Detection of Spatial Dependence using Symbolic Analysis

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

2011

Detección de Dependencia Espacial mediante Análisis Simbólico

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

[(1)]Universidad de Zaragoza; [(2)]Universidad Politécnica de Cartagena

Resumen

Contrastar el supuesto de independencia entre variables espaciales es uno de los primeros pasos en econometría espacial. Usualmente, la literatura especializada utiliza una generalización bivariante del estadístico de Moran, especificando a priori una matriz de contactos. Dicho estadístico es solo válido para relaciones lineales entre pares de variables. En este trabajo desarrollamos un nuevo estadístico no-paramétrico, basado en dinámica simbólica, que no presenta estas limitaciones. El nuevo estadístico es consistente, computacionalmente simple de obtener y con buen comportamiento en muestras finitas como se muestra en los resultados del experimento Monte Carlo.

Abstract

Testing for the assumption of independence between spatial variables is an important first step in spatial econometrics. Usually the researchers use the bivariate generalization of the Moran's statistic, specifying a spatial matrix a priori. This test is applicable only to detect linear relations in pairs of variables, which must be spatially non-autocorrelated. We develop a new non-parametric test, based on symbolic dynamics, that is free of these shortcomings. The test is consistent, computationally simple to obtain and powerful as shown in our Monte Carlo experiment.

Palabras Claves: *Contraste Bivariante de Moran, Dinámica Simbólica, Entropía Simbólica.*

Código JEL: ***C14, C21.***

*Autor para correspondencia: Departamento de Análisis Económico, Universidad de Zaragoza. Gran Vía 2-4 (50005). Zaragoza (España). Mail: mherreragomez@gmail.com

1. Introducción

La dependencia es una de las características distintivas de los datos espaciales. La noción de que observaciones espacialmente cercanas se encuentran altamente correlacionadas es bastante natural para diferentes áreas científicas. Este es un problema bien estudiado en la literatura de estadística espacial conocido como autocorrelación espacial y ha sido ampliamente analizado en el pasado. Entre las principales referencias podemos citar a Cliff y Ord (1981), Cressie (1993), Haining (2003) o LeSage y Pace (2009).

La correlación espacial entre variables puede ser positiva debido a un comportamiento similar ante factores ambientales o por compartir determinados recursos, por ejemplo. Por otra parte, la correlación espacial negativa entre dos o más variables puede provenir de la competencia sobre los mismos recursos escasos o por diferencias en la respuesta a los factores ambientales. En ambos casos, las variables no son independientes y es importante obtener información sobre el patrón que genera dicha dependencia.

Contrastar el supuesto de independencia entre diferentes variables es un importante primer paso en el análisis de datos espaciales. En primera instancia, puede pensarse en utilizar el coeficiente de correlación de Pearson o de Spearman. El problema es que, cuando se tienen datos espaciales, una aparentemente asociación significativa entre dos variables puede ser artificial. Como refleja Bivand (1980), la presencia de correlación espacial en los datos genera una subestimación de la varianza de los coeficientes de correlación habituales (Pearson o Spearman). Cliff y Richardson (1985) proponen ajustar la pérdida de información usando un 'número equivalente de observaciones' (similar a Cerioli, 1997, para datos cualitativos). La propuesta de Haining (1991) es 'à la Barlett': pre-blanquear los datos y luego contrastar la asociación entre las variables. Estas soluciones no son completamente satisfactorias, por ejemplo, debido a su arbitrariedad.

Un camino más efectivo es la utilización de contrastes desarrollados dentro de la estadística espacial. La principal característica de estos estadísticos es que consideran de manera explícita la dimensión espacial de los datos. Este es el caso del estadístico bivariante de Moran (Wartenberg, 1985) que es una generalización del ampliamente conocido estadístico I de Moran. La propuesta de Moran para contrastar la hipótesis de independencia es probablemente la más popular en el campo de la estadística espacial: es simple y potente en muchas circunstancias. Sin embargo, para nuestro caso, este contraste sufre de dos importantes restricciones. Primeramente, las dos series deben estar linealmente relacionadas. Segundo, debe mencionarse que el contraste bivariante de Moran contrasta la relación entre pares de variables y, en algunos casos, el interés se centra sobre la asociación entre grupos de más de dos variables.

Dadas estas limitaciones, proponemos un nuevo estadístico para contrastar la hipótesis de incorrelación espacial entre variables utilizando la técnica conocida como análisis simbólico. Mediante esta técnica obtenemos un contraste no-paramétrico, denominado $\Upsilon(m)$, que es libre supuestos distributivos incluyendo normalidad. Este estadístico puede ser fácilmente generalizado para analizar grupos de más de dos variables.

El resto del trabajo esta organizado de la siguiente manera. En la sección 2 se presentan algunas definiciones y conceptos básicos requeridos para simbolizar un determinado conjunto de datos. En la sección 3 presentamos el contraste de independencia. La sección 4 presenta los resultados del experimento Monte Carlo en el que se compara el estadístico $\Upsilon(m)$ con el Moran bivariante. Conclusiones aparecen en la 5 sección.

2. Herramientas de Análisis Simbólico

El análisis simbólico es una técnica que realiza una discretización de los datos originales dentro de una correspondiente secuencia de símbolos. La secuencia de símbolos permite capturar información estadísticamente útil pero no directamente observable. Una ventaja práctica de trabajar mediante símbolos es la mayor eficiencia

numérica en su cómputo a medida que los datos originales tienden a incrementarse. Por otra parte, el análisis simbólico conserva características esenciales sobre las series analizadas tales como la dependencia y es, a menudo, menos sensible a errores de medida.

En el caso de datos espaciales, es decir, georeferenciados, la idea es considerar un espacio en el que todos los posibles estados del sistema se encuentran representados. Este espacio puede ser particionado en un número finito de regiones, y cada región es representada por un símbolo. En este sentido, la dinámica simbólica es una descripción segmentada de un sistema dinámico (Hao y Zheng, 1998, para mayores detalles).

2.1. Proceso de Simbolización

Esta sección presenta un procedimiento de simbolización para tratar con series espaciales. Dependiendo de la disponibilidad de información estadística, la simbolización propuesta puede ser mejorada por el investigador.

Sean $\{x_s\}_{s\in S}$ e $\{y_s\}_{s\in S}$ dos procesos espaciales reales, donde S es un conjunto de puntos o localizaciones sobre el espacio. Con el objetivo de simbolizar la serie, tenemos que definir un conjunto finito no vacío de símbolos que sean capaces de recoger la información necesaria del proceso espacial. Denotaremos a este conjunto por $\Gamma_n = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$; y a cada uno de sus elementos σ_i lo llamaremos *símbolo* para $i = 1, 2, \ldots, n$.

Entonces, simbolizar una serie es definir una función

$$f : \{x_s\}_{s\in S} \to \Gamma_n, \tag{1}$$

tal que a cada elemento x_s se le asocia un único símbolo $f(x_s) = \sigma_{i_s}$ con $i_s \in \{1, 2, \ldots, n\}$. Diremos que la localización $s \in S$ es de tipo σ_i, con respecto a la serie $\{x_s\}_{s\in S}$, si y solo si $f(x_s) = \sigma_{i_s}$. Llamaremos a f *función de simbolización*. El mismo proceso se puede repetir para la serie y_s.

A continuación, introducimos el proceso bivariante $\{Z_s\}_{s\in S}$ como:

$$Z_s = \{x_s, y_s\}, \tag{2}$$

donde x_s e y_s son los procesos espaciales univariantes antes definidos. Para este proceso bivariante definimos el conjunto de símbolos Ω_n como el producto directo de los dos conjuntos Γ_n, es decir, $\Omega_n^2 = \Gamma_n \times \Gamma_n$ y sus elementos son de la forma $\eta_{ij} = \left(\sigma_i^x, \sigma_j^y\right)$. La función de simbolización del proceso bivariante será

$$g : \{Z_s\}_{s\in S} \to \Omega_n^2 = \Gamma_n \times \Gamma_n, \tag{3}$$

definida por

$$g(Z_s = (x_s, y_s)) = (f(x_s), f(y_s)) = \eta_{ij} = \left(\sigma_i^x, \sigma_j^y\right). \tag{4}$$

Diremos que s es de tipo η_{ij} para $Z = (x, y)$ o simplemente que s es de tipo η_{ij} , si y solo si s es tipo σ_i^x para x y de tipo σ_j^y para y.

En este trabajo nos restringiremos a procesos bivariantes, aunque el esquema puede generalizarse fácilmente para procesos multivariantes de la siguiente forma. Consideremos un proceso espacial $k-dimensional$, $\{Z_s\}_{s\in S} = \{x_{1s}, x_{2s}, \ldots, x_{ks}\}$. Sea $\Omega_n^k = \Gamma_n \times \Gamma_n \cdots \times \Gamma_n$ el producto directo de k copias de Γ_n y sea $\eta_{i_1,i_2,\ldots i_k} = (\sigma_{i_1}, \sigma_{i_2}, \ldots, \sigma_{i_k}) \in \Omega_n^k$. Entonces, diremos que s es de tipo $\eta_{i_1,i_2,\ldots i_k}$ si y solo si s es de tipo σ_{i_j} para x_{j_s} para todo $j = 1, 2, \ldots, k$.

Dependiendo del problema, podemos definir diferentes funciones de simbolización. En nuestro caso, definiremos una función de simbolización f utilizando la mediana, M_e^x, del proceso espacial univariante $\{x_s\}_{s\in S}$. Sea la función indicadora

$$\tau_s = \begin{cases} 1, & si \quad x_s \geq M_e^x, \\ 0, & en\ cualquier\ otro\ caso. \end{cases} \tag{5}$$

Para cada $s \in S$, sea n_s el conjunto formado por los $(m-1)$ vecinos de s. Denominaremos $m-entorno$ al conjunto formado por cada s y el conjunto n_s, tal que el $m-entorno\ x_m(s) = \left(x_s, x_{s_1}, \ldots, x_{s_{m-1}}\right) = (x_s, x_{n_s})$. Con $m \geq 2$ nos referiremos a la *dimensión de encaje*. Definimos para cada s_i con $i = 1, 2, \ldots, m-1$ la función indicadora:

$$\iota_{ss_i} = \begin{cases} 0, & si \quad \tau_s \neq \tau_{s_i}, \\ 1, & en\ cualquier\ otro\ caso. \end{cases} \tag{6}$$

Finalmente, podemos establecer una función de simbolización para el proceso espacial $\{x_s\}_{s\in S}$ como $f : \{x_s\}_{s\in S} \to \Gamma_m$, definida por:

$$f(x_s) = \sum_{i=1}^{m-1} \iota_{ss_i}, \tag{7}$$

donde $\Gamma_m = \{0, 1, \ldots, m-1\}$.

El proceso de simbolización consiste en comparar, para cada localización s, el valor τ_s con τ_{s_i} recorriendo s_i el conjunto de los $m-1$ vecinos más próximos a la localización s. Esta simbolización nos permite capturar la información relevante del vecindario de la observación s.

Para facilitar la interpretación del proceso de simbolización, proponemos el siguiente ejemplo. Supongamos que tenemos dos procesos espaciales distribuidos en un mapa regular de dimensión 3×3. Su representación espacial se muestra en la Figura 1.

Figura 1: Ejemplo de Mapa Regular 3×3 para x_s e y_s

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

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

Considerando $m = 4$, podemos representar a la serie x_s, tal que $x_4(s_1) = (x_{s_1} = 4, x_{s_2} = 1, x_{s_4} = 6, x_{s_5} = 2)$ representa al $4-entorno$ de s_1 que está formado por los 3 vecinos más cercanos a la localización s_1.

Para cada localización de la serie x_s debemos formar los $4-entorno$ restantes: $x_4(s_2), x_4(s_3), \ldots, x_4(s_9)$.

Denotemos por n_{s_i} al conjunto formado por los 3 vecinos más próximos s_i. Entonces tenemos que: $(n_{s_2} = \{s_3, s_1, s_5\})$, $(n_{s_3} = \{s_2, s_6, s_5\})$, $(n_{s_4} = \{s_5, s_1, s_7\})$, $(n_{s_5} = \{s_6, s_2, s_4\})$, $(n_{s_6} = \{s_3, s_5, s_9\})$, $(n_{s_7} = \{s_8, s_4, s_5\})$, $(n_{s_8} = \{s_9, s_5, s_7\})$, $(n_{s_9} = \{s_6, s_8, s_5\})$.

Este proceso se puede aplicar de manera similar a y_s.

En el ejemplo anterior, la mediana de x_s es 3. En consecuencia, el símbolo asociado a la localización s_1 es: $f(x_{s_1}) = (\iota_{s_1 s_2} = 0) + (\iota_{s_1 s_4} = 1) + (\iota_{s_1 s_5} = 0) = 1$.

De igual manera, podemos obtener los símbolos asociados al resto de las localizaciones: $f(x_{s_2}) = 1$; $f(x_{s_3}) = 1$; $f(x_{s_4}) = 1$; $f(x_{s_5}) = 1$; $f(x_{s_6}) = 2$; $f(x_{s_7}) = 2$; $f(x_{s_8}) = 2$; $f(x_{s_9}) = 1$.

Procediendo de igual forma, podemos obtener los símbolos asociados a la serie y_s tal que: $f(y_{s_1}) = 0$; $f(y_{s_2}) = 1$; $f(y_{s_3}) = 1$; $f(y_{s_4}) = 1$; $f(y_{s_5}) = 2$; $f(y_{s_6}) = 2$; $f(y_{s_7}) = 1$; $f(y_{s_8}) = 2$; $f(y_{s_9}) = 2$.

Nótese que si cada proceso espacial es independiente sobre el espacio (en el sentido que su distribución espacial es aleatoria), para la función de simbolización propuesta, la probabilidad de ocurrencia de cada símbolo viene dada por $p(\sigma) = {}^{C_\sigma^{m-1}}/_{2^{(m-1)}}$, donde $C_\sigma^{m-1} = {}^{(m-1)!}/_{[(m-1-\sigma)!\sigma!]}$ denota las combinaciones de $m-1$ elementos tomados de σ en σ para todo símbolo $\sigma \in \{0, \ldots, m-1\}$. Es decir, para $m = 4$, bajo la hipótesis de independencia espacial del proceso respectivo, las frecuencias relativas esperadas para cada símbolo son: $p(\sigma = 0) = 1/8$, $p(\sigma = 1) = 3/8$, $p(\sigma = 2) = 3/8$, $p(\sigma = 3) = 1/8$.

Una vez simbolizados los procesos univariantes, procedemos a emparejar, para cada localización, los símbolos obtenidos en los dos procesos para obtener la simbolización del proceso bivariante. Por ejemplo, la localización $s = 2$ es $tipo - (1, 1)$.

Ejemplos de diferentes funciones de simbolización pueden ser consultados en Matilla y Ruiz (2008, 2009), López *et al.* (2010) y Ruiz, López y Páez (2009). En este último trabajo la propuesta se aplica a datos discretos, mientras los anteriores se circunscriben al tratamiento de variables de tipo continuo.

Es importante mencionar que el procedimiento de simbolización es igualmente aplicable a estructuras espaciales regulares como irregulares, así también a puntos como a áreas.

2.2. Entropía: Definiciones y Conceptos

En esta sección brindamos algunos conceptos básicos de la Teoría de Información. Un tratamiento exhaustivo puede encontrarse en Cover y Thomas (1991).

El núcleo central de la Teoría de la Información es el concepto de entropía, como medida de la incertidumbre de un proceso estocástico. Sea x una variable aleatoria discreta que toma los valores $\{x_1, x_2, \ldots, x_N\}$ con probabilidades $p(x_i)$ para cada $i = 1, 2, \ldots, N$, respectivamente.

Definición 1: La entropía de Shannon, $h(x)$, de una variable aleatoria discreta x se define como:

$$h(x) = -\sum_{i=1}^{N} p(x_i) \ln(p(x_i)).$$

Usualmente, cuando la base del logaritmo es igual a 2, las unidades de medida se expresan en *bits*. Nosotros trabajaremos con la base neperiana, por lo que las unidades se expresan en *nats*. Se asume, por convención, que $0 \ln 0 = 0$, es decir añadir términos iguales a cero no modifica la entropía.

Sobre la base de la definición de la entropía individual, podemos considerar la entropía conjunta de un par de variables aleatorias.

Definición 2: La entropía $h(x, y)$ de un par de variables aleatorias discretas (x, y) con distribución conjunta $p(x, y)$ es:

$$h(x,y) = -\sum_x \sum_y p(x,y) \ln (p(x,y)).$$

A su vez podemos definir la entropía condicional.

Definición 3: La entropía condicional $h(x|y)$, correspondiente a la distribución $p(x,y)$, se define como:

$$h(x|y) = -\sum_x \sum_y p(x,y) \ln (p(x|y)).$$

En otras palabras, la entropía condicional $h(x|y)$ es la entropía de x que permanece cuando y se ha observado.

Es necesario destacar que las medidas de entropía son funciones de la distribución de probabilidad de las variables aleatorias. Es decir, no dependen del valor que dichas variables toman en un caso particular, solo de su probabilidad. En contraste, la varianza depende de los valores que asumen las variables y es sensible a las unidades de medida.

Estos términos, asociados al concepto de entropía pueden ser adaptados al caso de la distribución de probabilidad de los símbolos computados en la sección previa. Para ello necesitamos una serie de definiciones adicionales.

Obtenida la simbolización de la serie para una dimensión de encaje $m \geq 2$, fácilmente se puede calcular la frecuencia absoluta y relativa de las diferentes colecciones de símbolos $\sigma_{i_s}^x \in \Gamma_n$ y $\sigma_{j_s}^y \in \Gamma_n$.

Definimos la frecuencia absoluta del símbolo σ_i^x como:

$$n_{\sigma_i^x} = \#\left\{s \in S | s \quad es \quad de \quad tipo \quad \sigma_i^x \quad para \quad x\right\}. \tag{8}$$

Análogamente, para la serie $\{y_s\}_{s\in S}$ se define la frecuencia absoluta del símbolo σ_j^y como

$$n_{\sigma_j^y} = \#\left\{s \in S | s \quad es \quad de \quad tipo \quad \sigma_j^y \quad para \quad y\right\}. \tag{9}$$

Una vez obtenidas las frecuencias absolutas, se pueden calcular las frecuencias relativas:

$$p(\sigma_i^x) \equiv p_{\sigma_i^x} = \frac{\#\left\{s \in S | s \quad es \quad de \quad tipo \quad \sigma_i^x \quad para \quad x\right\}}{|S|} = \frac{n_{\sigma_i^x}}{|S|}, \tag{10}$$

$$p\left(\sigma_j^y\right) \equiv p_{\sigma_j^y} = \frac{\#\left\{s \in S | s \quad es \quad tipo \quad \sigma_j^y \quad para \quad y\right\}}{|S|} = \frac{n_{\sigma_j^y}}{|S|}, \tag{11}$$

donde $|S|$ denota el cardinal del conjunto S; en general $|S| = N$.

De manera similar, calculamos para $\eta_{ij} \in \Omega_n^2$ su frecuencia relativa:

$$p(\eta_{ij}) \equiv p_{\eta_{ij}} = \frac{\#\left\{s \in S | s \quad es \quad de \quad tipo \quad \eta_{ij}\right\}}{|S|} = \frac{n_{\eta_{ij}}}{|S|}. \tag{12}$$

Usando estas definiciones, podemos desarrollar el concepto de *entropía simbólica* para una serie espacial $bidimensional$ $\{Z_s\}_{s\in S}$. Esta entropía es la entropía de Shannon para los m^2 símbolos distintos

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

La entropía simbólica es un indicador de la información contenida en los m^2 símbolos utilizados en la simbolización.

De manera similar se pueden definir las entropías simbólicas marginales como

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

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

Obsérvese que las entropías marginales y la conjunta satisfacen que $0 \leq h(m) \leq \ln(n)$. La cota inferior se alcanza cuando sólo aparece un único símbolo y la cota superior cuando todos los símbolos tienen igual probabilidad de ocurrencia.

A su vez, podemos obtener la entropía simbólica de y condicionada a la ocurrencia del símbolo σ^x en x como:

$$h_{y|\sigma^x}(m) = -\sum_{\sigma^y \in \Gamma_m} p(\sigma^y|\sigma^x) \ln(p(\sigma^y|\sigma^x)). \tag{16}$$

Podemos calcular, además, la entropía simbólica condicional de y_s dado x_s:

$$h_{y|x}(m) = -\sum_{\sigma^x \in \Gamma_m} \sum_{\sigma^y \in \Gamma_m} p(\sigma^x, \sigma^y) \ln(p(\sigma^y|\sigma^x)). \tag{17}$$

Quisiéramos destacar que la entropía, entendida como medida informativa, no es una medida nueva en la literatura especializada en econometría y estadística espacial. Uno de los textos de referencia en estas disciplinas es el de Cressie (1993), que inicia el primer capítulo planteando el concepto de entropía como medida del desorden de un sistema para motivar la discusión estadística en datos con naturaleza espacial. El mismo Cressie pone de manifiesto la dicotomía existente entre entropía y varianza. Gracias a la simplicidad en la estimación, la varianza ha tenido un rol casi excluyente como medida de información.

3. Independencia en Procesos Espaciales

Las herramientas desarrolladas nos permiten abordar el análisis de datos espaciales usando un enfoque libre de hipótesis a priori.

Un buen ejemplo del potencial de esta línea la encontramos en el trabajo de López *et al.* (2010). Los autores desarrollan un estadístico, denominado SG, para analizar el supuesto de independencia transversal en el proceso espacial univariante $\{x_s\}_{s \in S}$. Dicho estadístico utiliza una simbolización diferentes, pero puede ser adaptado a la simbolización propuesta en la sección anterior. Para una dimensión de encaje fija $m \geq 2$, su expresión final es:

$$SG(m) = 2N \left[2(m-1)\ln(2) - \left[\sum_{i=1}^{m} \ln\left(C_{\sigma_i}^{m-1}\right) \right] - h_x(m) \right]. \tag{18}$$

Bajo la hipótesis nula H_0 de independencia, el estadístico SG se distribuye como una Chi-Cuadrado con m grados de libertad.

3.1. Contraste de Independencia entre Procesos Espaciales

Vamos a considerar una serie espacial bidimensional $\{Z_s = \{x_s, y_s\}\}_{s \in S}$ y una dimensión de encaje fija, $m \geq 2$.

Para desarrollar el contraste de independencia entre las series $\{x_s\}_{s \in S}$ e $\{y_s\}_{s \in S}$, especificamos las siguientes hipótesis nula y alternativa:

$H_0 : \{x_s\}_{s\in S}\ e\ \{y_s\}_{s\in S}\quad son\quad i.i.d.\quad e\quad independientes\quad entre\quad si,$
$H_1 :\ No\quad H_0.$

Ahora, para un símbolo $\eta \in \Omega_n^2$ del proceso bivariante $\{Z_s\}_{s\in S}$, definimos la variable aleatoria τ_{η_s} como sigue:

$$\tau_{\eta_s} = \begin{cases} 1, & si\ \ s\ es\ tipo-\eta, \\ 0, & en\ cualquier\ otro\ caso. \end{cases} \tag{19}$$

Entonces, τ_{η_s} es una variable Bernoulli con probabilidad de "éxito" p_η, donde "éxito" significa que s es de $tipo-\eta$.

Es sencillo ver que

$$\sum_{\eta\in\Omega_n^2} p_\eta = 1. \tag{20}$$

Ahora supondremos que el conjunto de localizaciones S es finito; N denota el número de elementos de S. Estamos interesados en saber cuántos elementos del conjunto S son de $tipo-\eta$, para todo símbolo $\eta \in \Omega_n^2$. Para avanzar en esta cuestión, construimos la siguiente variable aleatoria

$$Q_\eta = \sum_{s\in S} \tau_{\eta_s}. \tag{21}$$

Nótese que no todas las variables son independientes (debido al solapamiento de los m entornos), provocando que Q_η no sea exactamente una variable aleatoria Binomial. Sin embargo, la suma de variables dependientes Bernoulli puede ser aproximada como una variable aleatoria Binomial si se cumplen las siguientes condiciones (Soon, 1996):

1. La dependencia entre los indicadores es débil; y
2. La probabilidad de ocurrencia de los indicadores es pequeña.

La segunda condición se satisface por la forma en que los símbolos han sido construidos. Bajo la hipótesis nula, la probabilidad de éxito de un indicador τ_{η_s} es pequeña ($p(\sigma_i) = {}^{C_{\sigma_i}^{m-1}}/_{2^{(m-1)}}$) para la mayoría de los símbolos. La primera condición puede ser satisfecha solo si la distribución de las localizaciones sobre el mapa es regular y la dimensión de encaje es relativamente pequeña. Si la dimensión de encaje es grande, o el sistema espacial es irregular, esta condición se vuelve más difícil de sostener.

Para asegurar que la dependencia entre los indicadores τ_{η_s} sea lo suficientemente débil es posible controlar el grado de solapamiento de los $m-entornos$. El solapamiento ocurre cuando los $m-entornos$ de localizaciones distintas comparten vecinos comunes. Puede obtenerse buena aproximación a la distribución Binomial utilizando un subconjunto de localizaciones $\tilde{S} \subseteq S$ con solapamiento controlado, de forma que la dependencia entre los indicadores τ_{η_s} sea débil para $s \in \tilde{S}$.

Obviamente, una buena aproximación Binomial conlleva pérdida de información muestral, existiendo un relación negativa que debe ser tenida en cuenta. Esta estrategia permite obtener un índice de solapamiento que al momento de implementarse el contraste puede ser establecido. Un método para construir el conjunto de localizaciones $\tilde{S}$ con solapamiento controlado puede encontrarse en Ruiz, López y Páez (2009).

Por lo tanto, bajo las condiciones enunciadas, la variable Q_η puede ser aproximada como una variable aleatoria Binomial:

$$Q_\eta \approx B(N, p_\eta). \tag{22}$$

Bajo la hipótesis nula H_0, la función de probabilidad conjunta de las n^2 variables $(Q_{\eta_{11}}, Q_{\eta_{12}}, \ldots, Q_{\eta_{nn}})$ es una multinomial con función de probabilidad:

$$P\left(Q_{\eta_{11}} = a_1, \ldots, Q_{\eta_{nn}} = a_{nn}\right) = \frac{(a_1 + a_2 + \cdots + a_{nn})!}{a_1! a_2! \cdots a_{nn}!} p_{\eta_{11}}^{a_1} p_{\eta_{12}}^{a_2} \cdots p_{\eta_{nn}}^{a_{nn}}, \tag{23}$$

donde $a_1 + a_2 + \cdots + a_{nn} = N$.

La función de verosimilitud de la distribución (23) es:

$$L\left(p_{\eta_{11}}, p_{\eta_{12}}, \ldots, p_{\eta_{nn}}\right) = \frac{N!}{n_{\eta_{11}}! n_{\eta_{12}}! \cdots n_{\eta_{nn}}!} p_{\eta_{11}}^{n_{\eta_{11}}} p_{\eta_{12}}^{n_{\eta_{12}}} \cdots p_{\eta_{nn}}^{n_{\eta_{nn}}}, \tag{24}$$

y como $\sum_{i,j} p_{\eta_{ij}} = 1$, se sigue que

$$\begin{aligned} L\left(p_{\eta_{11}}, p_{\eta_{12}}, \ldots, p_{\eta_{nn}}\right) &= \frac{N!}{n_{\eta_{11}}! n_{\eta_{12}}! \cdots n_{\eta_{nn}}!} p_{\eta_{11}}^{n_{\eta_{11}}} p_{\eta_{12}}^{n_{\eta_{12}}} \cdots \\ &\quad \cdots p_{\eta_{nn-1}}^{n_{\eta_{nn-1}}} \left(1 - p_{\eta_{11}} - \ldots - p_{\eta_{nn-1}}\right)^{n_{\eta_{nn}}}. \end{aligned} \tag{25}$$

El logaritmo de la función de verosimilitud es

$$\begin{aligned} l\left(p_{\eta_{11}}, p_{\eta_{12}}, \ldots, p_{\eta_{nn}}\right) &= \ln\left(\frac{R!}{n_{\eta_{11}}! n_{\eta_{12}}! \cdots n_{\eta_{nn}}!}\right) + \sum_{i=1}^{n} \sum_{j=1}^{n-1} n_{\eta_{ij}} \ln\left(p_{\eta_{ij}}\right) \\ &\quad + n_{\eta_{nn}} \ln\left(1 - p_{\eta_{11}} - p_{\eta_{12}} - \ldots - p_{\eta_{nn-1}}\right). \end{aligned}$$

Para obtener los estimadores máximo verosímiles $\hat{p}_{\eta_{ij}}$ de $p_{\eta_{ij}}$ para todo $i, j = 1, 2, \ldots, n$, recurrimos al gradiente

$$\frac{\partial l\left(p_{\eta_{11}}, p_{\eta_{12}}, \ldots, p_{\eta_{nn}}\right)}{\partial p_{\eta_{ij}}} = 0,$$

de modo que

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

Entonces el estadístico de razón de verosimilitud es (Lehmann, 1986):

$$\begin{aligned} \lambda\left(Q\right) &= \frac{\frac{N!}{n_{\eta_{11}}! n_{\eta_{12}}! \cdots n_{\eta_{nn}}!} p_{\eta_{11}}^{(0) n_{\eta_{11}}} p_{\eta_{12}}^{(0) n_{\eta_{12}}} \cdots p_{\eta_{nn}}^{(0) n_{\eta_{nn}}}}{\frac{N}{n_{\eta_{11}}! n_{\eta_{12}}! \cdots n_{\eta_{nn}}!} \hat{p}_{\eta_{11}}^{n_{\eta_{11}}} \hat{p}_{\eta_{12}}^{n_{\eta_{12}}} \cdots \hat{p}_{\eta_{nn}}^{n_{\eta_{nn}}}} \\ &= N^N \prod_{i=1}^{n} \prod_{j=1}^{n} \left(\frac{p_{\eta_{ij}}^{(0)}}{n_{\eta_{ij}}}\right)^{n_{\eta_{ij}}}, \end{aligned}$$

donde $p_{\eta_{ij}}\left(0\right)$ denota la probabilidad del símbolo η_{ij} bajo la hipótesis nula.

Bajo el supuesto de independencia intra y entre las series, $\Upsilon\left(m\right) = -2\ln\left(\lambda\left(Q\right)\right)$ sigue asintóticamente una distribución Chi-Cuadrado con k grados de libertad, donde k es igual al número de parámetros desconocidos bajo

H_1 menos el número de parámetros desconocidos bajo H_0 (Lehmann, 1986). Luego,

$$\begin{aligned}\Upsilon(m) &= -2\ln(\lambda(Q)) \\ &= -2\left[N\ln N + \sum_{i=1}^{n}\sum_{j=1}^{n} n_{\eta_{ij}} \ln\left(\frac{p_{\eta_{ij}}^{(0)}}{n_{\eta_{ij}}}\right)\right] \sim \chi_k^2. \end{aligned} \tag{26}$$

Si, como indica la hipótesis nula, los procesos $\{x_s\}_{s\in S}$ e $\{y_s\}_{s\in S}$ son independientes entre si se cumplirá que

$$p_{\eta_{ij}}^{(0)} = p_{\sigma_i^x}^{(0)} p_{\sigma_j^y}^{(0)}.$$

Utilizando este resultado y teniendo en cuenta que

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{n_{\eta_{ij}}}{N} = 1,$$

podemos deducir que

$$\begin{aligned}\Upsilon(m) &= -2N\left[\ln N + \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{n_{\eta_{ij}}}{N} \ln\left(\frac{p_{\sigma_i^x}^{(0)} p_{\sigma_j^y}^{(0)}}{n_{\eta_{ij}}}\right)\right] \\ &= -2N\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{n_{\eta_{ij}}}{N} \ln\left(p_{\sigma_i^x}^{(0)} p_{\sigma_j^y}^{(0)}\right) - \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{n_{\eta_{ij}}}{N} \ln\left(\frac{n_{\eta_{ij}}}{N}\right)\right].\end{aligned}$$

Con la simbolización propuesta en la Sección 2.1 y teniendo en cuenta que la hipótesis nula implica que los procesos espaciales son $i.i.d.$ e independientes entre si, obtenemos que $p_{\sigma_i^x}^{(0)} = p_{\sigma_i^y}^{(0)} = C_{\sigma_i}^{m-1}/2^{(m-1)}$ para todo $i = 1, 2, \ldots, m$.

Por otro lado, teniendo en cuenta que $h_Z(m) = -\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{n_{\eta_{ij}}}{N} \ln\left(\frac{n_{\eta_{ij}}}{N}\right)$, el resultado es:

$$\Upsilon(m) = 2N\left\{2(m-1)\ln(2) - \left[\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{n_{\eta_{ij}}}{N} \ln\left(C_{\sigma_i^x}^{m-1} C_{\sigma_j^y}^{m-1}\right)\right] - h_Z(m)\right\}. \tag{27}$$

Por lo tanto hemos demostrado lo siguiente:

Teorema 1:

Sean $\{x_s\}_{s\in S}$ e $\{y_s\}_{s\in S}$ dos procesos espaciales estacionarios con $|S| = N$. Supongamos que dichos procesos han sido simbolizados con la aplicación de simbolización definida en la Sección 4.1.1. Denotemos por $h_Z(m)$ la entropía definida en (13) para una dimensión de encaje fija $m \geq 2$, con $m \in \mathbb{N}$. Si las series espaciales $\{x_s\}_{s\in S}$ e $\{y_s\}_{s\in S}$ son i.i.d. e independientes entre si, entonces el estadístico

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

se distribuye asintóticamente como una $\chi^2_{m^2+1}$.

Sea α un número real con $0 \leq \alpha \leq 1$ donde
$P\left(\chi_k^2 > \chi_\alpha^2\right) = \alpha$.

Entonces para contrastar

$H_0 : \{x_s\}_{s\in S}\ e\ \{y_s\}_{s\in S} \quad son \quad i.i.d. \quad e \quad independientes \quad entre \quad si,$

La regla de decisión en la aplicación del contraste $\Upsilon(m)$ con un nivel de confianza del $100(1-\alpha)\ \%$ es:

$Si \quad 0 \leq \Upsilon(m) \leq \chi^2_\alpha, \quad No \quad podemos \quad rechazar \quad H_0,$

$En \ caso \quad contrario, \quad Rechazamos \quad H_0.$

El contraste planteado puede generalizarse para k procesos espaciales. La estructura final del contraste multivariante es:

$$\Upsilon(m) = 2N\left\{k(m-1)\,ln(2) - \left[\sum_{i_1=1}^{n}\cdots\sum_{i_k=1}^{n}\frac{n_{\eta_{ij}}}{N}ln\left(\prod_{j=1}^{k}C_{\sigma_{ij}^{x_i}}^{m-1}\right)\right] - h_Z(m)\right\}, \tag{28}$$

siendo Z un proceso conjunto $k-dimensional$ que se distribuye asintóticamente como una $\chi^2_{m^k+1}$.

3.1.1. Consistencia del Contraste $\Upsilon(m)$

En la sección previa hemos obtenido el contraste de dependencia demostrando que su distribución asintótica se encuentra bien definida bajo la hipótesis nula. Ahora añadiremos la propiedad de consistencia bajo condiciones débiles y para una amplia gama de procesos espaciales.

La consistencia es una propiedad de suma importancia para un contraste estadístico, dado que nos asegura, asintóticamente, que la hipótesis nula será rechazada con probabilidad uno si ésta es falsa. Para el caso específico del contraste $\Upsilon(m)$, podemos afirmar que este rechazará la hipótesis nula de independencia siempre que la estructura de dependencia (lineal o no lineal) sea de orden menor a m.

Teorema 2:

Sean $\{x_s\}_{s\in S}$ e $\{y_s\}_{s\in S}$ dos procesos espaciales estacionarios y $m \geq 2$ con $m \in \mathbb{N}$. Entonces, bajo dependencia de orden menor que m, $\lim_{N\to\infty} P\left(\widehat{\Upsilon}(m) > C\right) = 1$, para todo $0 < C < \infty$, $C \in \mathbb{R}$.[1]

Cabe destacar que el parámetro m garantiza la propiedad de consistencia para el estadístico $\Upsilon(m)$. Debido a que el investigador debe fijar previamente este valor, adelantamos las siguientes consideraciones:

1. El valor mínimo de m es igual a 2 para cada serie.
2. El valor máximo de m dependerá del tamaño muestral N. Notemos que N deberá ser más grande que el número de símbolos $((n \times n) \leq N)$ para disponer de, al menos, el mismo número de $m - entornos$ como símbolos posibles. Siendo conservadores, como sugiere Rohatgi (1976), y en orden de tener una buena buena aproximación al límite de los valores tabulados de la distribución χ^2, el valor de las frecuencias esperadas debería ser igual ó mayor a 5. Esto implica que la dimensión de encaje debería ser fijada tal que $(5(n \times n) \leq N)$.

Por ejemplo, en nuestra particular simbolización no estándar, si establecemos un $m = 5$, la distribución conjunta tendría 25 símbolos. Entonces, necesitaríamos de, al menos, una muestra de 125 datos para disponer de una adecuada aproximación a la distribución χ^2.

[1] La prueba del Teorema 2 puede ser consultada en Herrera (2011).

3.1.2. Alternativa de Permutación para el Contraste de Independencia

Como hemos mencionado, la variable aleatoria Q_η (ecuación 21) no siempre brinda una buena aproximación a la distribución Binomial. Si la dimensión de encaje es grande, o el sistema espacial es irregular, el grado de solapamiento de los $m - entornos$ tiende a ser importante, provocando que Q_η no sea una buena aproximación a una variable aleatoria Binomial. Ante estas situaciones, presentamos una estrategia alternativa para contrastar la independencia por permutación.

En primer lugar, proponemos las siguientes hipótesis nula y alternativa:

$$H_0 : \{x_s\}_{s\in S} \; e \; \{y_s\}_{s\in S} \quad son \quad i.i.d. \quad e \quad independientes \quad entre \quad si, \tag{29}$$

$$H_1 : \; No \quad H_0. \tag{30}$$

Denotando por $\Psi_1 = \frac{1}{2N}\Upsilon$. El procedimiento del contraste permutado, con un número B de réplicas de permutación, esta compuesto por los siguientes pasos:

1. Computar el valor del estadístico $\hat{\Psi}_1$ para la muestra original $\{x_s\}_{s\in S}$ e $\{y_s\}_{s\in S}$.
2. Remuestreando $\{x_s\}_{s\in S}$ e $\{y_s\}_{s\in S}$, obtener dos series permutadas $\{x_s\,(b)\}_{s\in S}$ e $\{y_s\,(b)\}_{s\in S}$, donde b indica el número de repeticiones de permutación.
3. Para las series $\{x_s\,(b)\}_{s\in S}$ e $\{y_s\,(b)\}_{s\in S}$ estimar el estadístico:

$$\hat{\Psi}_1^{(b)} \tag{31}$$

4. Repetir $B-1$ veces los pasos 2 y 3 para obtener B realizaciones permutadas del estadístico, $\left\{\hat{\Psi}_1^{(b)}\right\}_{b=1}^{B}$.
5. Computar el $p_{boots} - valor$ permutado:

$$p_{permuts} - valor\left(\hat{\Psi}_1\right) = \frac{1}{B}\sum_{b=1}^{B}\tau\left(\hat{\Psi}_1^{(b)} > \hat{\Psi}_1\right) \tag{32}$$

   donde $\tau\,(\cdot)$ es una función indicadora que asigna 1 si la desigualdad es verdad y 0 en otro caso.

6. Rechazar la hipótesis nula de independencia intra y entre $\{x_s\}_{s\in S}$ e $\{y_s\}_{s\in S}$ si

$$p_{permuts} - valor\left(\hat{\Psi}_1\right) < \alpha \tag{33}$$

   para un tamaño nominal α.

4. Comportamiento en Muestras Finitas de $\hat{\Psi}_1$

En esta sección, presentamos el tamaño y potencia estimada del nuevo contraste, usando un experimento de simulación Monte Carlo.Para la hipótesis nula de $i.i.d.$ y series inter-dependientes, comparamos el contraste $\hat{\Psi}_1$ con el bien conocido contraste de autocorrelación espacial bivariante de Moran, I_{yx}.

En cada simulación, se utiliza una distribución espacial aleatoria de N localizaciones. Esto significa que la estructura espacial no posee una distribución regular y la aproximación bootstrap puede ser preferible, debido al problema de traslapamiento, para garantizar la robustez de la inferencia de los contrastes de independencia.

Los parámetros globales que utilizamos son los siguientes:

$$N \in \{100, 400, 1000\}\,,\ m \in \{4, 6, 8\}\,, \tag{34}$$

donde N es el tamaño muestral, m es la dimensión de encaje.

En el experimento, hemos buscado simular relaciones lineales y no lineales entre las variables x e y. En el primer caso, linealidad, controlamos la relación mediante el coeficiente de determinación esperado basado en una especificación como la siguiente:

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

La fortaleza de la relación puede ser deducida por medio del coeficiente R^2 esperado. Bajo la anterior ecuación, el coeficiente de determinación esperado entre las variables es igual a (asumiendo varianza unitaria de x y de ε así como incorrelación entre las dos variables):

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

Para analizar el tamaño empírico utilizamos un Proceso Generador de Datos simple (PGD desde ahora):

$$\begin{aligned} y &\sim \mathcal{N}(0,1)\,, \\ x &\sim \mathcal{N}(0,1)\,. \end{aligned} \tag{35}$$

Para los casos de potencia, hemos considerado procesos bivariantes del siguiente tipo:

$$\begin{gathered} y = F\left[x, Wy, Wx, \varepsilon\right];\ \varepsilon \backsim \mathcal{N}(0,1)\,, \\ x \backsim \mathcal{N}(0,1)\,, \end{gathered} \tag{36}$$

donde W es la usual matriz de contactos estandarizada por filas, especificada usando los $(m-1)$ vecinos más cercanos, y F es la correspondiente forma funcional. Los procesos simulados son tres de tipo lineal y tres de tipo no lineal denotados como **PGD1** hasta **PGD6**. Ellos son los siguientes:

PGD1: Intra-dependencia e inter-independencia. Proceso Lineal,

$$y = \rho W y + \varepsilon. \tag{37}$$

Los valores de ρ, denominado coeficiente de autocorrelación espacial, son $\rho \in \{0{,}4; 0{,}7; 0{,}9\}$.

PGD2: Intra-independencia e inter-dependencia. Proceso Lineal,

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

El parámetro β se ha considerado fijo en el valor 0.5. Los valores de θ han sido obtenidos como:

$$\theta = \sqrt{\frac{(m-1)\left(\beta^2\left(1-R^2\right)-R^2\right)}{R^2-1}}, \tag{39}$$

en orden de asegurar un determinado coeficiente R^2esperado. Hemos simulado tres valores para el R^2 esperado: $R^2 \in \{0{,}4; 0{,}6; 0{,}8\}$.

PGD3: Intra-dependencia e inter-dependencia. Proceso Lineal,

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

El coeficiente ρ se ha fijado en 0.4 y θ ha sido definido como en (39), donde $\beta = 0$, usando los mismos valores para el R^2 esperado: $R^2 \in \{0{,}4; 0{,}6; 0{,}8\}$.

PGD4: Intra-dependencia e inter-dependencia. Proceso No-lineal,

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

PGD5: Intra-independencia e inter-dependencia. Proceso No-lineal,

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

PGD6: Intra-dependencia e inter-dependencia. Proceso No-lineal,

$$y = {}^{1}\!/\!\left[(I-\rho W)^{-1}(\theta W x+\varepsilon)\right]. \tag{43}$$

Es obvio que en los casos **PGD4-PGD5-PGD6** se han invertido los valores obtenidos por los correspondientes procesos lineales **PGD1-PGD2-PGD3**. En todos los casos Cov $(x, \varepsilon) = 0$.

Los parámetros más importantes del ejercicio de simulación son m, que define la dimensión de encaje, ρ, que mide la intensidad de la intra-dependencia espacial y el R^2 esperado que define la intensidad de la inter-dependencia espacial. Finalmente, cada experimento ha sido repetido 1000 veces.

4.1. La Hipótesis Nula Conjunta: $i.i.d.$ más inter-independencia

Como hemos mencionado, en este caso haremos uso de contraste $\hat{\Psi}_1$ junto al estadístico de autocorrelación espacial de Moran, I_{yx}.

Brevemente, el contraste bivariante de Moran es un coeficiente tipo Mantel (Mantel, 1967) adaptado por Wartenberg (1985) como un índice para medir la correlación espacial cruzada entre dos variables. La expresión es la siguiente:

$$I_{xy} = \frac{\sum_{i=1}^{R}\sum_{\substack{j=1 \\ i\neq j}}^{R} y_i w_{ij} x_j}{S_0\sqrt{\mathrm{Var}(y)\mathrm{Var}(x)}}, \tag{44}$$

donde w_{ij} es el $(i,j) - \acute{e}simo$ elemento de la matriz de contactos W y S_0 es la suma de todos los elementos de W; Var(y) y Var(x) se refieren a las varianzas (estimadas) de las series x e y. La hipótesis nula de interés es que las variables son espacialmente inter-independientes. Czaplewski y Reich (1993) obtienen los momentos

del estadístico I_{xy} sobre todas las posibles $n!$ permutaciones aleatorias de los pares de valores $\{x_s; y_s\}_{s\in S}$. La expresión de estos momentos más bien complicada, como puede verse en las ecuaciones (25) y (60) del trabajo de Czaplewski y Reich (1993), aunque son computacionalmente manejables. Además, ellos muestran que, para tamaños muestrales desde moderados a grandes (es decir, N debe ser mayor a 40), el contraste bivariante de Moran, I_{yx}, se encuentra normalmente distribuido. Sin embargo, en orden de hacer comparaciones entre Ψ_1 e I_{xy} de manera simple, nosotros proponemos aproximar la distribución de este último contraste mediante bootstrap de acuerdo a la siguiente secuencia:

1. Computar el valor del estadístico $\hat{I}_{yx}$ desde la muestra original $\{x_s\}_{s\in S}$ e $\{y_s\}_{s\in S}$.
2. Remuestreando $\{x_s\}_{s\in S}$ e $\{y_s\}_{s\in S}$, obtenemos dos series bootstrapeadas $\{x_s(b)\}_{s\in S}$ e $\{y_s(b)\}_{s\in S}$, donde b es el número de la muestra bootstrapeada.
3. Para las series $\{x_s(b)\}_{s\in S}$ e $\{y_s(b)\}_{s\in S}$, estimar el estadístico: $\hat{I}_{yx}^{(b)}$.
4. Repetir $B-1$ veces los pasos 2 y 3 para obtener B realizaciones bootstrapeadas del estadístico $\left\{\hat{I}_{yx}^{(b)}\right\}_{b=1}^{B}$.
5. Computar el $p-valor$ del bootstrap estimado:

$$p_{boots}-valor\left(\hat{I}_{yx}\right)=\frac{1}{B}\sum_{b=1}^{B}1\left(\hat{I}_{yx}^{(b)}>\hat{I}_{yx}\right), \tag{45}$$

donde $1(\cdot)$ es una función indicadora que asigna 1 si la desigualdad es verdadera y 0 en cualquier otro caso.

6. Rechazar la hipótesis nula H_0 si

$$p_{boots}-valor\left(\hat{I}_{yx}\right)<\alpha, \tag{46}$$

para un tamaño nominal α.[2]

La Tabla[3] 1 muestra el tamaño estimado para los dos contrastes de independencia, $\hat{I}_{xy}$ y $\hat{\Psi}_1$. En general, los resultados son bastantes similares aunque hay una leve tendencia a sobre-estimar el tamaño en ambos casos, más claramente para el contraste $\hat{\Psi}_1$.

Cuadro 1: Tamaño Estimado de $\hat{\Psi}_1$ e $\hat{I}_{xy}$ el nivel del 5 %

	$\hat{\Psi}_1$			$\hat{I}_{xy}$		
m	4	6	8	4	6	8
$N=100$	4,9	–	–	6,0	-	-
$N=400$	6,3	6,2	6,4	5,2	5,6	5,4
$N=1000$	4,8	6,2	6,1	5,4	5,9	5,3

Nota: Permuts: 200. Nº de repeticiones: 1000.

$PGD1$ presenta únicamente intra-dependencia y los resultados aparecen en la Tabla 2. Nótese que este caso cae bajo la hipótesis nula de $\hat{I}_{xy}$ de Moran de no inter-dependencia.

[2]La aproximación asintótica de este contraste es la habitual: Sea α un número real con $0\leq\alpha\leq1$ y sea $N_{\varepsilon/2}$ tal que $Pr\left(|N(0;1)|>N_{\varepsilon/2}\right)=\alpha$. Para contrastar la hipótesis de (29) la regla de decisión del contraste I_{xy} con un $100(1-\alpha)$ % del nivel de confianza es: $Si\quad 0\leq z_{xy}\leq N_{\varepsilon/2}\quad No\quad rechazar\quad H_0, en\quad otro\ caso\ rechazar\quad H_0$, siendo z_{xy} el valor estandarizado del Moran bivariante, esto es $z_{xy}=\frac{|I_{xy}-E(I_{xy})|}{\sqrt{Var(I_{xy})}}$. En general, los resultados de la versión asintótica del estadístico de Moran son peores que la versión bootstrapeada particularmente, como es de esperar, para tamaños muestrales reducidos.

[3]Siguiendo la regla que cada símbolo debe tener al menos una frecuencia esperada de 5, para un tamaño muestral de $N=100$, nosotros solo consideramos el caso de $m=4$. La misma regla ha sido aplicada en los demás experimentos.

Cuadro 2: Potencia estimada del contraste $\hat{\Psi}_1$ e $\hat{I}_{xy}$ de Moran al nivel del 5 %

$PGD1$	$N = 100$	$N = 400$			$N = 1000$		
m	4	4	6	8	4	6	8
	$\hat{\Psi}_1$						
$\rho = 0{,}4$	39,0	96,0	85,0	71,5	100	99,0	96,0
$\rho = 0{,}7$	94,5	100,0	100,0	100,0	100,0	100,0	100,0
$\rho = 0{,}9$	100,0	100,0	100,0	100,0	100,0	100,0	100,0
$PGD1$	$\hat{I}_{xy}$						
$\rho = 0{,}4$	5,9	6,0	5,1	6,0	5,2	6,2	5,1
$\rho = 0{,}7$	10,1	9,8	13,4	14,2	9,1	12,7	11,7
$\rho = 0{,}9$	13,9	13,7	17,6	17,0	14,3	17,8	20,7

Nota: Permuts.: 200. Nº de repeticiones: 200. Ver nota al pie 3.

Para $PGD1$, nosotros esperamos una tasa de rechazo $\hat{I}_{xy}$ cercana al tamaño empírico de la Tabla 1. Sin embargo esto es solo verdad para pequeños valores del coeficiente de autocorrelación espacial, ρ. Claramente la presencia de la intra-correlación en una variable tiende a contaminar la conducta del contraste bivariante. Sin embargo, los resultados son satisfactorios para el contraste $\hat{\Psi}_1$: la potencia incrementa rápidamente cuando el tamaño muestral, la dependencia espacial o la dimensión de encaje se incrementan.

La Tabla 3 muestra la potencia estimada para ambos contrastes bajo el $PGD2$ (inter-dependencia únicamente). La potencia estimada de ambos contrastes es muy alta en casi todos los casos, alcanzando un valor máximo de 100 % para tamaños muestrales intermedios. La potencia estimada para el contraste $\hat{I}_{xy}$ es, por lo general, más alta que para el contraste $\hat{\Psi}_1$, especialmente para tamaños muestrales pequeños, independientemente del coeficiente R^2 esperado.

Para el $PGD3$ (intra- e inter-dependencia), los resultados son aún mejores, con valores de potencia estimada prácticamente del 100 % en todos los casos (Tabla 4).

Cuadro 3: Potencia estimada del contraste $\hat{\Psi}_1$ e $\hat{I}_{xy}$ de Moran al nivel del 5 %

$PGD2$	$N = 100$	$N = 400$			$N = 1000$		
m	4	4	6	8	4	6	8
	$\hat{\Psi}_1$						
$R^2 = 0{,}4$	46,0	98,5	100,0	98,5	100,0	100,0	100,0
$R^2 = 0{,}6$	76,5	100,0	100,0	100,0	100,0	100,0	100,0
$R^2 = 0{,}8$	83,5	100,0	100,0	100,0	100,0	100,0	100,0
$PGD2$	$\hat{I}_{xy}$						
$R^2 = 0{,}4$	99,7	100,0	100,0	100,0	100,0	100,0	100,0
$R^2 = 0{,}6$	100,0	100,0	100,0	100,0	100,0	100,0	100,0
$R^2 = 0{,}8$	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Nota: Permuts.: 200. Nº de repeticiones: 200. Ver nota al pie 3.

Cuadro 4: Potencia estimada del contraste $\hat{\Psi}_1$ e $\hat{I}_{xy}$ de Moran al nivel del 5 %

$PGD3$	$N = 100$	$N = 400$			$N = 1000$		
m	4	4	6	8	4	6	8
	$\hat{\Psi}_1$						
$R^2 = 0{,}4$	95,0	100,0	100,0	100,0	100,0	100,0	100,0
$R^2 = 0{,}6$	98,0	100,0	100,0	100,0	100,0	100,0	100,0
$R^2 = 0{,}8$	99,5	100,0	100,0	100,0	100,0	100,0	100,0
$PGD3$	$\hat{I}_{xy}$						
$R^2 = 0{,}4$	99,5	100,0	100,0	100,0	100,0	100,0	100,0
$R^2 = 0{,}6$	100,0	100,0	100,0	100,0	100,0	100,0	100,0
$R^2 = 0{,}8$	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Nota: Permuts: 200. Nº de repeticiones: 200. Ver nota al pie 3.

Las potencias estimadas son más bajas en los procesos no-lineales. Las Tablas 5, 6 y 7 muestran los resultados para $PGD4$ (intra-dependencia únicamente), $PGD5$ (inter-dependencia únicamente) y $PGD6$ (intra- e inter-dependencia), respectivamente. En todos los casos, la potencia estimada aumenta cuando el tamaño muestral se incrementa.

Cuadro 5: Potencia estimada del contraste $\hat{\Psi}_1$ e $\hat{I}_{xy}$ de Moran al nivel del 5 %

$PGD4$	$N = 100$	$N = 400$			$N = 1000$		
m	4	4	6	8	4	6	8
	$\hat{\Psi}_1$						
$\rho = 0{,}4$	20,5	75,5	57,5	45,5	100,0	95,0	93,0
$\rho = 0{,}7$	65,0	100,0	97,5	96,0	100,0	100,0	100,0
$\rho = 0{,}9$	99,0	100,0	100,0	100,0	100,0	100,0	100,0
$PGD4$	$\hat{I}_{xy}$						
$\rho = 0{,}4$	4,8	6,1	6,3	5,1	5,7	6,1	5,2
$\rho = 0{,}7$	7,2	5,9	5,7	5,3	5,9	6,3	6,2
$\rho = 0{,}9$	6,1	5,4	6,6	7,2	6,5	6,1	6,5

Nota: Permuts.: 200. Nº de repeticiones: 200. Ver nota al pie 3.

Los valores de R^2 esperado que aparecen en las Tablas 6 y 7 se presentan a modo informativo (la relación es no-lineal). Igual que en el caso lineal, la potencia estimada del contraste $\hat{\Psi}_1$, para inter- e intra-dependencia (Tabla 7), es claramente satisfactoria aún para tamaños muestrales reducidos.

Es importante destacar el deficiente desempeño del contraste bivariante de Moran. El $PGD4$ presenta intra-dependencia no-lineal en la variable y siendo ambas variables, x e y, independientes. El porcentaje de rechazos de este contraste es cercano al nominal pero los resultados de los otros dos $PGD's$ son, simplemente, inaceptables. Las Tablas 5 a 7 confirman que la $\hat{I}_{xy}$ de Moran es un contraste de correlación espacial inadecuado cuando la relación es no-lineal.

Cuadro 6: Potencia estimada del contraste $\hat{\Psi}_1$ e $\hat{I}_{xy}$ de Moran al nivel del 5 %

$PGD5$	$N = 100$	$N = 400$			$N = 1000$		
m	4	4	6	8	4	6	8
	$\hat{\Psi}_1$						
$R^2 = 0{,}4$	20,5	82,5	92,5	86,0	99,0	100,0	100,0
$R^2 = 0{,}6$	34,5	95,0	98,0	99,0	100,0	100,0	100,0
$R^2 = 0{,}8$	49,5	100,0	100,0	100,0	100,0	100,0	100,0
$PGD5$	$\hat{I}_{xy}$						
$R^2 = 0{,}4$	9,9	10,6	9,4	12,2	7,9	10,8	11,9
$R^2 = 0{,}6$	8,9	9,7	8,8	11,5	10,0	10,0	11,4
$R^2 = 0{,}8$	11,7	9,5	11,4	11,8	9,5	9,8	12,3

Nota: Permuts.: 200. Nº de repeticiones: 200. Ver nota al pie 3.

Cuadro 7: Potencia estimada del contraste $\hat{\Psi}_1$ e $\hat{I}_{xy}$ de Moran al nivel del 5 %

$PGD6$	$N = 100$	$N = 400$			$N = 1000$		
m	4	4	6	8	4	6	8
	$\hat{\Psi}_1$						
$R^2 = 0{,}4$	67,0	100,0	99,5	99,5	100,0	100,0	100,0
$R^2 = 0{,}6$	82,5	100,0	100,0	100,0	100,0	100,0	100,0
$R^2 = 0{,}8$	88,5	100,0	100,0	100,0	100,0	100,0	100,0
$PGD6$	$\hat{I}_{xy}$						
$R^2 = 0{,}4$	10,0	10,3	11,8	8,0	11,0	10,6	9,3
$R^2 = 0{,}6$	15,8	15,1	16,6	17,6	12,8	16,6	19,2
$R^2 = 0{,}8$	19,2	16,3	23,2	24,1	19,1	22,8	25,5

Nota: Permuts.: 200. Nº de repeticiones: 200. Ver nota al pie 3.

5. Conclusiones

Existe una incipiente literatura dedicada a la detección de autocorrelación espacial. Como menciona Anselin (1988), este es uno de los tópicos más importante en econometría espacial. Sin embargo, hemos detectado un importante déficit con respecto al análisis de independencia espacial entre pares o grupos de variables. El contraste de Moran bivariante, aunque muy intuitivo, es claramente insuficiente debido a que sufre severas limitaciones: linealidad y solo aplicable entre pares de ellas. En este trabajo se presenta una versión mejorada del mismo utilizando su versión bootstrapeada pero ésta continua siendo insuficiente para detectar dependencia espacial en relaciones no lineales.

Nuestra propuesta, el contraste $\Upsilon(m)$ (en su versión de contraste de permutación) no sufre de estas limitaciones: no se encuentra restringido a una forma funcional específica y puede ser generalizado al caso de más de dos variables o a variables de naturaleza diferentes (continuas, discretas). El estadístico es consistente y computacionalmente simple de obtener. Los resultados Monte Carlo presentados, desde nuestro punto de vista, son claramente satisfactorios.

Referencias

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

[2] Bivand, R. (1980). "A Monte Carlo Study of Correlation Coefficient Estimation with Spatially Correlated Observations", Quaestiones Geographicae, 6, pp.5-10.

[3] Cerioli, A. (1997). "Modified Tests of Independence in 2x2 Tables with Spatial Data", Biometrics, 53, pp. 619-628.

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

[5] Clifford, P. y S. Richardson (1985). "Testing for the Association between Two Spatial Processes", Statistics and Decissions, Suppl., 2, pp. 155-160.

[6] Cressie, N. (1993). Statistics for Spatial Data (revised version). John Wiley & Sons, New York.

[7] Czaplewski, R. y 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.

[8] Haining, R. (1991). "Bivariate Correlation with Spatial Data", Geographical Analysis, 23, pp. 210-227.

[9] Haining, R. (2003). Spatial Data Analysis. Theory and Practice. Cambridge University Press, Cambridge.

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

[11] Herrera, M. (2011). Causality. Contributions to Spatial Econometrics. Ph.D Thesis, Universidad de Zaragoza (España). Disponible en https://sites.google.com/site/spatialcausality/

[12] Hong, Y. y H. White (2005). "Asymptotic distribution theory for nonparametric entropy measures of serial dependence", Econometrica, 73, pp. 837-901.

[13] LeSage, J. y K. Pace (2009). Introduction to Spatial Econometrics. Chapman & Hall/CRC, Boca Raton.

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

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

[16] Matilla-García, M. y M. Ruiz Marín (2008). "A Non-parametric Independence Test Using Permutation Entropy", Journal of Econometrics, 144, pp. 139-155.

[17] Matilla-García, M. y M. Ruiz Marín (2009). "Detection of Non-linear Structure in Time Series", Economics Letters, 105, pp. 1-6.

[18] Ruiz, M., López, F. y A. Páez (2009). "Testing for Spatial Association of Qualitative Data Using Symbolic Dynamics", Journal of Geographical Systems, 10.1007/s10109-009-0100-1.

[19] Soon, S. (1996). "Binomial Approximation for Dependent Indicators", Statistica Sinica, 6, pp. 703–714.

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