

This article was downloaded by: [Marcos Herrera]

On: 18 December 2013, At: 13:01

Publisher: Routledge

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



## Spatial Economic Analysis

Publication details, including instructions for authors and subscription information:

<http://www.tandfonline.com/loi/rsea20>

### Detecting Dependence Between Spatial Processes

Marcos Herrera, Manuel Ruiz & Jesús Mur

Published online: 16 Dec 2013.

To cite this article: Marcos Herrera, Manuel Ruiz & Jesús Mur (2013) Detecting Dependence Between Spatial Processes, *Spatial Economic Analysis*, 8:4, 469-497, DOI: [10.1080/17421772.2013.835437](http://dx.doi.org/10.1080/17421772.2013.835437)

To link to this article: <http://dx.doi.org/10.1080/17421772.2013.835437>

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at <http://www.tandfonline.com/page/terms-and-conditions>

## Detecting Dependence Between Spatial Processes

MARCOS HERRERA, MANUEL RUIZ & JESÚS MUR

(Received March 2012; accepted February 2013)

**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,  $\Upsilon(\mathbf{m})$ , based on symbolic dynamics with better properties. We show that the  $\Upsilon(\mathbf{m})$  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.*

### Détection de la dépendance entre procédés spatiaux

**RÉSUMÉ** *Il est important de tester l'hypothèse de l'indépendance entre variables spatiales. Toutefois, dans les ouvrages existants, nous ne pouvons mentionner que les statistiques de Moran à deux variables, qui font l'objet de plusieurs restrictions. Applicable à des paires de variables: nous avons besoin de linéarité et d'une matrice de pesage; l'hypothèse nulle n'est pas entièrement claire. Nous présentons un nouvel essai non paramétrique,  $\Upsilon(\mathbf{m})$ , présentant de meilleures propriétés: il peut être étendu à un cadre à variables multiples, il est résistant aux écarts par rapport à la linéarité, et il est flexible au nul. Le test présente de bonnes caractéristiques de taille et puissance. Une application au cas de notre productivité illustre notre approche.*

Marcos Herrera, University of Zaragoza, Economic Analysis, Zaragoza, Spain. Email: [mherrera@unizar.es](mailto:mherrera@unizar.es).  
Manuel Ruiz, Technical University of Cartagena, Computing and Quantitative Methods, Cartagena, Spain. Email: [manuel.ruiz@upct.es](mailto:manuel.ruiz@upct.es).  
Jesús Mur (to whom correspondence should be sent), University of Zaragoza, Economic Analysis, Gran Via, 2-4, Zaragoza, 50005 Spain. Email: [jmur@unizar.es](mailto:jmur@unizar.es). We highly appreciate the constructive comments received from many colleagues, especially Jeff Racine, Antonio Aznar, Cem Ertur, Jean Paelinck, Roberto Basile, Mariano Matilla, Fernando López and two anonymous reviewers. This article has been supported by the projects ECO2012-36032-C03-01 and ECO2012-36032-C03-02 of the Ministerio de Educación, Cultura y Deporte del Reino de España.

## Detección de dependencia entre procesos espaciales

EXTRACTO Comprobar la hipótesis de independencia entre variables espaciales es importante. No obstante, en la literatura solo podemos mencionar la estadística bivariada de Moran que sufre de varias restricciones: es aplicable a pares de variables y necesitamos linealidad y una matriz de ponderaciones; la hipótesis nula no está totalmente clara. Presentamos un nuevo test no paramétrico,  $\Upsilon(m)$ , con mejores propiedades: puede extenderse a un marco multivariante, es robusto frente a incumplimientos del supuesto de linealidad y se adapta bien a diferentes especificaciones de la hipótesis nula. La prueba tiene un tamaño y una potencia satisfactorias. Una aplicación al caso de la productividad ilustra este planteamiento.

### 检测空间过程之间的依赖性

摘要：检验空间变量之间的独立性假设非常重要。但是，在文献中我们只能找到二元 Moran 统计数据，并且它受到以下几种限制：要适用于成对的变量，需要建立线性和加权矩阵；零假设不完全清晰。本文提出了新的非参数检验函数  $\Upsilon(m)$ ，它具有更优的属性：它可以扩展为多元框架，也很稳健，可脱离线性，并且对于零假设具有灵活性。该检验具有合适的规模和功效。本文还通过生产力案例的应用阐明了这种方法。

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 & 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 other 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 & 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:

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

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' (1985, p. 264). Assuming that  $k$  different variables are observed in locations 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 & Ord, 1981; Hall, 1985; Upton & 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 & 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 forms one (or separate identical) population(s), and the joint distribution function(s) for  $Y$  and  $X$  are unknown' (1993, 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:

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

$\rho_{yx}$  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 & 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 (see 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 & 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 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  $Y(m)$ . The  $Y(m)$  test is free from distributional assumptions, it does not 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 & 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  $\Gamma_n = \{\sigma_1, \sigma_2, \dots, \sigma_n\}$ .

Symbolizing a process means defining a map

$$f_x : \{x_s\}_{s \in S} \rightarrow \Gamma_n, \quad (3)$$

such that each element of the series,  $x_s$ , is associated with a unique symbol  $f_x(x_s) = \sigma_{i_s}$  with  $i_s \in \{1, 2, \dots, n\}$ . 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  $\{Z_s\}_{s \in S}$ :

$$Z_s = \{x_s, y_s\}, \quad (4)$$

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

$$g : \{Z_s\}_{s \in S} \rightarrow \Omega_n^2 = \Gamma_n \times \Gamma_n, \quad (5)$$

where

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

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,  $\{Z_s\}_{s \in S} = \{x_{1s}, x_{2s}, \dots, x_{ks}\}$ . Let

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

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

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 & Ruiz, 2008, 2009; López et al., 2010; Ruiz et al., 2010). Let  $M_e^x$  be the median of the spatial series  $\{x_s\}_{s \in S}$ . Define the indicator function

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

For each  $s \in S$ , let  $N_s$  be the set formed by the  $(m - 1)$  neighbours of  $s$ ,  $N_s = \{s_1, \dots, s_{m-1}\}$ , 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:

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

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

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

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:

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

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

In sum, this symbolization process consists of comparing, for each location  $s$ , the value  $\tau_s$  with  $\tau_{s_i}$  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  $\sigma_i^x$  is simply:

$$n_{\sigma_i^x} = \# \{s \in S \mid s \text{ is } \sigma_i^x \text{-type for } x\}, \tag{10}$$

whose relative frequencies can be estimated:

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

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

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

We introduce the concept of symbolic entropy for a *two-dimensional* spatial series  $\{Z_s\}_{s \in S}$  through the Shannon statistic:

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

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

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

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

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

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

where  $C_{\sigma}^{m-1} = (m-1)! / [(m-1-\sigma)! \sigma!]$  denotes the number of  $\sigma$ -combinations of symbols,  $\sigma \in \{0, \dots, m-1\}$ , 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  $3 \times 3$  regular lattice as shown in [Figure 1](#).

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

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

Figure 1. Example of regular lattice  $3 \times 3$  for  $x_s$  and  $y_s$ .

In the first place, we must define the value of  $m$ , the embedding dimension, for example,  $m = 4$ . Then we identify the three nearest neighbours 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 the symbols. In accordance with (7)–(8)–(9), and for variable  $x$  in location  $s_1$ :

$$f_x(x_{s_1}) = (I_{s_1s_2} = 0) + (I_{s_1s_4} = 1) + (I_{s_1s_5} = 0) = 1.$$

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:

$$\left. \begin{aligned} H_0 &: \{x_s\}_{s \in S} \text{ and } \{y_s\}_{s \in S} \text{ are i.i.d. and independent.} \\ H_1 &: \{x_s\}_{s \in S} \text{ and } \{y_s\}_{s \in S} \text{ are not i.i.d. and independent.} \end{aligned} \right\} \quad (17)$$

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  $J_{\eta_s}$  as follows:

$$J_{\eta_s} = \begin{cases} 1 & \text{if location } s \text{ is } \eta\text{-type.} \\ 0 & \text{otherwise.} \end{cases} \quad (18)$$

$J_{\eta_s}$  is a Bernoulli variable with probability of ‘success’  $p_\eta$ , where ‘success’ means that  $s$  is  $\eta$ -type.

This restriction is obvious:

$$\sum_{\eta \in \Omega_n^2} p_\eta = 1. \quad (19)$$

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:

$$Q_\eta = \sum_{s \in S} J_{\eta_s}. \quad (20)$$

$Q_\eta$  is a random variable whose values are integers in the range  $\{0, 1, 2, \dots, L\}$ . Notice that not all  $J_\eta$  variables are independent (due to the possible overlapping of the  $m$ -surroundings) and, therefore,  $Q_\eta$  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 achieved by controlling 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  $J_{\eta_s}$  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  $J_\eta$  indicators is weak, the variable  $Q_\eta$  can be approximated to a Binomial random variable:

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

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

$$P(Q_{\eta_{11}} = a_{11}, \dots, Q_{\eta_{mm}} = a_{mm}) = \frac{(a_{11} + a_{12} + \dots + a_{mm})!}{a_{11}! a_{12}! \dots a_{mm}!} p_{\eta_{11}}^{a_{11}} p_{\eta_{12}}^{a_{12}} \dots p_{\eta_{mm}}^{a_{mm}}, \tag{22}$$

where  $a_{11} + a_{12} + \dots + a_{mm} = L$ .

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

**Theorem 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  $J_\eta$  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*

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

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

Let  $\alpha$  be a real number with  $0 \leq \alpha \leq 1$  and  $\chi_\alpha^2$  is such that  $Pr(\chi_k^2 > \chi_\alpha^2) = \alpha$ . To test

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

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

$$\begin{aligned} \text{If } 0 \leq Y(m) \leq \chi_\alpha^2, & \quad \text{Do not reject } H_0. \\ \text{Otherwise,} & \quad \text{Reject } H_0. \end{aligned}$$

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

$$Y(m) = 2L \left[ k(m-1) \ln(2) - \left[ \sum_{i_1=1}^n \dots \sum_{i_k=1}^n \frac{n_{\eta_{i_1 \dots i_k}}}{L} \ln \left( \prod_{j=1}^k C_{\sigma_{i_j}^{x_j}}^{m-1} \right) \right] - h_Z(m) \right], \tag{25}$$

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

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, secondly, 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 Appendix C for proofs).

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

$$\lim_{L \rightarrow \infty} \Pr(\hat{\Upsilon}(m) > C) = 1,$$

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  $\chi^2$  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  $m \leq \sqrt{L}/5$ . 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  $Q_\eta$  of (20) will be a poor approximation to the Binomial distribution and, consequently, the convergence of  $\Upsilon(m)$  to a  $\chi^2$  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  $\Psi_1 = \frac{1}{L} \Upsilon$ . 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  $p_{\text{permutation-value}}$ :

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

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

$$p_{\text{permutation}} - \text{value}(\hat{\Psi}_1) < \alpha, \quad (27)$$

for a nominal size  $\alpha$ .

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  $p_{\text{permutation}} - \text{value}$ :

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

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

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

for a nominal size  $\alpha$ .

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 enables 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 null that preserves the spatial structure of each series.

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

$$p_{\eta_{ij}} = p_{\sigma_i} p_{\sigma_j}. \quad (30)$$

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:

$$\Psi_2 = h_x(m) + h_y(m) - h_Z(m), \quad (31)$$

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

$$H'_0 : \{x_s\}_{s \in S} \text{ and } \{y_s\}_{s \in S} \text{ are independent.} \quad (32)$$

The distribution function of the  $\Psi_2$  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 & Slade, 1998; Fingleton 2005, 2008; Fingleton & LeGallo, 2008; Monchuk et al., 2011) and many of them focus on the case of the  $J$  test (Burrige & Fingleton, 2010; Burrige, 2012; Han & 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:

- (i) 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 cover the whole spatial system.
- (ii) Divide the two spatial series according to (i).
- (iii) 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  $L$ .
- (iv) 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 (i) 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*: the observations of each block

were ordered according to the distance to the *buoy* of the block; then, the blocks are resampled; finally, 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 2(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 2(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,  $b(0,1)$ .

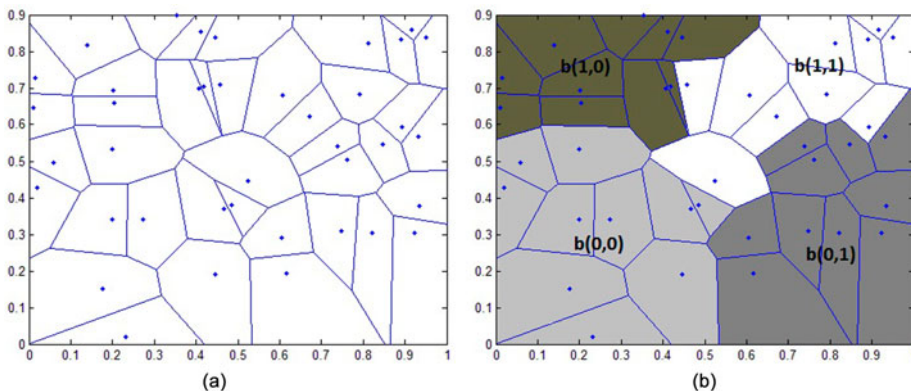
The *SSB* adapts well to problems where the  $W$  matrix has been built according to some measure of distance, as in the  $k$ -nearest neighbours criterion. Obviously, changes in the contiguity criterion should also involve changes in the spatial bootstrapping scheme. The *SSB* will, inevitably, produce some mismatches 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



**Figure 2.** Spatial block bootstrapping. (a) Original sample. (b) Non-overlapping blocks.

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:

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

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

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

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:

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

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.

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

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

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

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

$$y = \rho Wy + \theta Wx + \varepsilon. \tag{38}$$

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

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

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

$$y = \frac{1}{(\beta x + \theta Wx + \varepsilon)}. \tag{40}$$

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

$$y = \frac{1}{[(I - \rho W)^{-1}(\theta Wx + \varepsilon)]}. \tag{41}$$

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

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

In the case of **DGP3**, coefficient  $\rho$  has been fixed to 0.5 and  $\theta$  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,  $\rho$ , 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  $m = 4$ .

**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

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

| $m$        | $\hat{\Psi}_1$ |      |      | $\hat{I}_{xy}$ |      |      |
|------------|----------------|------|------|----------------|------|------|
|            | 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.

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 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.

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

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  $\rho$  and of the  $R^2$ .

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.

**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

| <i>DGP3</i> | <i>L</i> = 100 | <i>L</i> = 400 |                |        | <i>L</i> = 1000 |        |        |  |
|-------------|----------------|----------------|----------------|--------|-----------------|--------|--------|--|
| <i>M</i>    | 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

| <i>DGP4</i>  | <i>L</i> = 100 | <i>L</i> = 400 |                |        | <i>L</i> = 1000 |        |        |  |
|--------------|----------------|----------------|----------------|--------|-----------------|--------|--------|--|
| <i>m</i>     | 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

| <i>DGP5</i> | <i>L</i> = 100 | <i>L</i> = 400 |                |       | <i>L</i> = 1000 |        |        |  |
|-------------|----------------|----------------|----------------|-------|-----------------|--------|--------|--|
| <i>m</i>    | 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.

Conversely, the behaviour of the bivariate Moran’s test is very poor. *DGP4* presents nonlinear intra-dependence in  $\gamma$ ,  $x$  and  $\gamma$  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 *DGP* s 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.

**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      |  |
|             |         |        |        | $\hat{\Psi}_1$ |        |        |          |        |  |
| $R^2 = 0.4$ | 61.50   | 100.00 | 99.25  | 98.25          | 100.00 | 100.00 | 100.00   | 100.00 |  |
| $R^2 = 0.6$ | 79.25   | 100.00 | 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   | 100.00 |  |
|             |         |        |        | $\hat{I}_{xy}$ |        |        |          |        |  |
| $R^2 = 0.4$ | 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.

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  $\gamma$  and white noise process in  $x$ , the empirical size is slightly higher than the 5% rate.

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.

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  $L = 100$  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.

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 8.** Estimated size of  $\hat{\Psi}_2$  test at 5% level

| DGP      | $\gamma \sim \mathcal{N}(0, 1);$<br>$x \sim \mathcal{N}(0, 1).$ |      |      | $\gamma = 0.5\gamma + \varepsilon;$<br>$\varepsilon \sim \mathcal{N}(0, 1);$<br>$x \sim \mathcal{N}(0, 1).$ |      |      |      |
|----------|---|------|------|---|------|------|------|
|          | M   | 4    | 6    | 8   | 4    | 6    | 8    |
| L = 400  |   | 5.00 | 3.50 | 5.50  | 5.25 | 6.25 | 9.0  |
| L = 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.

Downloaded by [Marcos Herrera] at 13:01 18 December 2013

**Table 9.** Estimated power of  $\hat{\Psi}_2$  test at 5% level

| <i>DGP2</i> | <i>L</i> = 100 |       |       | <i>L</i> = 400 |       |        | <i>L</i> = 1000 |       |  |
|-------------|----------------|-------|-------|----------------|-------|--------|-----------------|-------|--|
|             | <i>m</i>       | 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  $L = 100$  and  $b = 8$  if  $L > 100$ .  
 Number of bootstraps: 399. Replications: 400.

**Table 10.** Estimated power of  $\hat{\Psi}_2$  test at 5% level

| <i>DGP3</i> | <i>L</i> = 100 |       |       | <i>L</i> = 400 |       |       | <i>L</i> = 1000 |       |  |
|-------------|----------------|-------|-------|----------------|-------|-------|-----------------|-------|--|
|             | <i>m</i>       | 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  $L = 100$  and  $b = 8$  if  $L > 100$ .  
 Number of bootstraps: 399. Replications: 400.

**Table 11.** Estimated power of  $\hat{\Psi}_2$  test at 5% level

| <i>DGP5</i> | <i>L</i> = 100 |       |       | <i>L</i> = 400 |       |       | <i>L</i> = 1000 |       |  |
|-------------|----------------|-------|-------|----------------|-------|-------|-----------------|-------|--|
|             | <i>m</i>       | 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  $L = 100$  and  $b = 8$  if  $L > 100$ .  
 Number of bootstraps: 399. Replications: 400.

**Table 12.** Estimated power of  $\hat{\Psi}_2$  test at 5% level

| <i>DGP6</i> | <i>L</i> = 100 |      |       | <i>L</i> = 400 |       |       | <i>L</i> = 1000 |       |  |
|-------------|----------------|------|-------|----------------|-------|-------|-----------------|-------|--|
|             | <i>m</i>       | 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  $L = 100$  and  $b = 8$  if  $L > 100$ .  
 Number of bootstraps: 399. Replications: 400.

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 & Strange, 2004) although it does seem to be so in relative terms. Ciccone & 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 & Lindgren (2009) stressed that a good balance between the two profiles is fundamental for achieving permanent improvements in productivity.

We are interested in verifying whether, effectively, the productivity data are related to the habitual indicators of these externalities. 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<sup>2</sup> 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<sup>2</sup> in municipality *i*. This is an indicator of urban externalities.
- *QL*. Location coefficient for the manufacturing sector in municipality *i* obtained as:

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

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:

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

where  $\omega_i^M$  is the number of firms of the manufacturing sector in municipality *i*. This is an indicator of Porter externality.

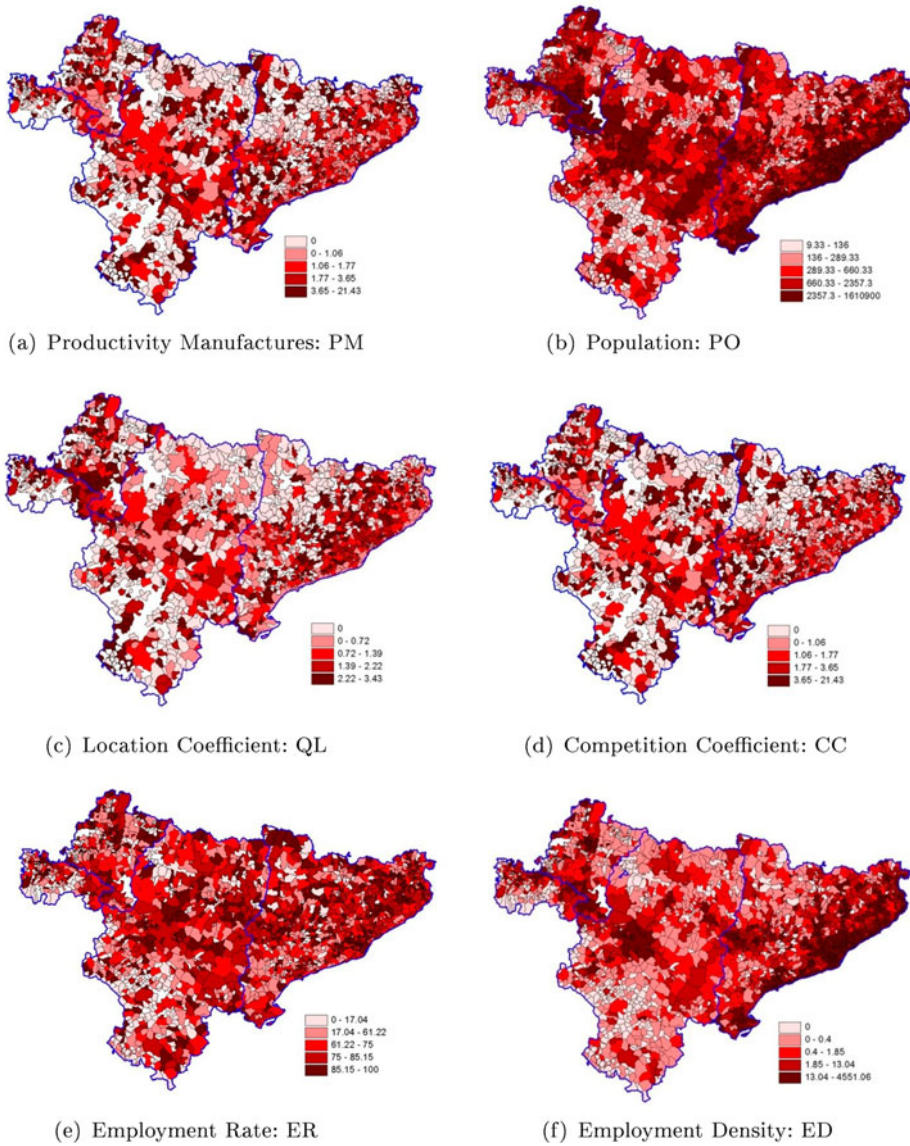
- *ER*. Employment rate, obtained as the quotient between total employment and working population in municipality *i*. This is an indicator of labour externalities.

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.

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 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 tests 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.

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 extent, spurious. In fact, only for the case of the location coefficient, *QL*, a measure of



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

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

**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 significant 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 significant 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.

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  $Y(m)$  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  $Y(m)$  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.

## References

- Angulo, A., Mur, J., López, F. & Herrera, M. (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.
- Anselin, L. (1988) *Spatial Econometrics. Methods and Models*, Dordrecht, Kluwer Academic.
- Anselin, L., Syabri, I. & Smirnov, O. (2002) Visualizing multivariate spatial correlation with dynamically linked windows, in: L. Anselin & S. Rey (eds), *New Tools for Spatial Data Analysis: Proceedings of the Specialist Meeting. Center for Spatially Integrated Social Science (CSISS)*, CD-ROM, Santa Barbara, University of California.
- Arrow, K. (1962) The economic implications of learning by doing, *Review of Economic Studies*, 29, 155–173. doi:10.2307/2295952
- Bureau van Dijk (2012) *Sistema de análisis de balances ibéricos*, SABI, available from: <http://sabi.bvdep.com>
- Burrige, P. (2012) Improving the J test in the SARAR model by likelihood based estimation, *Spatial Economic Analysis*, 7, 75–107 doi:10.1080/17421772.2011.647055
- Burrige, P. & Fingleton, B. (2010) Bootstrap inference in spatial econometrics: the J-test, *Spatial Economic Analysis*, 5, 93–119. doi:10.1080/17421770903511346
- 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. doi:10.1214/aos/1176350057
- Ciccone, A. (2002) Agglomeration effects in Europe, *European Economic Review*, 46, 213–228. doi:10.1016/S0014-2921(00)00099-4
- Ciccone, A. & Hall, R. (1996) Productivity and the density of economic activity, *American Economic Review*, 86, 54–70.
- Cliff, A. & Ord, K. (1981) *Spatial Processes: Models and Applications*, London, Pion.
- Czaplewski, R. & Reich, R. (1993) *Expected Value and Variance of Moran's I Bivariate Spatial Autocorrelation Statistic for a Permutation Test*, USDA, Forest Service, Research Paper RM-309.
- Eriksson, R. & Lindgren, U. (2009) Localized mobility clusters: impacts of labour market externalities on firm performance. *Journal of Economic Geography*, 9, 33–53. doi:10.1093/jeg/lbn025
- 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. doi:10.1111/1467-9787.00121
- Fingleton, B. (2003) Increasing returns; evidence from local wage rates in Great Britain, *Oxford Economic Papers*, 55, 716–739. doi:10.1093/oenp/55.4.716
- 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. doi:10.1111/j.1435-5957.2005.00039.x
- 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. doi:10.1007/s00181-007-0151-4
- Fingleton, B. & LeGallo, J. (2008) Estimating spatial models with endogenous variables, a spatial lag and spatially dependent disturbances: finite sample properties, *Papers in Regional Science*, 87, 319–339. doi:10.1111/j.1435-5957.2008.00187.x
- GeoDa (2012) *GeoDa Center for Geospatial Analysis and Computation*, available from: <https://geodacenter.asu.edu/>

- Glaeser, E., Kallal, H., Scheinkman, J. & Shleifer, A. (1992) Growth in cities, *The Journal of Political Economy*, 100, 1126–1152. doi:10.1086/261856
- Goodchild, M. & Janelle, D. (2003) *Spatially Integrated Social Science*, Oxford, Oxford University Press.
- Hall, P. (1985) Resampling a coverage pattern, *Stochastic Processes and Their Applications*, 20, 231–246. doi:10.1016/0304-4149(85)90212-1
- Han, X. & 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.
- Hao, B. & Zheng, W. (1998) *Applied Symbolic Dynamics and Chaos*, Singapore, World Scientific.
- Instituto Nacional de Estadística, INE (2012) *Estimaciones de Población Actual*, available from: <http://www.ine.es>
- Jacobs, J. (1984) *Cities and the Wealth of Nations: Principles of Economic Life*, New York, Vintage.
- Krugman, P. (1997) *The Age of Diminished Expectations. U.S. Economic Policy in the 1990s*, 2nd edn, Cambridge, MA, MIT Press.
- Lawson, A. (2006) *Statistical Methods in Spatial Epidemiology*, 2nd edn, Berlin, Springer.
- Lesage, J. & Pace, K. (2009) *Introduction to Spatial Econometrics*, Boca Raton, Chapman & Hall/CRC.
- López, F., Matilla-García, M., Mur, J. & Ruiz, M.M. (2010) A non-parametric spatial independence test using symbolic entropy, *Regional Science and Urban Economics*, 40, 106–115. doi:10.1016/j.regsciurbeco.2009.11.003
- Mantel, N. (1967) The detection of disease clustering and a generalized regression approach, *Cancer Research*, 27, 209–220.
- Marshall, A. (1890) *Principles of Economics*, 8th edn, London, Macmillan.
- Matilla-García, M. & Ruiz, M.M. (2008) A non-parametric independence test using permutation entropy, *Journal of Econometrics*, 144, 139–155. doi:10.1016/j.jeconom.2007.12.005
- Matilla-García, M. & Ruiz, M.M. (2009) Detection of non-linear structure in time series, *Economics Letters*, 105, 1–6. doi:10.1016/j.econlet.2009.04.014
- Monchuk, D., Hayes, D., Miranowski, J. & Lambert, D. (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. doi:10.1111/j.1467-9787.2011.00716.x
- Mur, J. & Trivez, J. (2003) Unit roots and deterministic trends in spatial econometric models, *International Regional Science Review*, 26, 289–312. doi:10.1177/0160017603253790
- Pinkse, J. & Slade, M. (1998) Contracting in space: an application of spatial statistics to discrete-choice models, *Journal of Econometrics*, 85, 125–154. doi:10.1016/S0304-4076(97)00097-3
- Porter, M. (1998) *The Competitive Advantage of Nations*, London, Free Press.
- Reich, R., Czaplewski, R. & Bechtold, W. (1994) Spatial cross-correlation in growth of undisturbed natural shortleaf pine stands in northern Georgia. *Journal of Environmental and Ecological Statistics*, 1, 201–217. doi:10.1007/BF00571392
- Rice, P., Venables, A. & Patacchini, E. (2006) Spatial determinants of productivity: analysis for the regions of Great Britain, *Regional Science and Urban Economics*, 36, 727–752. doi:10.1016/j.regsciurbeco.2006.03.006
- Rohatgi, V. (1976) *An Introduction to Probability Theory and Mathematical Statistics*, New York, Wiley.
- Romer, P. (1990) Endogenous technological change, *Journal of Political Economy*, 98, 71–102. doi:10.1086/261725
- Rosenthal, S. & 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, pp. 2063–2117, Amsterdam, Elsevier.
- Ruiz, M., López, F. & Páez, A. (2010) Testing for spatial association of qualitative data using symbolic dynamics, *Journal of Geographical Systems*, 12, 281–309. doi:10.1007/s10109-009-0100-1
- Schmitz, O. (2010) *Resolving Ecosystem Complexity*, Oxford, Oxford University Press.
- Soon, S. (1996) Binomial approximation for dependent indicators, *Statistica Sinica*, 6, 703–714.
- Upton, G. & Fingleton, B. (1985) *Spatial Data Analysis By Example*, New York, Wiley.
- Wartenberg, D. (1985) Multivariate spatial correlation: a method for exploratory geographical analysis, *Geographical Analysis*, 17, 263–283. doi:10.1111/j.1538-4632.1985.tb00849.x

## Appendix

### Appendix A: A bound for the Binomial approximation of a sum of dependent indicators

If variables  $J_{\eta_i}$  and  $J_{\eta_j}$  are not independent for all  $i \neq j$  for some  $\eta \in \Gamma$ , is possible that  $Q_{\eta}$  does not follow a Binomial distribution. In this case, an interesting question is how far is  $Q_{\eta}$  from  $B(L, p_{\eta})$ . Under certain assumptions, the variable  $Q_{\eta}$  can be approximated to a Binomial random variable

$$Q_{\eta} \approx B(L, p_{\eta}). \quad (45)$$

In fact, it is possible to compute a bound for this binomial approximation.

Denote by  $\mathcal{L}(Y_\eta)$  the distribution of  $Y_\eta$ . We are interested in the bound of the Binomial approximation of the distribution of  $Y_\eta$  in terms of total variation distance,  $d_{TV}$ , which measures the distance between two probability distributions  $P$  and  $Q$ :

$$d_{TV}(P, Q) = \sup_A |P(A) - Q(A)|,$$

where the supremum is taken over all measurable sets of the real line.

Following Theorem 1.1 in Soon (1996) and after a few calculations, a bound can be given as follows:

For each  $i, j \in S$  let  $J_{\eta i}, J_{\eta j}$  and  $J_{\eta ij}$  be defined in the same probability space where  $M_{\eta ij} = (J_{\eta i} J_{\eta j} = 1)$ , let

$$W_i = \sum_{j \neq i} J_{\eta j} V_i = \sum_{j \neq i} M_{\eta ij},$$

and

$$C_{L,p\eta} = \frac{1 - p_\eta^{L+1} - q_\eta^{L+1}}{(L+1)p_\eta q_\eta}.$$

$W_i$  counts the number of indexes that are of  $\eta$ -type and  $q_\eta = 1 - p_\eta$ . On the other hand

$$V_i = \sum_{j \neq i} (J_{\eta j} | J_{\eta i} = 1),$$

counts the number of indexes that are of  $\eta$ -type conditioned to location  $i$  is of  $\eta$ -type. Then

$$d_{TV}(\mathcal{L}(Y_\eta), B(L, p)) \leq C_{L,p\eta} \sum_{i \in S} E(|W_i - V_i|).$$

Therefore in order to get a bound for the binomial approximation we have to bound

$$\sum_{i \in S} E(|W_i - V_i|). \tag{46}$$

On the other hand we have that  $E(|W_i - V_i|) \leq \sum_{j \neq i} E(|J_{\eta j} - M_{\eta ij}|)$ . Now denote by  $B_i^j$  a subset of indexes such that  $J_{\eta i}$  is independent of  $\{J_{\eta j} | j \notin B_i^j\}$ . We will call overlapping degree of location  $i$  to  $r_i = |B_i^j| - 1$ . Therefore we obtain that

$$\begin{aligned} d_{TV}(L(Q_\eta), B(L, p)) &\leq C_{L,p\eta} \sum_{i \in S} E(|W_i - V_i|) \leq \\ &\leq C_{L,p\eta} \sum_{i \in S} \sum_{j \in B_i^j \setminus \{i\}} E(|J_{\eta j} - M_{\eta ij}|). \end{aligned} \tag{47}$$

Since  $\lim_{L \rightarrow \infty} C_{L,p} = 0$ , we have shown that the sum of dependent indicators can be approximated to a Binomial random variable when

$$\lim_{L \rightarrow \infty} \sum_{i \in S} \sum_{j \in B_i^j \setminus \{i\}} E(|J_{\eta j} - J_{\eta ij}|) = C \tag{48}$$

for some real positive constant  $C$ . A natural way for (48) to hold is to control for the overlapping degree by selecting a subset of symbolized locations  $S'$ .

By simplicity we will call  $L = |S'|$ . Therefore if the set of symbolized locations  $S'$  is such that

$$\lim_{L \rightarrow \infty} \sum_{i \in S'} r_i = K, \tag{49}$$

for some positive constant  $K$ , then (48) holds. Therefore by (47) we have that  $Q_\eta$  asymptotically follows a Binomial distribution.

We will say that the dependence of the indicators  $J_{\eta s}$  is weak if  $Q_\eta$  asymptotically follows a Binomial distribution  $B(L, p_\eta)$ .

**Appendix B: Proof of Theorem 3.1.**

The likelihood function of the multinomial distribution of (22) is:

$$F(p_{\eta_{11}}, p_{\eta_{12}}, \dots, p_{\eta_{mm}}) = \frac{L!}{n_{\eta_{11}}! n_{\eta_{12}}! \dots n_{\eta_{mm}}!} p_{\eta_{11}}^{n_{\eta_{11}}} p_{\eta_{12}}^{n_{\eta_{12}}} \dots p_{\eta_{mm}}^{n_{\eta_{mm}}}, \tag{50}$$

and as  $\sum_{i,j} p_{\eta_{ij}} = 1$ , it follows that

$$F(p_{\eta_{11}}, p_{\eta_{12}}, \dots, p_{\eta_{mm}}) = \frac{L!}{n_{\eta_{11}}! n_{\eta_{12}}! \dots n_{\eta_{mm}}!} p_{\eta_{11}}^{n_{\eta_{11}}} p_{\eta_{12}}^{n_{\eta_{12}}} \dots (1 - p_{\eta_{11}} - \dots - p_{\eta_{mm-1}})^{n_{\eta_{mm}}}. \tag{51}$$

The logarithm of the likelihood function is

$$\begin{aligned} \ln(F(p_{\eta_{11}}, p_{\eta_{12}}, \dots, p_{\eta_{mm}})) &= \ln\left(\frac{L!}{n_{\eta_{11}}! n_{\eta_{12}}! \dots n_{\eta_{mm}}!}\right) + \sum_{i=1}^n \sum_{j=1}^{n-1} n_{\eta_{ij}} \ln(p_{\eta_{ij}}) \\ &\quad + n_{\eta_{mm}} \ln(1 - p_{\eta_{11}} - p_{\eta_{12}} - \dots - p_{\eta_{mm-1}}). \end{aligned}$$

To obtain the maximum likelihood estimators  $\hat{p}_{\eta_{ij}}$  of  $p_{\eta_{ij}}$  for all  $i, j = 1, 2, \dots, n$ , we need the score

$$\frac{\partial \ln(F(p_{\eta_{11}}, p_{\eta_{12}}, \dots, p_{\eta_{mm}}))}{\partial p_{\eta_{ij}}} = 0,$$

so that

$$\hat{p}_{\eta_{ij}} = \frac{n_{\eta_{ij}}}{L}.$$

The likelihood ratio statistic is:

$$\begin{aligned} \lambda(Q) &= \frac{\frac{R!}{n_{\eta_{11}}! n_{\eta_{12}}! \dots n_{\eta_{mm}}!} p_{\eta_{11}}^{(0)n_{\eta_{11}}} p_{\eta_{12}}^{(0)n_{\eta_{12}}} \dots p_{\eta_{mm}}^{(0)n_{\eta_{mm}}}}{\frac{R!}{n_{\eta_{11}}! n_{\eta_{12}}! \dots n_{\eta_{mm}}!} \hat{p}_{\eta_{11}}^{n_{\eta_{11}}} \hat{p}_{\eta_{12}}^{n_{\eta_{12}}} \dots \hat{p}_{\eta_{mm}}^{n_{\eta_{mm}}}} = \frac{\prod_{i=1}^n \prod_{j=1}^n p_{\eta_{ij}}^{(0)n_{\eta_{ij}}}}{\prod_{i=1}^n \prod_{j=1}^n \binom{n_{\eta_{ij}}}{R} n_{\eta_{ij}}} \\ &= L \sum_{i=1}^n \sum_{j=1}^n \frac{\prod_{i=1}^n \prod_{j=1}^n p_{\eta_{ij}}^{(0)n_{\eta_{ij}}}}{\prod_{i=1}^n \prod_{j=1}^n n_{\eta_{ij}}^{n_{\eta_{ij}}}} = L \prod_{i=1}^n \prod_{j=1}^n \left(\frac{p_{\eta_{ij}}^{(0)}}{n_{\eta_{ij}}}\right)^{n_{\eta_{ij}}}, \end{aligned}$$

where  $p_{\eta_{ij}}^{(0)}$  denotes the probability of symbol  $\eta_{ij}$  under the null hypothesis.

Under the assumption of independent iid series,  $Y(m) = -2 \ln(\lambda(Q))$  asymptotically follows a Chi-square distribution with  $k$  degrees of freedom, where  $k$  is the number of unknown parameters under  $H_1$  less the number of unknown parameters under  $H_0$  (Lehmann, 1986). Then,

$$Y(m) = -2 \ln(\lambda(Q)) = -2 \left[ L \ln L + \sum_{i=1}^n \sum_{j=1}^n n_{\eta_{ij}} \ln \left( \frac{p_{\eta_{ij}}^{(0)}}{n_{\eta_{ij}}} \right) \right] \sim \chi_k^2. \tag{52}$$

The spatial processes  $\{x_s\}_{s \in S}$  and  $\{y_s\}_{s \in S}$  are independent under the null hypothesis, so we may write

$$P_{\eta_{ij}}^{(0)} = P_{\sigma_i'}^{(0)} P_{\sigma_j'}^{(0)}.$$

Using this result, and considering that  $\sum_{i=1}^n \sum_{j=1}^n \frac{n_{ij}}{R} = 1$ , we obtain that

$$\begin{aligned} Y(m) &= -2L \left[ \ln L + \sum_{i=1}^n \sum_{j=1}^n \frac{n_{ij}}{L} \ln \left( \frac{p_{\sigma_i}^{(0)} p_{\sigma_j}^{(0)}}{n_{ij}} \right) \right] \\ &= -2L \left[ \ln L + \sum_{i=1}^n \sum_{j=1}^n \frac{n_{ij}}{L} \ln(p_{\sigma_i}^{(0)} p_{\sigma_j}^{(0)}) - \sum_{i=1}^n \sum_{j=1}^n \frac{n_{ij}}{L} \ln \left( n_{ij} \frac{L}{L} \right) \right] \\ &= -2L \left[ \ln L + \sum_{i=1}^n \sum_{j=1}^n \frac{n_{ij}}{L} \ln(p_{\sigma_i}^{(0)} p_{\sigma_j}^{(0)}) - \sum_{i=1}^n \sum_{j=1}^n \frac{n_{ij}}{L} \ln \left( \frac{n_{ij}}{L} \right) \right. \\ &\quad \left. - \sum_{i=1}^n \sum_{j=1}^n \frac{n_{ij}}{R} \ln(L) \right] \\ &= -2L \left[ \sum_{i=1}^n \sum_{j=1}^n \frac{n_{ij}}{L} \ln(p_{\sigma_i}^{(0)} p_{\sigma_j}^{(0)}) - \sum_{i=1}^n \sum_{j=1}^n \frac{n_{ij}}{L} \ln \left( \frac{n_{ij}}{L} \right) \right]. \end{aligned}$$

So, with the symbolization proposed in Section 2.1, and considering that, under the null hypothesis, spatial processes are independent and i.i.d., then  $p_{\sigma_i}^{(0)} = p_{\sigma_j}^{(0)} = C_{\sigma_i}^{m-1} / 2^{(m-1)}$  for all  $i = 1, 2, \dots, m$ . On the other hand,  $h_Z(m) = -\sum_{i=1}^n \sum_{j=1}^n \frac{n_{ij}}{L} \ln \left( \frac{n_{ij}}{L} \right)$ . We can easily obtain the following result:

$$Y(m) = 2L[2(m-1)\ln(2) - \left[ \sum_{i=1}^n \sum_{j=1}^n \frac{n_{ij}}{L} \ln(C_{\sigma_i}^{m-1} C_{\sigma_j}^{m-1}) - h_Z(m) \right]], \tag{53}$$

as desired. ■

### Appendix C: Proof of Theorem 3.2.

First note that  $\text{plim}_{L \rightarrow \infty} \hat{p}_\sigma = p_\sigma$  exists for every spatial process, where  $\hat{p}_\sigma = \frac{n_\sigma}{L}$ .

The logarithm function is continuous so it follows that

$$\text{plim}_{L \rightarrow \infty} \hat{h}(m) = h(m). \tag{54}$$

Let  $\hat{Y}(m)$  represent the estimator of  $Y(m)$ :

$$\hat{Y}(m) = 2L[2(m-1)\ln(2) - \left[ \sum_{i=1}^n \sum_{j=1}^n \frac{n_{ij}}{L} \ln(C_{\sigma_i}^{m-1} C_{\sigma_j}^{m-1}) - h_Z(m) \right]].$$

The measure  $H(m) = 2(m-1)\ln(2) \left[ \sum_{i=1}^n \sum_{j=1}^n \frac{n_{ij}}{L} \ln(C_{\sigma_i}^{m-1} C_{\sigma_j}^{m-1}) \right] - h_Z(m)$  can be expressed as:

$$H(m) = -\sum_{i=1}^m \sum_{j=1}^m \frac{n_{ij}}{L} \ln \left( \frac{C_{\sigma_i}^{m-1} C_{\sigma_j}^{m-1}}{2^{2(m-1)}} \right). \tag{55}$$

Furthermore, for every positive real number  $x \in \mathbb{R}$ ,  $-\ln(x) \geq 1-x$  (the equality will hold when  $x = 1$ ). Let us remember that  $2^{m-1} = (1+1)^{m-1} = \sum_{i=0}^{m-1} C_i^{m-1}$ . Using these results, it is easy to see that the statistics of expression (55) verifies that:

$$H(m) > \sum_{i=1}^m \sum_{j=1}^m \frac{n_{ij}}{L} \left( 1 - \frac{C_{\sigma_i}^{m-1} C_{\sigma_j}^{m-1}}{2^{2(m-1)}} \right) = 0.$$

Therefore, we have obtained that, under dependence of order less than  $m$ :

$$H(m) > 0. \tag{56}$$

Let  $0 < C < \infty$  with  $C \in \mathbb{R}$  and  $R$  with large enough so that

$$\frac{C}{2L} < H(m). \tag{57}$$

Then

$$\begin{aligned} \Pr\left[\hat{\Upsilon}(m) > C\right] &= \Pr[2LH(m) > C] \\ &= \Pr\left[H(m) > \frac{C}{2L}\right]. \end{aligned}$$

So, for equation (37) we have

$$\text{plim}_{L \rightarrow \infty} \left(\hat{\Upsilon}(m) > C\right) = 1,$$

as desired. ■