Estadística Española • Volumen 54, número 177 / 2012, pp. 135-156

Distribution-free inference for *Q(m)* based on permutational bootstrapping: an application to the spatial co-location pattern of firms in Madrid

Fernando A. López[(*)]
Department of Quantitative Methods and Computing
Technical University of Cartagena

Antonio Páez[(**)]
Centre for Spatial Analysis
School of Geography and Earth Sciences, McMaster University

Abstract

The objective of this paper is to present a distribution-free inferential framework for the *Q(m)* statistic based on permutational bootstrapping. *Q*(*m*) was introduced in the literature as a tool to test for spatial association of qualitative variables, or more precisely, patterns of co-location/co-occurrence. The existing inferential framework for this statistic is based on asymptotic results. A challenge for these results is the need to limit the overlap in the neighborhoods of proximate observations, which tends to reduce the size of the sample, with consequent impacts on the size and power of the statistic. A computationally intensive inferential framework, such as presented in this paper, allows for greater versatility of *Q(m)*. We show that under the bootstrap version the issues with size are ameliorated and the test is more powerful. Furthermore, in this framework there is no longer the need to control for overlap, which allows for applications to variables with more categories and smaller sample sizes. The proposed approach is demonstrated empirically using a case study of co-location of business establishments in Madrid.

Key words: categorical data, spatial independence, distribution-free, business establishments, firm micro-data, Madrid.

JEL Classification: C21, R12

AMS Classification: 62M30, 62H11

(*) Fernando.lopez@upct.es

(**) Paezha@mcmaster.ca

Inferencia con *Q(m)* basada en bootstrappermutacional: explorando modelos de co-localización espacial de empresas en Madrid

Resumen

El objetivo de este artículo es evaluar el comportamiento del estadístico *Q(m)* bajo un marco inferencial basado en remuestreo permutacional. El estadístico *Q(m)*fue introducido en la literatura para contrastar la independencia de la distribución espacial de variables categóricas/cualitativas, siendo además un instrumento útil para explorar patrones de co-localización o co-ocurrencia de eventos. El marco inferencial bajo el que originalmente fue desarrollado está basado en el comportamiento asintótico del estadístico y requiere limitar el número de observaciones que de forma efectiva se utilizan en la muestra con el fin de evitar el solapamiento de observaciones espacialmente próximas. Esta reducción de la información que se suministra al contraste tiene importantes consecuencias sobre el tamaño y potencia del test en muestras pequeñas. El marco inferencial basado en bootstrap permutacional permite una mayor versatilidad de *Q(m)*. En este nuevo marco no es necesario controlar el grado de solapamiento de las *m*-historias y por tanto puede ser aplicado en aquellas situaciones en las que se disponga de un elevado número de categorías aunque el tamaño muestral sea pequeño. Un ejemplo del funcionamiento del contrate se utiliza para explorar los patrones de co-localización de las empresas en Madrid.

Palabras Clave: Datos Categóricos, Independencia Espacial, *bootstrap*, empresas, Madrid.

Clasificación JEL: C21, R12

Clasificación AMS: 62M30, 62H11

1. Introduction

A foundational problem in spatial analysis is the detection of spatial pattern in georeferenced variables. Initial work in the analysis of statistical maps was concerned with the case of qualitative (categorical) data or *k*-color maps, where *k* refers to the number of colors or categories (Moran 1948, Dacey 1968). The need to test for independence in the residuals obtained from linear regression models quickly drew attention to the analysis of continuous variables (Geary 1954; particularly p. 144). Subsequently, most work to date has been concerned with the analysis of spatial pattern for continuous variables (Anselin 1988, Griffith 1988, Haining 1990, Cressie 1993).

As reviewed by Ruiz et al. (2010), a burgeoning field of applications concerned with the analysis of qualitative data in a spatial context has been observed in recent years. This has prompted the development of tools that extend the analytical potential of the classical join-count statistic developed for *k*-color maps (Dacey 1968). These developments include the Local Indicators for Categorical Data (LICD) of Boots (2003, 2006), which are currently limited to binary data and regular lattices. Transiograms are similar to variograms, but instead of capturing the change in a continuous variable at

increasing spatial separations, measure the change in transition probability to a new categorical state as a function of distance (e.g. Li 2006).The Co-location Quotient (CLQ) of Leslie and Kronenfeld (2011), in contrast, is defined on the basis of the number of first neighbors that correspond to a certain category, say a_j, with respect to a certain reference category, say a_i (where i is possibly equal to j). Thus, although the classification system can contain k different categories, the CLQ is implemented for pairwise comparison of classes between first-order neighbors.

Another recent development, and the focus of this paper, is the *Q(m)* statistic of Ruiz et al. (2010). *Q(m)* is designed for the exploratory spatial analysis of qualitative variables, and can be used to contrast an empirical spatial distribution of values against the null hypothesis of a spatially random (i.e. independent) sequence. The statistic is defined for an arbitrary number of categories k, and can be implemented using a neighborhood (called an m-surrounding) that includes the m-1 nearest neighbors of an observation.The basis for *Q(m)* is the fact that there is a finite number of unique ways of arranging k categorical outcomes in a surrounding of size m (i.e. of patterns of co-location). Given an empirical sequence of values, it is possible to tally the frequency of appearance of each unique configuration, which is denoted by a *symbol*. A measure of symbolic entropy can then be calculated to assess the degree of randomness or ordering in the sequence: in a highly ordered map, a small number of symbols (unique configurations of map elements) will appear with high frequency. On the contrary, in a spatially random map, all symbols will tend to appear with similar relative frequencies.

Q(m) has features that make it an attractive alternative for the exploratory analysis of georeferenced qualitative variables. The statistic is intuitive and easily interpretable. The relative frequency of the symbols at the core of the statistic can be retrieved and used to further explore the characteristics of the spatial distribution. Also, the use of symbols presents additional opportunities for spatial analysis (see Páez, Ruiz, López and Logan 2012).

The inferential framework for the statistic, based on asymptotic results, was presented in Ruiz et al. (2010). It was shown there that the statistic is asymptotically χ^2 distributed, which allows for testing departures from the null hypothesis at a desired level of significance α. The finite sample properties of the test were explored by Ruiz et al. (2010) by means of Monte Carlo simulations, the results of which provide evidence that as the values of k (number of categories) and/or m (size of the neighborhood) increase, convergence to asymptotic results can be slower, in particular for smaller samples. This issue is compounded by the need to limit the overlap of m-surroundings of proximate observations in order to achieve good conformity with the underlying assumptions used to derive the asymptotic results. A consequence of this is that the sample size (S) that can effectively be used in the analysis is often only a subset of the number of observations actually available (N). This also has an impact on the size and power of the statistic.

The objective of this paper is to present an alternative inferential framework for the *Q(m)* statistic that does not depend on asymptotic results. Use of distribution-free approaches – for instance, based on permutational boostrapping – is appropriate in situations where asymptotic or exact results are not available (e.g. Anselin 1995). In

some applications the underpinnings of the asymptotic theory may not be present. Lin et al.(2011), for example, investigated the performance of Moran's *I* using asymptotic results and bootstrapping, and demonstrated that under ideal conditions both approaches are equivalent. However, when there are departures from these conditions (i.e. non-normality), or for densely connected systems, the distribution-free version is superior. Yet another reason for investigating a distribution-free approach in the present case is that larger values of *k* and *m* impose important limitations for *Q(m)* in applications to smaller samples. Use of a computational approach may provide a valuable alternative for inference in such situations.

The structure of the paper is as follows. In the following section we introduce the *Q(m)* statistic. An inferential framework based on permutational bootstrapping is described next. This is followed by the results of numerical experiments that provide useful information to decide which inferential framework might be more appropriate given the characteristics of the data. An empirical case study, of co-location patterns of firms in Madrid, with a focus on industrial activities, helps to demonstrate the application of the statistic and inferential approaches. Finally, we summarize our findings and suggest directions for further research.

2. The *Q(m)* statistic

Q(m) was introduced by Ruiz et al. (2010) as a tool to explore geographical co-location/co-occurrence of qualitative data. Consider a spatial variable $\boldsymbol{X}$ which is the result of a qualitative process with a set number of categorical outcomes a_j (j=1,...,k). The spatial variable is observed at a set of fixed locations indexed by their coordinates $\boldsymbol{s}_i$ (i=1,..., N), so that at each location $\boldsymbol{s}_i$ where an event is observed, X_i takes one of the possible values a_j.

Since the observations are georeferenced, a spatial embedding protocol can be devised to assess the spatial property of co-location. Let us define, for an observation at a specified location, say s_0, a surrounding of size m, called an m-surrounding. The m-surrounding can be constructed in various ways – the one proposed by Ruiz et al. (2010) beingbased on a distance criterion. Using this criterion, the m-surrounding is the set of m-1 nearest neighbors from the perspective of location $\boldsymbol{s}_0$. In the case of distance ties, a secondary criterion can be invoked based on direction. Various embedding protocols are discussed in Ruiz et al. (2010, 2011) and Páez et al. (2012). A general rule for different protocols is that they must satisfy the uniqueness of observations within a given m-surrounding (i.e. the same observation should not be present more than once). The analyst can define rules of proximity in other ways if desired.

Once that an embedding protocol is adopted and the elements of the m-surrounding for location $\boldsymbol{s}_0$ have been determined, a string can be obtained that collects the elements of the local neighborhood (the m-1 nearest neighbors) of the observation at s_0. The m-surrounding can then be represented in the following way:

$$X_m(\boldsymbol{s}_0) = \left(X_{s_0}, X_{s_1}, ..., X_{s_{m-1}}\right) \quad [1]$$

Since each observation X_s takes one of k possible values, and there are m observations in the m-surrounding, there are exactly k possible unique ways in which those values can co-locate. This is the number of permutations with replacement. For instance, if k=2 (e.g. the possible outcomes are a_1=0 and a_2=1) and m=3, the following eight unique patterns of co-location are possible (the number of symbols is n_σ=8): {0,0,0}, {1,0,0}, {0,1,0}, {0,0,1}, {1,1,0}, {1,0,1}, {0,1,1}, and {1,1,1}. Each unique co-locationtype can be denoted in a convenient way by means of a symbol σ_i (i=1, 2,...,k^m). It follows that each site can be uniquely associated with a specific symbol, in a process termed *symbolization*. In this way, we say that a location s is of type σ_i if and only if $X_m(\mathbf{s})=\sigma_i$. Equivalent symbols (see Páez, et al. 2012) can be obtained by counting the number of occurrences of each category within an m-surrounding. This surrenders some topological information (ordering within the m-surrounding is lost) in favor of a more compact set of symbols, since the number of combinations with replacement, which is calculated as $k(k+1)...(k+m-1)/m!$, is in general < k^m. The equivalent symbols (denoted by σ^*) for k=2 and m=3 are as follows: {3,0}, {2,1}, {1,2}, {0,3}. The symbols in this case denote the number of zeros and ones in the m-surrounding (i.e. {1,2} means that one observation is of category a_1, and two are of category a_2).

It is straightforward, once that a set of locations $\mathbf{s}$ where observations have been recorded is symbolized (i.e. the unique symbol for each location has been determined), to calculate the relative frequency of each symbol. This is simply the number of times that σ_i is observed (n_{σ_i}), divided by the number of symbolized locations (S):

$$p_{\sigma_i} = \frac{n_{\sigma_i}}{S} \qquad [2]$$

These frequencies can be used to calculate a measure of symbolic entropy as follows:

$$h(m) = -\sum_j p_{\sigma_j} \ln\left(p_{\sigma_j}\right) \qquad [3]$$

The entropy function, it can be easily ascertained, has a theoretical lower bound of zero when only one symbol is observed in the empirical series. Accordingly, as the symbolic entropy approaches zero, this tends to indicate that the map is highly organized. The value of the entropy function under the null hypothesis of a spatial random sequence, call it $\eta(m)$, is derived by Ruiz et al. (2010). The *Q(m)* statistic, finally, is a likelihood ratio test that contrasts the symbolic entropy of the empirical sequence, to the symbolic entropy of a random sequence:

$$Q(m) = 2S\left(\eta(m) - h(m)\right) \qquad [4]$$

Q(m) is asymptotically χ^2 distributed, with degrees of freedom equal to the number of symbols minus one. In order to derive the asymptotic distribution, a random variable $Z_{\sigma_i s}$ is defined as follows:

$$Z_{\sigma_i s} = \begin{cases} 1 \text{ if } X_m(s) = \sigma_i \\ 0 \text{ otherwise,} \end{cases} \quad [5]$$

It can be seen that $Z_{\sigma_i s}$ is a Bernoulli variable with probability of "success" p_{σ_i}, where "success" means that location ***s*** corresponds to symbol σ_i. As is always the case $\sum_j p_{\sigma_j} = 1$. The total number of cases where *s* is of type σ_i is:

$$Y_{\sigma_i} = \sum_{s \in S} Z_{\sigma_i s} \quad [6]$$

which is a sum of Bernoulli variables, and is bounded between 0 (when the symbol is not observed) and *S* (when every symbolized location is of type σ_i).

In order to derive the test under asymptotic conditions (see proof of Theorem 1 in Ruiz, et al. 2010), we must assume that Y_{σ_i} is a binomial random variable. The sum of Bernoulli variables can be approximated to a binomial random variable under two conditions(see Soon 1996): (i) the probability of success p_{σ_i} is small for all *i*; and (2) the dependency between $Z_{\sigma_i s}$ and $Z_{\sigma_i t}$ is weak, between locations *s* and *t* ($s \neq t$). Condition (i) is satisfied by the way the symbols are constructed (the number of symbols tends to be large, and therefore the probability of success for each symbol is low under the null). In contrast, condition (ii) is more challenging. Clearly, $Z_{\sigma_i s}$ and $Z_{\sigma_i t}$ will not be independent if the *m*-surroundings of locations *s* and *t* overlap. In order to address this issue, Ruiz et al. (2010) proposed to conduct the analysis using a sub-sample of observations (size *S*), selected in such a way that the maximum degree of overlap between proximate *m*-surroundings is controlled. This approach seems to work reasonably well, based on the evidence of numerical experiments reported in Ruiz et al. (2010). On the other hand, the need to control the degree of overlap has the unfortunate consequence of reducing the size of the sample that effectively can be used for analysis (i.e. $S<N$ since not all observations are symbolized). This, in turn, can impair the applicability of *Q(m)* in situations where *k* and/or *m* are large, or where *N* is small. These situations (combined or even in isolation), tend to increase the number of symbols (n_σ) relative to the size of the sample. Greater values of *m*, as well, mean that the potential for overlap, and consequently dependencies, is greater. Hence the motivation for introducing an alternative inferential framework for the statistic based on a distribution-free approach.

3. Distribution-free inference for *Q(m)*

The principle behind distribution-free approachesis described by Bivand (2009): "[these] procedures [provide] a way to examine the distribution of the statistic of interest by exchanging at random the observed values between observations, and then comparing the simulated distribution under the null hypothesis of no spatial patterning with the observed value of the statistic in question." This description highlights the

computationally intensive nature of bootstrapping, particularly the random exchange of values between locations, and the recalculation of the statistic of interest. Computationally intensive approaches have been used for inference in previous work in spatial data analysis, including in applications of LISA (Anselin 1995), the Lagrange Multiplier test for spatial error autocorrelation (Born and Breitung 2011), the join-count statistic (Stevens and Jenkins 2000), Ripley's *k*-function (Marcon and Puech 2003), and CLQ (Leslie, et al. 2011).

Distribution-free approaches, in particular using bootstrapping, are attractive in situations where exact or asymptotic results are not available. It is possible as well that the conditions required for asymptotic results are difficult to attain. Or, as is the present case, the conditions can be met, but at a certain cost, for instance reduced sample size. A distribution-free approach can be computationally expensive, but is still manageable for moderately sized samples. The simulation procedure for *Q(m)*, given a fixed embedding dimension $m \geq 2$ with a number *r* of replications, is composed of the following steps:

– Compute the value of the statistic *Q(m)* for the original samples $\{X_{\mathbf{s}}\}_{\mathbf{s}\in S}$.

– Re-label the set of coordinates by randomly drawing from the list of outcomes without replacement, to obtain the series $\left\{X_s^r\right\}_{s\in S}$ where *r* is the index of the replication.

– Calculate the bootstrapped statistic $Q_B^r(m)$ for the simulated sample $\left\{X_s^r\right\}_{s\in S}$.

– Repeat steps 2 and 3 T-1 times to obtain T realizations of the bootstrapped statistic $\left\{Q_B^r(m)\right\}_{r=1}^{T}$.

– Compute the pseudo-probability as: $p_b = \frac{1}{T}\sum_{r=1}^{R} I\left(Q_B^r(m) > Q(m)\right)$ where $I(\bullet)$ is the indicator function which assigns a value of 1 to a true statement and 0 otherwise.

– Reject the null hypothesis if $p_b < \alpha$ for a nominal size α.

It is worth noting that *Q(m)* is a global independence indicator. In addition to this indicator of co-location, one might be interested in the empirical distribution of each symbol – that is, the frequency of different type of co-location – and its relationship to the expected distribution under the null. Using a procedure similar to the one described above, it is possible to compute the 100(1-α)% confidence interval for the relative frequency of a symbol σ_i, p_{σ_i}, by computing its bootstrapped realization $p_{\sigma_i}^r$ for each of the permuted series $\left\{X_s^r\right\}_{s\in S}$, for every symbol as follows:

$$I_\sigma(\alpha) = \left(I_{\alpha/2}, I_{1-\alpha/2}\right) \qquad [7]$$

where I_α is the percentile α of the distribution of $\left\{p_{\sigma_i}^r\right\}_{r=1}^T$. Accordingly, we will say that the null is rejected at the α level for symbol σ_i whenever the relative frequency of the symbol falls outside of the confidence interval, i.e. when $p_{\sigma_i} \notin I_{\sigma_i}(\alpha)$. The interval of confidence can be used to determine, separately from the global significance of the statistic, whether a specific symbol is observed more or less frequently than expected by chance.

4. Numerical experiments

In order to assess the performance of the permutational bootstrapping inferential approach, in this section we report the results of a series of numerical experiments. The experiments have two objectives. First, we compare the performance of the test under asymptotic results and bootstrapping, in terms of size and power for the case of small samples. And secondly, we explore the behavior of the test in extreme situations where the number of observations is small relative to the number of symbols.

4.1 Data generation process

With respect to the first objective stated above, we use the same experimental design described in Ruiz et al. (2010). However, for brevity of exposition, we consider only situations where the distribution of events is irregular and categorical outcomes are not equally probable (categories are not observed with equal frequency).

In order to obtain categorical random variables with controlled degrees of spatial dependence, we have designed a two-stage data generating process. Firstly, we simulate autocorrelated data using the following model:

$$\mathbf{Y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon} \qquad [8]$$

where $\varepsilon \sim N(0,1)$, $\mathbf{I}$ is the identity matrix, ρ is a parameter of spatial dependence, and $\mathbf{W}$ is a connectivity matrix that determines the set of spatial relationships among points. In the second step of the data generation process, the continuous spatially autocorrelated variable $\boldsymbol{Y}$ is used to define a discrete spatial process as follows. Let b_{ij} be defined by:

$$P(Y \le b_{ij}) = \frac{i}{j} \text{ with } i < j \qquad [9]$$

Let $A = \{a_1, a_2, ..., a_k\}$ and define the discrete spatial process as:

$$X_s = \begin{cases} a_1 & \text{if } Y_s \le b_{1k} \\ a_i & \text{if } b_{i-1k} < Y_s \le b_{ik} \\ a_k & \text{if } Y_s > b_{k-1k} \end{cases} \qquad [10]$$

The data are generated using equation (8) with the connectivity matrix defined in terms of first-order contiguity. Matrix **W** is row-standardized for the calculations.The number of replications is 1,000 for the experiments on size and 200 for the experiments on power.

4.2 Size and power of tests: Comparison between asymptotic and bootstrapped tests

The experimental design in this case covers three values of k (k=2, 3, and 4) and six values of N (N=100, 400, 900, 1,600, 2,500 and 3,600). The results of the experiments are reported in Tables 1, 2, and 3 for different values of k. It is important to note that the degree of overlap (o_d) has been set to a small value, to try to approximate the theoretical conditions as closely as possible (i.e. by reducing the dependencies between locations) for the asymptotic implementation of the test. The degree of overlap determines the number of observations S that can be symbolized for testing under asymptotic results. This is not a consideration in the case of testing under bootstrapping, and all observations are symbolized in every case (i.e. S=N).

With respect to the size of the test, the results indicate that bootstrapping provides more stable results than the asymptotic version, with only minor deviations from the nominal level of 0.05, irrespective of the size of m. The size is not affected by the number of observations in the sample N or the number of categories k. This is true for the calculation of *Q(m)* using both standard and equivalent symbols.

With regards to the power of the tests, comparison between the asymptotic and bootstrapped implementations clearly shows that bootstrapping has a clear advantage with greater power, for all combinations of parameters m, k, and N used in the experiment. It can also be observed that the power of the bootstrapped version of the test is slightly superior for lower values of m when N is small. As N increases the effect of m disappears. Use of equivalent symbols results in a slight increase in power.

Table 1

Size and Power of *Q(m)* for *k*=2. Irregular lattice and p_1=1/4; p_2=3/4

				Asymptotic test (see Ruiz et al. 2010)				Permutational bootstrapping test (*S*=*N*)							
				standard symbols				standard symbols				equivalent symbols			
N	*S*	m	o_d	ρ=0	ρ=0.2	ρ=0.5	ρ=0.9	ρ=0	ρ=0.2	ρ=0.5	ρ=0.9	ρ=0	ρ=0.2	ρ=0.5	ρ=0.9
100	49	3	1	0.027	0.042	0.090	0.704	0.051	0.099	0.489	0.998	0.052	0.111	0.509	1.000
	97	4	3	0.053	0.054	0.170	0.897	0.056	0.087	0.457	0.996	0.046	0.086	0.498	0.996
400	199	3	1	0.025	0.059	0.354	1.000	0.055	0.102	0.408	0.992	0.042	0.233	0.970	1.000
	199	4	2	0.036	0.062	0.423	1.000	0.052	0.227	0.970	1.000	0.049	0.204	0.967	1.000
900	449	3	1	0.038	0.078	0.764	1.000	0.050	0.165	0.946	1.000	0.050	0.496	1.000	1.000
	449	4	2	0.048	0.113	0.817	1.000	0.050	0.161	0.862	1.000	0.050	0.447	1.000	1.000
	299	5	2	0.056	0.113	0.685	1.000	0.051	0.462	1.000	1.000	0.044	0.387	1.000	1.000
1600	799	3	1	0.036	0.116	0.950	1.000	0.056	0.377	1.000	1.000	0.049	0.754	1.000	1.000
	799	4	2	0.046	0.144	0.977	1.000	0.043	0.304	0.999	1.000	0.065	0.647	1.000	1.000
	532	5	2	0.067	0.126	0.942	1.000	0.049	0.720	1.000	1.000	0.054	0.596	1.000	1.000

Table 2

Size and Power of *Q(m)* for *k*=3. Irregular lattice and p_1=1/8; p_2=3/8; p_3=4/8

				Asymptotic test (see Ruiz et al. 2010)				Permutational bootstrapping test (*S*=*N*)							
				standard symbols				standard symbols				standard symbols			
N	*S*	m	o_d	ρ=0	ρ=0.2	ρ=0.5	ρ=0.9	ρ=0	ρ=0.2	ρ=0.5	ρ=0.9	ρ=0	ρ=0.2	ρ=0.5	ρ=0.9
400	199	3	1	0.027	0.042	0.090	0.704	0.050	0.258	0.995	1.000	0.056	0.304	0.998	1.000
900	449	3	1	0.025	0.059	0.354	1.000	0.053	0.550	1.000	1.000	0.046	0.637	1.000	1.000
	449	4	2	0.036	0.062	0.423	1.000	0.056	0.434	1.000	1.000	0.051	0.472	1.000	1.000
1600	799	3	1	0.038	0.078	0.764	1.000	0.052	0.840	1.000	1.000	0.052	0.906	1.000	1.000
	799	4	2	0.048	0.113	0.817	1.000	0.055	0.704	1.000	1.000	0.058	0.729	1.000	1.000
2500	1249	3	1	0.036	0.116	0.950	1.000	0.049	0.972	1.000	1.000	0.043	0.986	1.000	1.000
	1249	4	2	0.046	0.144	0.977	1.000	0.054	0.914	1.000	1.000	0.056	0.907	1.000	1.000

Table 3

Size and Power of the *Q*(*m*) test for *k*=4. p_1=1/12; p_2=2/12; p_3=3/12; p_4=6/12

				Asymptotic test (see Ruiz et al. 2010)				Permutational bootstrapping test (*S*=*N*)							
				standard symbols				standard symbols				standard symbols			
N	*S*	m	o_d	ρ=0	ρ=0.2	ρ=0.5	ρ=0.9	ρ=0	ρ=0.2	ρ=0.5	ρ=0.9	ρ=0	ρ=0.2	ρ=0.5	ρ=0.9
900	449	3	1	0.039	0.085	0.755	1.000	0.057	0.875	1.000	1.000	0.054	0.872	1.000	1.000
1600	799	3	1	0.052	0.103	0.965	1.000	0.054	0.980	1.000	1.000	0.058	0.983	1.000	1.000
2500	1249	3	1	0.036	0.149	0.999	1.000	0.044	1.000	1.000	1.000	0.039	1.000	1.000	1.000
3600	1799	3	1	0.038	0.240	1.000	1.000	0.047	1.000	1.000	1.000	0.050	1.000	1.000	1.000
	1799	4	2	0.098	0.399	1.000	1.000	0.051	0.990	1.000	1.000	0.049	1.000	1.000	1.000

4.3 *Q(m)* with small sample size relative to number of symbols

Another situation of interest is when the sample size (*N*) and/or the number of symbolized locations (*S*) is small relative to the number of symbols (n_σ), since this affects convergence of the distribution of *Q*(*m*) to the asymptotic χ^2 distribution. A problem in this case is that the size of the test, which indicates the probability of detecting false positives, increases. Since the number of symbols standard (n_σ) and equivalent ($n_{\sigma*}$) depends on *k* and *m*, this may impose some limitations to the asymptotic approach when working with multiple categories, or restrict the analysis to smaller *m*-surroundings. In this section we report the results of experiments designed to evaluate the size and power of *Q*(*m*) when testing is conducted using bootstrapping, and N/n_σ is small (<5). The experimental design in this case covers three values of *k* (*k*=5, 7, and 10), four values of *N* (*N*=50, 100, 200, and 400), and three values of *m* (*m*=2, 3, and 4), using standard and equivalent symbols.

The results of the experiment are reported in Table 4. With respect to size, the values obtained are slightly higher to those reported in the previous set of experiments, but still close to the nominal value of α=0.05, and less than α=0.10 in every case. The power of the test indicates that it might be difficult to identify patterns with moderately strong association in the case of very small samples ($N\approx50$). In contrast, the power is high for stronger patterns of spatial association in small samples, and for moderately sized samples ($N\geq200$) even under extreme conditions ($N/n_\sigma<2$).

Table 4

Size and Power of *Q(m)*: bootstrapped test and $N/n_\sigma<5$

					Permutational bootstrapping test							
					standard symbols				equivalent symbols			
	N	m	n_σ	n_{σ^*}	ρ=0	ρ=0.2	ρ=0.5	ρ=0.9	ρ=0	ρ=0.2	ρ=0.5	ρ=0.9
k=5	50	2	25	15	0.058	0.087	0.235	0.871	0.060	0.097	0.287	0.922
		3	125	35	0.055	0.076	0.190	0.911	0.066	0.069	0.233	0.938
		4	625	70	0.063	0.062	0.155	0.702	0.066	0.074	0.163	0.805
	100	2			0.072	0.107	0.479	1.000	0.068	0.118	0.521	0.999
		3			0.062	0.085	0.376	0.998	0.062	0.092	0.437	0.998
		4			0.070	0.093	0.269	0.996	0.066	0.082	0.320	1.000
	200	2			0.069	0.152	0.741	1.000	0.064	0.161	0.776	1.000
		3			0.056	0.131	0.683	1.000	0.063	0.138	0.734	1.000
		4			0.054	0.116	0.550	1.000	0.061	0.120	0.614	1.000
	400	2			0.075	0.198	0.949	1.000	0.071	0.215	0.954	1.000
		3			0.064	0.194	0.932	1.000	0.058	0.200	0.942	1.000
		4			0.063	0.171	0.842	1.000	0.062	0.180	0.881	1.000
k=7	50	2	49	28	0.064	0.089	0.210	0.852	0.062	0.115	0.308	0.938
		3	343	84	0.068	0.071	0.178	0.864	0.054	0.087	0.276	0.951
	100	2			0.053	0.096	0.396	0.997	0.062	0.129	0.594	1.000
		3			0.051	0.077	0.283	0.993	0.049	0.097	0.488	0.999
	200	2			0.061	0.126	0.759	1.000	0.066	0.204	0.913	1.000
		3			0.055	0.132	0.592	1.000	0.055	0.165	0.854	1.000
	400	2			0.071	0.244	0.983	1.000	0.067	0.344	0.999	1.000
		3			0.062	0.169	0.943	1.000	0.065	0.257	0.994	1.000
k=10	50	2	100	55	0.053	0.066	0.153	0.725	0.065	0.103	0.330	0.960
		3	1000	220	0.066	0.064	0.147	0.573	0.068	0.092	0.216	0.846
	100	2			0.049	0.073	0.284	0.997	0.061	0.139	0.637	1.000
		3			0.056	0.079	0.232	0.983	0.058	0.108	0.385	1.000
	200	2			0.059	0.095	0.599	1.000	0.069	0.245	0.948	1.000
		3			0.053	0.097	0.447	1.000	0.055	0.132	0.680	1.000
	400	2			0.065	0.146	0.949	1.000	0.051	0.398	1.000	1.000
		3			0.044	0.148	0.830	1.000	0.048	0.217	0.967	1.000

n_σ=number of standard symbols; n_{σ^*}=number of equivalent symbols;
for k=5 p_1=0.1; p_2=0.4;p_3=0.2;p_4=0.2;p_5=0.1;
with k=7 p_1=0.1; p_2=0.2 ;p_3=0.2;p_4=0.2;p_5=0.1; p_6=0.1; p_7=0.1;
for k=10 p_1=0.1; p_2=0.05 ;p_3=0.15; p_4=0.15 ;p_5=0.1 ; p_6=0.1; p_7=0.1; p_8=0.1 ; p_9=0.1; p_{10}=0.05

4.4 Practical recommendations

The experiments reported above suggest some practical recommendations for the selection of testing approach, using asymptotic results or bootstrapping. Use of asymptotic testing is preferred to bootstrapping when the sample size is large and the number of categories k is low, as long as the degree of overlap between m-surroundings is kept low. In a situation like this, bootstrapping can be onerous in terms of computational load. Bootstrapping would be appropriate in situations where the sample size is small and the number of

categories is large. In this case, the computational requirements are manageable, and the bootstrapped version of the test provides good power with reasonable size. In the case of larger samples, if the number of categories is high, bootstrapping might be the only alternative for testing, if for the asymptotic test $S/n_{\sigma}<5$.

With respect to the size of the *m*-surrounding, lower values of *m* lead to fewer symbols, which, as suggested by the numerical experiments, give a more powerful test. Ideally, the selection of *m* must be reflective of the true scale of the spatial process (Matilla-García and Marin 2011); in practice, this scale is not always evident, and other practical considerations may be relevant. The selection between standard and equivalent symbols depends on the objective of the analysis. Standard symbols retain more information about the pattern of co-location which allows the analyst to detect, for instance, asymmetric patterns. For instance, the following symbols {0,0,1} and {1,0,0} are equivalent as far as the number of categories is concerned: {2,1}. When standard symbols are used, bootstrapped-based testing can provide an advantage since overlapping between *m*-surroundings is not a consideration, and the full sample can be used. Equivalent symbols can be used if detailed information on the ordering of co-located events is not essential.

5. Application: Co-location of business establishments in Madrid

A broad consensus exists about the competitive advantage of spatial proximity between businesses, due to reduced transfer costs, the emergence of economies of agglomeration and scale, and the ease of exchange of information, among other factors (e.g. Hoover and Giarratani 1984, Sorenson 2003, Jimenez and Junquera 2010). An important body of research is concerned with the detection of geographical patterns in the distribution of businesses and industrial activities (e.g. Ellison and Glaeser 1997, Arbia 2001, Espa, Arbia and Giuliani). Traditionally, this research has relied on information aggregated at the level of administrative units (e.g. counties or regions), but increasingly use is made of micro-data in a trend that Arbia (2011) identifies as the emergence of spatial micro-econometrics. The availability of micro-data creates an ideal opportunity for the application of techniques that rely on information at the level of individual firms (Albert, Casanova and Orts 2011, Leslie, et al. 2011, Paez, Trepanier and Morency 2011). In this section, we are interested in the patterns of co-location of firms in Madrid. Analysis using *Q(m)* helps to address the following research questions: (1) are firms of different types co-located in non-random ways? (2) If so, which types of business tend to co-locate? Does co-location tend to happen with firms within the same or different sectors? Analysis is conducted with asymptotic or bootstrapped tests, as appropriate given the characteristics of the data.

Two sources of data form the basisfor our application.

The first one is the Central Business Directory (DIRCE for the acronym in Spanish) compiled by Instituto Nacional de Estadística (INE; the National Statistics Institute). DIRCE is a census of all businesses located in Spain, classified according to their economic activity (CNAE-93 groups: National Classification of Economic Activities). Georeferencing in this directory is relatively coarse, and the location of firms is

identified at the level of Spanish provinces (equivalent to NUTS 3 level). From DIRCE we are able to extract all businesses located in Madrid (ES300 in Eurostat terminology). These business records include the type of economic activity at the 2-digit level according to CNAE-93. The database corresponding to the year 2009 contains 511,804 records, with distribution by activity code as shown in the Appendix.

The second source of data is the database SABI (acronym for Iberian Balance Analysis System) which collects economic information, with a focus on accounting information, for an extensive list of businesses in Spain and Portugal. The database corresponding to 2009 contains 213,282 records for businesses in Madrid. Unlike DIRCE, SABI is not a census and coverage is uneven for different regions in Spain – Madrid is one of the regions with the best coverage, and approximately 40% of all businesses listed in DIRCE are also contained in SABI. An important advantage of working with SABI is that individual firms are georeferenced at the level of x and y coordinates.

In order to explore the patterns of co-location of businesses in Madrid, we take a stratified sample from records in SABI, maintaining the known composition, in terms of percentage of firms in 2-digit level sectors, from the census information in DIRCE. In this way we obtain 51,183 records. These records cover 52 different categories of economic activity, corresponding to four primary classes of activities[1]: (M)anufacturing (5.66% of all firms), (C)onstruction (11.75%), (T)rade (26.43%), and (O)ther Services (56.16%).These businesses are located in 31,897 unique coordinates, since some firms are in the same building. These locations are shown in Figure 1, for the province and in the inset the central part of the city.

In some cases two or more firms shared a coordinate, typically by sharing the same outside address (offices in the same building). In these cases, in order to determine the m-surroundings, first the firms with the same coordinates were selected as nearest neighbors. If the number of firms with the same coordinates exceeded the number required to complete the m-surrounding, a random selection was made of m-1 firms.

[1] DIRCE does not include agricultural activities, and therefore we exclude firms of this type from the sample.

Figure 1

Spatial distribution of firms in Madrid (with central part of city in inset)

5.1 Co-location of business establishments in Madrid: Analysis and results

Analysis with k=4

The first analysis is conducted for k=4, corresponding to the broad activity classes previously identified (Manufacturing, Construction, Trade, and Other Services). The results of applying *Q(m)* to the data when we consider 4 categories are summarized in Table 1. We calculate the statistic using standard and equivalent symbols for different values of m. The results are highly significant, and indicate that the spatial pattern is not random.

Table 5

Results for k=4 (Manufacturing, Construction, Trade, and Other Services)

		Asymptotic test							Permutational bootstrapping test	
		standard symbols				equivalent symbols			standard symbols	equivalent symbols
m	o_d	S	n_σ	S/n_σ	$Q(m)$	n_{σ^*}	S/n_{σ^*}	$Q(m)$	$Q(m)$	$Q(m)$
2	1	51182	16	3198.8	839.3*	10	5118.2	836.4*	1054.2*	1052.4*
3	1	25591	64	399.8	1190.1*	20	1279.5	1149.3*	2826.4*	2792.0*
4	1	17060	256	66.6	1630.5*	35	487.4	1388.7*	5241.3*	5031.0*
4	2	25590	256	100.0	2309.9*	35	731.1	2079.3*		
5	1	12795	1024	12.5	2826.5*	56	228.5	1721.4*	8466.9*	7364.7*
5	2	17060	1024	16.6	3197.7*	56	304.7	2132.7*		

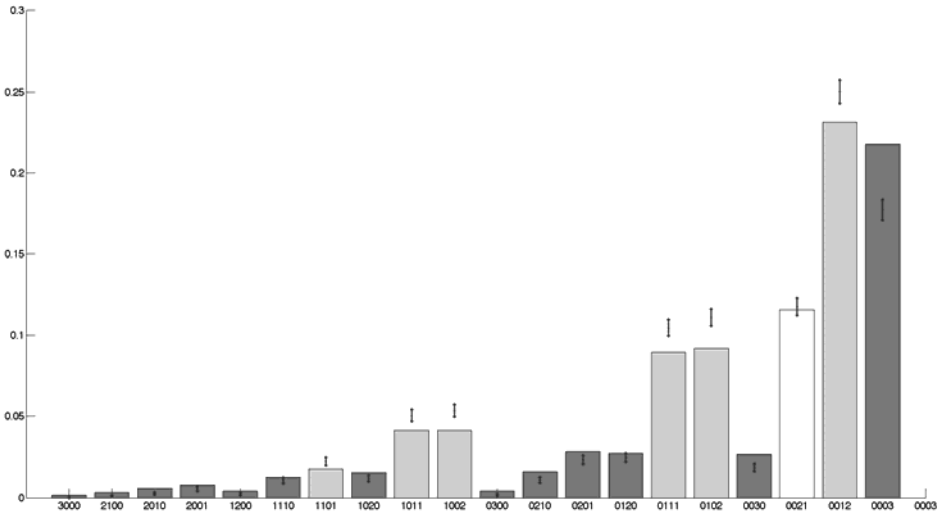
* p-value or p_b-value < 0.001 (999 iterations)

As seen in the table, the null hypothesis can be rejected with a high level of confidence for all selected values of m and r. In this case, given the large ratio of S/n_{σ} and S/n_{σ^*} (>5), we are confident that the asymptotic distribution of the statistic is valid.

It is possible to further explore the reasons why the null hypothesis has been rejected, by comparing the empirical frequency of each symbol to the expected frequency under the null.

Figure 2.

Relative symbol frequency for *Q*(3) and economic activity Manufacture; Construction; Trade and Other Services.



The sequence of numbers on the x axis denote $n_M n_c n_T n_O$ in an m-surrounding of 3 (e.g. 0111 is 0 Manufacture; 1 Construction; 1 Trade; 1 Other Services).

As an illustration, Figure 2 shows the empirical frequency of equivalent symbols for the case of m=3 and o_d=1. Each bar represents the frequency of one specific symbol, and is accompanied by its corresponding 99% confidence interval. As seen in the figure, 13 of 20 symbols appear significantly more frequently than expected. Included in this group are the symbols corresponding to three firms of the same category, for every category. This is a result that echoes the findings of Leslie and Kronenfeld (2011), who report that businesses, irrespective of their category, display a strong geographical affinity with other businesses of the same class. It bears remarking that the symbol for three co-located manufacturing firms {3M}, despite its low relative frequency, is the one that is seen with greater probability than expected under the null. Empirically, this symbol is observed 43 times, or in 0.168% of cases, whereas under the null hypothesis if would have been observed only 4.6 times, or in 0.018% of cases. Similar results are obtained for m=4 and m=5. In every case there is significant evidence of a tendency towards co-location by firms of the same activity category. A question is whether the same pattern

of co-location can be observed under a more detailed classification of firms. This is explored next.

Analysis with k=12

In order to refine the analysis, we obtain a more detailed disaggregation of firms. Construction and Manufacturing are retained as separate classes. Firms in the category Trade are further categorized as follows: Wholesale Trade (8.41% of the total of firms), Retail Trade (11.90%), and Hostels and Restaurants (6.12%). Other Services are further categorized as: Transport (7.52%), Financial Intermediation (2.07%), Real Estate Activities (11.09%), Other Business Activities (21.33%), Education (2.79%), Health (4.42%), and Other Services Activities (6.94%).

The results of applying *Q(m)* to the data when we consider 12 categories are summarized in Table 6. We calculate the statistic using standard and equivalent symbols for different values of m. The first thing to note is the large number of standard symbols when k=12, even for relatively small values of m. The ratio S/n_σ is less than 5 only when $m \leq 3$, and the use of asymptotic results for testing is not recommended otherwise. Use of equivalent symbols allows us to use $m \leq 4$, although in this case we lose the ability to explore asymmetric co-location patterns, since position of the events within the m-surrounding is lost.

As seen in Table 6, the null hypothesis is rejected in some cases when $S/n_\sigma > 5$ and the asymptotic test is used; these results are suspect due to the known issues with the size of the statistic (i.e. the probability of false positives). In order to use standard symbols, the best alternative is to use the distribution-free version of the test, which, as it can be ascertained from the table, rejects the null hypothesis at a high level of confidence (99%) for all values of m tested, including those where the asymptotic testing framework yields results that are questionable.

Table 6

Results for k=12

			Asymptotic test						Permutational bootstrapping test	
			Standard symbols			equivalent symbols			standard symbols	equivalent symbols
m	o_d	S	n_σ	S/n_σ	$Q(m)$	n_{σ^*}	S/n_{σ^*}	$Q(m)$	$Q_B(m)$	$Q_B(m)$
2	1	51182	144	355.4	2487.9*	78	656.2	2450.4*	58559.4*	2857.5*
3	1	25591	1728	14.8	5021.8*	364	70.3	3557.3*	9223.9*	8023.8*
4	1	17060	20736	0.8	21015.2	1365	12.5	5079.5*	34510.4*	15777.2*
4	2	25590	20736	1.2	24336.2*	1365	18.7	7114.1*		
5	1	12795	248832	0.0	59807.8*	4368	2.9	8318.4	136686.4*	27760.1*
5	2	17060	248832	0.1	71914.6		3.9	10016.5		

* p-value or p_b-value < 0.001 (999 iterations)

6. Conclusions

Analysis of spatial qualitative/categorical data is receiving renewed attention in the literature. A recent addition to the toolbox for the spatial exploratory analysis of this type of data is the *Q(m)* statistic. This statistic, which is based on the principle of symbolic entropy, can be used to test the hypothesis that the spatial distribution of values of a qualitative spatial variable is random, or contrariwise, displays patterns of co-location/co-occurrence. As an additional exploratory tool, the empirical frequency of specific co-location patterns can be investigated and contrasted against the expected frequency under the null.

The inferential framework introduced by Ruiz et al. (2010) for $Q(m)$ is based on asymptotic results. These results are applicable in a broad range of situations. In other cases, the asymptotic results can be less appropriate. For instance, when the number of categories k, or the desired size for the m-surroundings is large, the number of symbols can become large relative to the number of symbolized observations, even for relatively large sample sizes. Equivalently, the sample size may be small to begin with. In situations like this, a relatively small ratio of symbolized observations to symbols can overwhelm the ability of the asymptotic test to reliably discriminate between random and non-random patterns.

The objective of this paper has been to propose an alternative inferential framework for situations where asymptotic results may be suspect. Use of a distribution-free approach for testing, based on permutational bootstrapping, can be computationally expensive. It does, on the other hand, circumvent the need to reduce the overlap between proximate m-surroundings (thus increasing the number of effectively usable observations to N), and can be used in situations where the ratio of symbolized observations to symbols is critically low for the application of the asymptotic test. A set of numerical experiments show that the distribution-free approach does not present problems with the size of the test, and is in general at least as powerful, and in some situations more powerful, than the asymptotic test. A distribution-free approach is found to be in general more reliable, especially in limit situations with large k and/or m, and/or small sample size. Application of the testing approach proposed in this paper to a sample of firms in Madrid indicates that even in the case of large samples, a distribution-free approach can expand the potential range of applications of *Q(m)* and increase the reliability of the findings.

The experiments suggest some guidelines for the selection of the testing approach. The computational load is likely manageable for smaller samples with large k and/or m where asymptotic results are in doubt. For larger samples, a possible avenue for reducing the computational cost would be the application of methods based on sequential Monte Carlo (Silva, Assuncao and Costa 2009). This is an avenue for further research.

Acknowledgements: The authors would like to express their thanks to the project ECO2009-10534 of the Ministerio de Ciencia e Innovación del Reino de España.

Appendix

(Continue)

	Codes of activity (CNAE-93) and number of firms	DIRCE	Sample SABI	%
10	Mining of coal and lignite; extraction of peat	15	2	0.003%
11	Extraction of crude petroleum and natural gas	17	2	0.003%
13	Mining of metal ores	16	2	0.003%
14	Other mining and quarrying	179	18	0.035%
15	Manufacture of food products and beverages	1555	156	0.304%
16	Manufacture of tobacco products	2	0	0.000%
17	Manufacture of textiles	569	57	0.111%
18	Manufacture of wearing apparel and dressing	1744	174	0.341%
19	Tanning and dressing of leather	224	22	0.044%
20	Manufacture of wood and of products of wood	967	97	0.189%
21	Manufacture of pulp, paper and paper products	367	37	0.072%
22	Publishing, printing and recorded media	7158	716	1.399%
23	Manufacture of coke, refined petroleum products	10	1	0.002%
24	Manufacture of chemicals and chemical products	586	59	0.114%
25	Manufacture of rubber and plastic products	614	61	0.120%
26	Manufacture of other non-metallic mineral products	644	64	0.126%
27	Manufacture of basic metals	220	22	0.043%
28	Manufacture of fabricated metal products	4445	445	0.868%
29	Manufacture of other machinery and equipment	1637	164	0.320%
30	Manufacture of office machinery and computers	262	26	0.051%
31	Manufacture of electrical machinery	390	39	0.076%
32	Manufacture of radio, television and other appliances	254	25	0.050%
33	Manufacture of medical, precision and instruments	992	99	0.194%
34	Manufacture of motor vehicles, trailers	206	21	0.040%
35	Manufacture of other transport equipment	170	17	0.033%
36	Manufacture of furniture and other products	2952	295	0.577%
37	Recycling	24	2	0.005%
40	Electricity, gas, steam and hot water supply	2679	268	0.523%
41	Collection, purification and distribution of water	65	7	0.013%
45	Construction	60143	6014	11.751%
50	Sale, maintenance and repair of motor vehicles	9712	971	1.898%
51	Wholesale trade and commission trade	33316	3332	6.510%
52	Retail trade, except of motor vehicles and motorcycles	60919	6092	11.903%
55	Hotels and restaurants	31327	3133	6.121%
60	Land transport; transport via pipelines	31000	3100	6.057%
61	Water transport	52	5	0.010%
62	Air transport	87	9	0.017%
63	Supporting and auxiliary transport activities	4813	481	0.940%

(Conclusion)

	Codes of activity (CNAE-93) and number of firms	DIRCE	Sample SABI	%
64	Post and telecommunications	2540	254	0.496%
65	Financial intermediation (exc. insurance/pension funding)	696	70	0.136%
66	Insurance and pension funding	272	27	0.053%
67	Activities auxiliary to financial intermediation	9645	965	1.885%
70	Real estate activities	38603	3860	7.543%
71	Renting of machinery and equipment without operator	3208	321	0.627%
72	Computer and related activities	11244	1124	2.197%
73	Research and development	3721	372	0.727%
74	Other business activities	2679	268	21.329%
80	Education	109163	10916	2.786%
85	Health and social work	14261	1426	4.415%
90	Sewage and refuse disposal, sanitation and similar activities	22595	2260	0.182%
91	Activities of membership organizations n.e.c.	930	93	1.072%
92	Recreational, cultural and sporting activities	5485	549	3.406%
93	Other service activities	17431	1743	2.282%
	Total:	511804	51183	100%

References

ALBERT, J. M., CASANOVA, M. R., AND ORTS, V. (2011), «Spatial Location Patterns of Spanish Manufacturing Firms» *Papers in Regional Science*, (in press; 10.1111/j.1435-5957.2011.00375.x).

ANSELIN, L. (1988), *Spatial Econometrics: Methods and Models*, Dordrecht: Kluwer

ANSELIN, L. (1995), «Local Indicators of Spatial Association - Lisa» *Geographical Analysis*, 27, 93-115.

ARBIA, G. (2001), «The Role of Spatial Effects in the Empirical Analysis of Regional Concentration» *Journal of Geographical Systems*, 3, 271-281.

ARBIA, G. (2011), «A Lustrum of Sea: Recent Research Trends Following the Creation of the Spatial Econometrics Association (2007-2011)» *Spatial Economic Analysis*, 6, 377-395.

BIVAND, R. (2009),«Applying Measures of Spatial Autocorrelation: Computation and Simulation» *Geographical Analysis*, 41, 375-384.

BOOTS, B. (2003), «Developing Local Measures of Spatial Association for Categorical Variables» *Journal of Geographical Systems*, 5, 139-160.

BOOTS, B. (2006), «Local Configuration Measures for Categorical Spatial Data: Binary Regular Lattices» *Journal of Geographical Systems*, 8, 1-24.

BORN, B., AND BREITUNG, J. (2011), «Simple Regression-Based Tests for Spatial Dependence» *Econometrics Journal*, 14, 330-342.

CRESSIE, N. A. C. (1993), *Statistics for Spatial Data*, New York: John Wiley & Sons.

DACEY, M. F. (1968), «A Review on Measures of Contiguity for Two and K-Color Maps» in *Spatial Analysis: A Reader in Statistical Geography*, eds. B. J. L. Berry and D. F. Marble, Englewood Cliffs, NJ: Prentice Hall, pp. 479-495.

ELLISON, G., AND GLAESER, E. L. (1997), «Geographic Concentration in Us Manufacturing Industries: A Dartboard Approach» *Journal of Political Economy*, 105, 889-927.

ESPA, G., ARBIA, G., AND GIULIANI, D. (2012), «Conditional Versus Unconditional Industrial Agglomeration: Disentangling Spatial Dependence and Spatial Heterogeneity in the Analysis of ≪I≫Ict≪/I≫ Firms' Distribution in Milan» *Journal of Geographical Systems*, 1-20.

GEARY, R. C. (1954),«The Contiguity Ratio and Statistical Mapping» *The Incorporated Statistician*, 5, 115-145.

GRIFFITH, D. A. (1988), *Advanced Spatial Statistics: Special Topics in the Exploration of Quantitative Spatial Data Series*, Dordrecht: Kluwer.

HAINING, R. (1990), *Spatial Data Analysis in the Social and Environmental Sciences*, Cambridge: Cambridge University Press.

HOOVER, E. M., AND GIARRATANI, F. (1984), *An Introduction to Regional Economics*, New York: Knopf.

JIMENEZ, K. P., AND JUNQUERA, B. (2010),«Why Are Clusters Beneficial? A Review of the Literature» *Human Factors and Ergonomics in Manufacturing & Service Industries*, 20, 161-173.

LESLIE, T. F., AND KRONENFELD, B. J. (2011),«The Colocation Quotient: A New Measure of Spatial Association between Categorical Subsets of Points» *Geographical Analysis*, 43, 306-326.

LI, W. D. (2006),«Transiogram: A Spatial Relationship Measure for Categorical Data» *International Journal of Geographical Information Science*, 20, 693-699.

LIN, K. P., LONG, Z. H., AND OU, B. L. (2011),«The Size and Power of Bootstrap Tests for Spatial Dependence in a Linear Regression Model» *Computational Economics*, 38, 153-171.

MARCON, E., AND PUECH, F. (2003),«Evaluating the Geographic Concentration of Industries Using Distance-Based Methods» *Journal of Economic Geography*, 3, 409-428.

MATILLA-GARCIA, M., AND RUIZ, M. (2011),«Spatial Symbolic Entropy: A Tool for Detecting the Order of Contiguit» *Geographical Analysis*, 43, 228-239.

MORAN, P. A. P. (1948),«The Interpretation of Statistical Maps» *Journal of the Royal Statistical Society.Series B (Methodological)*, 10, 243-251.

PÁEZ, A., RUIZ, M., LÓPEZ, F. A., AND LOGAN, J. (2012),«Measuring Ethnic Clustering and Exposure with the Q Statistic: An Exploratory Analysis of Irish, Germans, and Yankees in 1880 Newark,» *Annals of the Association of American Geographers*, 102, 84-102.

PÁEZ, A., TREPANIER, M., AND MORENCY, C. (2011),«Geodemographic Analysis and the Identification of Potential Business Partnerships Enabled by Transit Smart Cards» *Transportation Research Part A-Policy and Practice*, 45, 640-652.

RUIZ, M., LÓPEZ, F., AND PÁEZ, A. (2010),«Testing for Spatial Association of Qualitative Data Using Symbolic Dynamics» *Journal of Geographical Systems*, 12, 281-309.

RUIZ, M., LÓPEZ, F.A., AND PÁEZ, A. (2011),«Comparison of Thematic Maps Using Symbolic Entropy» *International Journal of Geographical Information Science*, 1-27.

SILVA, I. R., ASSUNCAO, R., AND COSTA, M. (2009),«Power of the Sequential Monte Carlo Test» *Sequential Analysis*, 28, 163-174.

SOON, S. Y. T. (1996),«Binomial Approximation for Dependent Indicators» *Statistica Sinica*, 6, 703-714.

SORENSON, O. (2003),«Social Networks and Industrial Geography» *Journal of Evolutionary Economics*, 13, 513-527.

STEVENS, P. H., AND JENKINS, D. G. (2000),«Analyzing Species Distributions among Temporary Ponds with a Permutation Test Approach to the Join-Count Statistic» *Aquatic Ecology*, 34, 91-99.