$ (intra- and inter-dependence), the results of Table 4 are even better, with an estimated power of practically 100% in all cases.

[Table 4 about here]

The estimated power worsens for the non-linear processes. Tables 5, 6 and 7 show the results for $DGP4$ (intra-dependence only), $DGP5$ (inter-dependence only) and $DGP6$ (intra-

and inter-dependence), respectively. In all cases, the estimated power reacts positively to the sample size and to higher values of ρ and of the R^2.

[Table 5 about here]

[Table 6 about here]

[Table 7 about here]

The values of the R^2 included in Tables 6 and 7 are only for informative purposes (the relationship is non-linear); in fact, they measure the relation between the elements of the equation before inverting the right-hand side. As in the linear case, the estimated power of the $\hat{\Psi}_1$ test for non-linear, intra- and interdependence (Table 7), is fully satisfactory even with small sample sizes.

On the contrary, the behaviour of the bivariate Moran's test is very poor. $DGP4$ presents nonlinear intra-dependence in y, x and y being independent. The percentage of rejections is close to the nominal size. The case of the $DGP4$ corresponds also to the alternative hypothesis of Moran's $\hat{I}_{xy}$. The results for the other two nonlinear DGPs are also unacceptable. Tables 5 to 7 confirm that Moran's $\hat{I}_{xy}$ is a spatial correlation coefficient that is not adequate when the relationship is nonlinear.

4.2.2 Null hypothesis: independence

This section discusses the results of the $\hat{\Psi}_2$ bootstrap test for the null hypothesis of independence of (32).

Table 8 shows the rejection frequencies of the $\hat{\Psi}_2$ test for two different specifications of the null hypothesis. When the processes are $i.i.d.$ $\mathcal{N}(0,1)$, the estimated size fluctuates around 5%, with a maximum of 5.75% and a minimum of 3.5%. For the case of a spatial autoregressive (SAR) in y and white noise process in x, the empirical size is slightly higher that the 5% rate.

[Table 8 about here]

The estimated power for the case of linear inter-dependence, $DGP2$, appears in Table 9. The performance of the $\hat{\Psi}_2$ test is not satisfactory for the small sample case, $L = 100$, although there is a clear improvement for $L = 400$. As expected, the power is close to 100% for large sample sizes, $L = 1000$. A similar pattern emerges for the case of inter-and intra-dependence, $DGP3$, in Table 10.

[Table 9 about here]

[Table 10 about here]

Tables 11 and 12 show the results for non-linear models. The estimated power of the $\hat{\Psi}_2$ test decreases in all the cases. $DGP5$ in Table 11 corresponds to the case of inter-dependence between the series which are intra-independent. The estimated power for small sample sizes is very low, always below the 20% rate of correct decisions. The power increases for $L = 400$ and attains acceptable values for large sample sizes. In all cases, a stronger relation between the variables (measured through the R^2 coefficient) improves the power of the test.

[Table 11 about here]

Finally, Table 12 summarizes the behaviour of the test for the case of nonlinear inter- and intradependence, $DGP6$. The situation does not change: bad results for small samples which improve as the sample size increases. The impact of the R^2, as a measure of association between the variables, is not very clear. It appears to be a kind of trade-off: if the R^2 increases there is a beneficial effect, but the length of the m-surrounding, m, has a clear negative impact.

[Table 12 about here]

In order to better appreciate these results, we would like to add that the inverse nonlinear transformations of DGPs 4 to 6 are the worst options for our testing procedure (the same applies for the case of the bivariate Moran's $\hat{I}_{xy}$). Better results were obtained for all the tests with other nonlinear specification, that are closer to linearity.

5 An application to the Ebro Valley

In this section, we focus on the productivity of the labour factor which, in the long run, '*is almost everything*', according to Krugman (1997, p 9). The literature on Regional Economics highlights the importance of externalities, as interaction mechanisms which affect the evolution of productivity in different and, sometimes distant, areas. In general terms, the literature distinguishes four wide categories of external effects:

(i)- Urban externalities, which have to do with the role of cities in today's economy. Jacobs (1984) points to the importance of the diversity of the urban network. The evidence is not conclusive with respect to size (Rosenthal and Strange, 2004) although it does seem to be so in relative terms. Ciccone and Hall (1996), Ciccone (2002) and Fingleton (2003) find a positive effect of the density of employment on productivity.

(ii)- Location externalities or MAR externalities (in honour of Marshall, 1890, Arrow, 1962, Romer, 1990). Part of the literature emphasizes the importance, for a particular industry, of the spatial concentration: physical concentration generates knowledge externalities, reduces transport and communication costs, stimulates competition in the supply markets and improves the qualification in local labour markets.

(iii)- Competition externalities. Porter (1998) claims that the need to remain competitive is the main incentive for firms to innovate. Pressure from competitors leads firms to better absorb new technologies, increase their innovative capacity and raise their productivity. The evidence provided by Glaeser *et al.* (1992) seems robust.

(iv)- Labour externalities, alluding to the need for the employment demand profile of a

territory to match the supply profile. Rice *et al.* (2006) and Erikson and Lindgren (2009) stressed that a good balance between the two profiles is fundamental for achieving permanent improvements in productivity.

We are interested in verifiying whether, effectively, the productivity data are related to the habitual indicators of these extenalities. We will use data from the manufacturing sector of the municipalities of the Ebro valley, in the Iberian Peninsula. We limit the study to the period 2007-2009; we use the three-years average.

The Ebro valley is a depression situated in the northeast of the Iberian Peninsula that follows the course of the River Ebro from its central part to its mouth in the delta of Amposta. In this macro-region, there are 2,125 municipalities distributed among four Autonomous Communities: Aragón, Cataluña, La Rioja and Navarra. They make up a little more than 20% of the Spanish economy, with a GDP per capita that is 15% above the average. The municipalities of the valley tend to be small with a surface area of 44.8 Km^2 and a population of 4,500 on average (for a more detailed analysis, see Angulo *et al.*, 2012).

Most of the information comes from SABI, *Sistema de análisis de balances ibéricos* (Bureau van Dijk, 2012), complemented with other sources. SABI is a database that contains information of a representative directory of firms that operate in the Iberian Peninsula. The information in which we are interested (basically, figures about sales and employment) refer to firms from the manufacturing sector which, later, have been grouped by municipality. The variables that we are going to employ are the following:

- PM. The labour productivity of the manufacturing sector of municipality i, obtained by dividing the sales (in real terms, base 100=2008) declared by the firms of the sector established in municipality i, in thousands of euros, by the total number of jobs in the same firms (Source SABI, 2012).

- PO. Population residing in municipality i (Source Instituto Nacional de Estadística, INE, 2012). This is an indicator of urban externalities.

- ED. Employment density, measured as the number of jobs per Km^2 in municipality i. This is an indicator of urban externalities.

- QL. Location coefficient for the manufacturing sector in municipality i obtained as:

$$QL_{ir} = \frac{\left(\frac{e_i^M}{e_\bullet^M}\right)}{\left(\frac{e_i^\bullet}{e_\bullet^\bullet}\right)}, \tag{43}$$

where e_i^M is the employment of the manufacture sector in municipality i. A dot, $\bullet$, means aggregate. This is an indicator of location externalities.

- CC. Competition coefficient defined as:

$$QL_{ir} = \frac{\left(\frac{\omega_i^M}{e_\bullet^M}\right)}{\left(\frac{\omega_i^\bullet}{e_\bullet^\bullet}\right)}, \tag{44}$$

where ω_i^M is the number of firms of the manufacturing sector in municipality i. This is an indicator of Porter externality.

- ER. Employment rate, obtained as the quotient between total employment and working population in municipality i. This is an indicator of labour externalities.

Figure 3 shows the municipal distribution of the indicators. There is a strong structure in these data, as the information in Table 13 corroborates: 10 of the 15 linear correlation coefficients and the six Moran coefficients of spatial correlation are significant. The relationships that we detect in all cases are positive, as theory predicts.

[FIGURE 3 about here]

In sum, space is a relevant factor to explain the municipal distribution of the productivity data in the manufacturing sector in the Ebro valley and externalities appear to play a determinant role.

[Table 13 about here]

Table 14 shows the results of the tests of Independence between spatial series carried out in Sections 2 and 3. It should be remembered that not all the test have the same null hypothesis. The tests $\hat{\Psi}_1$ and the bivariate Moran's $\hat{I}_{yx}$, Geoda version, have a composite hypothesis of intra-independence (for each of the two series) and inter-independence (between the two series). The null hypothesis of test $\hat{\Psi}_2$ is that the two series are independent, whatever the spatial structure of them.

[Table 14 about here]

These results are interesting in the sense that the composite tests ($\hat{\Psi}_1$ and $\hat{I}_{yx}$) detect a strong dependence between data on productivity and the indicators of externalities. However, if we take out the spatial structure of both group of variables we conclude, with the test $\hat{\Psi}_2$, that this dependence is, to a great extend, spurious. In fact, only for the case of the location coefficiente, QL, a measure of location or MAR externalities, we obtain a statistically significant relation with labour productivity.

For the manufacturing sector in the Ebro valley, neither urban externalities nor labour externalities nor competition externalities a la Porter seem to have a significant relation with the productivity data. Obviously, if they have no relation, beyond the similarity of their spatial distributions, they could hardly affect (in the sense of *causing*) the productivity results. The hypothesis of dependence cannot be rejected for the productivity-spatial concentration case of the manufacturing sector.

We believe that this result is important in itself because it indicates that not all the externalities have equal importance: to explain the productivity of the manufacturing sector in the Ebro valley, we should focus on location externalities. Logically, the next step is to try to identify the direction of the causal mechanism between the two variables (if it exists). In fact, either interpretation is possible: higher productivity in a municipality may help attract manufacturing firms or, alternatively, a concentration of firms in a municipality may improve the productivity of the labour factor due to specialization.

6 Conclusions

There is an abundant literature devoted to the issue of spatial autocorrelation, a '*hot topic*' in Spatial Econometrics as stated by Anselin (1988). However, the references for the case of cross-correlation analysis between variables are very limited. To our knowledge, we can only cite the bivariate Moran's I_{xy}. This is a good statistic, simple, intuitive and powerful, but has several limitations: it is restricted to pairs of series, the relation must be linear and a weighting matrix is needed to solve the test. Moreover, the exact meaning of the null hypothesis of the test is not totally clear.

The $\Upsilon(m)$ statistic developed in this paper does not suffer from these shortcomings: it is not restricted to a specific functional form, it can be generalized to the case of more than two variables, these variables may be of a discrete nature and it does not need a weighting matrix. The test is consistent and computationally simple to obtain.

In our view, one of the most important properties of the $\Upsilon(m)$ test is that it can be easily adapted to the null hypothesis of interest. As shown in the paper, it is possible to test for three different hypotheses within the same framework: (i) the series are *i.i.d.* and independent (ii) the bivariate series are non spatially autocorrelated and (iii) the series are independent. The Monte Carlo results that we report are convincing in favour of the $\Upsilon(m)$ test, especially when nonlinearities can be present in the models.

Our impression is that the $\Upsilon(m)$ test could be a useful technique for applied research with spatial data. By way of example, we have discovered that the productivity of the manufacturing sector in the Ebro Valley, during the last decade, is only related to location externalities. In fact, urban, labour or externalities 'a-la-Porter' have been discarded by our test. Finally, we must acknowledge that there remain aspects that need further attention, including a more thorough study of non-linear patterns.

Supplementary Materials Web Appendices A, B and C referred to in Section 3 will be available with this paper at the journal website

References

[1] Angulo, A., J. Mur, F. López and M. Herrera (2012): Un análisis de concentración geográfica de las empresas del valle del Ebro atendiendo a sus niveles de productividad. Working Paper. Fundación de Economía Aragonesa.

[2] Anselin, L. (1988). Spatial Econometrics. Methods and Models. Kluwer Academic, Dordrecht.

[3] Anselin, L., Syabri, I., and Smirnov, O. (2002). Visualizing multivariate spatial correlation with dynamically linked windows. In Anselin, L. and Rey, S., editors, New Tools for Spatial Data Analysis: Proceedings of the Specialist Meeting. Center for Spatially Integrated Social Science (CSISS), University of California, Santa Barbara. CD-ROM.

[4] Arrow, K. (1962): The economic implications of learning by doing. Review of Economic Studies, 29, 155-173.

[5] Bureau van Dijk (2012): *Sistema de análisis de balances ibéricos,* SABI. http://sabi.bvdep.com

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

[7] Burridge, P (2012): Improving the J Test in the SARAR Model by Likelihood based Estimation. Spatial Economic Analysis, 7, 75-107

[8] Carlstein, E. (1986): The use of subseries methods for estimating the variance of a general statistic from a stationary time series. The Annals of Statistics, 14, 1171-1179.

[9] Ciccone, A. (2002): Agglomeration effects in Europe, European Economic Review. 46, 213–228.

[10] Ciccone, A. and R. Hall (1996): Productivity and the density of economic activity. American Economic Review, 86, 54–70.

[11] Cliff, A. and K. Ord (1981): Spatial Processes: Models and Applications. Pion, London.

[12] Czaplewski, R. and R. Reich (1993): Expected Value and Variance of Moran's I Bivariate Spatial Autocorrelation Statistic for a Permutation Test. USDA, Forest Service. Research Paper RM-309.

[13] Eriksson, R. and Lindgren, U. (2009): Localized mobility clusters: impacts of labour market externalities on firm performance. Journal of Economic Geography, 9 33-53.

[14] Fingleton, B. (2001): Spurious Spatial Regression: Some Monte Carlo Results with a Spatial Unit Root and Spatial Cointegration. Journal of Regional Science, 39 1-19.

[15] Fingleton, B. (2003): Increasing returns; evidence from local wage rates in Great Britain. Oxford Economic Papers, 55, 716–739.

[16] Fingleton, B. (2005): Beyond neoclassical orthodoxy: A view based on the new economic geography and UK regional wage data. Papers in Regional Science, 84 351–375

[17] Fingleton, B. (2008:) A Generalized Method of Moments Estimator for a Spatial Model with Moving Average Errors, with Application to Real Estate Prices. Empirical Economics, 34, 35-57.

[18] Fingleton, B. and J. LeGallo (2008): Estimating Spatial Models with Endogenous Variables, a Spatial Lag and Spatially Dependent Disturbances: Finite Sample Properties. Papers in Regional Science, 87, 319-339.

[19] GeoDa (2012): GeoDa Center for Geospatial Analysis and Computation. https://geodacenter.asu.edu/

[20] Glaeser, E., Kallal, H., Scheinkman J. and Shleifer, A. (1992): Growth in Cities. The Journal of Political Economy, 100, 1126-1152.

[21] Goodchild, M. and D. Janelle (2003). Spatially Integrated Social Science. Oxford University Press, Oxford.

[22] Hall, P. (1985): Resampling a coverage pattern. Stochastic Processes and Their Applications, 20, 231-246.

[23] Han, X. and Lee, L. (2012): Model selection using J-test for the spatial autoregressive model vs. the matrix exponential spatial model. Regional Science and Urban Economics (forthcoming)

[24] Hao, B. and W. Zheng (1998): Applied symbolic dynamics and chaos. World Scientific, Singapore.

[25] Instituto Nacional de Estadística, INE (2012): Estimaciones de Población Actual. http://www.ine.es

[26] Jacobs, J. (1984): Cities and the wealth of nations: principles of economic life, Vintage, New York.

[27] Krugman, P. (1997): The Age of Diminished Expectations. U.S. Economic Policy in the 1990s. Second Edition, MIT Press, Cambridge.

[28] Lawson, A. (2006): Statistical Methods in Spatial Epidemiology, 2nd Edition. Springer, Berlin

[29] Lesage, J. and K. Pace (2009). Introduction to Spatial Econometrics. Chapman & Hall/CRC, Boca Raton.

[30] López, F., Matilla-García, M., Mur, J., and M. Ruiz Marín (2010): A non-parametric spatial independence test using symbolic entropy. Regional Science and Urban Economics, 40, 106-115.

[31] Mantel, N. (1967): The Detection of Disease Clustering and a Generalized Regression Approach. Cancer Research, 27, 209-220.

[32] Marshall, A. (1890): Principles of Economics, 8th ed. Macmillan, London.

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

[34] Matilla-García, M. and M. Ruiz Marín (2009): Detection of non-linear structure in time series. Economics Letters, 105, 1-6.

[35] Monchuk, D., D. Hayes, J. Miranowski, and D. Lambert. (2011): Inference Based on Alternative Bootstrapping Methods in Spatial Models with an Application to County Income Growth in the United States. Journal of Regional Science, 51, 880-896.

[36] Mur, J. and J. Trivez (2003): Unit Roots and Deterministics Trends in Spatial Econometric Models. International Regional Science Review, 26, 289-312.

[37] Pinkse, J. and M. Slade (1998): Contracting in space: An application of spatial statistics to discrete-choice models. Journal of Econometrics, 85, 125-154

[38] Porter, M. (1998): The Competitive Advantage of Nations, Free Press, London.

[39] Reich, R., R. Czaplewski, and W. Bechtold (1994): Spatial cross-correlation in growth of undisturbed natural shortleaf pine stands in northern Georgia. Journal of Environmental and Ecological Statistics. 1, 201-217.

[40] Rice, P, Venables, A. and Patacchini, E. (2006): Spatial determinants of productivity: analysis for the regions of Great Britain. Regional Science and Urban Economics, 36, 727-752.

[41] Rohatgi, V. (1976): An introduction to probability theory and mathematical statistics. Wiley, New York.

[42] Romer, P. (1990): Endogenous Technological Change, Journal of Political Economy, 98, 71-102.

[43] Rosenthal, S. and Strange, W. (2004): Evidence on the nature and sources of agglomeration economies. In Henderson and Thisse (Eds.), Handbook of Urban and Regional Economics, vol. 4, 2063-2117.

[44] Ruiz, M., López, F. and A. Páez (2010); Testing for spatial association of qualitative data using symbolic dynamics. Journal of Geographical Systems, 12, 281-309.

[45] Schmitz, O. (2010): Resolving Ecosystem Complexity. Oxford University Press, Oxford.

[46] Soon, S. (1996): Binomial approximation for dependent indicators. Statistica Sinica, 6, 703–714.

[47] Upton, G. and B. Fingleton (1985): Spatial data analysis by example. Wiley, New York.

[48] Wartenberg, D. (1985): Multivariate Spatial Corrrelation: A Method for Exploratory Geographical Analysis. Geographical Analysis, 17, 263-283.

Table 1: Estimated Size of $\hat{\Psi}_1$ and $\hat{I}_{xy}$ tests at 5% level

	$\hat{\Psi}_1$			$\hat{I}_{xy}$		
m	4	6	8	4	6	8
$L = 100$	4.25	–	–	3.50	-	-
$L = 400$	4.25	4.50	5.00	3.50	5.00	5.75
$L = 1000$	3.00	4.25	5.50	4.25	4.50	5.75

Note: Permutations: 399. Replications: 400.

Table 2: Estimated Power of $\hat{\Psi}_1$ and $\hat{I}_{xy}$ tests at 5% level

$DGP1$	$L=100$	$L=400$			$L=1000$		
m	4	4	6	8	4	6	8
$\hat{\Psi}_1$							
$\rho=0.4$	31.25	95.25	78.00	71.75	100.00	98.50	97.75
$\rho=0.7$	96.25	100.00	100.00	100.00	100.00	100.00	100.00
$\rho=0.9$	100.00	100.00	100.00	100.00	100.00	100.00	100.00
$\hat{I}_{xy}$							
$\rho=0.4$	8.50	10.50	12.75	11.75	13.00	11.25	13.75
$\rho=0.7$	19.75	19.50	24.75	29.00	18.50	25.75	31.00
$\rho=0.9$	30.75	27.25	34.25	44.75	32.25	33.75	36.25

Note: Permutations: 399. Replications: 400.

Table 3: Estimated Power of $\hat{\Psi}_1$ and $\hat{I}_{xy}$ tests at 5% level

$DGP2$	$L = 100$	$L = 400$			$L = 1000$		
m	4	4	6	8	4	6	8
			$\hat{\Psi}_1$				
$R^2 = 0.4$	39.00	97.00	99.00	99.00	100.00	100.00	100.00
$R^2 = 0.6$	70.50	100.00	100.00	100.00	100.00	100.00	100.00
$R^2 = 0.8$	83.50	100.00	100.00	100.00	100.00	100.00	100.00
			$\hat{I}_{xy}$				
$R^2 = 0.4$	100.00	100.00	100.00	100.00	100.00	100.00	100.00
$R^2 = 0.6$	100.00	100.00	100.00	100.00	100.00	100.00	100.00
$R^2 = 0.8$	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Note: Permutations: 399. Replications: 400.

Table 4: Estimated Power of $\hat{\Psi}_1$ and $\hat{I}_{xy}$ tests at 5% level

$DGP3$	$L=100$	$L=400$				$L=1000$			
m	4	4	6	8		4	6	8	
$\hat{\Psi}_1$									
$R^2=0.4$	67.50	100.00	100.00	100.00		100.00	100.00	100.00	
$R^2=0.6$	92.75	100.00	100.00	100.00		100.00	100.00	100.00	
$R^2=0.8$	95.75	100.00	100.00	100.00		100.00	100.00	100.00	
$\hat{I}_{xy}$									
$R^2=0.4$	100.00	100.00	100.00	100.00		100.00	100.00	100.00	
$R^2=0.6$	100.00	100.00	100.00	100.00		100.00	100.00	100.00	
$R^2=0.8$	100.00	100.00	100.00	100.00		100.00	100.00	100.00	

Note: Permutations: 399. Replications: 400.

Table 5: Estimated Power of $\hat{\Psi}_1$ and $\hat{I}_{xy}$ tests at 5% level

$DGP4$	$L=100$	$L=400$			$L=1000$		
m	4	4	6	8	4	6	8
$\hat{\Psi}_1$							
$\rho=0.4$	14.75	72.00	59.25	49.75	99.50	93.75	89.50
$\rho=0.7$	68.50	99.75	97.50	94.25	100.00	100.00	100.00
$\rho=0.9$	99.50	100.00	100.00	100.00	100.00	100.00	100.00
$\hat{I}_{xy}$							
$\rho=0.4$	4.25	4.00	6.00	3.25	5.75	5.25	6.25
$\rho=0.7$	4.75	5.25	5.00	4.50	3.00	4.50	4.75
$\rho=0.9$	6.75	6.75	4.75	7.00	6.00	3.00	5.50

Note: Permutations: 399. Replications: 400.

Table 6: Estimated Power of $\hat{\Psi}_1$ and $\hat{I}_{xy}$ tests at 5% level

$DGP5$	$L=100$	$L=400$			$L=1000$		
m	4	4	6	8	4	6	8
$\hat{\Psi}_1$							
$R^2=0.4$	18.25	82.00	84.75	87.75	98.50	99.50	99.75
$R^2=0.6$	36.50	96.00	97.50	97.00	100.00	100.00	100.00
$R^2=0.8$	48.75	99.25	100.00	99.75	100.00	100.00	100.00
$\hat{I}_{xy}$							
$R^2=0.4$	5.00	4.50	4.50	6.75	4.50	5.75	5.50
$R^2=0.6$	4.50	6.25	5.50	4.50	4.75	5.75	4.50
$R^2=0.8$	4.25	5.25	5.75	5.75	5.00	6.25	5.75

Note: Permutations: 399. Replications: 400.

Table 7: Estimated Power of $\hat{\Psi}_1$ and $\hat{I}_{xy}$ tests at 5% level

$DGP6$	$L = 100$	$L = 400$			$L = 1000$		
m	4	4	6	8	4	6	8
			$\tilde{\Psi}_1$				
$R^2 = 0.4$	61.50	100.00	99.25	98.25	100.00	100.00	100.00
$R^2 = 0.6$	79.25	100.00	100.00	100.00	100.00	100.00	100.00
$R^2 = 0.8$	85.25	100.00	100.00	100.00	100.00	100.00	100.00
			$\tilde{I}_{xy}$				
$R^2 = 0.4$	4.00	4.50	5.00	6.75	5.25	6.25	5.50
$R^2 = 0.6$	5.25	5.50	4.50	6.00	4.75	6.75	6.00
$R^2 = 0.8$	5.75	5.50	6.50	7.50	4.25	4.25	6.00

Note: Permutations: 399. Replications: 400.

Table 8: Estimated Size of $\hat{\Psi}_2$ test at 5% level

D.G.P.	$y \sim \mathcal{N}(0,1)$; $x \sim \mathcal{N}(0,1)$.			$y = 0.5y + \varepsilon$; $\varepsilon \sim \mathcal{N}(0,1)$; $x \sim \mathcal{N}(0,1)$.		
m	4	6	8	4	6	8
$R = 400$	5.00	3.50	5.50	5.25	6.25	9.0
$R = 1000$	5.75	4.50	4.50	5.00	5.25	7.75

Note: Number of blocks in the SSB $b = 8$

Number of bootstraps: 399. Replications: 400.

Table 9: Estimated Power of $\hat{\Psi}_2$ test at 5% level

$DGP2$	$R = 100$	$R = 400$			$R = 1000$		
m	4	4	6	8	4	6	8
$R^2 = 0.4$	18.50	59.00	62.25	66.75	94.50	96.00	97.00
$R^2 = 0.6$	26.75	75.50	76.75	74.25	100.00	99.75	98.00
$R^2 = 0.8$	32.25	89.50	81.50	73.75	99.75	99.75	99.50

Note: Number of blocks in the SSB, $b = 4$ if $R = 100$ and $b = 8$ if $R > 100$.

Number of bootstraps: 399. Replications: 400.

Table 10: Estimated Power of $\hat{\Psi}_2$ test at 5% level

$DGP3$	$L = 100$	$L = 400$			$L = 1000$		
m	4	4	6	8	4	6	8
$R^2 = 0.4$	7.75	21.75	30.25	33.75	56.00	69.75	74.50
$R^2 = 0.6$	12.50	51.00	48.50	46.50	88.75	92.25	89.75
$R^2 = 0.8$	15.75	70.25	54.75	45.25	99.50	97.00	94.50

Note: Number of blocks in the SSB, $b = 4$ if $R = 100$ and $b = 8$ if $R > 100$.
Number of bootstraps: 399. Replications: 400.

Table 11: Estimated Power of $\hat{\Psi}_2$ test at 5% level

$DGP5$	$L = 100$	$L = 400$			$L = 1000$		
m	4	4	6	8	4	6	8
$R^2 = 0.4$	12.75	33.75	42.75	49.00	81.50	82.25	86.50
$R^2 = 0.6$	16.75	42.75	45.75	48.25	86.00	87.50	88.00
$R^2 = 0.8$	12.00	52.75	51.50	46.00	96.25	91.00	88.25

Note: Number of blocks in the SSB, $b = 4$ if $R = 100$ and $b = 8$ if $R > 100$.

Number of bootstraps: 399. Replications: 400.

Table 12: Estimated Power of $\hat{\Psi}_2$ test at 5% level

$DGP6$	$L = 100$	$L = 400$			$L = 1000$		
m	4	4	6	8	4	6	8
$R^2 = 0.4$	9.00	25.75	28.00	33.25	57.25	64.00	67.75
$R^2 = 0.6$	6.00	27.00	23.50	27.25	53.50	56.75	63.00
$R^2 = 0.8$	8.00	26.50	23.50	21.25	66.25	56.50	55.25

Note: Number of blocks in the SSB, $b = 4$ if $R = 100$ and $b = 8$ if $R > 100$.

Number of bootstraps: 399. Replications: 400.

Table 13: Productivity and externalities in the Ebro Valley. (2007-2009)

Linear Correlation Coefficients						
	PM	PO	ER	ED	QL	CC
PM	1	$0.101^{(*)}$	$0.109^{(*)}$	0.158	0.652	$0.3101^{(*)}$
PO		1	$0.993^{(*)}$	0.410	0.016	-0.010
ER			1	$0.457^{(*)}$	$0.019^{(*)}$	-0.012
ED				1	$0.040^{(*)}$	0.011
QL					1	$0.206^{(*)}$
CC						1
Moran's I coefficient of spatial dependence						
Moran's I	$7.425^{(*)}$	$20.905^{(*)}$	$8.859^{(*)}$	$36.131^{(*)}$	$18.480^{(*)}$	$3.363^{(*)}$

(*) An asterisk means that the statistic is significative at a 5% level.

Table 14: Bivariate measures of spatial correlation of the productivity in the Manufacturing sector in the Ebro valley. Period: 2007-2009.

	$\hat{\Psi}_1$	$\hat{\Psi}_1^b(0.05)$	$\hat{\Psi}_2$	$\hat{\Psi}_2^b(0.05)$	$\hat{I}_{yx}$	$\hat{I}_{yx}^b(0.025)$	$\hat{I}_{yx}^b(0.975)$
PM-PO	$0.572^{(*)}$	0.231	5.338	14.760	$0.069^{(*)}$	0.033	-0.029
PM-ER	$0.579^{(*)}$	0.332	5.380	14.865	$0.077^{(*)}$	0.040	-0.039
PM-ED	$0.503^{(*)}$	0.145	4.530	14.630	$0.199^{(*)}$	0.058	-0.078
PM-QL	$1.619^{(*)}$	0.230	$15.654^{(*)}$	12.725	$0.173^{(*)}$	0.044	-0.045
PM-CC	$1.572^{(*)}$	0.188	5.514	12.787	$0.039^{(*)}$	0.022	-0.023

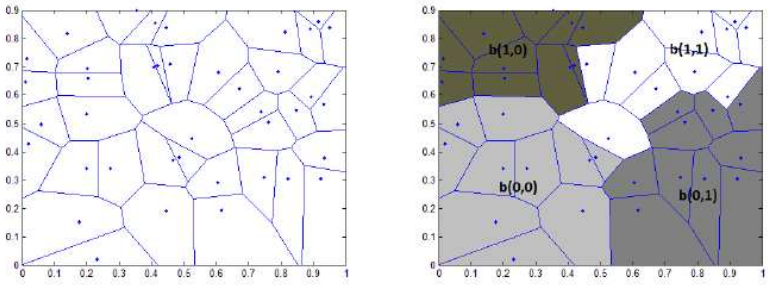
Number of blocks in the SSB, $b = 8$.The embedding dimension used is equal to 6, $m = 6$.
An asterisk, (*), means that the statistic is significative at a 5% level. b means bootstrapped critical values, after 199 boots, at the significance value indicated between brackets. Note that $\hat{\Psi}_1$ and $\hat{\Psi}_2$ are one-sided test whereas Moran's bivariate $\hat{I}_{yx}$ is a two-sided test.

Figure 1: Example of Regular Lattice 3×3 for x_s and y_s

$X_{s_1} = 4$	$X_{s_2} = 1$	$X_{s_3} = 3$
$X_{s_4} = 6$	$X_{s_5} = 2$	$X_{s_6} = 5$
$X_{s_7} = 1$	$X_{s_8} = 2$	$X_{s_9} = 4$

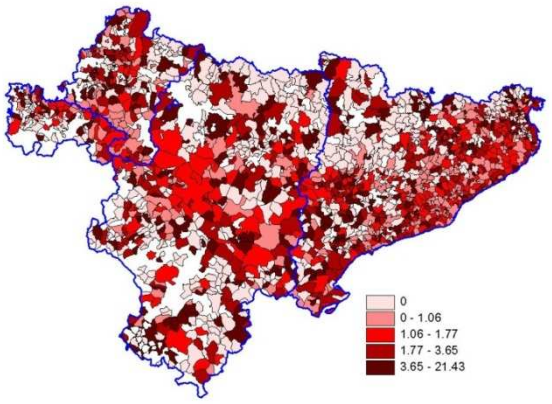
$Y_{s_1} = 5$	$Y_{s_2} = 2$	$Y_{s_3} = 4$
$Y_{s_4} = 0$	$Y_{s_5} = 2$	$Y_{s_6} = 3$
$Y_{s_7} = 7$	$X_{s_8} = 9$	$Y_{s_9} = 3$

Figure 2: Spatial Block Bootstrapping

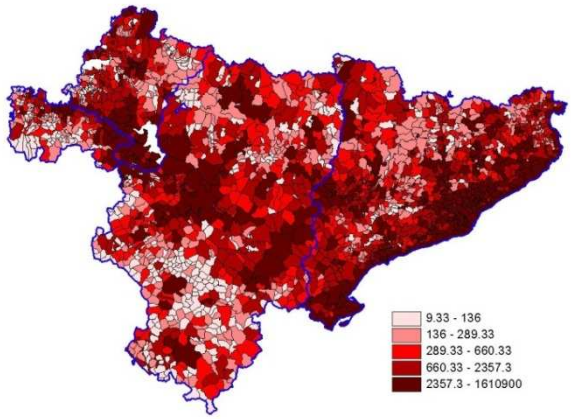


(a) Original Sample *(b) Non-Overlapping Blocks*

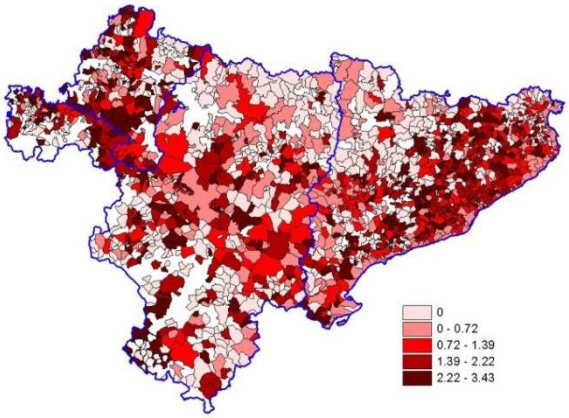
Figure 3: Productivity and externality indicators by municipalities in the Ebro valley. Period 2007-2009.



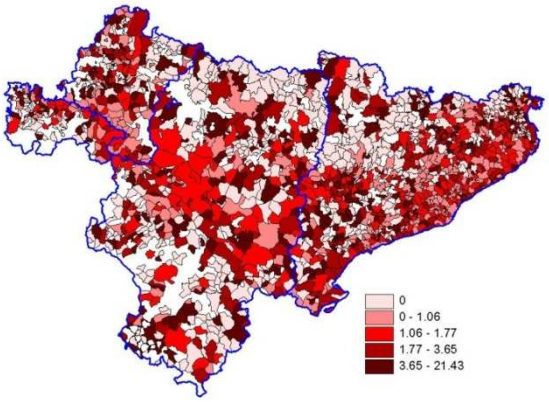
Productivity Manufactures : PM



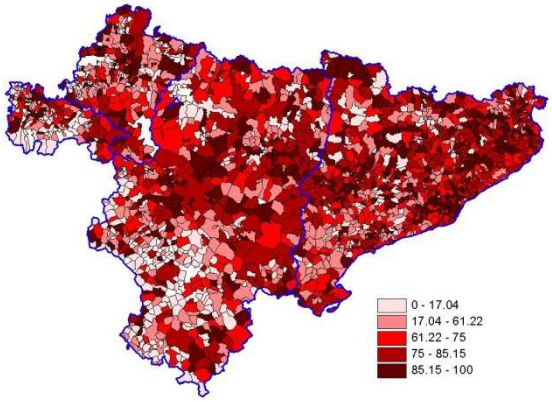
Population : PO



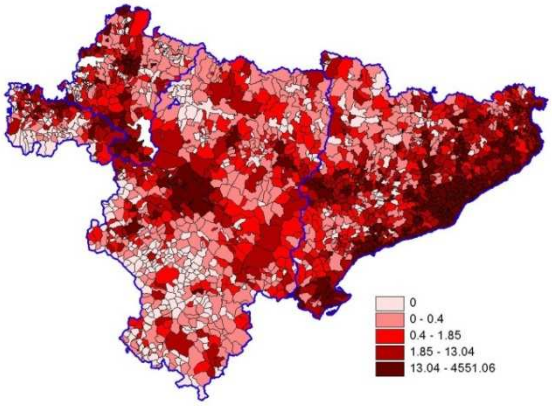
Location Coefficient : QL



Competition Coefficient : CC



Employment Rate : ER



Employment Density : ED