A note on the $SG(m)$ Test

Fernando A. López Mariano Matilla-García Jesús Mur Antonio Paez
Manuel Ruiz Marín.

Abstract

López *et al.* (2010) introduce a nonparametric test of spatial dependence, called $SG(m)$. The test is claimed to be consistent and asymptotically chi-squared distributed. Elsinger (2013) raises doubts about the two properties. Using a particular counterexample, he shows that the asymptotic distribution of the $SG(m)$ test may be far from the chi-square family; the property of consistency is also questioned. In this note the authors want to clarify the properties of the $SG(m)$ test. We argue that the cause of the conflict is in the specification of the symbolization map. The discrepancies can be solved by adjusting some of the definitions made in the original paper. Moreover, we introduce a permutational bootstrapped version of the $SG(m)$ test, which is powerful and robust to the underlying statistical assumptions. This bootstrapped version may be very useful in an applied context.

Keywords: $SG(m)$ Test; Spatial Dependence; Nonparametric Tests; Symbolic Dynamics; Entropy; Bootstrap

JEL-Classification: C4

1 Introduction

In López *et al.* (2010) a new nonparametric test for spatial independence, called $SG(m)$, is introduced using symbolic entropy. The idea is simple: spatial dependence brings order to the data, which means that the entropy should decrease as dependence in the spatial data increases. The test measures how much order exists in the series, in relation to the case of randomness.

A crucial point in relation to the $SG(m)$ test is the symbolization of the series, which depends on the test designer. The symbolization map in López *et al.* (2010) is simple and efficient, but other alternatives are also possible. It is worth remembering that simulations in this study reveal a balanced empirical size and considerable power against different types of spatial dependence processes.

Two points of concern, in relation to the symbolization process, were discussed by López *et al.* (2010) and Ruiz *et al.* (2010), namely (i)- the overlapping of the m-surroundings, which results in a problem of dependent indicators and (ii)- the symbolization map should be standard, implying the *i.i.d.* of the symbols distribution under the null of independence for the series.

Elsinger (2013) raises doubts in relation to the fundaments of the $SG(m)$ test: the asymptotic distribution, for the case of overlapping m-surroundings, is not standard and the test is not consistent. The purpose of this paper is to cast light on this discussion and to clarify the use and interpretation of the $SG(m)$ test.

Section 2 slightly reformulates the concept of a *standard* symbolization map. Section 3 focuses on the asymptotic distribution of the $SG(m)$ statistic, with a brief remark on the result of consistency. Section 4 presents a small Monte Carlo experiment for a bootstrapped version of the test. Section 5 offers our conclusions.

2 The *standard* symbolization map

In the same vein as in López *et al.* (2010) and Elsinger (2013), we define $\{X_s\}_{s\in S}$ as a real stochastic process with a spatial domain. Let us introduce a non-empty, finite set of n symbols denoted by $\Gamma = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$. Let $m \in \mathbb{N}$ with $m \geq 2$. For $s_0 \in S$, define $Z_m(s_0)$ to be the $m-$vector

$$Z_m(s_0) = (X_{s_0}, X_{s_1}, \ldots, X_{s_{m-1}})$$

where $s_1, s_2, \ldots, s_{m-1}$ are $m-1$ neighbors to s_0. Symbolizing a process requires defining a map

$$f : \{X_s\}_{s\in\widehat{S}} \longrightarrow \Gamma \tag{1}$$

such that each element X_s is associated to a unique symbol, $f(X_s) = \sigma$ where $\widehat{S}$ is a subset of S (eventually $\widehat{S} = S$). The proposed symbolization procedure given in López et al. (2010) was given in terms of m-surroundings $Z_m(s)$, more concretely

$$\begin{array}{ccccc} f : \{X_s\}_{s\in\widehat{S}} & \rightarrow & \mathbb{R}^m & \rightarrow & \Gamma \\ X_s & \hookrightarrow & Z_m(s) & \mapsto & f(X_s) = \sigma \end{array}$$

The definition of a symbolization map is a prerequisite for inference based on symbolic entropy. Indeed this is a crucial decision for a proper functioning of the procedure, as will be shown later.

López *et al.* (2010, p.106) introduce the concept of *standard* symbolization map:

Definition 1. *[López et al. (2010)] Let f be a symbolization map based on m-surroundings. If under the null of independence of the spatial process $\{X_s\}_{s\in S}$, all the symbols are equally probable, then f is called a standard symbolization map.*

The property of equiprobability is used in the main theorem in López *et al.* (2010, p.109).

Theorem 2.1. *[López et al. (2010)] Let $\{X_s\}_{s\in S}$ be a real-valued spatial process with $|S| = R$. Assume that there exists a standard symbolization map f for $\{X_s\}_{s\in S}$. Denote by $h(m)$ the symbolic entropy for a fixed embedding dimension $m \geq 2$, with $m \in \mathbb{N}$. If the spatial process $\{X_s\}_{s\in S}$ is independent, then the affine transformation of the symbolic entropy*

$$SG(m) = 2R[Ln(n) - h(m)] \tag{2}$$

is asymptotically χ^2_k distributed.

k refers to the number of unknown parameters under the alternative hypothesis minus the number of unknown parameters under the null hypothesis.

A requisite in the proof of Theorem 2.1 is that the absolute frequency of the symbols (the so-called Y_σ variables) follow a binomial distribution, for every $\sigma \in \Gamma$. However, a *standard* symbolization map, as in Definition 1, does not guarantee this result because of the possible overlappings of the m-surroundings that introduce dependence in the distribution of the symbols.

López *et al.* (2010) warned of this problem (see footnote 4, p.108). Later, Ruiz *et al.* (2010) proposed a solution aimed at controlling the degree of overlapping, in which case both the binomial and the chi-squared distributions can be maintained as fair approximations. Elsinger (2013) continues in the same vein through a counterexample, which includes a peculiar symbolization map.

Its overlapping degree is so high (50% in this case) that prevents the Y_σ variables to attain a binomial distribution; therefore, even asymptotically the $SG(m)$ statistic is far from the chi-squared distribution.

This is a problem of terminology that can be solved by introducing an additional condition in the definition of *standard* symbolization map, as follows:

Definition 2. *Let $f : \{X_s\}_{s\in S} \to \Gamma$ be a symbolization map. If under the null of independence of the spatial process $\{X_s\}_{s\in S}$ the symbols are equidistributed and the corresponding indicator variables are independent in s, then f is called a standard symbolization map.*

Notice that with our new Definition 2 we have that if $f : \{X_s\}_{s\in S} \to \Gamma$ is a *standard* symbolization map, then the indicator variables, called $Z_{\sigma s}$ in López *et al.* (2010), are *i.i.d.* Bernoulli $B(\frac{1}{n})$ variables for all $\sigma \in \Gamma$ and hence Y_σ is binomial $B(R, \frac{1}{n})$. Therefore Theorem 2.1 remains valid.

López *et al.* (2010, p.109) also evaluate the consistency of the $SG(m)$ statistic. An alternative result for which the same proof as the one given in the Consistency Theorem in López *et al.* (2010) applies, is the following:

Theorem 2.2. *Let $\{X_s\}_{s\in S}$ be a spatial process. Assume that there exists a standard symbolization map f for $\{X_s\}_{s\in S}$. Then, for all $0 < C < \infty, C \in \mathbb{R}$*

$$\lim_{R\to\infty} \Pr(SG > C) = 1$$

under any alternative for which the distribution of the symbols is not uniform.

Elsinger (2013) shows with a simple example that, in fact, the consistency of the SG test is not guaranteed against some alternative to the null hypothesis. However, Theorem 2.2 remains valid for the case of alternatives which produce a non-uniform distribution of the symbols. Indeed the proof of Theorem 2.2 in López *et al.* (2010, p.114) is based on the fact that symbolic entropy $h(\Gamma)$ is bounded by $0 \leq h(\Gamma) \leq \ln(n)$, taking the lower bound when only one symbol occurs and the upper bound when the n symbols are equally probable. Therefore, if under the alternative the symbols are not uniformly distributed, then $\ln(n) - h(\Gamma) > \varepsilon$ with $\varepsilon \in \mathbb{R}^+$ allowing for

$$\lim_{R\to\infty} \Pr(SG > C) = 1$$

This result holds for every real positive number C. In short, the SG test will be consistent provided that, under the alternative, the distribution of the symbols is not uniform. This is generally the case of spatial dependence, main interest of the paper.

3 An alternative version of the SG statistic.

The issue raised in footnote 4 of López *et al.* (2010, p.109), and also noted by Elsinger (2013), can be described as follows. Let $Z_{\sigma i}$ and $Z_{\sigma j}$ represent two indicators, which are not independent for all $i \neq j$ and some $\sigma \in \Gamma$. Variable Y_σ is obtained by accumulating the indicators on the set of locations. Due to their dependence, the distribution of Y_σ is not a binomial. Ruiz *et al.* (2010) propose controlling for the overlapping of the m-surroundings in order to ensure a good binomial approximation of the dependent indicators. Indeed, it can be proved that controlling the overlap, according to Ruiz *et al.* (2010), such that the number of symbolized overlapping m-surroundings is of the order R^α, with $0 < \alpha < 1$, will provide that Y_σ is asymptotically binomial distributed. The benefits of controlling the overlap are clear and very promising. However, this is an aspect that is still under research considering a more general scenario, applicable both to spatial processes and spatio-temporal processes.

The procedure proposed by Ruiz et al. (2010) does not symbolizes all the observations, and this may result in a loss of power and, moreover, there might be situations where the sample size is too small to impede its application. Because of this, we now propose an alternative that might be useful to practitioners.

3.1 A bootstrapped procedure

We propose a bootstrap approach (namely, permutation bootstrap) which, as demostrated among others by Skaug and Tjostheim (1996) or Lahiri (2003), is very adequate for testing the assumption of independence. Considering a number B of bootstrap replications, the procedure is as follows:

1. Compute the value of the statistic $\widehat{SG}$ for the original sample $\{X_s\}_{s\in S}$.
2. Resample $\{X_s\}_{s\in S}$, to obtain the bootstrapped spatial series $\{X_s(b)\}_{s\in S}$ with b the number of the bootstrapped sample.
3. For the bootstrapped sample, estimate the bootstrap realization of the statistic denoted by $\widehat{SG}^b$
4. Repeat $B-1$ times steps 2 and 3 to obtain B bootstrap realizations of the statistic, $\{\widehat{SG}^b\}_{b=1}^B$.
5. Compute the bootstrap p_{boots}-value:

$$p_{boots} - value(\widehat{SG}) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}(\widehat{SG}^b > \widehat{SG}) \tag{3}$$

where $\mathbf{1}(\cdot)$ is the indicator function, which assigns 1 to a true statement and 0 otherwise.

6. Reject the null hypothesis of independence in the spatial process $\{X_s\}_{s\in S}$ if

$$p_{boots} - value(\widehat{SG}) < \alpha$$

for a nominal size α.

4 Monte Carlo results

4.1 Empirical size and power of the SG^{boot} test

This section presents the results of a Monte Carlo experiment for the bootstrapped version of the SG test, using the same four spatial processes as in López *et al.* (2010). We have considered regular lattices with a small, (7×7), medium, (20×20) and large, (71×71), sample sizes. The number of bootstraps is $B = 999$ and each experiment has been repeated 1,000 times. Table 1 shows the results for the two linear models and Table 2 for their nonlinear version.

Nonparametric tests tend to be very sensitive to sample size, which is also true for the SG test in what respects to power, but not to size. Overall, our results are very encouraging: the empirical size of the test is correct, attains good power with samples of medium size and/or strong symptoms of spatial dependence and seems robust to functional form.

TABLE 1

TABLE 2

4.2 Comparison among the tests

Elsinger (2013, Section 2, entitled "The New Test") suggests a new test based on a Pearson's chi-squared approximation to the problem of dependent indicators. Given that no indication about the behaviour of this proposed test appears, we think that it is interesting to compare both proposals on equal basis.

Following Examples 1 and 2 in Elsinger (2013), assume a real valued spatial process $\{X_s\}_{s\in\mathbb{N}}$ whose domain is $\mathbb{N}$. Each X_s is drawn *i.i.d.* from a continuous distribution. Let $S_R = \{1, 2, \ldots, R\}$. For $m = 2$, the neighborhood pattern is '*one ahead*', so $N(s; S_R) = \{s, s+1\}$ for $s \in \{1, 2, \ldots, R-1\}$ and $N(R; S_R) = \{R, R-1\}$. There are two symbols $\Gamma = \{\sigma_1, \sigma_2\}$ whereas the symbolization map is $f(X_s) = \sigma_1$ if $X_s < X_{s+1}$ and $f(X_s) = \sigma_2$ otherwise (for $s = R$, $f(X_R) = \sigma_1$ if $X_R < X_{R-1}$).

TABLE 3

As said, Elsinger (2013) obtains the asymptotic distribution of the new test, called $3X^2(S_R)$ in Example 2, which is χ^2_1.

We have simulated a linear SAR process $X = (I - \rho W)^{-1}\varepsilon$ with the weighting matrix corresponding to the example above; ε is *i.i.d.* $N(0,1)$ and $\rho = \{0.1, 0.5, 0.9\}$. In all cases, the $3X^2(S_R)$ has been calculated using the symbolization procedure suggested by Elsinger (2013). Besides, the SG test has been calculated using the symbolization proposed by López *et al.* (2010), for $m = 2$. Consequently, the differences between the two tests are due exclusively to differences in their symbolization map.

The results appear in Table 3. The $'boot'$ superscript means that the results correspond to the bootstrapped version of the test; otherwise, the theoretical asymptotic distribution has been used.

The empirical size of the four tests is close to the nominal value, $\alpha = 0.05$. The estimated power for the $3X^2(S_R)$, in the two versions, remains at nominal levels, while the two versions of the SG test react adequately to sample size and/or cross-sectional dependence. Our impression is that the poor performance of $3X^2(S_R)$ is attributable to an improper selection of the symbols for the problem at hand. It is not a question of the test in itself but of the symbols defined.

5 Closing Remarks

The scientific method often advances by incrementally refining our understanding of method and subject matter. Critical reviews are one of the best ways to learn from the past and amend mistakes. We sincerely thank Elsinger (2013) for monitoring of our work. Indeed, there were some inaccuracies in López et al. (2010) that have been identified and corrected. The additional results included in this paper offer an improved version of the $SG(m)$ test, including an amended definition for a standard symbolization map. The counterexamples of Elsinger (2013) have been useful, showing how a test statistic, consistently built, may act totally wrong if it is not wisely interpreted. The results of the Monte Carlo experiment corroborate our approach.

Acknowledgments: The authors grateful for the financial support offered by the projects ECO2012- 36032-C03-01 and EC02012-36032-C03-03 from the Spanish Ministry of Economía y Competitividad; the COST Action IS1104, The EU in the new economic complex geography: models, tools and policy evaluation; Departamento de Industria e Innovación of the Goverment of Aragon and from the European Social Fund; and Fundación Séneca (Comunidad Autónoma de

Murcia).

References

[1] Elsinger, H., 2013. *Comment on: A non-parametric spatial independence test using symbolic entropy*, Preprint.

[2] Lahiri, S.N., 2003. *Resampling Methods for Dependent Data.* Springer-Verlag, New York.

[3] López, F. Matilla-García, M., Mur, J. and Ruiz, M., 2010. Non-parametric spatial independence test using symbolic entropy. *Regional Science and Urban Economics*, 40, 106-115.

[4] Ruiz, M., López, F. Páez, A., 2010. Testing for spatial association of qualitative data using symbolic dynamics. *Journal of Geographical Systems*, 12, 281-309.

[5] Skaug, H.J. and Tjøstheim, D., 1996. Measures of distance between densities with application to testing for serial independence. In: Robinson, P.M., Rosenblatt, M. (Eds.), *Time Series Analysis in Memory of E.J. Hannan*. Springer-Verlag, New York, pp. 363–377.

Table 1: Estimated Size and Power of the SG^{boot} test. LINEAR MODELS

		Size	Estimated Power. SAR case			Estimated Power. SMA case		
	ρ	0	0.1	0.5	0.9	0.1	0.5	0.9
R = 49	m = 3	0.046	0.040	0.295	0.904	0.039	0.201	0.420
	m = 4	0.046	0.056	0.251	0.898	0.060	0.175	0.331
	m = 5	0.054	0.043	0.189	0.875	0.044	0.122	0.273
R = 400	m = 3	0.063	0.113	0.991	1.000	0.117	0.971	1.000
	m = 4	0.051	0.117	0.995	1.000	0.089	0.944	1.000
	m = 5	0.040	0.082	0.980	1.000	0.071	0.906	1.000
R = 5041	m = 3	0.045	0.794	1.000	1.000	0.795	1.000	1.000
	m = 4	0.052	0.834	1.000	1.000	0.735	1.000	1.000
	m = 5	0.050	0.703	1.000	1.000	0.669	1.000	1.000

SAR: $X = (I_R - \rho W)^{-1} \varepsilon$; SMA: $X = (I_R + \rho W) \varepsilon$

Table 2: Estimated Power of theSG^{boot} test. NONLINEAR MODELS

		Estimated Power. NL1 case			Estimated Power. NL2 case		
	ρ	0.1	0.5	0.9	0.1	0.5	0.9
R = 49	m = 3	0.053	0.116	0.603	0.043	0.327	0.893
	m = 4	0.044	0.113	0.552	0.053	0.263	0.901
	m = 5	0.047	0.094	0.487	0.051	0.214	0.891
R = 400	m = 3	0.073	0.843	1.000	0.097	0.989	1.000
	m = 4	0.099	0.857	1.000	0.123	0.997	1.000
	m = 5	0.051	0.756	1.000	0.068	0.976	1.000
R = 5041	m = 3	0.706	1.000	1.000	0.797	1.000	1.000
	m = 4	0.775	1.000	1.000	0.848	1.000	1.000
	m = 5	0.615	1.000	1.000	0.716	1.000	1.000

NL1: $X = 1/\left(I_R - \rho W\right)^{-1}\varepsilon$; NL2: $X = \left[\left(I_R - \rho W\right)^{-1}\varepsilon\right]^5$

Table 3: Estimated Size and Power for the $SG(m)$, $SG(m)^{boot}$, $3X^2(Sr)$, $3X^2(Sr)^{boot}$

		Size	Estimated Power: SAR case		
$m=2$	ρ	0	0.1	0.5	0.9
R = 49	$3X^2(Sr)$	0.047	0.050	0.054	0.054
	$SG(m)$	0.050	0.066	0.345	0.880
	$3X^2(Sr)^{boot}$	0.106	0.111	0.124	0.114
	$SG(m)^{boot}$	0.060	0.068	0.565	0.978
R = 400	$3X^2(Sr)$	0.050	0.051	0.050	0.062
	$SG(m)$	0.059	0.161	1.000	1.000
	$3X^2(Sr)^{boot}$	0.060	0.074	0.065	0.082
	$SG(m)^{boot}$	0.060	0.257	1.000	1.000
R = 5041	$3X^2(Sr)$	0.053	0.051	0.046	0.061
	$SG(m)$	0.045	0.885	1.000	1.000
	$3X^2(Sr)^{boot}$	0.055	0.053	0.050	0.061
	$SG(m)^{boot}$	0.046	0.994	1.000	1.000

SAR: $X = (I_R - \rho W)^{-1}\varepsilon$