Spatial Association of Qualitative Data:
A New Test 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

Working Paper
Do not cite without permission from the authors

First Draft:
October 2008

Revisions History:
December 2008 – January 2009
Major revision: Results for irregular point distributions incorporated

Spatial Association of Qualitative Data: A New Test Using Symbolic Dynamics

Abstract. In the present paper, we construct a simple, consistent and powerful test for qualitative spatial independence, called the *TG* test, by using symbolic dynamics and symbolic entropy as a measure of spatial dependence. We also give a standard asymptotic distribution of an affine transformation of the symbolic entropy under the null hypothesis of independence in the spatial qualitative process. Given that the test is based on entropy measures, it avoids smoothed nonparametric estimation. We include numerical experiments to demonstrate the finite sample behaviour of the test, and an empirical application of spatial association of fast food establishments in the city of Toronto, Canada.

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

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 and spatial analysis (Getis 2008). It is chiefly 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), passing through the landmark texts of Anselin (1988) and Griffith (1988), and from there eventually expanding to find applicability 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 regression residuals (see Geary 1954; pp. 115-116 and again p. 144). 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 eyeball 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 numerous situations where nominal variables, as opposed to continuous variables (for which existing statistics would be appropriate), are indeed the focus of the analysis. 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 is 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, Manelli 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 11, 22, 33). In separate research, Epperson and Alvarez-Buylla (1997)

also investigate the spatial structure of nominal variables based on joins for two genotypes. And Arbia et al. (1998) are interested in error propagation when maps are overlaid that contain data collection flaws. In this investigation, join count statistics are used to describe spatial error, and analysis of the variance is deployed to discern how much of the spatial error can be attributed to the measurement process and different components of the map.

As the examples cited above indicate, the hypothesis of spatial independence is important in a number of research endeavours that routinely deal with qualitative data. Beside join-count statistics, however, not much research has been devoted to this class of problems in an exploratory fashion, even if spatial modeling techniques for qualitative data have seen significant progress in recent years (Dubin 1995; McMillen 1992; Paez 2006). The objective of this paper is to propose a new statistic for the exploratory analysis of spatial qualitative/nominal data. The approach proposed to test the hypothesis of independence of spatial qualitative variables is based on principles drawn from the field of symbolic dynamics. Symbolic dynamics have been recently applied to the study of spatial processes, and provide an ideal set of tools for representing discrete processes. We use these tools to propose a statistic defined 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 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 city of Toronto in Canada. The results indicate that in addition to spatial clustering in the locational pattern of events, the classes of establishments also display a particular type of association, with spatial clusters typically displaying heterogeneity (a mixture of various types of establishments) as opposed to homogeneity. In the concluding section, we discuss a number of valuable features of our statistic, and directions for future research.

2. Symbolization of a Spatial Process with Discrete Outcomes

Symbolic dynamics is an approach, developed in the field of mathematics for the study of dynamical systems. More concretely, symbolic dynamics is the practice of modelling a dynamic system by means of a discrete set consisting of (in)finite sequences of abstract symbols obtained for a suitable partition of the state space. The

basic idea behind symbolic dynamics is to divide the phase space into a finite number of regions and label each region 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 should be 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 possible locations of events. 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 "yes"/"no" (k=2), "11"/"22"/"33" if there are three genetic markers (k=3), and so on. In other words, the process is spatially discrete, and the outcomes of the process are discrete as well. A natural way to symbolize such a process is to consider it embedded 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 the polar coordinates (ρ_i^0,θ_i^0) of s_i, taking

as the origin the location of the event at s_0. In our procedure, the m-1 nearest neighbours will be those events satisfying the following two conditions, which 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 events, and there are k possible outcomes for each event, there are k^m distinct combinations of events or possible configurations for an m-surrounding. We will denote each of these unique m-surroundings categories 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 Figure 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 the m-surrounding, this gives a total of 2^6=64 different combinations of eventual outcomes, or symbols (σ_1 through σ_{64}), as listed in Table 1. Please note that a hexagonal array is used only for illustrative purposes, and that the symbolization procedure is equally applicable to other regular, as well as irregular distributions of events.

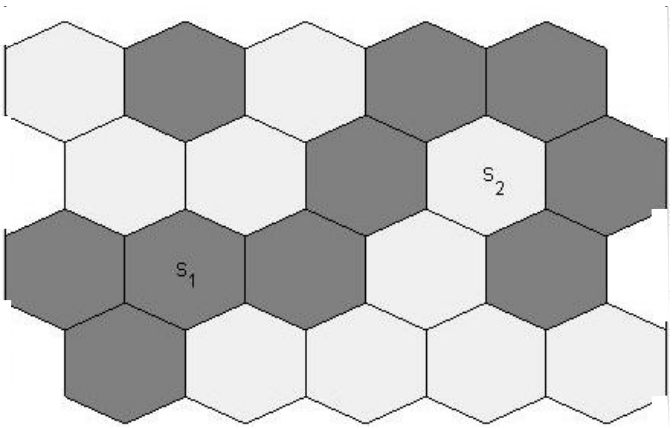

Figure 1. Simple spatial system and process with two types of outcomes.

Since in a hexagonal array 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 Figure 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 $X_m(s_2) = (2,1,1,1,1,2)$.

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)

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.

Once the symbolization of the process has been defined, it is possible to denote the cardinality of the subset of S formed by all the elements of σ_i -type by:

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

The cardinality is simply the number of locations *s* belonging to the set of all possible locations *S*, that are of symbol *i*. Since this cardinality 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:

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

where by $|S|$ we denote the cardinality of the set S (the number of events).

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 generated by the spatial process. Notice that when one symbol, say symbol *i*, tends to dominate the process, $p_{\sigma_i} \to 1$ and $p_{\sigma_j} \to 0$ for all $j \neq i$, which implies that $p_{\sigma_i} \ln(p_{\sigma_i}) \to 1$ and $p_{\sigma_j} \ln(p_{\sigma_j}) \to 0$. Furthermore, when all symbols appear with equal frequency $p_{\sigma_i} = 1/k^m$ for all *i*. The entropy function is then bounded between $0 < h(m) \leq \ln(1/k^m)$, where the lower bound indicates a tendency for only one symbol to occur (i.e. there is a tendency towards perfect regularity in co-location patterns), and the upper bound corresponds to a completely random system (i.i.d. spatial sequence) where all k^m possible symbols appear with identical frequency.

3. Construction of the Independence Test

In this section, we construct a spatial independence test for a discrete spatial process using the machinery defined in Section 2. We also prove that an affine transformation of the symbolic entropy defined in (2) 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:

$$H_0 : \{X_s\}_{s\in S} \quad \text{i.i.d} \tag{5}$$

against any other alternative.

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

$$Z_{\sigma_i s} = \begin{cases} 1 & \text{if } X_m(s) = \sigma_i \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

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:

$$\sum_{i=1}^{n} p_{\sigma_i} = 1 \tag{7}$$

Let us assume that set S is finite and of order R. Then we are interested in knowing how many s's are of σ_i-type for all symbols $\sigma_i \in S$. In order to answer the question, we construct the following variable

$$Y_{\sigma_i} = \sum_{s \in S} Z_{\sigma_i s} \tag{8}$$

The variable Y_{σ_i} can take the values $\{0,1,2,...,R\}$. 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 indicators can be approximated to a Binomial random variable whenever (see Soon 1996):

(1) Dependencies among the indicators are weak; and

(2) The probability of the indicators to occur is small.

Condition (2) is satisfied by the way the symbols have been constructed, since in this case, under the null hypothesis of i.i.d., the probability of success of the indicators $Z_{\sigma s}$ is small ($p_\sigma = 1/k^m$). Condition (1), 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 the lattice 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, one could consider a non-overlapping subset of locations $\tilde{S} \subseteq S$, so that the dependencies among the indicators $Z_{\sigma s}$ are weak for $s \in \tilde{S}$, while keeping in mind that use of a subset of locations will cause loss of information. Indeed, this loss of information will be greater in the measure that set $\tilde{S}$ is smaller, and hence the power of the test will decrease. A reasonable balance therefore must be struck between strongly dependent indicators and too much loss of information. In order to reduce the amount of overlap between indicators, we can take as $\tilde{S}$ those coordinates in S such that for any two coordinates $s, s' \in \tilde{S}$ the sets of nearest neighbors of s and s' have only a small (or even empty) intersection, that is:

$$| N_s \cap N_{s'} | = r \tag{9}$$

for a small enough positive integer r. Integer r is denominated the *degree* of *overlap* of the spatial process $\{X_s\}_{s \in S}$. We now turn to a method to select the set $\tilde{S}$ satisfying the above condition.

Let us define the set $\tilde{S}$ 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 be symbolize. Therefore we have constructed a set of locations

$$\tilde{S} = \{\tilde{s}_0, \tilde{s}_1, ..., \tilde{s}_t\} \tag{10}$$

such that the variable $Y_{\sigma_i} = \sum_{\tilde{s} \in \tilde{S}} Z_{\sigma_i \tilde{s}}$ can be approximated to a binomial distribution for a suitable choice of r (r small enough).

Notice that the maximum number of locations that can be symbolized with an overlapping degree r is $t = \left[\frac{R-m}{m-r}\right] + 1$, where the operator $[x]$ denotes the integer part of

a real number x. Also a different selection of $\tilde{s}_0$ gives a different set $\tilde{S}$, but under the null of i.i.d. the distribution of the random variable $Y_{\sigma_i} = \sum_{\tilde{s} \in \tilde{S}} Z_{\sigma_i \tilde{s}}$ for all the symbols $\sigma_i \in \Gamma$. In the particular case of a regular lattice it is possible to take $S = \tilde{S}$, that is an overlapping degree $r = 0$, because the dependencies introduced in this case are weak and balanced through the lattice.

Given the above considerations, we can now finally state the following result which proof can be found in the Appendix.

Theorem 1 *Let* $\{X_s\}_{s \in S}$ *be a 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. *Denote by* $h(m)$ *the symbolic entropy defined in (2) for a fixed embedding dimension* $m \geq 2$, *with* $m \in N$. *If the spatial process* $\{X_s\}_{s \in S}$ *is i.i.d., then the affine transformation of the symbolic entropy*

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

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

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

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

Then, to test

$$H_0 : \{X_s\}_{s \in S} \quad \text{i.i.d.} \tag{13}$$

the decision rule in the application of the $TG(m)$ test at a $100(1-\alpha)\%$ confidence level is:

$$\begin{cases} \text{If } TG(m) > \chi^2_\alpha \text{then reject } H_0 \\ \text{Otherwise do not reject } H_0 \end{cases} \tag{14}$$

4. Properties of the *TG*(*m*) Test

Next, we prove that the $TG(m)$ 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* $\leq m$ 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 $\widehat{TG}(m)$ the estimator of $TG(m)$. The proof of the following theorem can be found in Appendix.

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

$$\lim_{R\to\infty} \Pr(\widehat{TG}(m) > C) = 1 \tag{15}$$

under spatial dependence of order $\leq m$ *for all* $0 < C < \infty, C \in R$.

Thus, the test based on $TG(m)$ is consistent against all spatial dependence of order $\leq m$ alternatives to the null of spatial independence. Since Theorem 1 implies $TG(m) \to +\infty$ with probability approaching 1 under spatial dependence of order $\leq m$, 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 $TG(m)$ statistic. While this affords some flexibility, there are also some conditions that must be observed in order to guide a decision. Note that R should be larger than the number of symbols (k^m) in order to have at least the same number of *m*-surroundings as possible symbols (events) $\sigma_i, i = 1, ..., k^m$. When the χ^2 distribution is applied in practice, and all the expected frequencies are 5, 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 datasets containing *at least* $5k^m$ observations. Alternatively, this can be stated as constraining the size of m so that it is at most the nearest integer $\leq \log(N/5)/\log(k)$.

As indicated before, the possible dependence detected by the $TG(m)$ test has to be of order $\leq m$. This is due to the fact that, if the dependence structure of the process is of order $> m$, then this dependence is not present in every m-surrounding and therefore the symbols may not capture it.

5. Finite Sample Behavior of *TG*(*m*)

In this section, we examine the finite sample behavior of the $TG(m)$ test. It is worth making some considerations about the data generating process that we will use for the numerical 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:

$$Y = (I - \rho W)^{-1}\varepsilon \tag{16}$$

In equation (16) I is the $N \times N$ identity matrix, ρ is a parameter of spatial dependence, and W is a connectivity matrix that determines the set of spatial relationships among points. Secondly, we define the discrete spatial process as follows. Let b_{ij} be defined by:

$$p(Y \le b_{ij}) = \frac{i}{j} \tag{17}$$

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

$$X_s = \begin{cases} a_1 & \text{if } Y_s \le b_{1k} \\ a_i & \text{if } b_{i-1k} < Y_s \le b_{ik} \\ a_k & \text{if } Y_s > b_{k-1k} \end{cases} \tag{18}$$

The last item that needs to be determined before the data can be generated is a specific arrangement of spatial events, so that matrix *W* can be defined. In this regards, we note that Farber et al. (2009) indicate 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*=900 and 3600. In addition to these regular tessellations, we simulate two irregular, but not random, distributions of events with *N*=900 and 3600. The data are generated using equation (16), with the connectivity matrix defined in terms of first-order contiguity (which is based on Voronoi tessellations for the irregular point distributions), and standard normal random numbers. Matrix *W* is row-standardized for the calculations.

Figure 2 shows examples the different spatial distributions of *N*=900 events

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

- Autocorrelation parameter ρ: 0.0, 0.1, 0.3, 0.5, 0.7, and 0.9
- Sample size *N*: 900 and 3600
- Number of outcomes *k*: 2, 3, and 4
- Size of *m*-surrounding: 4 (self + 3 neighbours), 5 (self + 4 neighbours), 6 (self + 5 neighbours)
- Degree of overlap *r*: 1, 2,..., *m*-1, for a given *m*

Data is simulated 100 times for each combination of parameters (number of replications), the test was applied to each generated datasets at a 0.05 level of significance, and the number of times that the probability value of the statistic exceeded 0.05 was recorded. Following the decision rule posed in equation (14), this would indicate rejection of the null hypothesis. We would expect the statistic to reject the null hypothesis more frequently as the level of autocorrelation goes up. At the same time, we would expect it to fail to reject the null hypothesis most of the time when the level of autocorrelation is zero.

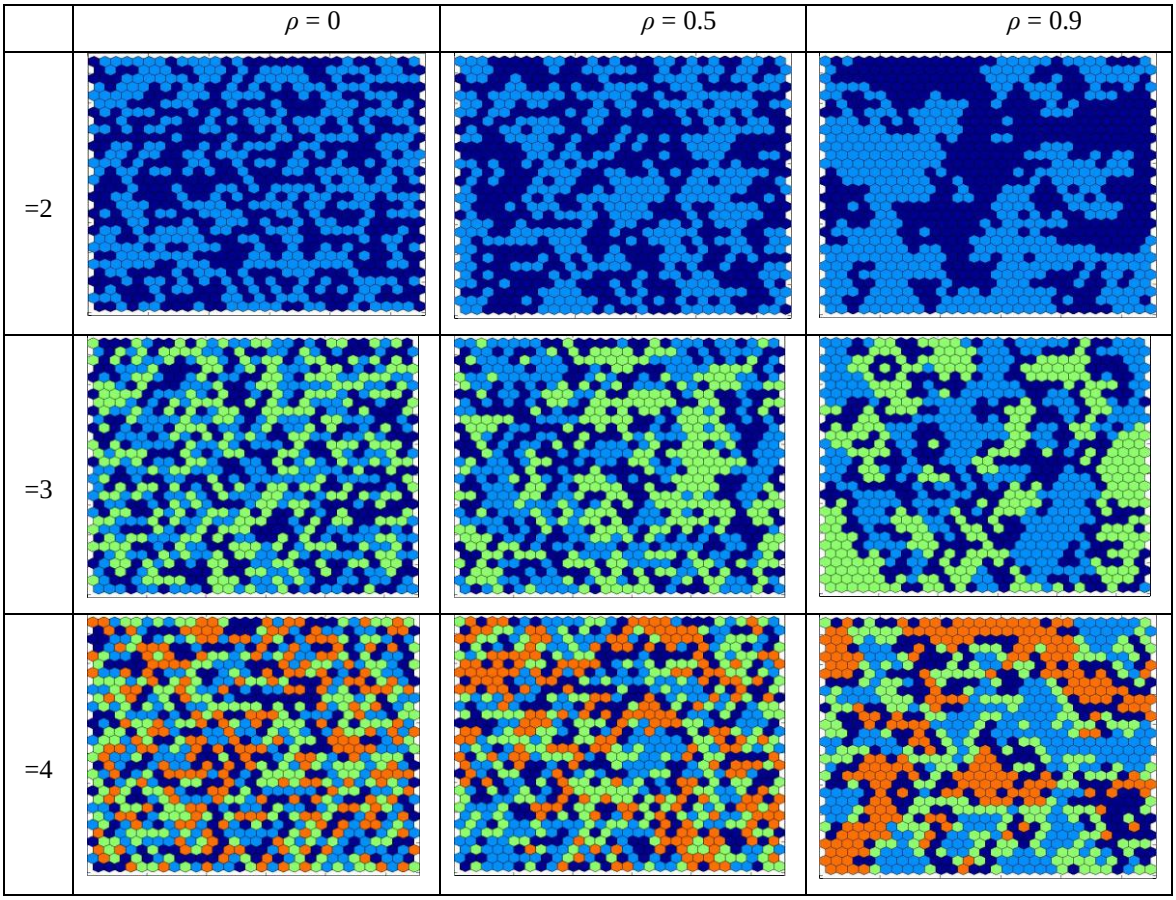

Figure 2. Examples of distributions of events on a regular hexagonal lattice (N=900), for different number of outcomes k and levels of ρ.

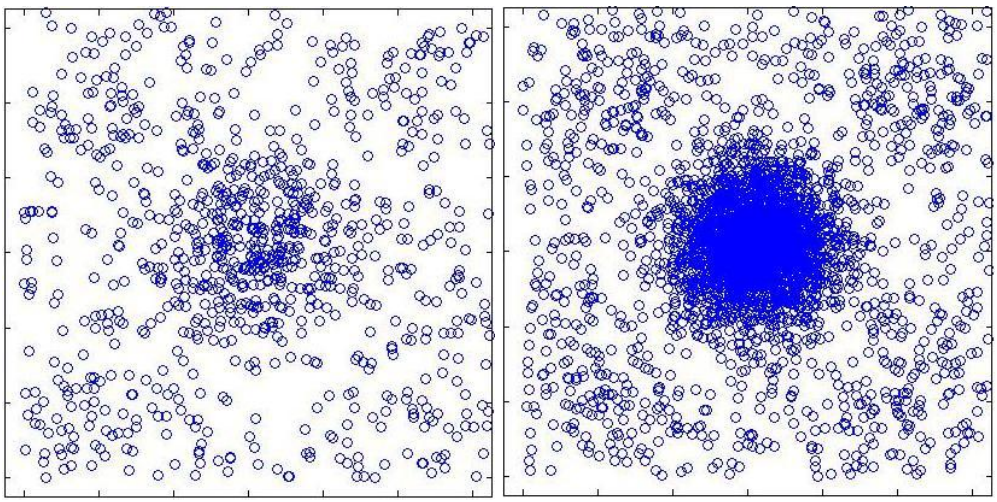

Figure 3. Irregular distributions of evens N=900 and N=3600

The results of the numerical experiments appear in Tables 1 and 2, for the regular and irregular settings respectively. Please note that combinations of parameters are selected that satisfy the rule that the sample must contain *at least* $5k^m$ observations. For this reason, while it is possible to explore k=3 and m=4 when the number of points is N=900, an m-surrounding of 5 when k=3 would necessitate at least N=1215 observations. We conducted additional simulations for this and other cases that do not meet the number of observations criterion (not reported, but available upon request), which only confirm that the power of the statistic is exceedingly low when this condition is not satisfied.

As is usually the case, the power of the statistic increases with the number of observations. When N=900, the statistic demonstrates high power (>0.90) for moderate to high levels of autocorrelation, however depending on the various combinations of parameters. When N=3600, the statistic achieves high levels of power more rapidly, being able to identify spatial association in processes with moderate levels of autocorrelation (ρ=0.3) around 90% of the time at the 0.05 significance level. As anticipated, the power of the statistic suffers in the measure that the degree of overlap becomes more restrictive. Higher degrees of overlap, excluding the case of perfectly overlapping locations (r=m), generally lead to more powerful performance of the statistic.

In a majority of cases examined the risk of false positives is relatively low (<0.10), but there are a few combinations of parameters that result in higher levels of false positives (>0.10). Although no clear-cut pattern emerges, these situations tend to be such that the overlapping degree is either low or high relative to the size of the m-surrounding, which suggests a golden mean in terms of the degree of overlap for which the risk of false positives is better controlled.

It is interesting to remark that the results, both in terms of power to detect spatial association as well as resistance to false positives, are noticeably better for the case where events are irregularly distributed compared to the regular tessellation, 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 1. Size and Power of the TG-test for regular lattices

			N=900					
			ρ=0.0	ρ=0.1	ρ=0.3	ρ=0.5	ρ=0.7	ρ=0.9
k=2	m=4	r=1	0.00	0.00	0.16	0.64	1.00	1.00
		r=2	0.05	0.03	0.23	0.76	1.00	1.00
		r=3	0.17	0.10	0.34	0.94	1.00	1.00
k=2	m=5	r=1	0.06	0.08	0.16	0.50	0.98	1.00
		r=2	0.03	0.07	0.18	0.65	1.00	1.00
		r=3	0.11	0.04	0.19	0.83	1.00	1.00
		r=4	0.13	0.15	0.46	0.97	1.00	1.00
k=2	m=6	r=1	0.11	0.14	0.19	0.51	0.96	1.00
		r=2	0.09	0.16	0.22	0.59	0.99	1.00
		r=3	0.09	0.13	0.21	0.64	1.00	1.00
		r=4	0.08	0.10	0.21	0.80	1.00	1.00
		r=5	0.14	0.18	0.44	0.95	1.00	1.00
k=3	m=4	r=1	0.11	0.09	0.20	0.58	1.00	1.00
		r=2	0.09	0.04	0.18	0.77	1.00	1.00
		r=3	0.10	0.07	0.28	0.94	1.00	1.00
			N=3600					
			ρ=0.0	ρ=0.1	ρ=0.3	ρ=0.5	ρ=0.7	ρ=0.9
k=2	m=4	r=1	0.04	0.04	0.56	1.00	1.00	1.00
		r=2	0.03	0.09	0.69	1.00	1.00	1.00
		r=3	0.15	0.19	0.85	1.00	1.00	1.00
k=2	m=5	r=1	0.03	0.07	0.40	1.00	1.00	1.00
		r=2	0.06	0.03	0.53	1.00	1.00	1.00
		r=3	0.02	0.09	0.65	1.00	1.00	1.00
		r=4	0.08	0.20	0.97	1.00	1.00	1.00
k=2	m=6	r=1	0.03	0.06	0.38	0.98	1.00	1.00
		r=2	0.04	0.11	0.33	0.99	1.00	1.00
		r=3	0.07	0.06	0.50	1.00	1.00	1.00
		r=4	0.03	0.08	0.66	1.00	1.00	1.00
		r=5	0.11	0.19	0.95	1.00	1.00	1.00
k=3	m=4	r=1	0.02	0.06	0.44	1.00	1.00	1.00
		r=2	0.05	0.05	0.62	1.00	1.00	1.00
		r=3	0.08	0.11	0.86	1.00	1.00	1.00
k=3	m=5	r=1	0.18	0.25	0.53	0.98	1.00	1.00
		r=2	0.08	0.09	0.43	1.00	1.00	1.00
		r=3	0.09	0.11	0.61	1.00	1.00	1.00
		r=4	0.12	0.18	0.93	1.00	1.00	1.00
k=4	m=4	r=1	0.11	0.15	0.39	1.00	1.00	1.00
		r=2	0.07	0.05	0.56	1.00	1.00	1.00
		r=3	0.07	0.13	0.85	1.00	1.00	1.00

Table 2. Size and Power of the TG-test for irregular lattices

			N=900					
			ρ=0.0	ρ=0.1	ρ=0.3	ρ=0.5	ρ=0.7	ρ=0.9
k=2	m=4	r=1	0.05	0.01	0.10	0.69	1.00	1.00
		r=2	0.03	0.05	0.27	0.96	1.00	1.00
		r=3	0.06	0.10	0.56	0.99	1.00	1.00
k=2	m=5	r=1	0.05	0.04	0.11	0.61	1.00	1.00
		r=2	0.05	0.03	0.16	0.67	1.00	1.00
		r=3	0.03	0.06	0.30	0.90	1.00	1.00
		r=4	0.09	0.11	0.55	0.97	1.00	1.00
k=2	m=6	r=1	0.11	0.10	0.24	0.59	1.00	1.00
		r=2	0.14	0.05	0.20	0.58	1.00	1.00
		r=3	0.07	0.02	0.21	0.64	1.00	1.00
		r=4	0.07	0.11	0.35	0.87	1.00	1.00
		r=5	0.10	0.13	0.55	0.96	1.00	1.00
k=3	m=4	r=1	0.09	0.10	0.25	0.79	1.00	1.00
		r=2	0.07	0.10	0.33	0.88	1.00	1.00
		r=3	0.05	0.12	0.51	0.98	1.00	1.00
			N=3600					
			ρ=0.0	ρ=0.1	ρ=0.3	ρ=0.5	ρ=0.7	ρ=0.9
k=2	m=4	r=1	0.05	0.05	0.84	1.00	1.00	1.00
		r=2	0.02	0.12	0.91	1.00	1.00	1.00
		r=3	0.03	0.16	1.00	1.00	1.00	1.00
k=2	m=5	r=1	0.05	0.03	0.64	1.00	1.00	1.00
		r=2	0.05	0.05	0.78	1.00	1.00	1.00
		r=3	0.06	0.13	0.94	1.00	1.00	1.00
		r=4	0.08	0.16	0.99	1.00	1.00	1.00
k=2	m=6	r=1	0.05	0.08	0.47	1.00	1.00	1.00
		r=2	0.09	0.08	0.57	1.00	1.00	1.00
		r=3	0.07	0.03	0.73	1.00	1.00	1.00
		r=4	0.06	0.16	0.88	1.00	1.00	1.00
		r=5	0.11	0.17	0.99	1.00	1.00	1.00
k=3	m=4	r=1	0.08	0.05	0.62	1.00	1.00	1.00
		r=2	0.05	0.06	0.86	1.00	1.00	1.00
		r=3	0.05	0.14	0.98	1.00	1.00	1.00
k=3	m=5	r=1	0.16	0.29	0.64	1.00	1.00	1.00
		r=2	0.18	0.21	0.70	1.00	1.00	1.00
		r=3	0.08	0.15	0.82	1.00	1.00	1.00
		r=4	0.08	0.15	0.98	1.00	1.00	1.00
k=4	m=4	r=1	0.09	0.12	0.65	1.00	1.00	1.00
		r=2	0.08	0.10	0.73	1.00	1.00	1.00
		r=3	0.07	0.14	0.99	1.00	1.00	1.00

6. Illustration: Fast food establishments in Toronto

In this section we illustrate the use of the new *TG* 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 consider the possibility of spatial clustering, which would indicate the existence of economies of agglomeration, and of co-locations patterns of three types of fast food establishments, to explore the question of whether establishments tend to attract similar establishments, or repel them.

6.1 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.

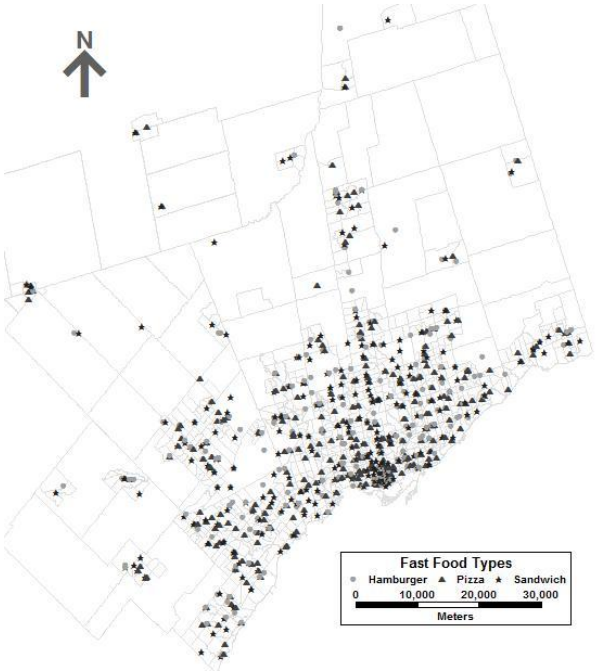

Figure 4. Fast food establishments in Toronto.

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 data

points is extracted from the file corresponding to 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 Figure 4.

6.2 Analysis and Results

The first step in our analysis of fast food places in Toronto is to verify the eyeball appraisal of clustering 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:

$$G(d) = \frac{\#\left(\min\left(d_i\right) \le d\right)}{N} \tag{19}$$

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 Figure 5, 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.

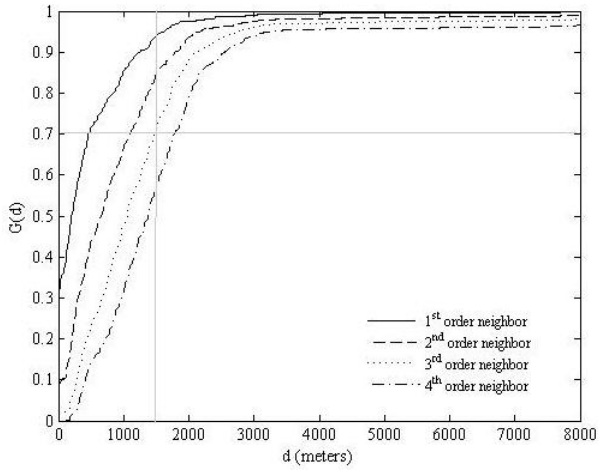

Figure 5. Event-to-event k[th] nearest neighbour distance analysis

We now turn to the question of whether there are patterns of co-location within clusters. 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 $\ln(N/5)/\ln(k) = 4.7033$, meaning that we can explore m-surroundings of size 2 (self and one nearest neighbour), 3 (self and two nearest neighbours) and 4 (self and three nearest neighbours). On the other hand, we are prevented by the sample size to explore m-surroundings of size 5 or larger. Based on our previous application of the G function, there appears to be only relatively a minor difference, in terms of the spatiality of clusters, between m=3 or 4. Since a property of the statistic is that it detects spatial dependence of order $\leq m$, it seems sensible to select m=4 for our analysis. If dependence is detected, it would carry for the cases of m=3 and 2. One additional decision to make concerns the degree of overlap. Seeing as the simulations for an irregular distribution of points with similar number of events as our case study suggest that an overlap of r=3 is both more powerful and more resistant to false positives (see Table 2), we select this value for our application. The value of the statistic calculated using these parameters is $TG(m=4,r=3)=167.9564$. This value is 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 statistic exceeds, and therefore, according to our decision rule, leads to rejection of the hypothesis of independence. Alternatively, the probability value of the TG statistic is 3.3541e-008.

The test rejects the hypothesis of independence. However, dependence could take different forms for different patterns of co-location. An attractive feature of the $TG(m)$ 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 877 points, each symbol would appear approximately 11 times. It is possible to plot a histogram with the actual frequency of the 81 symbols (see Figure 5). 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 directionality from s_0.

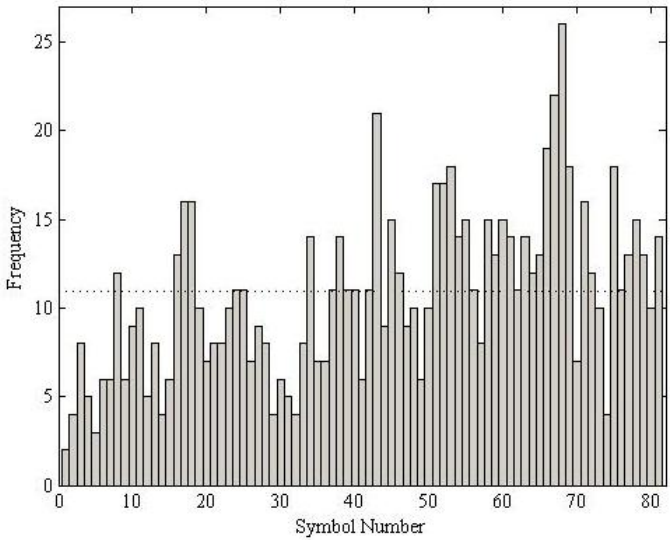

Figure 5. Frequency of fast food type co-location in Toronto (*m*=4 and *r*=2)

In Figure 6, 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. This allows us to discern that four establishments of a kind (4 pizza, 4 sandwich, or 4 hamburger places) seldom co-locate. Much more common is the case where spatial clusters will 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 – that within clusters there is a pattern of competition or repulsion between establishments of the same type. There is a slightly higher frequency of two pizza establishments in a cluster, although this may be due to the fact that pizza establishments are more numerous than other types of establishments, and we would not attach much weight to the relatively small difference seen in the histogram between (HPPS), (HHPS), and (HPSS).

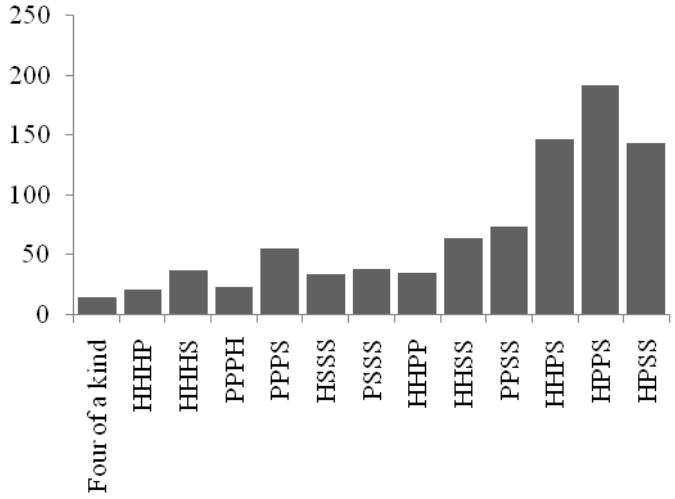

Figure 6. Co-location of events, condensed histogram

As a final experiment, and to verify the robustness of these findings, we also randomly re-label the process. Random re-labelling works by reassigning the categories of events to locations, sampling randomly and without replacements from the list of actual categories. This is done 10000 times and the statistic is calculated and recorded every time. Figure 7 shows the distribution of the statistic after random re-labelling. Clearly, with a value of 167.9564 the statistic for the actual distribution of events is extremely infrequent.

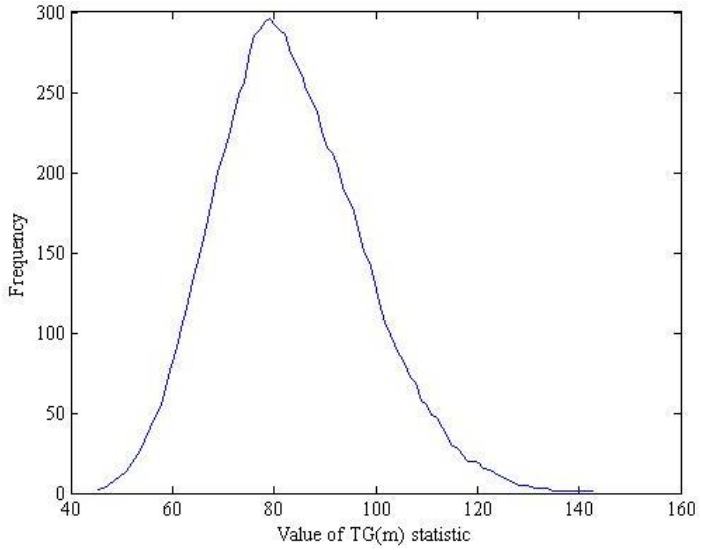

Figure 7. Smoothed histogram of statistics after randomly re-labelling series

7. Conclusions

In this paper, we have proposed a new statistic (*TG*) useful to test the hypothesis of independence among spatially distributed qualitative data. Qualitative data are receiving increased attention from a number of disciplinary perspectives, but the spatial analysis of qualitative data has tended to lag behind the development of methods and techniques useful to study continuous variables. Our statistic therefore comes to fill a research gap, in terms of providing a new tool to further enrich the diversity of the spatial analysis toolbox.

The statistic proposed 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 limited (qualitative) 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

power of the statistic to identify spatial association under a range of different conditions. In addition, and example illustrates, in combination with other spatial analytical techniques, 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.

There are a number of additional points that we suggest as requiring additional research. First, the statistic depends to some still unknown degree on the condition that k events (but not their spatial arrangement, which is captured by the symbols) appear with more or less similar frequencies. This currently limits the ability of the statistic to deal with situations where one class of events is relatively rare compared to others. Furthermore, the number of outcomes k typically depends on the nature of the process, and is beyond the control of the analyst. The number of observations needed to conduct analysis can quickly explode. For example, if k=4, and one desires to examine m-surroundings of size 3, would require at least 625 points. In contrast, an m-surrounding of 5 would require 3125 observations. A topic for further investigation is whether different symbolization schemes can help to maintain data needs under control. A final idea would be to test, using the histogram of frequency of symbols whether each symbol departs significantly from the expected frequency. 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).

8. 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:

$$P\left(Y_{\sigma_1} = a_1, Y_{\sigma_2} = a_2, ..., Y_{\sigma_{k^m}} = a_{k^m}\right) = \frac{\left(a_1 + a_2 + ... + a_{k^m}\right)!}{a_1! a_2! \cdot ... \cdot a_{k^m}!} p_{\sigma_1}^{a_1} p_{\sigma_2}^{a_2} \cdots p_{\sigma_{k^m}}^{a_{k^m}} \tag{20}$$

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 (20) is:

$$L(p_{\sigma_1}, p_{\sigma_2}, ..., p_{\sigma_{k^m}}) = \frac{R!}{n_{\sigma_1}! n_{\sigma_2}! \cdot ... \cdot n_{\sigma_{k^m}}!} p_{\sigma_1}^{n_{\sigma_1}} p_{\sigma_2}^{n_{\sigma_2}} \cdots p_{\sigma_{k^m}}^{n_{\sigma_{k^m}}} \tag{21}$$

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

$$L\left(p_{\sigma_1}, p_{\sigma_2}, ..., p_{\sigma_{k^m}}\right) = \frac{R!}{n_{\sigma_1}! n_{\sigma_2}! \cdot ... \cdot n_{\sigma_{k^m}}!} p_{\sigma_1}^{n_{\sigma_1}} p_{\sigma_2}^{n_{\sigma_2}} \cdots \left(1 - p_{\sigma_1} - p_{\sigma_2} - ... - p_{\sigma_{k^m - 1}}\right)^{n_{\sigma_{k^m}}} \tag{22}$$

Then the logarithm of this likelihood function remains as

$$\begin{aligned} \ln\left[L\left(p_{\sigma_1}, p_{\sigma_2}, ..., p_{\sigma_{k^m}}\right)\right] = \ln\left(\frac{R!}{n_{\sigma_1}! n_{\sigma_2}! \cdot ... \cdot n_{\sigma_{k^m}}!}\right) + \sum_{i=i}^{k^m - 1} n_{\sigma_i} \ln\left(p_{\sigma_i}\right) \\ + n_{\sigma_{k^m}} \ln\left(1 - p_{\sigma_1} - p_{\sigma_2} - ... - p_{\sigma_{k^m - 1}}\right) \end{aligned} \tag{23}$$

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

$$\frac{\partial}{\partial p_{\sigma_i}} \ln\left[L\left(p_{\sigma_1}, p_{\sigma_2}, ..., p_{\sigma_n}\right)\right] = 0 \tag{24}$$

to get that:

$$\hat{p}_{\sigma_i} = \frac{n_{\sigma_i}}{R} \tag{25}$$

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

$$\lambda(Y) = \frac{\frac{R!}{n_{\sigma_1}!n_{\sigma_2}!\cdots n_{\sigma_{k^m}}!} p_{\sigma_1}^{n_{\sigma_1}} p_{\sigma_2}^{n_{\sigma_2}} \cdots p_{\sigma_{k^m}}^{n_{\sigma_{k^m}}}}{\frac{R}{n_{\sigma_1}!n_{\sigma_2}!\cdots n_{\sigma_{k^m}}!} \hat{p}_{\sigma_1}^{n_{\sigma_1}} \hat{p}_{\sigma_2}^{n_{\sigma_2}} \cdots \hat{p}_{\sigma_{k^m}}^{n_{\sigma_{k^m}}}} = \frac{\prod_{i=1}^{k^m} p_{\sigma_i}^{n_{\sigma_i}}}{\prod_{i=1}^{k^m} \left(\frac{n_{\sigma_i}}{R}\right)^{n_{\sigma_i}}} = \tag{26}$$

$$R^{\sum_{i=1}^{k^m} n_{\sigma_i}} \prod_{i=1}^{k^m} \left(\frac{p_{\sigma_i}}{n_{\sigma_i}}\right)^{n_{\sigma_i}} = R^R \prod_{i=1}^{k^m} \left(\frac{p_{\sigma_i}}{n_{\sigma_i}}\right)^{n_{\sigma_i}}$$

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

$$TG(m) = -2\ln\left(\lambda(Y)\right) = -2\left[R\ln(R) + \sum_{i=1}^{k^m} n_{\sigma_i} \ln\left(\frac{p_{\sigma_i}}{n_{\sigma_i}}\right)\right] \sim \chi^2_{k^m-1} \tag{27}$$

Now, under the null hypothesis, all the symbols have the same probability of occurring, $p_{\sigma_i} = \frac{1}{k^m}$ for all $i = 1, 2, ..., k^m$, then it follows that

$$\begin{aligned} TG(m) &= -2R\left[\ln(R) + \sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{R} \ln\left(\frac{p_{\sigma_i}}{n_{\sigma_i}}\right)\right] \\ &= -2R\left[\ln(R) + \sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{R}\left(\ln\left(\frac{1}{k^m}\right) - \ln\left(n_{\sigma_i}\right)\right)\right] \\ &= -2R\left[\ln(R) + \sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{R}\left(\ln\left(\frac{1}{k^m}\right) - \ln\left(\frac{n_{\sigma_i}}{R}\right) - \ln(R)\right)\right] \end{aligned} \tag{28}$$

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

$$TG(m) = -2R\left[\ln\left(\frac{1}{k^m}\right) + h(m)\right] = -2R\left[h(m) - \ln\left(k^m\right)\right] = 2R\left[\ln\left(k^m\right) - h(m)\right] \tag{29}$$

and the proof finishes. □

Proof of Theorem 2

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

$$p\lim_{R\to\infty} \hat{h}(m) = h((m) \tag{30}$$

Recall that $TG(m) = 2R\left[\ln\left(k^m\right) - h(m)\right]$ and $0 \le h(m) \le \ln\left(k^m\right)$. Denote by $H(m) = \ln\left(k^m\right) - h(m)$, so $TG(m) = 2RH(m)$. Then by (21) it follows that

$$p\lim_{R\to\infty} \widehat{H}(m) = H(m) \tag{31}$$

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

$$\frac{C}{2R} < H(m) \tag{32}$$

Then, under the spatial dependence of order $\le m$ it follows that $H(m) \neq 0$ and, thus,

$$\begin{aligned} &\Pr\left[\widehat{SG}(m) > C\right] = \Pr\left[2R\widehat{H}(m) > C\right] \\ &= \Pr\left[2R\left(\widehat{H}(m) - H(m)\right) > C - 2RH(m)\right] \\ &= \Pr\left[2R\left(H(m) - \widehat{H}(m)\right) < 2RH(m) - C\right] \\ &= \Pr\left[H(m) - \widehat{H}(m) < H(m) - \frac{C}{2R}\right] \end{aligned} \tag{33}$$

Therefore, by (22), (23) and (24) we have that:

$$\lim_{R\to\infty} \Pr(\widehat{TG}(m) > C) = 1 \tag{34}$$

as desired. □

Acknowledgments

References

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

Anselin L (1995) Local Indicators of Spatial Association - LISA. Geographical

Analysis 27:93-115

Arbia G, riffith D, aining R (1998) Error propagation modelling in raster GIS: overlay operations. International Journal of Geographical Information Science 12:145-167

Austin SB, Melly SJ, Sanchez BN et al (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 95:1575-1581

Bailey TC, Gatrell AC (1995) Interactive Spatial Data Analysis. Addison Wesley Longman, Essex

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 37:357-369

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

Cliff AD, Ord JK (1981) Spatial Processes: Models and Applications. Pion, London

Dacey MF (1968) A Review on Measures of Contiguity for Two and k-Color Maps. In: Berry, B. J. L. and MArble, D. F. (ed) Spatial Analysis: A Reader in Statistical Geography. Prentice Hall, Englewood Cliffs, NJ, 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 101:133-137

Dubin R (1995) Estimating Logit Models with Spatial Dependence. In: Anselin, L. and Florax, R. J. G. M. (ed) New Directions in Spatial Econometrics. Springer-Verlag, Berlin, 229-242

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

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

Geary RC (1954) The Contiguity Ratio and Statistical Mapping. The Incorporated Statistician 5:115-145

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

Getis A, Ord JK (1993) The Analysis of Spatial Association by Use of Distance Statistics. Geographical Analysis 25:276-276

Ghent AW, Warner RE, Mankin PC (1992) Accurate Counts for Moran Joins Tests in Ecological-Studies. American Midland Naturalist 128:366-376

Goldsborough LG (1994) Heterogeneous Spatial Distribution of Periphytic Diatoms on Vertical Artificial Substrata. Journal of the North American Benthological Society 13:223-236

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 78:21-45

Haining RP (1978) Spatial Model for High-Plains Agriculture. Annals of the Association of American Geographers 68:493-504

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 36:135-141

Lehman EL (1986) Testing Statistical Hypothesis. John Wiley and Sons, New York

McMillen DP (1992) Probit with Spatial Autocorrelation. Journal of Regional Science 32:335-348

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

Moran PAP (1948) The Interpretation of Statistical Maps. Journal of the Royal Statistical Society Series B (Methodological) 10:243-251

Moran PAP (1950) Notes on Continuous Stochastic Phenomena. Biometrika 37:17-23

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

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 30:309-317

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

Soon SYT (1996) Binomial Approximation for Dependent Indicators. Statistica Sinica 6:703-714

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

Taam W, Hamada M (1993) Detecting Spatial Effects from Factorial-Experiments - An Application from Integrated-Circuit Manufacturing. Technometrics 35:149-160