Testing for Spatial Association of Qualitative Data Using Symbolic Dynamics

Manuel Ruiz, Fernando López
Facultad de C.C. de la Empresa
Dpto. Métodos Cuantitativos e Informáticos
Universidad Politécnica de Cartagena

Antonio Páez
Centre for Spatial Analysis
School of Geography and Earth Sciences
McMaster University

Published in Journal of Geographical Systems
doi:10.1007/s10109-009-0100-1

Abstract. Qualitative spatial variables are important in many fields of research. However, unlike the decades-worth of research devoted to the spatial association of quantitative variables, the exploratory analysis of spatial qualitative variables is relatively less developed. The objective of the present paper is to propose a new test (Q) for spatial independence. This is a simple, consistent, and powerful statistic for qualitative spatial independence that we develop using concepts from symbolic dynamics and symbolic entropy. The Q test can be used to detect, given a spatial distribution of events, patterns of spatial association of qualitative variables in a wide variety of settings. In order to enable hypothesis testing, we give a standard asymptotic distribution of an affine transformation of the symbolic entropy under the null hypothesis of independence in the spatial qualitative process. We include numerical experiments to demonstrate the finite sample behaviour of the test, and show its application by means of an empirical example that explores the spatial association of fast food establishments in the Greater Toronto Area in Canada.

Keywords. Spatial independence, qualitative variables, symbolic dynamics, entropy, fast food

1 Introduction

The concept of spatial autocorrelation is central to any effort to understand the spatiality of phenomena, and to build spatial theory and models (Griffith 1999; Miller 2004). From its origins in mathematical statistics (Geary 1954; Krishna Iyer 1949; Moran 1948) the notion of autocorrelation has animated, and in turn been given lasting currency by, quantitative geography, spatial analysis, and spatial statistics (Getis 2008). It is from these disciplines that the analysis of map patterns has diffused throughout, starting with the work of quantitative geographers (e.g., Dacey 1968), to Cliff and Ord (1973, 1981) and Ripley (1981), through the texts of Anselin (1988), Griffith (1988), Haining (1990), and Cressie (1993). Now, spatial autocorrelation analysis is used to support research in an ever increasing sphere of cogent disciplines.

A vast majority of work in spatial analysis has historically been concerned with the analysis of variables of a continuous and interval nature. It is thus interesting to note that in fact the first attempt to describe maps from a statistical point of view, was made in reference to qualitative variables (Dacey 1968; Moran 1948), specifically black and white colored (or later *k*-colored) maps, and only in second place to continuous variables (Cliff and Ord 1973; Geary 1954; Moran 1950). The reason for this historical development seems clear. Linear regression for the multivariate analysis of continuous variables was, until relatively recent times, the instrument of choice for statistical analysis of spatial data. In turn, the analysis of map patterns was, almost from the beginning, meant to serve as a diagnostic tool for the analysis of residuals in linear regression (see Geary 1954, pp. 115-116 and again p. 144).

Despite the traditional focus on continuous variables in spatial analysis, there are numerous situations where qualitative variables are the focus of research, and it is in this context that the hypothesis of spatial independence of qualitative data is important. Besides early work with join count statistics (e.g., Cliff and Ord 1981; Dacey 1968; Upton and Fingleton 1985), and some more recent work by Boots (2003), not much research has been devoted to this class of problems in an exploratory setting, even if spatial modeling techniques for qualitative data have seen significant progress in recent years (e.g., Bhat and Sener 2009; Chakir and Parent 2009; Dubin 1995; McMillen 1992; Paez 2006; Robertson et al. 2009; Wang and Kockelman 2009). The objective of this paper is to propose a new statistic for the exploratory analysis of spatial qualitative/nominal data. The statistic is meant to identify whether neighbouring values of a spatial qualitative variable tend to be more similar or dissimilar than would be expected by chance.

The approach proposed to test this hypothesis of spatial independence for qualitative variables is based on principles drawn from symbolic dynamics. Symbolic dynamics has been used for the investigation of non-linear dynamic systems (Hao and Zheng 1998) and provide an ideal set of tools for representing discrete processes. We use these tools to derive a new statistic, termed *Q*, parting from a function of symbolic entropy. In addition, we discuss the theoretical properties of the proposed statistic and investigate its finite sample behaviour by means of an extensive set of numerical experiments. Finally, we illustrate the usefulness of the *Q* statistic empirically with a case study that explores the spatial association of various fast food establishment types, namely Pizza, Hamburger, and Sandwich establishments, in the Greater Toronto Area eein Canada. In the concluding section, we discuss a number of valuable features of our statistic, and directions for future research.

2 Background

As noted above, the study of autocorrelation of qualitative variables was among the earliest forms of spatial analysis, but from the start meant to support the use of linear regression for continuous variables. Some early applications confirm this connection, as for example the analysis that Haining (1978) conducted for crop failures in Nebraska and Kansas. While the premise that crop failures formed one or more regional clusters had been previously advanced (e.g., Hewes 1965), application of a contiguity measure by Haining (1978) provided the statistical evidence necessary to confirm the visual appraisal of crop failure patterns. An intriguing feature of this study is the conversion of an interval variable (percentage failure) to a nominal variable by taking values below or above the mean, or in other words, the categorization of a continuous variable. This is not a lonely example of such practice of discretizing continuous variables, and other instances include Chuang and Huang's (1992) assessment of the level of noise in digital images that converted grey scale radiological images to black and white patterns, or Goldsborough's (1994) study of algal enumeration, whereby overall mean density was used to classify units as "dense" or "sparse". One can only speculate as to the reasons why continuous variables were converted to nominal variables in these studies, since the fact that reduction to a nominal variable involves some serious information loss was not lost in these authors (see Chuang and Huang 1992, p. 367). From a computational standpoint, there are indications that as late as 1992, the process of counting joins required to calculate autocorrelation statistics was still fraught with difficulties and plagued with errors (Ghent et al. 1992). Relative simplicity may have also been a factor. In any case, it is clear that a vast majority of research efforts were indeed devoted to the development of statistics for continuous variables to serve the needs imposed by the extended use of regression analysis. As a result, it is conventional in contemporary spatial analytical practice to use statistics appropriate for continuous variables at the global (Moran's *I*, Geary's *c*, variographic analysis) or local level (Anselin 1995; Getis and Ord 1993).

There are multiple examples of research where the focus is in fact a qualitative variable. In integrated chip manufacturing, for instance, the spatial structure of non-functional chips in wafers is recognized as a way to provide useful information about the manufacturing process. In this case, chips in a wafer are classified as "good" or "bad" (e.g., Taam and Hamada 1993, p. 150), and the objective is to determine whether defects are randomly or non-randomly scattered. Nominal data are also found in plant pathology, as in De Jong and De Bree's (1995) study of spatial patterns of disease in commercial fields of leek, where the variable of interest is a health status binary classification ("healthy" and "infected"). Likewise, Real and McElhany (1996) discuss the use nominal variables when these are the disease status of plants. In veterinary science, Mannelli et al. (1998) have studied swine fever in Sardinia using municipal level data following a binary classification scheme defined as "outbreak" and "unaffected". In evolutionary biology, spatial variation in fitness was examined by Stratton and Bennington (1996) in an experiment implemented to infer natural selection processes that operate in space through the assessment of spatial variations in genotype distribution. In this experiment, data collected after a random initial distribution of seeds was analyzed to elucidate whether plants that carry identical genetic markers are spatially associated, and the classification was defined by means of identity, that is, patterns of association of plants with the same genetic markers (e.g., if there are three markers, then AA, Aa, aa). In separate research, Epperson and Alvarez-Buylla (1997) also investigate the spatial structure of nominal variables based on joins for two genotypes. Bell et al. (2008) are interested in spatial patterns of injury. In this

investigation, join count statistics are used to describe the spatial co-occurrence of injuries by assault or intentional self-harm, with the results suggesting that assault injuries sustained by males who resided in neighbouring areas were more frequent than expected purely by chance. Self-harm injuries did not display the same strength of spatial pattern.

The intention of the statistic proposed in this paper is to support analysis in research that makes use of qualitative variables, such as the examples above.

3 Symbolization of a spatial process with discrete outcomes

Development of the *Q* statistic is based on the application of symbolic dynamics concepts. Symbolic dynamics is an approach, developed in the field of mathematics for the study of dynamical systems (Hao and Zheng 1998) that consists of modelling a dynamic system by means of a discrete set consisting of sequences of abstract symbols obtained for a suitable partition of the state space. The basic idea behind symbolic dynamics is to consider a space in which the possible states of a system are represented, and each possible state corresponds to one unique point in the state space. This space can then be partitioned into a finite number of regions and each region can be labelled by an alphabetical letter. In this regard, symbolic dynamics is a coarse-grained description of dynamics. Even though coarse-grained methods lose a certain amount of detailed information, some essential features of the dynamics may be kept, including periodicity and dependencies, among others (for an overview of these concepts see Hao and Zheng 1998). If the process is inherently discrete to begin with, then symbolic dynamics provide an ideal tool for its study.

In order to implement symbolic dynamics concepts the symbols for a process must be defined, or in other words, the process needs to be symbolized. In principle, there is no reason to anticipate that symbolization procedures will be unique given a spatial process, and in fact it is possible to conceive of several possible ways to symbolize a process. Therefore the general framework proposed here can be adapted to the necessities of specific problems, and just as is the case with connectivity matrices in spatial modelling, it is generally possible to incorporate substantive understanding of the process of interest in order to refine the symbolization procedure. This is a feature that lends great flexibility to our approach. In order to ensure broad applicability of the statistic proposed, in this paper we propose a general, all-purpose symbolization procedure which allows us to capture the dependencies of a discrete process in geographical space.

Let us begin by defining a discrete spatial process $\{X_s\}_{s\in S}$, where *S* is a set of geographical coordinates that denote the locations of events. These locations are given and fixed. Further, denote by $A=\{a_1,a_2,...,a_k\}$ the set of possible values that X_s can take, for all $s\in S$. Clearly, there are *k* different categories in this notation, which could be "black"/"white" or "yes"/"no" (*k*=2), "AA"/"Aa"/"aa" if there are three genetic markers (*k*=3), and so on. In other words, observations are made at spatially discrete locations, and the outcome of the process is discrete as well. A natural way to symbolize such a process is to embed it in an *m*-dimensional space as follows:

$$X_m(s_0)=(X_{s_0},X_{s_1},...,X_{s_{m-1}}) \text{ for } s_0\in S \quad (1)$$

where s_1, s_2,..., s_{m-1} are the *m*-1 nearest neighbours of s_0. We will call this *m*-dimensional space an *m*-surrounding. A key to symbolizing the process is to define the criteria that determine which spatial events are the neighbours of s_0. To this end, we propose a definition of neighbours based on proximity (i.e. nearest neighbour criterion). Whenever

two neighbours are equidistant, then the polar coordinates (ρ_i^0, θ_i^0) of s_i are considered, taking s_0 as the origin. This implies that the m-1 nearest neighbours will be those events satisfying the following two conditions that ensure the uniqueness of $X_m(s)$ for all $s \in S$:

(a) The distance of the m-1 neighbours from s_0 satisfies the condition that $\rho_1^0 \le \rho_2^0 \le ... \le \rho_{m-1}^0$; and

(b) In the case of a tie in terms of the distance from s_0, (i.e. if $\rho_i^0 = \rho_{i+1}^0$) then precedence goes to the smaller angle (i.e. $\theta_i^0 < \theta_{i+1}^0$).

The set of the m-1 nearest neighbours is denoted as $N_s = \{s_1, s_2, ..., s_{m-1}\}$. Since an m-surrounding $X_m(s)$ consists of m observations, and there are k possible values that each observation can take, there are k^m distinct combinations of values for an m-surrounding. We will denote each of these unique combinations by an abstract symbol, say σ_i, and will define $\Gamma = \{\sigma_1, \sigma_2, ..., \sigma_{k^m}\}$ as the set of all possible symbols. Furthermore, we will say that a location s is of σ_i-type if and only if $X_m(s) = \sigma_i$.

As an illustration of the symbolization procedure, consider a simple spatial system consisting of a regular hexagonal tessellation as shown in Fig. 1, and a process with two possible outcomes (k=2). The outcomes are shown in the figure in dark color when they are class 1 and light color when they are of class 2. Taking m=6 as the size of the m-surrounding, this gives a total of 2^6=64 different combinations of values, or symbols (σ_1 through σ_{64}), as listed in Table 1. Please note that a hexagonal tessellation is used only for illustrative purposes. The symbolization procedure is equally applicable to regular and irregular distributions of observations, and to points as well as areas.

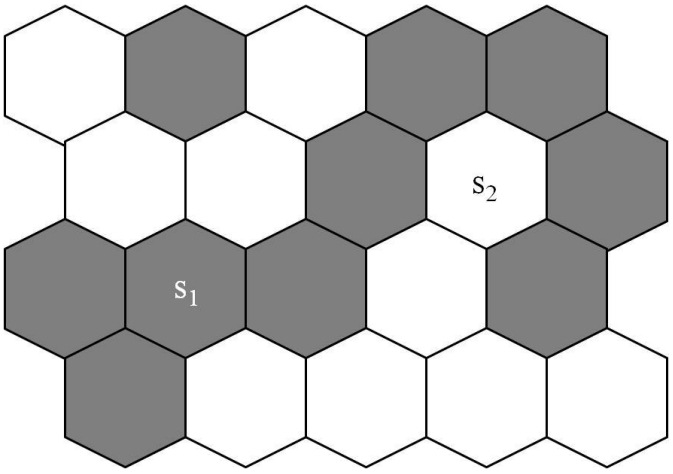

Fig. 1. Simple spatial system and process with two types of outcomes.

Since in a hexagonal tessellation the distance from s_0 is the same for all 6 contiguous spatial units, and keeping in mind that polar coordinates begin at an angle of 0 in the positive direction of the x axis in Cartesian coordinates, it should be clear that neighbours are arranged in order of increasing angle from the origin of the polar coordinate system. Then, referring again to Fig. 1, we say that location s_1 is of symbol σ_{13}, since $X_m(s_1) = (1,1,2,2,1,1)$, whereas location s_2 is of symbol σ_{34}, since $X_m(s_2) = (2,1,1,1,1,2)$.

It is important to note that while the number of classes k is determined by the

nature of the process, the size of the *m*-surrounding is not, which gives some flexibility to the analyst to explore various alternatives, however bounded by the necessity to satisfy some minimum conditions required to ensure desirable statistical properties, as discussed more fully below.

Table 1. List of symbols for k=2, m=6

σ_1 =(1,1,1,1,1,1)	σ_{17} =(1,2,1,1,1,1)	σ_{33} =(2,1,1,1,1,1)	σ_{49} =(2,2,1,1,1,1)
σ_2 =(1,1,1,1,1,2)	σ_{18} =(1,2,1,1,1,2)	σ_{34} =(2,1,1,1,1,2)	σ_{50} =(2,2,1,1,1,2)
σ_3 =(1,1,1,1,2,1)	σ_{19} =(1,2,1,1,2,1)	σ_{35} =(2,1,1,1,2,1)	σ_{51} =(2,2,1,1,2,1)
σ_4 =(1,1,1,1,2,2)	σ_{20} =(1,2,1,1,2,2)	σ_{36} =(2,1,1,1,2,2)	σ_{52} =(2,2,1,1,2,2)
σ_5 =(1,1,1,2,1,1)	σ_{21} =(1,2,1,2,1,1)	σ_{37} =(2,1,1,2,1,1)	σ_{53} =(2,2,1,2,1,1)
σ_6 =(1,1,1,2,1,2)	σ_{22} =(1,2,1,2,1,2)	σ_{38} =(2,1,1,2,1,2)	σ_{54} =(2,2,1,2,1,2)
σ_7 =(1,1,1,2,2,1)	σ_{23} =(1,2,1,2,2,1)	σ_{39} =(2,1,1,2,2,1)	σ_{55} =(2,2,1,2,2,1)
σ_8 =(1,1,1,2,2,2)	σ_{24} =(1,2,1,2,2,2)	σ_{40} =(2,1,1,2,2,2)	σ_{56} =(2,2,1,2,2,2)
σ_9 =(1,1,2,1,1,1)	σ_{25} =(1,2,2,1,1,1)	σ_{41} =(2,1,2,1,1,1)	σ_{57} =(2,2,2,1,1,1)
σ_{10} =(1,1,2,1,1,2)	σ_{26} =(1,2,2,1,1,2)	σ_{42} =(2,1,2,1,1,2)	σ_{58} =(2,2,2,1,1,2)
σ_{11} =(1,1,2,1,2,1)	σ_{27} =(1,2,2,1,2,1)	σ_{43} =(2,1,2,1,2,1)	σ_{59} =(2,2,2,1,2,1)
σ_{12} =(1,1,2,1,2,2)	σ_{28} =(1,2,2,1,2,2)	σ_{44} =(2,1,2,1,2,2)	σ_{60} =(2,2,2,1,2,2)
σ_{13} =(1,1,2,2,1,1)	σ_{29} =(1,2,2,2,1,1)	σ_{45} =(2,1,2,2,1,1)	σ_{61} =(2,2,2,2,1,1)
σ_{14} =(1,1,2,2,1,2)	σ_{30} =(1,2,2,2,1,2)	σ_{46} =(2,1,2,2,1,2)	σ_{62} =(2,2,2,2,1,2)
σ_{15} =(1,1,2,2,2,1)	σ_{31} =(1,2,2,2,2,1)	σ_{47} =(2,1,2,2,2,1)	σ_{63} =(2,2,2,2,2,1)
σ_{16} =(1,1,2,2,2,2)	σ_{32} =(1,2,2,2,2,2)	σ_{48} =(2,1,2,2,2,2)	σ_{64} =(2,2,2,2,2,2)

Once the symbolization of the process has been defined, it is possible to calculate the frequency of each symbol σ_i, which is simply the number of locations s that are of σ_i -type:

$$n_{\sigma_i} = \sharp\{s \in S \mid X_m(s) = \sigma_i\} \tag{2}$$

where $\sharp$ denotes the cardinality of a set. Since this frequency is defined for each of k^m symbols, under the conditions above, the relative frequency of a symbol $\sigma \in \Gamma$ can be easily computed as:

$$p(\sigma) = p_\sigma = \frac{\sharp\{s \in S \mid s \text{ is of } \sigma - \text{type}\}}{|S|} \tag{3}$$

whereby $|S|$ denotes the cardinality of the set S (the total number of symbolized observations).

Now, under this setting, we can define the *symbolic entropy* of the spatial process $\{X_s\}_{s \in S}$ for an embedding dimension $m \geq 2$. This entropy is defined as the Shanon's entropy of the k^m distinct symbols as follows:

$$h(m) = -\sum_{\sigma \in \Gamma} p_\sigma \ln(p_\sigma) \tag{4}$$

Symbolic entropy, or $h(m)$, is the information contained in comparing the *m*-surroundings defined for the spatial process. Notice that when one symbol, say σ_i, tends to dominate the process then $p_{\sigma_i} \to 1$ and $p_{\sigma_j} \to 0$ for all $j \neq i$, which implies that $p_{\sigma_i} \ln(p_{\sigma_i}) \to 0$ and $p_{\sigma_j} \ln(p_{\sigma_j}) \to 0$ and therefore $h(m) = 0$. Furthermore, when the values of the qualitative variable are identically and independently distributed, all k^m symbols should appear with equal frequency, in which case we have that $p_{\sigma_i} = 1/k^m$ for all *i*. The entropy function is then bounded between $0 < h(m) \leq \ln(k^m)$, where the lower bound indicates a tendency for only one symbol to occur (i.e. there is a tendency towards patterning in the distribution of the values of the qualitative variable), and the upper bound corresponds to a completely random system (i.i.d. spatial sequence). As an illustration, consider the situation illustrated in Fig. 2, with *k*=2 and *m*=3, which means that there are $2^3 = 8$ symbols. The left panel shows a random distribution of the values of the qualitative variable. The histogram of the frequency of each of eight symbols verifies that all symbols appear with similar frequency. The right panel shows the case where the values are distributed non-randomly and two symbols tend to appear with more frequency than the rest. Rarely will the frequency of symbols be identical, and the question that emerges is whether departures from this are significant. In other words, do some symbols appear with more or less frequency than what would be expected by chance alone? The results needed to statistically test this hypothesis are derived next.

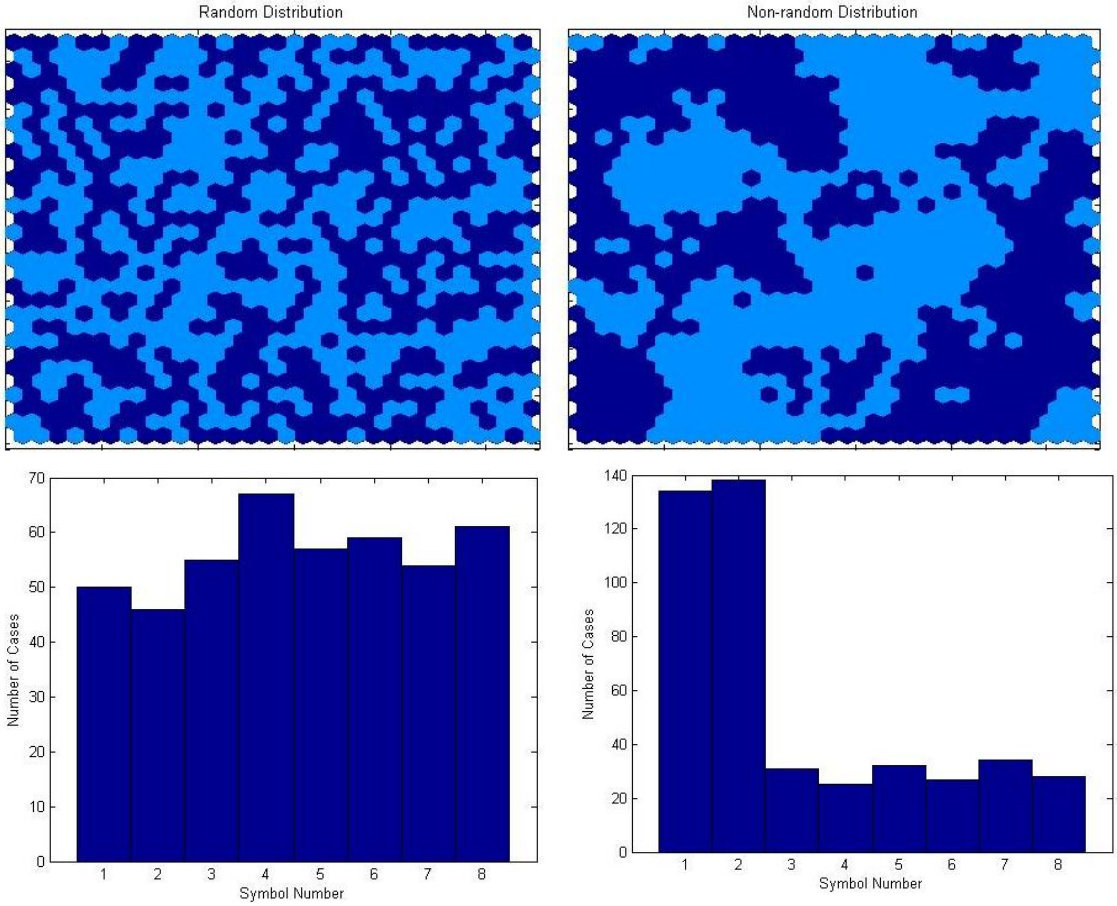

Fig. 2. Random and non-random distributions of values of qualitative variable (*k*=2) and frequency of symbols

4 Construction of the independence test

In this section, we construct a spatial independence test for a discrete qualitative spatial variable. We also prove that an affine transformation of the symbolic entropy defined in Eq. (4) is asymptotically χ^2 distributed.

Let $\{X_s\}_{s\in S}$ be a discrete spatial process and m be a fixed embedding dimension. In order to construct a test for spatial independence in $\{X_s\}_{s\in S}$, we consider the following null hypothesis: $H_0 : \{X_s\}_{s\in S}$ is spatially independent, against any other alternative.

Now, for a symbol $\sigma_i \in \Gamma$, we define the random variable $Z_{\sigma_i s}$ as follows:

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

that is, we have that $Z_{\sigma_i s} = 1$ if and only if s is of σ_i-type, $Z_{\sigma_i s} = 0$ otherwise. Then $Z_{\sigma_i s}$ is a Bernoulli variable with probability of "success" p_{σ_i}, where "success" means that s is of σ_i-type. It is straightforward to see that:

$$\sum_{i=1}^{n} p_{\sigma_i} = 1 \tag{6}$$

Let us assume that set S is finite and of order R (the number of symbolized locations). Then we are interested in knowing how many s's are of σ_i-type for all symbols $\sigma_i \in \Gamma$. We construct the following variable to this end:

$$Y_{\sigma_i} = \sum_{s\in S} Z_{\sigma_i s} \tag{7}$$

The variable Y_{σ_i} can take the values $\{0,1,2,\ldots,R\}$. Notice that not all the variables $Z_{\sigma_i s}$ are independent (due to the overlapping of some m-surroundings), and therefore Y_{σ_i} is not exactly a binomial random variable. Nevertheless, the sum of dependent Bernoulli variables can be approximated to a binomial random variable whenever (see Soon 1996):

(i) Dependencies among the indicators are weak; and
(ii) The probability of the indicators to occur is small.

Condition (ii) is satisfied by the way the symbols have been constructed, since in this case, under the null hypothesis, the probability of success of the indicators $Z_{\sigma s}$ is small ($p_\sigma = 1/k^m$). Condition (i), on the other hand can usually be satisfied only if the events are distributed in a regular array, and the size of the *m*-surrounding is relatively small, in which case the overlaps are minor. More generally, when the size of the *m*-surrounding is large, or when their spatial arrangement is irregular, this condition becomes more difficult to maintain, if we consider all the indicators $Z_{\sigma s}$ for all $s \in S$. Additional steps are therefore needed to ensure that the dependencies among the indicators $Z_{\sigma s}$ are weak.

In order to attain a good binomial approximation, we consider a subset of locations $S \subseteq S$ with controlled overlap, so that the dependencies among the indicators $Z_{\sigma s}$ are weak for $s \in S$. Use of a subset of locations will cause a loss of information, and this loss will be greater in the measure that set S is smaller. A reasonable balance therefore must be struck between strongly dependent indicators and too much loss of information. In order to control the amount of overlap among the Bernoulli variables, we can take as S those coordinates in S such that for any two coordinates $s_i, s_j \in S$ the sets of nearest neighbours of s_i and s_j are at most *r* (a small enough positive integer) if they intersect:

$$| N_{s_i} \cap N_{s_j} | = \begin{cases} 0 & \text{if non-overlapping} \\ r & \text{otherwise} \end{cases} \tag{8}$$

We call this integer r the *degree* of *overlap* of the spatial process $\{X_s\}_{s \in S}$. We now turn to a method to select the set S satisfying the above condition. Let us define the set S recursively as follows. First chose a location $\tilde{s}_0 \in S$ at random and fix an integer r with $0 \le r < m$. Let $N_{\tilde{s}_0} = \{s_1^0, s_2^0, ..., s_{m-1}^0\}$ be the set of nearest neighbours to $\tilde{s}_0$, where the s_i^0's are ordered by distance to $\tilde{s}_0$. Let us call $\tilde{s}_1 = s_{m-r-1}^0$ and define $A_0 = \{\tilde{s}_0, s_1^0, ..., s_{m-r-2}^0\}$. Take the set of nearest neighbours to $\tilde{s}_1$, namely $N_{\tilde{s}_1} = \{s_1^1, s_2^1, ..., s_{m-1}^1\}$, in the set of locations $S \setminus A_0$ and define $\tilde{s}_2 = s_{m-r-1}^1$. Now for $i > 1$ we define $\tilde{s}_i = s_{m-r-1}^{i-1}$ where s_{m-r-1}^{i-1} is in the set of nearest neighbours to $\tilde{s}_{i-1}$, $N_{\tilde{s}_{i-1}}^0 = \{s_1^{i-1}, s_2^{i-1}, ..., s_{m-1}^{i-1}\}$, of the set $S \setminus \{\cup_{j=0}^{i-1} A_j\}$. Continue this process while there are locations to symbolize. In the end, we have constructed a set of locations:

$$S = \{\tilde{s}_0, \tilde{s}_1, ..., \tilde{s}_R\} \tag{9}$$

such that the variable $Y_{\sigma_i} = \sum_{\tilde{s} \in S} Z_{\sigma_i \tilde{s}}$ can be approximated to a binomial distribution for a suitable choice of r. Notice that the maximum number of locations that can be symbolized with an overlapping degree r is $R = \left[\frac{N-m}{m-r}\right] + 1$, where the operator $[x]$ denotes the integer part of a real number x.

Given the above considerations, we can now state the following results (the proof can be found in the Appendix).

Theorem 1 *Let $\{X_s\}_{s \in S}$ be a qualitative discrete spatial process with $|S| = N$. Let $A = \{a_1, a_2, ..., a_k\}$ be the set of possible values that X_s can take, for all $s \in S$. Let r be the overlapping degree of $\{X_s\}_{s \in S}$ and $R = [\frac{N-m}{m-r}] + 1$, where $[x]$ denotes the integer part of a real number x. Let $\Gamma = \{\sigma_1, \sigma_2, ..., \sigma_{k^m}\}$ be the set of symbols defined in Section 2. Let α_{ij} the number of times that class a_j appears in symbol σ_i and $q_j = P(X = a_j)$. Denote by $h(m)$ the symbolic entropy defined in Eq. (2) for a fixed embedding dimension $m \ge 2$, with $m \in \mathbb{N}$. If the spatial process $\{X_s\}_{s \in S}$ is independent, then:*

$$Q(m) = -2R\left[\sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{R} \sum_{j=1}^{k} \alpha_{ij} \ln\left(q_j\right) + h(m)\right] \tag{10}$$

is asymptotically $\chi^2_{k^m-1}$ *distributed.*

Note that if $\{X_s\}_{s\in S}$ is also identically distributed, in other words, each value of the variable appears with equal frequency, then $q_j = \frac{1}{k}$ and therefore Eq. (10) reduces to:

$$Q(m) = 2R\left(Ln(k^m) - h(m)\right). \tag{11}$$

Let α be a real number with $0 \le \alpha \le 1$. Let χ^2_α be such that:

$$P(\chi^2_{k^m-1} > \chi^2_\alpha) = \alpha. \tag{12}$$

Then, to test: $H_0 : \{X_s\}_{s\in S}$ is spatially independent, the decision rule in the application of the $Q(m)$ test at a $100(1-\alpha)\%$ confidence level is:

$$\begin{cases} \text{If } Q(m) > \chi^2_\alpha \text{ then reject } H_0 \\ \text{Otherwise do not reject } H_0 \end{cases} \tag{13}$$

5 Properties of the *Q*(*m*) test

Next, we prove that the $Q(m)$ test is consistent for a wide variety of spatially dependent processes. This is a valuable property since the test will reject asymptotically the assumption of spatial independence whenever there is spatial dependence within the *m*-surrounding. By *spatial dependence of order less than or equal to m* we mean that, whatever the structure of the spatial process is, there exists dependence between the random variable located at point s and its m-surrounding *or a part of it*. We will denote by $Q(m)$ the estimator of $Q(m)$. The proof of the following theorem can be found in the Appendix.

Theorem 2 *Let* $\{X_s\}_{s\in S}$ *be a discrete spatial process, and* $m \ge 2$ *with* $m \in \mathbb{N}$. *Then:*

$$\lim_{R\to\infty} \Pr(Q(m) > C) = 1 \tag{14}$$

under spatial dependence of order smaller than or equal to m for all $0 < C < \infty, C \in \mathbb{R}$.

Thus, the test based on $Q(m)$ is consistent against all spatial dependence of order less than or equal to *m*. Conversely, since the dependence detected by the test is at most of order *m*, if the dependence structure of the process is of order larger than *m*, then it will not be present in every m-surrounding and therefore the symbols may not capture it. Since Theorem 2 implies $Q(m) \to +\infty$ with probability approaching one under spatial dependence of order less than or equal to *m*, then upper-tailed critical values are appropriate.

As previously noted, from a practical point of view, the researcher has to decide upon the embedding dimension m in order to compute symbolic entropy and therefore, to calculate the $Q(m)$ statistic. While this affords some flexibility, there are also some conditions that must be observed in order to guide a decision. Note that the number of locations that are symbolized (*R*) should be larger than the number of symbols (k^m) in order to have at least the same number of *m*-surroundings as symbols have been defined ($\sigma_i, i = 1,...,k^m$). When the χ^2 distribution is applied in practice, and all the expected frequencies are larger than or equal to five the limiting tabulated χ^2 distribution gives,

as a rule, the value χ^2_α with an approximation sufficient for ordinary purposes (see chapter 10 of Rohatgi 1976). For this reason, it is strongly advisable to work with *at least* $5k^m$ symbolized observations.

6 Finite sample behaviour of *Q*(*m*)

In this section, we examine the finite sample behaviour of the *Q*(*m*) test. This is to establish the power and size of the statistic under various levels of spatial association, sample size, size of the *m*-surrounding, and degree of overlap. In addition, we explore the potential impact of boundary effects.

6.1 Size and Power of the Test in Finite Samples

To investigate the power and size of the test, we conduct an extensive set of numerical experiments. Let us begin with some considerations regarding the data generating process used for the experiments. 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:

$$\boldsymbol{Y} = (I - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon} \tag{15}$$

where $\varepsilon \sim N(0,1)$, I is the $N \times N$ identity matrix, ρ is a parameter of spatial dependence, and $\boldsymbol{W}$ is a connectivity matrix that determines the set of spatial relationships among points. The process, therefore, is based on the auto-normal model. Alternative models for the data generation process were considered (e.g., the auto-logistic) but the auto-normal provides the best alternative for controlling the frequency of each categorical value in the simulations. 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:

$$p(Y \le b_{ij}) = \frac{i}{j} \text{ with i<j} \tag{16}$$

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

$$X_s = \begin{cases} a_1 & \text{if } Y_s \le b_{1k} \\ a_i & \text{if } b_{i-1k} < Y_s \le b_{ik} \\ a_k & \text{if } Y_s > b_{k-1k} \end{cases} \tag{17}$$

The last item that needs to be determined before data can be generated is a specific spatial arrangement of observations, so that matrix **W** can be defined. In this regards, we note that Farber et al. (2009) report that square tessellations provide poor approximates to the topology of real geographical systems. For this reason, we prefer to use for our experiments hexagonal tessellations, which more closely resemble the topology of Voronoi tessellations and administrative zoning systems used in many empirical applications. Two experiments use regular lattices of sizes *N*=100, 400, 900, 1,600, 2,500, and 3,600. In addition to these regular tessellations, we simulate irregular, but not random, spatial distributions of observations, with the same numbers of observations. The data are generated using Eq. (15), with the connectivity matrix defined in terms of first-order contiguity. Matrix **W** is row-standardized for the calculations.

Figure 3 shows examples of the different spatial distributions of *N*=900

observations generated for $\rho = 0$ (no spatial structure), $\rho = 0.5$ and $\rho = 0.9$, and for k=2, 3, 4 possible outcomes. As can be seen there, when the value of parameter ρ increases, more cells of the same colour cluster together. Examples of the irregular distributions of observations used in the second set of experiments are shown in Fig. 4. The following parameter space is explored, from no- to high-autocorrelation, six sample sizes, three classes for number of outcomes, three m-surrounding sizes, and overlap:

- Autocorrelation parameter ρ= 0.0, 0.2, 0.5, and 0.9
- Sample size N= 100, 400, 900, 1,600, 2,500 and 3,600.
- Number of outcomes k=2, 3 and 4
- Size of m-surrounding: three (self + 2 neighbours), four (self + 3 neighbours), five (self + 4 neighbours)
- Degree of overlap r=1, 2,..., as appropriate (see below)

Data are simulated 1,000 times for each combination of parameters (number of replications), and the test was applied to each generated dataset at level of significance $\alpha = 0.05$. The number of times that the probability value of the statistic exceeded 0.05 was recorded, which, following the decision rule posed in Eq. (13), would indicate rejection of the null hypothesis. We would expect the statistic to fail to reject the null hypothesis most of the time when the level of autocorrelation is zero (size of the statistic). At the same time, we would expect it to reject the null hypothesis more frequently as the level of autocorrelation goes up (power of the statistic).

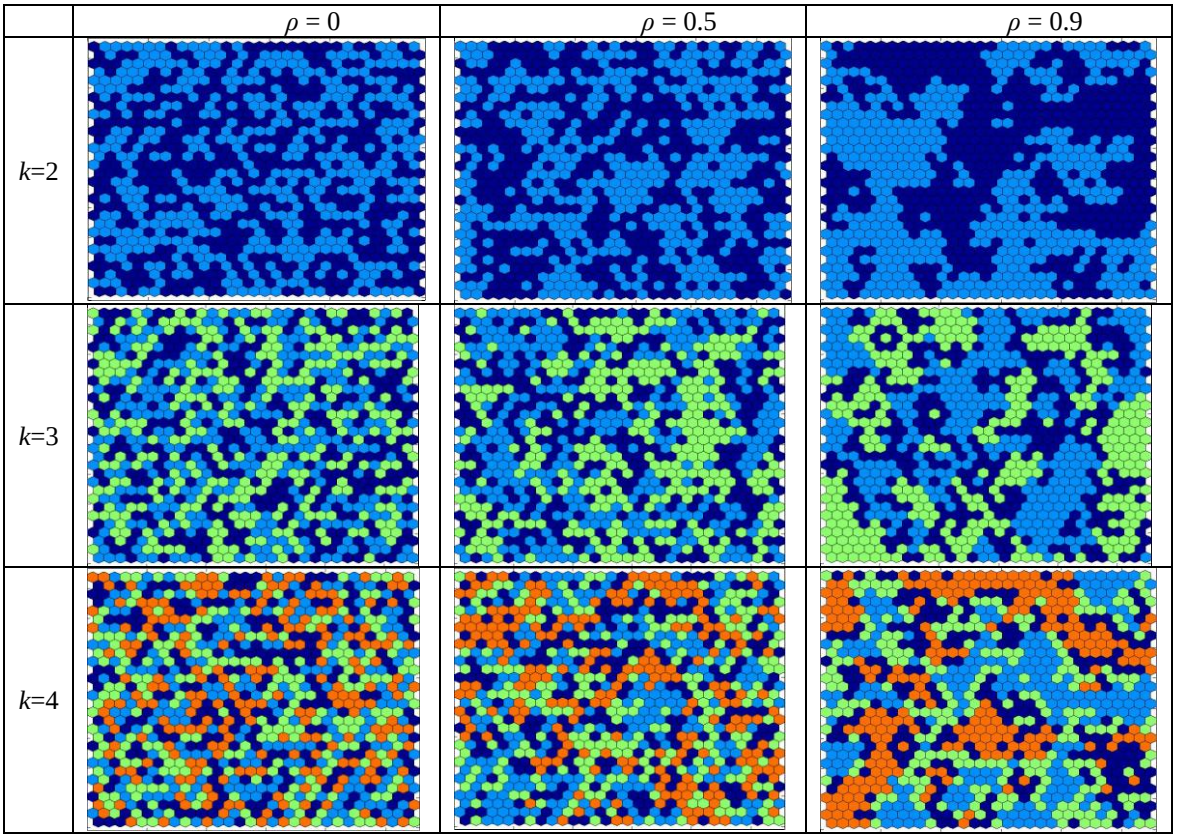

Fig. 3. Examples of distributions of observations on a regular hexagonal lattice (N=900), for different number of outcomes k and levels of ρ.

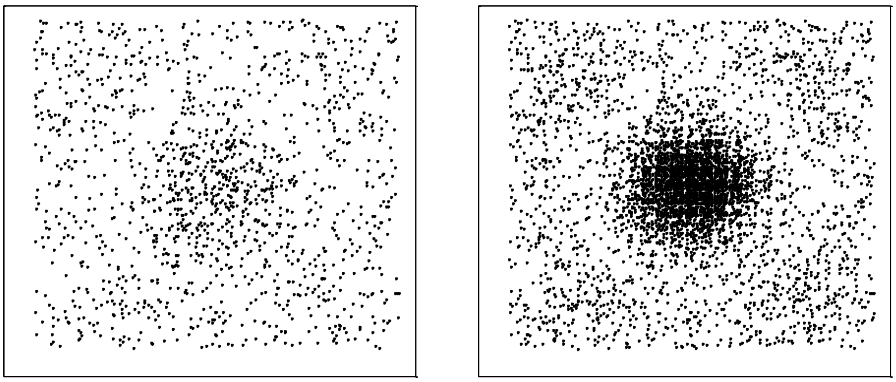

Fig. 4. Irregular distributions of observations N=900 and N=3,600

The results of the numerical experiments appear in Tables 2, 3, and 4, for the regular and irregular settings respectively. Please note that combinations of parameters are selected that satisfy the general rule that there must be *at least* $5k^m$ symbolized observations R. Since the number of symbolized locations depends on the degree of overlap, r must be selected so that the number of symbolized locations R is greater than $5k^m$.

The results of the experiment indicate that the size of the test (the rejection rate when the variable is independent) is typically higher for regular lattices, indicating a slightly greater risk for false positives, compared to irregular distributions of observations. Increasing the overlapping degree leads to more symbolized observations. This, as previously discussed, can result in non-independent Bernoulli variables, and therefore a higher size of the test, as seen in the tables. The increase in size is expected to carry over for higher levels of spatial dependence. The experiments are conducted for equal and unequal frequency of the values of the qualitative variable. The size of the test is not affected by changes in the proportionality of variable values in our experiments.

With regards to the power of the statistic, when the overlapping degree is high, the power will tend to be high as well. This is due to two effects: first, the starting size is higher, and secondly, the number of symbolized locations is greater. With regard to equal and unequal frequencies of the values of the variable, there is a slight loss in power when the values are not observed with identical frequencies. This loss is more marked for small sample sizes, which may make it difficult to identify even moderately strong spatially dependent processes in small sample situations. The loss in power becomes less relevant as the size of the sample increases, even for moderately large samples such as N=400. As usual, the power of the statistic tends to increase with increasing sample size. For a fixed sample size, the power is lower as the number of categories k increases, but this is due to the fact that the number of symbols will increase as well, and so the ratio of symbolized locations to symbols will decrease.

It is interesting to remark that the results are noticeably different for the case where observations are irregularly distributed compared to regular tessellations, when other parameters are comparable. This would suggest that the topology of the system to some extent can influence the performance of the statistic. While an in-depth investigation of the effect of topology is beyond the scope of this paper, this is suggested as a topic for future research along the lines of the studies by Páez et al. (2008) and Farber et al. (2009).

Table 2. Size and power of the Q test for k=2

Regular lattice

				p_1= p_2 =1/2				p_1=1/4; p_2=3/4			
N	R	m	r	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	0.051	0.032	0.088	0.736	0.034	0.041	0.069	0.640
	97	4	3	0.081	0.118	0.242	0.936	0.077	0.088	0.205	0.876
400	199	3	1	0.027	0.036	0.329	1.000	0.032	0.040	0.204	1.000
	199	4	2	0.045	0.052	0.370	1.000	0.052	0.065	0.268	1.000
	397	4	3	0.083	0.131	0.609	1.000	0.096	0.112	0.512	1.000
900	449	3	1	0.031	0.062	0.676	1.000	0.024	0.048	0.517	1.000
	449	4	2	0.038	0.070	0.792	1.000	0.028	0.051	0.621	1.000
	897	4	3	0.076	0.144	0.926	1.000	0.082	0.138	0.823	1.000
	299	5	2	0.062	0.074	0.610	1.000	0.062	0.105	0.498	1.000
1600	799	3	1	0.034	0.088	0.878	1.000	0.035	0.075	0.779	1.000
	799	4	2	0.041	0.118	0.974	1.000	0.046	0.102	0.902	1.000
	1597	4	3	0.083	0.259	0.997	1.000	0.082	0.203	0.977	1.000
	532	5	2	0.046	0.090	0.859	1.000	0.057	0.103	0.765	1.000

Irregular lattice

				p_1= p_2 =1/2				p_1=1/4; p_2=3/4			
N	R	m	r	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	0.029	0.037	0.086	0.811	0.027	0.042	0.090	0.704
	97	4	3	0.052	0.068	0.242	0.952	0.053	0.054	0.170	0.897
400	199	3	1	0.024	0.055	0.465	1.000	0.025	0.059	0.354	1.000
	199	4	2	0.036	0.061	0.539	1.000	0.036	0.062	0.423	1.000
	397	4	3	0.046	0.127	0.766	1.000	0.053	0.119	0.675	1.000
900	449	3	1	0.025	0.096	0.876	1.000	0.038	0.078	0.764	1.000
	449	4	2	0.037	0.125	0.924	1.000	0.048	0.113	0.817	1.000
	897	4	3	0.053	0.215	0.984	1.000	0.052	0.207	0.952	1.000
	299	5	2	0.053	0.102	0.800	1.000	0.056	0.113	0.685	1.000
1600	799	3	1	0.024	0.160	0.993	1.000	0.036	0.116	0.950	1.000
	799	4	2	0.038	0.199	0.998	1.000	0.046	0.144	0.977	1.000
	1597	4	3	0.039	0.383	1.000	1.000	0.052	0.308	1.000	1.000
	532	5	2	0.033	0.126	0.978	1.000	0.067	0.126	0.942	1.000

Table 3. Size and power of the Q test for k=3

Regular lattice

				$p_1= p_2= p_3 =1/3$				$p_1=1/8; p_2=3/8; p_3=4/8$			
N	R	m	R	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.9$
400	199	3	1	0.030	0.048	0.290	1.000	0.037	0.039	0.296	1.000
900	449	3	1	0.033	0.037	0.611	1.000	0.034	0.055	0.555	1.000
	449	4	2	0.064	0.077	0.690	1.000	0.058	0.098	0.726	1.000
	897	4	3	0.067	0.148	0.891	1.000	0.101	0.162	0.890	1.000
1600	799	3	1	0.020	0.072	0.879	1.000	0.029	0.056	0.839	1.000
	799	4	2	0.025	0.104	0.958	1.000	0.066	0.130	0.946	1.000
	1597	4	3	0.078	0.209	0.995	1.000	0.127	0.225	0.991	1.000
2500	1249	3	1	0.027	0.090	0.992	1.000	0.031	0.091	0.977	1.000
	1249	4	2	0.041	0.134	0.996	1.000	0.047	0.153	0.992	1.000
	2497	4	3	0.094	0.265	0.999	1.000	0.107	0.302	1.000	1.000

Irregular lattice

				$p_1= p_2= p_3 =1/3$				$p_1=1/8; p_2=3/8; p_3=4/8$			
N	R	m	R	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.9$
400	199	3	1	0.029	0.037	0.086	0.811	0.027	0.042	0.090	0.704
900	449	3	1	0.024	0.055	0.465	1.000	0.025	0.059	0.354	1.000
	449	4	2	0.036	0.061	0.539	1.000	0.036	0.062	0.423	1.000
	897	4	3	0.046	0.127	0.766	1.000	0.053	0.119	0.675	1.000
1600	799	3	1	0.025	0.096	0.876	1.000	0.038	0.078	0.764	1.000
	799	4	2	0.037	0.125	0.924	1.000	0.048	0.113	0.817	1.000
	1597	4	3	0.053	0.215	0.984	1.000	0.052	0.207	0.952	1.000
2500	1249	3	1	0.024	0.160	0.993	1.000	0.036	0.116	0.950	1.000
	1249	4	2	0.038	0.199	0.998	1.000	0.046	0.144	0.977	1.000
	2497	4	3	0.039	0.383	1.000	1.000	0.052	0.308	1.000	1.000

Table 4. Size and Power of the Q test for k=4

Regular lattice

				$p_1= p_2= p_3= p_4=1/4$				$p_1=1/12; p_2=2/12; p_3=3/12; p_4=6/12$			
N	R	m	r	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.9$
900	449	3	1	0.033	0.059	0.517	1.000	0.039	0.081	0.514	1.000
1600	799	3	1	0.026	0.043	0.788	1.000	0.040	0.069	0.725	1.000
2500	1249	3	1	0.026	0.076	0.971	1.000	0.026	0.086	0.927	1.000
3600	1799	3	1	0.031	0.099	0.997	1.000	0.037	0.099	0.995	1.000
	1799	4	2	0.070	0.185	1.000	1.000	0.097	0.271	0.998	1.000
	3597	4	3	0.077	0.280	1.000	1.000	0.147	0.400	0.999	1.000

Irregular lattice

				$p_1= p_2= p_3= p_4=1/4$				$p_1=1/12; p_2=2/12; p_3=3/12; p_4=6/12$			
N	R	m	r	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.9$
900	449	3	1	0.033	0.081	0.799	1.000	0.039	0.085	0.755	1.000
1600	799	3	1	0.031	0.111	0.991	1.000	0.052	0.103	0.965	1.000
	1597	4	3	0.098	0.308	1.000	1.000	0.098	0.366	1.000	1.000
2500	1249	3	1	0.036	0.167	1.000	1.000	0.036	0.149	0.999	1.000
	2497	4	3	0.086	0.395	1.000	1.000	0.112	0.490	1.000	1.000
3600	1799	3	1	0.017	0.251	1.000	1.000	0.038	0.240	1.000	1.000
	1799	4	2	0.062	0.302	1.000	1.000	0.098	0.399	1.000	1.000
	3597	4	3	0.059	0.521	1.000	1.000	0.122	0.576	1.000	1.000

6.2 Boundary effects

In this section, we are interested in assessing the potential effect of system boundaries when data points are not observed beyond an arbitrarily defined boundary. The usual question when considering boundary effects is whether the behaviour of a statistical estimator changes if variable X_i is influenced by X_j and j is a location outside of the study area (Haining 1990, p. 101; Upton and Fingleton 1985, p. 365). For us, the question is whether our ability to detect a spatially dependent process is influenced by unknown values of the variable at locations beyond the boundary of the study area. In other words, how frequently the statistic would agree to either reject or fail to reject the null hypothesis for two systems that are comparable (in size) if one is observed completely and the other is observed only partially.

Table 5. Size and power of $Q(m)$ in the presence of boundary effects

k=2					$p_1 \simeq p_2 \simeq 1/2$								
					Partial system				*Complete system*				
N_O	N_C	R	m	r	ρ = 0	ρ = 0.2	ρ = 0.5	ρ = 0.9	N_O=N_C	ρ = 0	ρ = 0.2	ρ = 0.5	ρ = 0.9
100	196	49	3	1	0.044	0.035	0.052	0,609	100	0.051	0.032	0.088	0.736
		97	4	3	0.091	0.111	0.178	0,876		0.081	0.118	0.242	0.936
400	576	199	3	1	0.024	0.047	0.248	1,000	400	0.027	0.036	0.329	1.000
		397	4	3	0.073	0.110	0.571	1,000		0.083	0.131	0.609	1.000
900	1156	449	3	1	0.029	0.043	0.570	1,000	900	0.031	0.062	0.676	1.000
		897	4	3	0.065	0.124	0.874	1,000		0.076	0.144	0.926	1.000
1600	1936	799	3	1	0.027	0.084	0.841	1,000	1600	0.034	0.088	0.878	1.000
		1597	4	3	0.083	0.236	0.989	1,000		0.083	0.259	0.997	1.000
k=3					$p_1 \simeq p_2 \simeq p_3 \simeq 1/3$								
					Partial system				*Complete system*				
N_O	N_C	R	m	r	ρ = 0	ρ = 0.2	ρ = 0.5	ρ = 0.9	N_O=N_C	ρ = 0	ρ = 0.2	ρ = 0.5	ρ = 0.9
400	576	199	3	1	0.025	0.038	0.223	0.999	400	0.030	0.048	0.290	1.000
		397	4	3	0.125	0.156	0.519	1.000		0.141	0.156	0.586	1.000
900	1156	449	3	1	0.023	0.041	0.502	1.000	900	0.033	0.037	0.611	1.000
		897	4	3	0.081	0.139	0.838	1.000		0.067	0.148	0.891	1.000
1600	1936	799	3	1	0.024	0.050	0.828	1.000	1600	0.020	0.072	0.879	1.000
		1597	4	3	0.077	0.192	0.984	1.000		0.078	0.209	0.995	1.000
2500	2916	1249	3	1	0.019	0.093	0.969	1.000	2500	0.027	0.090	0.992	1.000
		2497	4	3	0.097	0.267	0.999	1.000		0.094	0.265	0.999	1.000
k=4					$p_1 \simeq p_2 \simeq p_3 \simeq p_4 \simeq 1/4$								
					Partial system				*Complete system*				
N_O	N_C	R	m	r	ρ = 0	ρ = 0.2	ρ = 0.5	ρ = 0.9	N_O=N_C	ρ = 0	ρ = 0.2	ρ = 0.5	ρ = 0.9
900	1156	449	3	1	0.036	0.049	0.398	1.000	900	0.033	0.059	0.517	1.000
1600	1936	799	3	1	0.031	0.047	0.703	1.000	1600	0.026	0.043	0.788	1.000
2500	2916	1249	3	1	0.026	0.050	0.947	1.000	2500	0.026	0.076	0.971	1.000
3600	4096	1799	3	1	0.035	0.076	0.995	1.000	3600	0.031	0.099	0.997	1.000

To evaluate the effect of boundaries in the size and power of the test, we conduct a second simulation experiment. The data are generated using the same procedure described in the preceding section. In irregular lattices, *m*-surroundings are constructed based on proximity, only resorting to the direction criterion in the rare cases when two neighbours are equidistant. Boundary effects could be more critical in regular tessellations due to the way that the *m*-surroundings are constructed involving the direction of the neighbours. For this reason, we repeat the experiments with regular hexagonal tessellations only. The results can be compared to the experiments in the

preceding section (Tables 2, 3 and 4), conducted under the assumption that the system was completely observed. In the case of the new simulation, a complete system is simulated with N_C cases. In order to simulate the boundaries, we remove all hexagons in the periphery, to obtain an observed system with a total of N_O cases after omitting all observations in the boundaries. The simulation is done for 1000 replications, and the frequency of rejection of the statistic is recorded for each combination of parameters to calculate the size and power of the test when applied to a partial system.

The results of the experiment are presented in Table 5 for various values of N_O, ρ, and other parameters. The columns give the frequency of rejection of the null hypothesis. It is to be expected that the frequency of rejection in the case of a partial system will not be affected when the level of autocorrelation is zero. In this case, the data points are independent, and what happens beyond the boundary stays there. The results of the simulation confirm this, since it can be seen that the size is comparable for every case studied. The results indicate that boundary effects influence the power of the statistic when autocorrelation is present. The effect in general is to reduce the power of the statistic, although the reduction is relatively small in most cases, and the loss in power tends to be smaller as the number of observations and the level of autocorrelation increase, since this is naturally associated with higher power in any case.

Typical recommendations for the treatment of boundary effects include to collect more data points whenever possible, or to create an artificial boundary or buffer, and to consider the observations within the buffer as a known boundary. The first recommendation is sensible but frequently unfeasible. The second recommendation is of dubious merit in the case of our statistic, because any gains in power are bound to be minor as suggested by the simulation, and likely be offset by the loss of power associated with a reduced number of observations.

7 Illustration: Fast food establishments in Toronto

We now proceed to illustrate the use of the $Q(m)$ test by means of an empirical example concerning the spatial association of fast food establishments in the city of Toronto, Canada, specifically those offering primarily [P]izza, [S]andwich, and [H]amburger products. Use of spatial statistics has recently been applied to the study of food environments (Austin et al. 2005), and our example illustrates other ways in which the food landscape can be examined from a spatial perspective. In particular, we explore the question of whether the type of establishment is independent from its neighbours, whether establishments tend to attract or repel establishments of the same type.

7.1 Data

The analysis is based on business points, which record the location of different establishments in the city of Toronto, as well as their industrial codes and other characteristics, such as various categories of size, revenue, etc. The business directory is based on infoCanada data, which is compiled from over 200,000 sources, including telephone directories, annual reports, press releases, city and industrial directories, news items, and new business listings. The database is telephonically verified annually by infoCanada to ensure the accuracy of the information. This information is processed and packaged by Environics Analytics to produce a business profiles database.

The final database for analysis includes a custom Standard Industrial Classification code which allows for the identification of business groups. Location coordinates are coded by Environics Analytics to enable mapping applications of the businesses recorded in the database. For the purpose of this illustration, a subset of observations is extracted from the file corresponding to the region surrounding the city

of Toronto, to obtain a set of 877 businesses with Standard Industrial Codes 5812 classification ("Eating Places") that can be identified as offering primarily one of 3 types of fast food, including [P]izza (n_P=303), [S]andwich (n_S=299), and two major [H]amburger chains in the city (n_H=275). The spatial distribution of fast food places is shown in Fig. 5. It is reasonable in this case to think of the observations as a completely observed system, because development beyond the study area is sparse with the exception of the western boundary. As suggested by the simulations in Section 6.2, even if there are boundary effects, there is little risk that a false positive will be obtained, and the loss of power is bound to be relatively small.

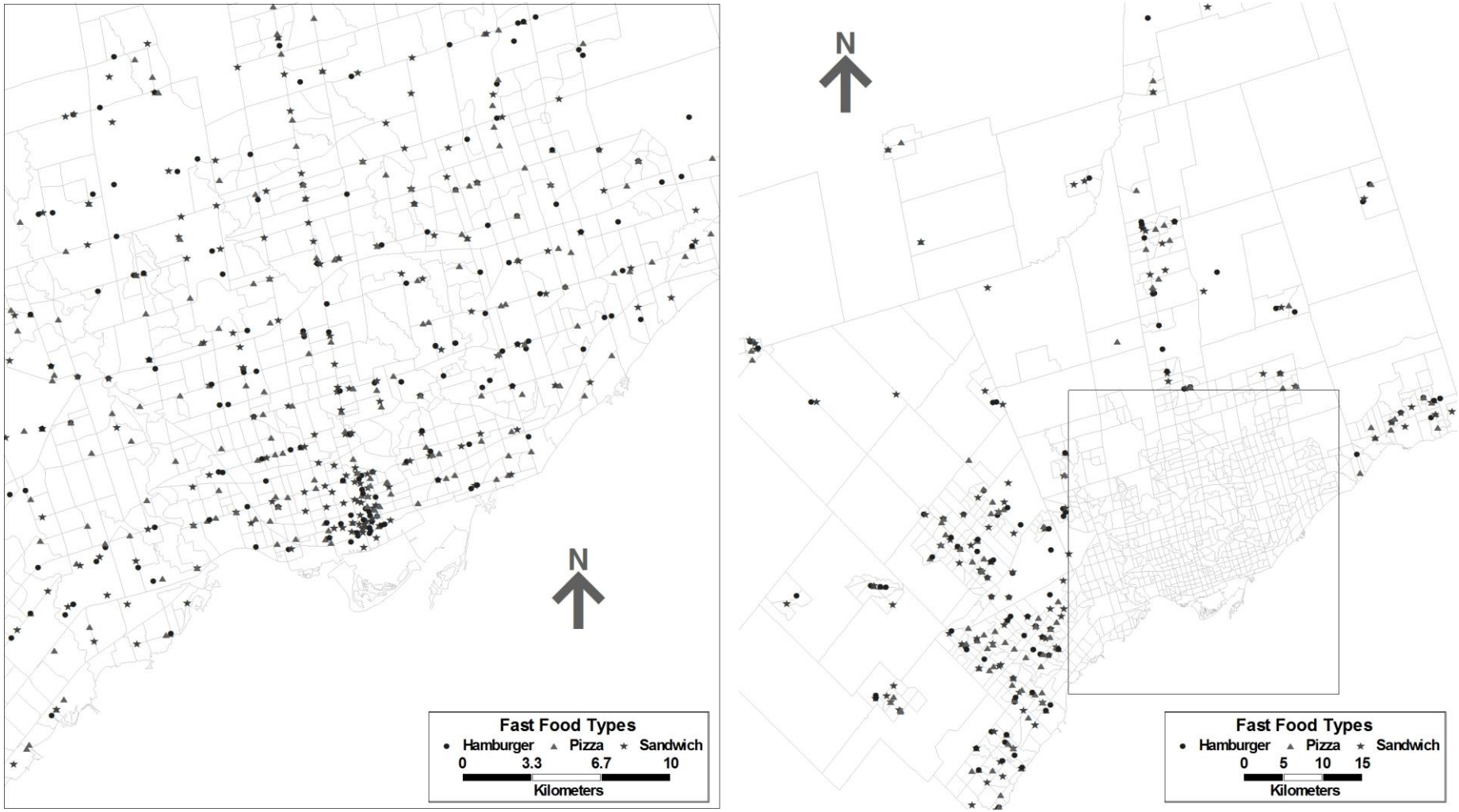

Fig. 5. Fast food establishments in Toronto and the Greater Toronto Area

7.2 Analysis and results

To support our analysis of spatial association of a qualitative variable (establishment type) we first analyze the point pattern of establishments. This is done by means of nearest neighbour analysis, an approach developed with the objective of measuring the degree of proximity between events and their nearest neighbours (Bailey and Gatrell 1995). The specific technique we use is the *G* function, a cumulative plot that shows the proportion of events that have a nearest neighbour at a distance of *d* or less:

$$G(d) = \frac{\#\left(\min\left(d_i\right) \le d\right)}{N} \tag{18}$$

The analysis can be performed for k^{th} order neighbours, that is, the proportion of events whose k^{th} nearest neighbour is at a distance *d* or less. A steep increase of the function indicates a tendency towards spatial clustering. The results of this analysis are shown in Fig. 6, where it can be seen that about 70% of events have a first order nearest neighbour within 500 m distance, and about 90% have first order neighbours within 1.2 km. About 70% of events have a second order neighbour within 1.1 km, and a third order neighbour within 1.5 km. This gives a stronger basis to the preliminary impression that there is a good deal of spatial clustering in the location pattern of fast food establishments.

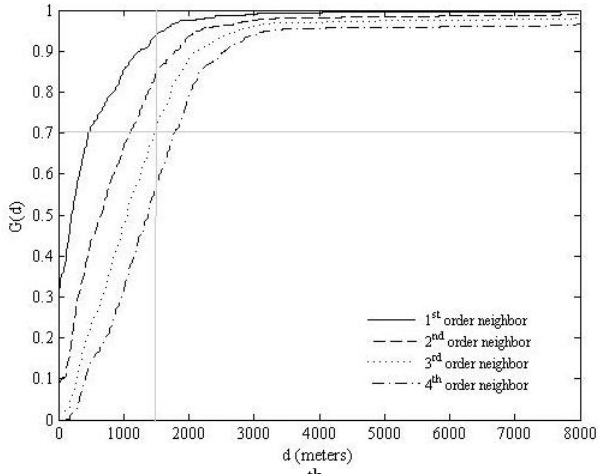

Fig. 6. Event-to-event k^{th} nearest neighbour distance analysis

We now turn to the question of whether there are patterns of association for the different types of establishments. Application of our statistic is straightforward. The number of possible event outcomes is k=3, and the number of observations is N=877. Given these values $\ln\left(N/5\right)/\ln\left(k\right) = 4.7033$, meaning that we can explore *m*-surroundings of size two (self and one nearest neighbour), three (self and two nearest neighbours) and four (self and three nearest neighbours). On the other hand, we are prevented by the sample size to explore *m*-surroundings of size five or larger. Based on our previous application of the *G* function, there appears to be only relatively a minor difference between m=3 and m=4, in terms of the spatial distribution of the locations of establishments. Since a property of the statistic is that it detects spatial dependence of order less than or equal to *m*, it seems sensible to select m=4 for our analysis. If dependence is detected, it would carry for the cases of m=2 and m=3. One additional decision to make concerns the degree of overlap. An overlap of r=1 does not satisfy the criterion that the number of symbolized locations be greater than $5k^{\text{m}}$=405. Overlaps of r=2 and r=3 result in R=437 and R=874 symbolized locations respectively. Since the simulation results indicate that higher values of R generally increase the power of the

statistic, we select r=3 for our application. The summary of these parameters and results of the test are shown in Table 6.

Table 6. Summary of parameters and results

Sample size (N)	877		
Symbolized observations (R)	874		
Number of classes (k)	3		
Size of m-surrounding (m)	4		
Degree of overlap (r)	3		
Number of symbols (n)	81		
Ratio R/n	10.79		
$5k^m$	405.00		
Frequency of classes	0.3136	0.3455	0.3409

Q test for spatial dependence in qualitative data			
Test	Value	DF	p-value
Q (equiprobab.)	177.27	80	$<10^{-5}$
Q (non-equiprob.)	166.38	80	$<10^{-5}$

The value of the statistic for the approximate case of equal frequency of classes is 177.27 (see Eq. (11)), and for non-equiprobability is 166.38 (see Eq. (10)). These values are tested using the χ^2 distribution with $k^m - 1 = 80$ degrees of freedom. The cut-off value for rejection at the 0.05 level of significance is 101.8795, which the values of the test exceed, and therefore, according to our decision rule, leads to rejection of the hypothesis of independence. Alternatively, the probability values in both cases are smaller than 10^{-5}.

The test rejects the hypothesis of independence. However, dependence could take different forms. An attractive feature of the $Q(m)$ statistic is that it is based on the frequency of different symbols being observed, which allows a more in-depth exploration of the patterns of association. Recall that the probability of each symbol appearing under the hypothesis of independence is $1/k^m$, so that in this case, since there are 874 symbolized locations, each symbol would appear approximately eleven times. It is possible to plot a histogram with the actual frequency of the 81 symbols (see Fig. 7). The expected frequency under the null is indicated by the dotted line in the figure, and it is possible to see which symbols deviate from this expectation, and in which direction (more frequent, less frequent). The symbols carry a fair amount of information, since each symbol represents a particular combination of events, and also their order of proximity and possibly directionality from s_0.

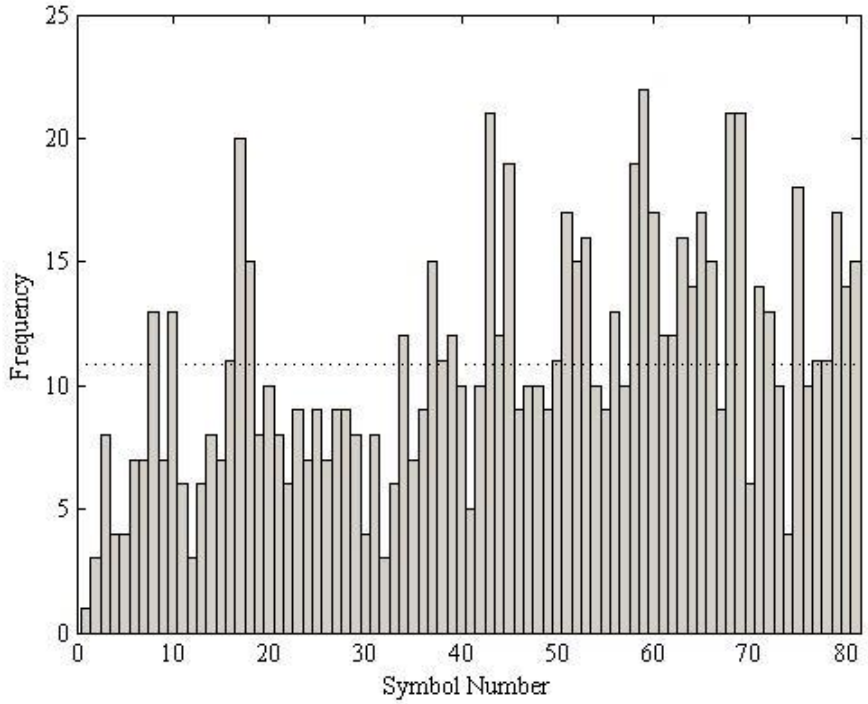

Fig. 7. Frequency of fast food type co-location symbols in Toronto (*m*=4 and *r*=3)

In Fig. 8, we condense the information contained in the histogram, in order to display only the types of events in *m*-surroundings, but not other features of the pattern, such as order of proximity. This allows us to discern that four establishments of a kind (four pizza, four sandwich, or four hamburger places) are seldom found together. Much more common is the case where neighbouring groups 4 establishments consist of a variety of establishments, with at most two of one class, and one each of the other two classes. This would tend to indicate – in addition to the evidence of economies of agglomeration evinced by the spatial clustering – that within clusters there is a pattern of competition or repulsion between establishments of the same type.

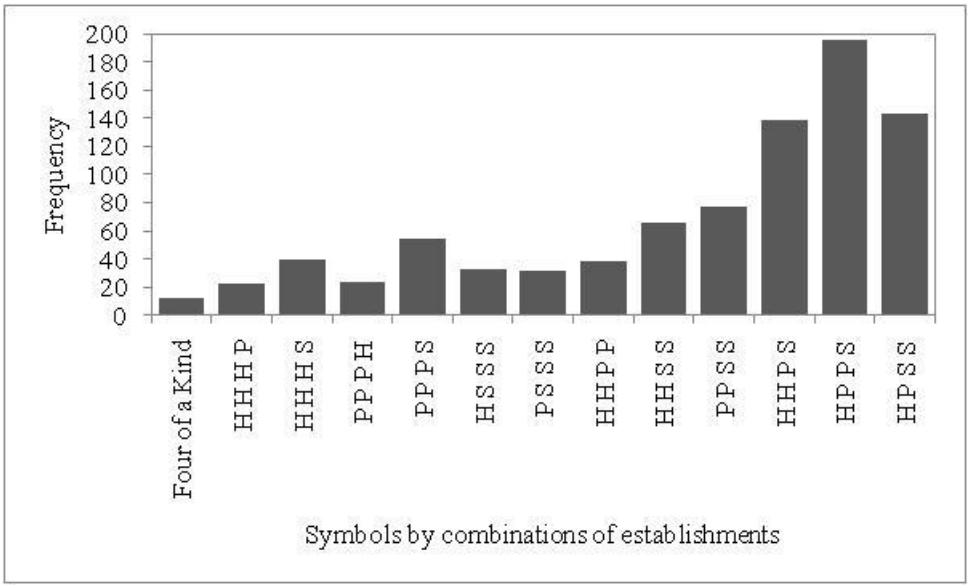

Figure 8. Co-location of events, condensed histogram. H,S, and P indicate the type of establishment, e.g., HHPP means two Hamburger and two Pizza in a group of four.

8 Conclusions

In this paper, we have proposed a new statistic $Q(m)$ useful to test the hypothesis of independence among spatially distributed qualitative data. Qualitative data are receiving increased attention from a number of disciplinary perspectives. However, besides black and white or k-coloured join count statistics (e.g., Cliff and Ord 1981; Dacey 1968; Upton and Fingleton 1985), and the work of Boots (2003), there has been only limited development in terms of spatial analysis of qualitative data, compared to the development of methods and techniques useful to study continuous variables. Our statistic therefore is proposed as a complement to further enrich the diversity of the spatial analysis toolbox.

The $Q(m)$ statistic is developed parting from concepts of symbolic dynamics. Symbolic dynamics provide an ideal set of tools to investigate discrete processes. Our statistic therefore is designed for the analysis of spatially discrete events, with qualitative/nominal outcomes. In this paper, we provide the inferential basis for conducting tests of hypothesis based on an affine transformation of the statistic, and a decision rule is proposed to reject or fail to reject the hypothesis of independence. We have also performed an extensive set of numerical experiments that demonstrate the size and power of the statistic to identify spatial association under a range of different conditions, and the potential effect of boundaries. In addition an example illustrates the usefulness of the statistic to address substantive research questions.

In addition to its ability to identify patterns of spatial association, an attractive feature of our statistic is that it is based on the frequency of occurrence of various abstract symbols that can be linked to meaningful states of the system. The frequency of the symbols can be examined to obtain in-depth information about departures from the expected frequencies under the null hypothesis of independence. The ability to do this is akin, if not identical, to that provided by Moran's scatterplot, in that it gives specific patterns of association that can be contrasted with different ideas about the substantive process. In our example, we discussed a condensed histogram of the symbols, which provides a simplified perspective on the patterns of association. However, it is not difficult to envision other questions of interest that could be explored using the full histogram, for example, concerning directional or proximity trends of other types of events (e.g., do sandwich establishments tend to be closer, or further away from pizza locations, relative to hamburger places?) In fact, the symbolization procedure can be modified in order to address specific research needs, for example to deal with questions of anisotropy or others. This is a matter for further research.

Two additional points are indicated as directions for additional research. First, the number of outcomes k typically depends on the nature of the process, and is therefore a parameter beyond the control of the analyst. The number of observations needed to conduct analysis can quickly explode depending on the size of the m-surroundings. For example, if k=4, and one desires to examine m-surroundings of size three, at least 625 points would be required. In contrast, an m-surrounding of five would require 3,125 observations. A topic for further investigation is whether different symbolization schemes can help to maintain data needs under control. Secondly, using the histogram of frequency of symbols to test whether each symbol departs significantly from the expected frequency appears as a promising direction for further research. The histogram already provides a decomposition of the statistic, and the ability to test deviations for specific symbols would further enhance the capabilities of the statistic, in the manner of various other local statistics of spatial association (Anselin 1995; Getis and Ord 1993).

9 Appendix: Proofs

Proof of Theorem 1

Under the null H_0, the joint probability density function of the n variables $(Y_{\sigma_1}, Y_{\sigma_2}, ..., Y_{\sigma_{k^m}})$ is:

$$P\left(Y_{\sigma_1} = a_1, Y_{\sigma_2} = a_2, ..., Y_{\sigma_{k^m}} = a_{k^m}\right) = \frac{\left(a_1 + a_2 + ... + a_{k^m}\right)!}{a_1! a_2! \cdot ... \cdot a_{k^m}!} p_{\sigma_1}^{a_1} p_{\sigma_2}^{a_2} \cdots p_{\sigma_{k^m}}^{a_{k^m}} \tag{A.1}$$

where $a_1 + a_2 + ... + a_n = R$. Consequently, the joint distribution of the n variables $(Y_{\sigma_1}, Y_{\sigma_2}, ..., Y_{\sigma_{k^m}})$ is a multinomial distribution.

The likelihood function of the distribution given by Eq. (A.1) is:

$$L(p_{\sigma_1}, p_{\sigma_2}, ..., p_{\sigma_{k^m}}) = \frac{R!}{n_{\sigma_1}! n_{\sigma_2}! \cdot ... \cdot n_{\sigma_{k^m}}!} p_{\sigma_1}^{n_{\sigma_1}} p_{\sigma_2}^{n_{\sigma_2}} \cdots p_{\sigma_{k^m}}^{n_{\sigma_{k^m}}} \tag{A.2}$$

and since, $\sum_{i=1}^{k^m} p_{\sigma_i} = 1$, it follows that

$$L\left(p_{\sigma_1}, p_{\sigma_2}, ..., p_{\sigma_{k^m}}\right) = \frac{R!}{n_{\sigma_1}! n_{\sigma_2}! \cdot ... \cdot n_{\sigma_{k^m}}!} p_{\sigma_1}^{n_{\sigma_1}} p_{\sigma_2}^{n_{\sigma_2}} \cdots \left(1 - p_{\sigma_1} - p_{\sigma_2} - ... - p_{\sigma_{k^m - 1}}\right)^{n_{\sigma_{k^m}}} \tag{A.3}$$

Then the logarithm of this likelihood function remains as

$$\begin{aligned} \ln\left[L\left(p_{\sigma_1}, p_{\sigma_2}, ..., p_{\sigma_{k^m}}\right)\right] = \ln\left(\frac{R!}{n_{\sigma_1}! n_{\sigma_2}! \cdot ... \cdot n_{\sigma_{k^m}}!}\right) + \sum_{i=i}^{k^m - 1} n_{\sigma_i} \ln\left(p_{\sigma_i}\right) \\ + n_{\sigma_{k^m}} \ln\left(1 - p_{\sigma_1} - p_{\sigma_2} - ... - p_{\sigma_{k^m - 1}}\right) \end{aligned} \tag{A.4}$$

In order to obtain the maximum likelihood estimators $\hat{p}_{\sigma_i}$ of p_{σ_i} for all $i = 1, 2, ..., n$, we solve the following equation

$$\frac{\partial}{\partial p_{\sigma_i}} \ln\left[L\left(p_{\sigma_1}, p_{\sigma_2}, ..., p_{\sigma_n}\right)\right] = 0 \tag{A.5}$$

to get that:

$$\hat{p}_{\sigma_i} = \frac{n_{\sigma_i}}{R} \tag{A.6}$$

Then the likelihood ratio statistic is (see for example Lehman 1986):

$$\begin{aligned} \lambda(Y) = \frac{\frac{R!}{n_{\sigma_1}! n_{\sigma_2}! \cdot ... \cdot n_{\sigma_{k^m}}!} p_{\sigma_1}^{(0) n_{\sigma_1}} p_{\sigma_2}^{(0) n_{\sigma_2}} \cdots p_{\sigma_{k^m}}^{(0) n_{\sigma_{k^m}}}}{\frac{R}{n_{\sigma_1}! n_{\sigma_2}! \cdot ... \cdot n_{\sigma_{k^m}}!} p_{\sigma_1}^{n_{\sigma_1}} p_{\sigma_2}^{n_{\sigma_2}} \cdots p_{\sigma_{k^m}}^{n_{\sigma_{k^m}}}} = \frac{\prod_{i=1}^{k^m} p_{\sigma_i}^{(0) n_{\sigma_i}}}{\prod_{i=1}^{k^m} \left(\frac{n_{\sigma_i}}{R}\right)^{n_{\sigma_i}}} = \\ = R^{\sum_{i=1}^{k^m} n_{\sigma_i}} \prod_{i=1}^{k^m} \left(\frac{p_{\sigma_i}^{(0)}}{n_{\sigma_i}}\right)^{n_{\sigma_i}} = R^R \prod_{i=1}^{k^m} \left(\frac{p_{\sigma_i}^{(0)}}{n_{\sigma_i}}\right)^{n_{\sigma_i}} \end{aligned} \tag{A.7}$$

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

On the other hand, $Q(m) = -2\ln\left(\lambda(Y)\right)$ asymptotically follows a Chi-squared distribution with $k^m - 1$ degrees of freedom (see Lehman 1986). Hence:

$$Q(m) = -2\ln\left(\lambda(Y)\right) = -2\left[R\ln(R) + \sum_{i=1}^{k^m} n_{\sigma_i} \ln\left(\frac{p_{\sigma_i}^0}{n_{\sigma_i}} \right) \right] \sim \chi^2_{k^m - 1} \tag{A.8}$$

Denote by α_{ij} the number of times that class a_j appears in symbol σ_i and by $q_j = P(X = a_j)$. Then under the null we have that $p_{\sigma_i}^{(0)} = \prod_{j=1}^{k} q_j^{\alpha_{ij}}$ and hence, it follows that

$$\begin{aligned} Q(m) &= -2R\left[\ln(R) + \sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{R} \ln\left(\frac{\prod_{j=1}^{k} q_j^{\alpha_{ij}}}{n_{\sigma_i}} \right) \right] \\ &= -2R\left[\ln(R) + \sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{R} \left(\ln\left(\prod_{j=1}^{k} q_j^{\alpha_{ij}} \right) - \ln\left(n_{\sigma_i} \right) \right) \right] \\ &= -2R\left[\ln(R) + \sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{R} \left(\ln\left(\prod_{j=1}^{k} q_j^{\alpha_{ij}} \right) - \ln\left(\frac{n_{\sigma_i}}{R} \right) - \ln(R) \right) \right] = \\ &= -2R\left[\ln(R) + \sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{R} \left(\sum_{j=1}^{k} \alpha_{ij} \ln\left(q_j \right) - \ln\left(\frac{n_{\sigma_i}}{R} \right) - \ln(R) \right) \right] = \end{aligned} \tag{A.9}$$

Now, taking into account that $h(m) = -\sum_{i=1}^{k^m} p_{\sigma_i} \ln\left(p_{\sigma_i} \right) = -\sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{R} \ln\left(\frac{n_{\sigma_i}}{R} \right)$, we have that

$$Q(m) = -2R\left[\sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{R} \sum_{j=1}^{k} \alpha_{ij} \ln\left(q_j \right) + h(m) \right] \tag{A.10}$$

Notice that if the spatial process is independent identically distributed then $q_j = \frac{1}{k}$ and therefore $Q(m) = 2R\left(Ln(k^m) - h(m) \right)$ which finishes the proof of the theorem. □

Proof of Theorem 2

First, notice that the estimator of $h(m)$, $h(m) = -\sum_{\sigma \in S_m} \hat{p}_\sigma \ln\left(\hat{p}_\sigma \right)$, where $\hat{p}_\sigma = n_\sigma / R$, is consistent because, $p\lim_{R \to \infty} \hat{p}_\sigma = p_\sigma$, and hence:

$$p\lim_{R \to \infty} h(m) = h(m) \tag{A.11}$$

Recall that:

$$Q(m) = 2R\left[-\sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{R} \sum_{j=1}^{k} \alpha_{ij} \ln\left(q_j\right) + \sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{R} \ln\left(\frac{n_{\sigma_i}}{R}\right)\right] \tag{A.12}$$

Now, let us call:

$$H(m) = -\sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{R} \sum_{j=1}^{k} \alpha_{ij} \ln\left(q_j\right) + \sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{R} \ln\left(\frac{n_{\sigma_i}}{R}\right) = \sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{R} \ln\left(\frac{\frac{n_{\sigma_i}}{R}}{\prod_{j=1}^{k} q_j^{\alpha_{ij}}}\right) \tag{A.13}$$

Also, since $\ln(x) \le x-1$ for all x with equality if and only if x=1, and under the alternative hypothesis of spatial dependence of order $\le m$ we have that:

$$\frac{\frac{n_{\sigma_i}}{R}}{\prod_{j=1}^{k} q_j^{\alpha_{ij}}} \neq 1 \tag{A.14}$$

it follows that:

$$H(m) = -\sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{R} \ln\left(\frac{\prod_{j=1}^{k} q_j^{\alpha_{ij}}}{\frac{n_{\sigma_i}}{R}}\right) > -\sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{R} \left(\frac{\prod_{j=1}^{k} q_j^{\alpha_{ij}}}{\frac{n_{\sigma_i}}{R}} - 1\right) = -\sum_{i=1}^{k^m} \prod_{j=1}^{k} q_j^{\alpha_{ij}} + 1 \ge 0 \tag{A.15}$$

Since, also $p\lim_{R\to\infty} \hat{q}_j = q_j$, then by Eq. (A.11) we have

$$p\lim_{R\to\infty} H(m) = H(m) \tag{A.16}$$

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

$$\frac{C}{2R} < H(m) \tag{A.17}$$

Then, under the spatial dependence of order less than or equal to m it follows that $H(m) \neq 0$ and, thus,

$$\begin{aligned} &\Pr\left[Q(m) > C \right] = \Pr\left[2RH(m) > C \right] \\ &= \Pr\left[2R\left(H(m) - H(m)\right) > C - 2RH(m) \right] \\ &= \Pr\left[2R\left(H(m) - H(m)\right) < 2RH(m) - C \right] \\ &= \Pr\left[H(m) - H(m) < H(m) - \frac{C}{2R} \right] \end{aligned} \tag{A.18}$$

Therefore, by Eqs. (A.16), (A.17) and (A.18) we have that:

$$\lim_{R\to\infty} \Pr(Q(m) > C) = 1 \tag{A.19}$$

as desired. □

Acknowledgments

The authors gratefully acknowledge for financial support of grant EC-02009-10534-ECON of Ministerio Español de Ciencia e Innovación and Fundación Séneca de la Región de Murcia. In preparing this paper we benefited from the comments of anonymous reviewers, and feedback received from participants in the 2009 Meetings of the AAG and the 2009 Spatial Econometric World Congress. In particular, we are grateful for useful discussions with Prof. Daniel A. Griffith and Ms. Melissa J. Rura. The authors alone are responsible for the contents of the paper.

References

Anselin L (1988) Spatial econometrics: Methods and models. Kluwer, Dordrecht

Anselin L (1995) Local indicators of spatial association - LISA. Geographical Analysis 27(2):93-115

Austin SB, Melly SJ, Sanchez BN, Patel A, Buka S, Gortmaker SL (2005) Clustering of fast-food restaurants around schools: A novel application of spatial statistics to the study of food environments. American Journal of Public Health 95(9):1575-1581

Bailey TC, Gatrell AC (1995) Interactive spatial data analysis. Addison Wesley Longman, Essex

Bell N, Schuurman N, Hameed SM (2008) Are injuries spatially related? Join-count spatial autocorrelation for small-area injury analysis. Injury Prevention 14(6):346-353

Bhat CR, Sener IN (2009) A copula-based closed-form binary logit choice model for accommodating spatial correlation across observational units. Journal of Geographical Systems 11(3):243-272

Boots B (2003) Developing local measures of spatial association for categorical variables. Journal of Geographical Systems 5(2):139-160

Chakir R, Parent O (2009) Determinants of land use changes: A spatial multinomial probit approach. Papers in Regional Science 88(2):327-344

Chuang KS, Huang HK (1992) Assessment of noise in a digital image using the join-count statistic and the Moran test. Physics in Medicine and Biology 37(2):357-369

Cliff AD, Ord JK (1973) Spatial autocorrelation. Pion, London

Cliff AD, Ord JK (1981) Spatial processes: Models and applications. Pion, London

Cressie NAC (1993) Statistics for spatial data. Wiley, New York

Dacey MF (1968) A review on measures of contiguity for two and k-color maps. In: Berry BJL and Marble DF (eds) Spatial Analysis: A Reader in Statistical Geography. Prentice Hall, Englewood Cliffs [NJ], pp 479-495

Dejong PD, Debree J (1995) Analysis of the spatial-distribution of rust-infected leek plants with the black-white join-count statistic. European Journal of Plant Pathology 101(2):133-137

Dubin R (1995) Estimating logit models with spatial dependence. In: Anselin L, Florax RJGM (eds) New directions in spatial econometrics. Springer, Berlin, Heidelberg, New York, pp 229-242

Epperson BK, AlvarezBuylla ER (1997) Limited seed dispersal and genetic structure in life stages of Cecropia obtusifolia. Evolution 51(1):275-282

Farber S, Páez A, Volz E (2009) Topology and dependency tests in spatial and network autoregressive models. Geographical Analysis 41(2):158-180

Geary RC (1954) The contiguity ratio and statistical mapping. The Incorporated Statistician 5(3):115-145

Getis A (2008) A history of the concept of spatial autocorrelation: A geographer's perspective. Geographical Analysis 40(3):297-309

Getis A, Ord JK (1993) The analysis of spatial association by use of distance statistics. Geographical Analysis 25(3):276-276

Ghent AW, Warner RE, Mankin PC (1992) Accurate counts for Moran joins tests in ecological studies. American Midland Naturalist 128(2):366-376

Goldsborough LG (1994) Heterogeneous spatial distribution of periphytic diatoms on vertical artificial substrata. Journal of the North American Benthological Society 13(2):223-236

Griffith DA (1988) Advanced spatial statistics: Special topics in the exploration of quantitative spatial data series. Kluwer, Dordrecht

Griffith DA (1999) Statistical and mathematical sources of regional science theory: Map pattern analysis as an example. Papers in Regional Science 78(1):21-45

Haining R (1990) Spatial data analysis in the social and environmental sciences. Cambridge University Press, Cambridge

Haining RP (1978) Spatial model for high-plains agriculture. Annals of the Association of American Geographers 68(4):493-504

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

Krishna Iyer PVA (1949) The first and second moments of some probability distributions arising from points on a lattice, and their applications. Biometrika 36(1/2):135-141

Lehman EL (1986) Testing statistical hypothesis. Wiley, New York

Mannelli A, Sotgia S, Patta C, Oggiano A, Carboni A, Cossu P, Laddomada A (1998) Temporal and spatial patterns of African swine fever in Sardinia. Preventive Veterinary Medicine 35(4):297-306

McMillen DP (1992) Probit with spatial autocorrelation. Journal of Regional Science 32(3):335-348

Miller HJ (2004) Tobler's first law and spatial analysis. Annals of the Association of American Geographers 94(2):284-289

Moran PAP (1948) The interpretation of statistical maps. Journal of the Royal Statistical Society Series B (Methodological) 10(2):243-251

Moran PAP (1950) Notes on continuous stochastic phenomena. Biometrika 37(1/2):17-23

Paez A (2006) Exploring contextual variations in land use and transport analysis using a probit model with geographical weights. Journal of Transport Geography 14(3):167-176

Paez A, Scott DM, Volz E (2008) Weight matrices for social influence analysis: An investigation of measurement errors and their effect on model identification and

estimation quality. Social Networks 30(4):309-317

Real LA, McElhany P (1996) Spatial pattern and process in plant-pathogen interactions. Ecology 77(4):1011-1025

Ripley BD (1981) Spatial statistics. Wiley, Hobroken [NJ]

Robertson RD, Nelson GC, De Pinto A (2009) Investigating the predictive capabilities of discrete choice models in the presence of spatial effects. Papers in Regional Science 88(2):367-388

Rohatgi VK (1976) An introduction to probability theory and mathematical statistics. Wiley, New York

Soon SYT (1996) Binomial approximation for dependent indicators. Statistica Sinica 6(3):703-714

Stratton DA, Bennington CC (1996) Measuring spatial variation in natural selection using randomly-sown seeds of Arabidopsis thaliana. Journal of Evolutionary Biology 9(2):215-228

Taam W, Hamada M (1993) Detecting spatial effects from factorial-experiments - An application from integrated-circuit manufacturing. Technometrics 35(2):149-160

Upton G, Fingleton B (1985) Spatial data analysis by example. Wiley, Chichester [NY]

Wang XK, Kockelman KM (2009) Application of the dynamic spatial ordered probit model: Patterns of land development change in Austin, Texas. Papers in Regional Science 88(2):345-365