A Test for Determinism in Spatial Processes

First Draft

Abstract

We propose the first statistical test to determine, given a spatial process, if the spatial structure is generated by a deterministic, rather than by a stochastic, process. The conditions under which the test can be applied are really weak. The advantages of the presented methods are simplicity, invariance respect to monotonic transformations and the applicability of the test regardless the discrete or continuous nature of the data generating process. We conduct several simulation studies to evaluate the performance of our test on well-known spatial processes.

1 Introduction

There should be an important interest in economics in measuring the stability of spatial processes, and hence to distinguishing between nonlinear deterministic processes and stochastic processes. It is central for economic analysis and economic forecasting to know whether the data at hand are essentially generated by a deterministic or stochastic spatial process, since the tools available in each case are certainly different. Precisely, this paper deals with this open problem, and puts forward the first test for deterministic complex spatial dynamics versus stochastic ones.

This paper presents a simple new test for determinism which rely, however, on symbolic dynamics and entropy. The symbolic approach was successfully and systematically introduced in Matilla and Ruiz (2008) and extended to the analysis of spatial processes in López et al (2009) and Ruiz et al. (2010). Symbolic dynamics studies dynamical systems on the basis of the symbol sequences 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 to label each region by an alphabetical *symbol*. One of the interesting properties of symbolic dynamics is that essential global features of the underlying process, like its deterministic or stochastic nature or its complexity, are kept. More precisely, our test relies on the concept of entropy, neither topological entropy nor Kolmogorov-Sinai entropy, but *symbolic entropy*, which is rooted on symbolic dynamics, as will be shown. Entropy has been widely used in econometrics; see Ullah (1996) for a review. Most of this research is related with establishing asymptotic distribution theory for certain entropy measures of serial dependence (Robinson 1991, Granger and Lin 1994, Maasoumi and Racine 2002, Granger et al. 2004 and Hong and

White 2005). We use another measure of entropy, *symbolic entropy*, to construct a test of determinism. Entropy gives a measure of the uncertainty or volatility of the process under study. As will be shown, the proposed new approach based on symbolic entropy benefits (for some symbolizations) from the property of invariance under any monotonous transformation of the data. Furthermore, its simplicity makes it easy to implement in programming language and hence portability is guaranteed.

In the paper, after presenting the basic new definitions and the theoretical results, the small sample properties of the test are studied using several spatial processes. We also conduct Monte Carlo experiments and we evaluate the performance of the new statistical procedures on deterministic noisy series.

The remainder of the paper is organized as follows: Definitions of 'symbol', 'symbolization map', 'probability of a symbol' and 'symbolic entropy' are presented in Section 2. Section 3 is devoted to the construction of a symbolization map. In Section 4 we present a test for determinism. Section 5 reports the results of applying our test for determinism to several spatial process and we report the results of Monte Carlo experiment. Some concluding remarks are made in the Conclusion.

2 Preliminaries and notation

Let $\{X_s\}_{s\in S}$ be a spatial process. Let $\Gamma = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ be a set of n symbols. Now assume that there is a map

$$f : \{X_s\}_{s\in S} \to \Gamma$$

defined by $f(X_s) = \sigma_{j_s}$ with $j_s \in \{1, 2, \ldots, n\}$. We will say that $s \in S$ is of σ_i-type if and only if $f(X_s) = \sigma_i$. We will call the map f a *symbolization map.* We will say that the symbol $\sigma \in \Gamma$ is *admissible* for the spatial process $\{X_s\}_{s\in S}$ if and only if $f(X_s) = \sigma$ for some $s \in S$.

Notice that the symbolization map, and thus the set of symbols Γ, have to be defined such that they are able to gather the spatial structure of $\{X_s\}_{s\in S}$.

Denote by

$$n_{\sigma_i} = \sharp\{s \in S |\, f(X_s) = \sigma_i\},$$

that is, the cardinality of the subset of S formed by all the elements of σ_i-type.

Also, under the conditions above, one could easily compute the relative frequency of a symbol $\sigma \in \Gamma$ by:

$$p(\sigma) := p_\sigma = \frac{\sharp\ \{s \in S \mid s \text{ is of } \sigma - \text{type}\}}{|S|} \tag{1}$$

where by $|S|$ we denote the cardinality of set S.

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

$$h(\Gamma) = -\sum_{\sigma\in\Gamma} p_\sigma \ln(p_\sigma). \tag{2}$$

Symbolic entropy, $h(\Gamma)$, is the information contained in comparing the m-surroundings generated by the spatial process. It is a measure of the degree of disorder in the spatial process. Notice that $0 \leq h(\Gamma) \leq \ln(n)$ where the lower bound is attained when only one symbol occurs, and therefore the spatial process posses high spatial structure, and the upper bound is reached for a completely random spatial system where all possible symbols appear with the same probability.

3 Example of symbolization

Now we propose a particular symbolization map f for the spatial process $\{X_s\}_{s\in S}$. There might be several possible symbolization maps. Therefore, this novel framework is adaptable to the necessities of the problem at hand, and so the procedure below can be refined in accordance with particular cases for which the researcher has a better understanding of the process to be studied. The proposed symbolization map f is defined as follows: denote by Me the median of the spatial process $\{X_s\}_{s\in S}$ and let

$$\gamma_s = \begin{cases} 0 & \text{if } X_s \leq Me \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

Now, define the indicator function

$$\mathcal{I}_{s_1 s_2} = \begin{cases} 0 & \text{if } \gamma_{s_1} \neq \gamma_{s_2} \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

Let $m \in \mathbb{N}$ with $m \geq 2$. Next, we consider that the spatial process $\{X_s\}_{s\in S}$ is embedded in an m-dimensional space as follows:

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

where $s_1, s_2, \ldots, s_{m-1}$ are the $m-1$ nearest neighbors to s_0, which are ordered from lesser to higher Euclidean distance with respect to location s_0. If two or more locations are equidistant to s_0 we choose them in an anticlockwise manner. In formal terms, $s_1, s_2, \ldots, s_{m-1}$ are the $m-1$ nearest neighbors to s_0 satisfying the following two conditions:

(a) $\rho_1^0 \leq \rho_2^0 \leq \cdots \leq \rho_{m-1}^0$,

(b) and if $\rho_i^0 = \rho_{i+1}^0$ then $\theta_i^0 < \theta_{i+1}^0$.

Notice that condition (b) is a technical condition that ensure the uniqueness of $X_m(s)$ for all $s \in S$ in the case in which two neighbors are at the same distance of s_0. We will call $X_m(s)$ an m-surrounding of s.

For any localization s, set $X_m(s) = (X_s, X_{s_1}, \ldots, X_{s_{m-1}})$ and denote by $N_s = \{s_1, \ldots s_{m-1}\}$ the $m-1$ nearest neighbors of s. This symbolization procedure consists of comparing at each localization s the value of γ_s with γ_{s_i} for all $s_i \in N_s$. Thus, that $\gamma_s = \gamma_{s_i}$ means that X_s and X_{s_i} are both less than, or greater than, Me.

Then, the symbolization map $f : \{X_s\}_{s\in S} \hookrightarrow \mathbb{R}^m \to \Gamma$ is defined as:

$$f(X_s) = (\mathcal{I}_{ss_1}, \mathcal{I}_{ss_2}, \ldots, \mathcal{I}_{ss_{m-1}}) \tag{5}$$

where $X_m(s) = (X_s, X_{s_1}, \ldots, X_{s_{m-1}})$ and Γ is the set of 2^{m-1} different vectors of dimension $m-1$ with entries in the set $\{0, 1\}$.

3.1 A Feasible Test for Spatial Determinism

The purpose of this subsection is to elaborate a test that allows us to discriminate between deterministic and stochastic (dependent or independent) processes. For a finite amount of data ($|S| < \infty$) we study, given a symbolization map $f : \{X_s\}_{s\in S} \to \Gamma$, the behavior of $h(\Gamma)$ when increasing the number of possible observable symbols.

Fix $w, k \in \mathbb{N}$ such that $w = \frac{n}{k}$ and $w << n$ where n is the cardinality of the set of symbols Γ. Let

$$\{\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_k\}$$

be k subsets of the set Γ. Each set $\mathcal{W}_j$ is a set of $j \cdot w$ symbols chosen at random in Γ.

We define the next *modified* symbolic entropy's function

$$h^{\mathcal{W}_j}(\Gamma) = -\sum_{\pi_i \in \mathcal{W}_j} p(\pi_i) \log p(\pi_i).$$

Now, (2) can be understood as the limit behavior of $h^{\mathcal{W}_j}(\Gamma)$ since, as can be easily verified, it holds that:

$$h(\Gamma) = \lim_{j\to k} h^{\mathcal{W}_j}(\Gamma) \tag{6}$$

We are not interested in $h(\Gamma)$ itself, but in the sequence $\left\{h^{\mathcal{W}_j}(\Gamma)\right\}_{j=1}^{k}$ which contains $k = n/w$ points.

Recall that symbolic entropy is a measure of the degree of disorder in the spatial process and therefore, the sequence of modify symbolic entropies are partial evaluations of such a disorder in the spatial process. Hence, for a stochastic process, the sequence of modify symbolic entropies will be increasing while for a deterministic process this might not happen.

Taking into account that $h^{\mathcal{W}_j}(\Gamma) \leq \log(jw)$, if the process is *stochastic*, the sequence $(h^{\mathcal{W}_j}(\Gamma))_j$ will scale with $(\log(jw))_j$, while this will not occur if the process is deterministic.

Summing up, the information (or complexity) measured via symbolic entropy reaches a level such that, in case of a deterministic spatial process, no more significant information is captured by increasing the number of symbols under consideration.

A new feature that makes a distinction between deterministic and nondeterministic spatial processes has been put forward. This new attribute will be used to construct a test for determinism.

As said before, the numerical slope of symbolic entropy of a random spatial process is increasing. This does not hold for the deterministic data set for which the slope varies in a non-increasing way. Denote by

$$dh^{\mathcal{W}_j}(\Gamma) = \frac{h^{\mathcal{W}_{j+1}}(\Gamma) - h^{\mathcal{W}_j}(\Gamma)}{\log \frac{j+1}{j}}$$

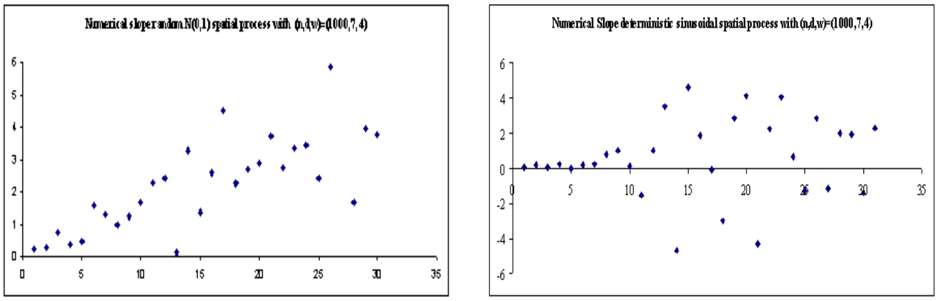

Figure 1: Numerical slope $dh^{\mathcal{W}}{}_j(\Gamma)$ for random $N(0,1)$ (left) and deterministic sinusoidal $y = 0.5Wy + sin(1+a+b)$ (right) where a and b are the vectors of coordinates of the locations in S (i.e. $s_i = (a_i, b_i)$ with $s_i \in S$) and W a weighting matrix.

$h^{\mathcal{W}_j}(\Gamma)$'s numerical slope. The non-increasing property of the modified symbolic entropy's slope ($h^{\mathcal{W}_j}(\Gamma)$) can be tested using a classical econometric test by performing the following regression:

$$dh^{\mathcal{W}_j}(\Gamma) = \alpha_0 + \alpha_1 j + \varepsilon_j, \quad \text{for} \quad j = 1, 2, ...k-1 \tag{7}$$

where ε_j is independent white noise with $E(\varepsilon_j^2) = \sigma^2$ and $E(\varepsilon_j^4) < \infty$. According to definition of $h^{\mathcal{W}_j}(\Gamma)$, regression (7) relates the mean gain of information due to the evaluation of a random set of symbols with the number of symbols. On the other hand, regression (10) can be understood as a Simple Symbol-Trend Model. As in the well-known simple time-trend model, the OLS estimates $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are so that asymptotically the usual t-test of H_0 is valid[1]. As a result, the estimated parameter $\hat{\alpha}_1$ can be used to test that $dh^{\mathcal{W}_j}(\Gamma)$ does not increase with j, which implies an underlying deterministic process.

The null hypothesis and the alternative hypothesis are then expressed as follows[2]:

$$\begin{aligned} H_0 &: \alpha_1 = 0 \text{ (deterministic process)} \\ H_1 &: \alpha_1 > 0 \text{ (stochastic process)} \end{aligned}$$

4 Monte Carlo simulations

In order to see the finite sample behavior of the dh-test we will use the following spatial processes. Let a and b be the vectors of coordinates of the locations in S (i.e. $s_i = (a_i, b_i)$ with $s_i \in S$), W a weighting matrix and ε an i.i.d. $N(0, \sigma)$.

1. Model 1: $y_1 = c_0 + c_1 a + c_2 b$

2. Model 2: $y_2 = c_0 + c_1 a + c_2 b + c_3 a^2 + c_4 b^2 + c_5 ab$

[1] See, for example, the proof given in Hamilton (1994), pags. 454-463.

[2] Notice that the test is one-sided since the numerical slope is never negative by construction of the test.

3. Model 3: $y_3 = sin(2\pi * y_1)$

4. Model 4: $y_4 = \rho W y_4 + y_1 + \varepsilon$

5. Model 5: $y_5 = \rho W y_5 + y_2 + \varepsilon$

6. Model 6: $y_6 = \rho W y_6 + y_3 + \varepsilon$

7. Model 7: $y_7 = \varepsilon$

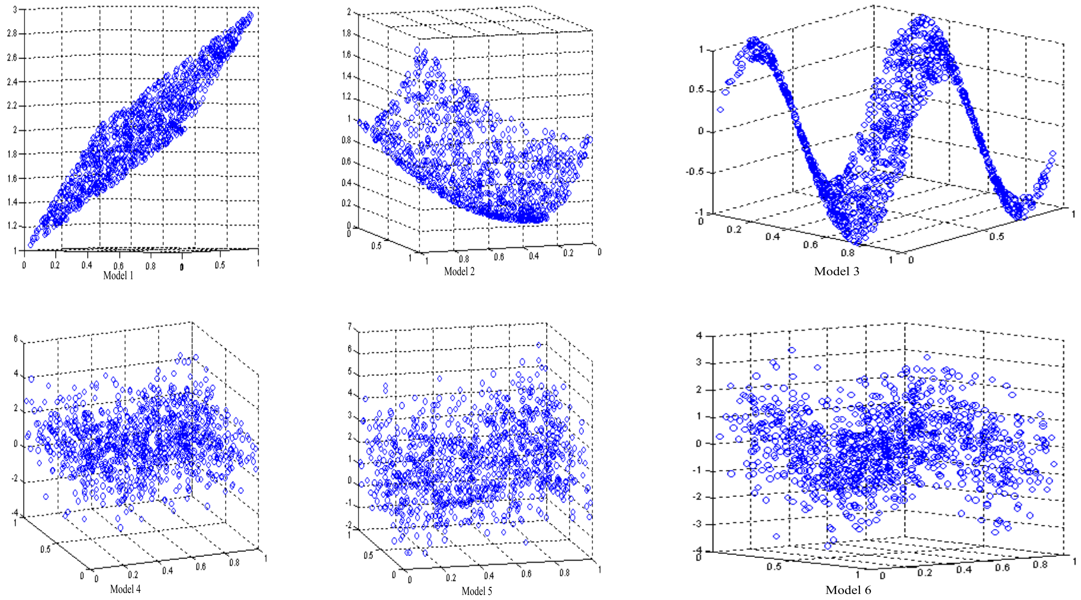

Figure 2: Scatter plot of Models 1-6. First row, from left to right Models 1-3. Second row from left to right Models 4-6.

As can be seen in Figure 2, y_1 is a deterministic linear process while y_2 and y_3 are deterministic nonlinear processes (quadratic and sinusoidal models respectively). Models 4-6 are SAR processes with linear, quadratic or sinusoidal deterministic trend and i.i.d $N(0, \sigma)$ respectively. Finally Model 7 is i.i.d white noise. Therefore Models 4-7 are not deterministic.

Table 1 and 2 show the results of the number of times that the test accepts determinism when a deterministic process is simulated and the number times that the test accepts determinism when a stochastic process is simulated respectively for 1000 Monte Carlo replications, with $R = 100, 400$ and 800 and embedding dimension $m = 5, 6$ and 7, using the symbolization procedure exposed in Section 3.

Table 1. Monte Carlo results for $P(\text{Accept determinism}|\text{deterministic process})$ in %.

		Model 1	Model 2	Model 3
R	m	Linear	Quadratic	Sinusoidal
100	5	90,7	90,9	87,4
400	5	88,4	88,8	85,3
	7	100	100	100
800	5	84,4	83,9	83,1
	7	100	100	100
	8	100	100	100

Table 2. Monte Carlo results for $P(\text{Accept determinism}|\text{non-deterministic process})$ in %.

		Model 4	Model 5	Model 6	Model 7
R	m	Linear	Quadratic	Sinusoidal	Random
100	5	20	19	16	23
400	5	0	2	1	0
	7	0	0	0	0
800	5	0	0	0	0
	7	0	0	0	0
	8	0	0	0	0

Finally and in order to see how sensible the test is to detect deterministic/stochastic processes we simulate the model $y = 0.5Wy + 2(p_0 + p_1 a + p_b) + \varepsilon$, which contains a strong deterministic skeleton, under different levels of noise and sample size $R = 800$ (see Figure 2). These results can be found in Table 3.

Table 3. Value of $\widehat{\alpha}_1$ and p-value in parenthesis.

	m=6	m=7	m=8
σ	$\widehat{\alpha}_1$	$\widehat{\alpha}_1$	$\widehat{\alpha}_1$
0.1	0.065 (0.13)	0.039 (0.19)	0.020 (0.28)
0.2	0.120 (0.14)	0.065 (0.17)	0.033 (0.18)
0.3	0.121 (0.12)	0.081 (0.09)	0.050 (0.21)
0.4	0.149 (0.15)	0.098 (0.16)	0.055 (0.23)
0.5	0.185 (0.07)	0.100 (0.04)	0.061 (0.15)
0.6	0.187 (0.09)	0.117 (0.08)	0.069 (0.13)
0.7	0.201 (0.05)	0.116 (0.01)	0.068 (0.08)
0.8	0.190 (0.02)	0.125 (0.002)	0.067 (0.03)
0.9	0.204 (0.003)	0.123 (<0.001)	0.070 (0.01)
1	0.208 (0.004)	0.122 (<0.001)	0.074 (0.003)

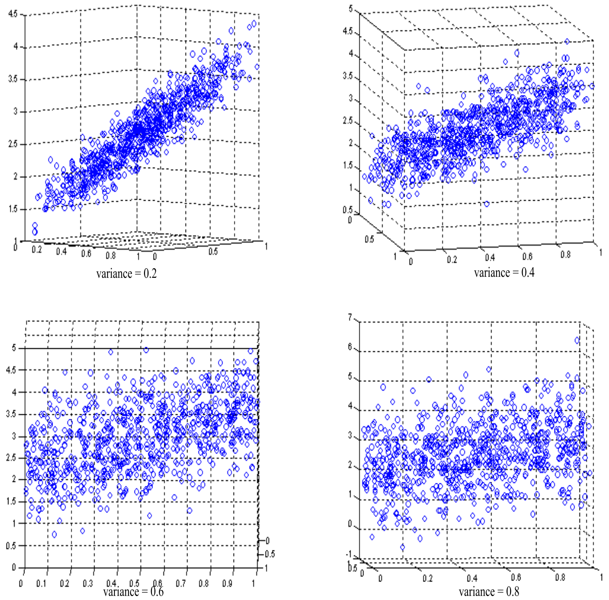

Figure 3: Scatter plot of $y = 0.5Wy + 2(1 + a + b) + \varepsilon$ where $\varepsilon = N(0, \sigma)$, for $\sigma = 0.2, 0.4, 0.6$ and 0.8

5 Conclusions

This paper has introduced a new way for testing for determinism in spatial processes. These statistical procedures critically rely on the concept of *entropy* which, as presented here, is formulated in terms of symbols. We emphasize that we do not work directly with the actual observed values (which are real numbers), rather we take the entropy of the distribution of the symbols as a potential measure of its complexity. Focusing on symbols, we are able to detect global properties of the data generating process such as determinism (stochastic nature). Importantly, we find an intrinsic common characteristic for any deterministic process that does not hold for a stochastic one. The statistical methods presented in this paper are fast in computing times, simple and powerful and invariant under monotonous transformations. Moreover the test is free regarding the discrete or continuous nature of the generating process. As a result the only parameter that has to be freely fixed by the researcher is the embedding dimension (m).

6 Bibliography

Granger, C.W.J. and Lin J.L. 1994. Using the mutual information coefficient to identify lags in nonlinear models. *Journal of Time Series Analysis*, 15, 371-84.

Granger, C.W.J., Maasoumi, E. and Racine, J. 2004. A Dependence Metric for possibly nonlinear processes. *Journal of Time Series Analysis* 23, 649-669.

Hong, Y. and White, H. 2005. Asymptotic distribution theory for nonparametric entropy measures of serial dependence. *Econometrica* 73, 837–901.

López, F., Mariano Matilla, Jesús Mur and Manuel Ruiz Marín, Non-Parametric Spatial Independence Test Using Symbolic Entropy, (2009), *Regional Science and Urban Economics*, doi: 10.1016/j.regsciurbeco.2009.11.003.

Maasoumi, E. and Racine, J. 2002. Entropy and predictability of stock market returns. *Journal of Econometrics*, 107, 191-312.

Matilla, M. y Manuel Ruiz, A non-parametric independence test using permutation entropy, *Journal of Econometrics*, Volume 144 (1), (2008), pp 139–155.

Robinson, P.M. 1991. Consistent nonparametric Entropy-Based testing. *Review of Economic Studies*, 58, 437-453.

Ruiz, M, Fernando López and Antonio Paez, Testing for spatial Association of Qualitative Data Using Symbolic Dynamics, *Journal of Geographical Systems*, doi: 10.1007/s10109-009-0100-1.

Ullah, A. 1996. Entropy, divergence and distance measures with econometric applications. *Journal of Statistical Planning and Inference*, 49, 137-162.