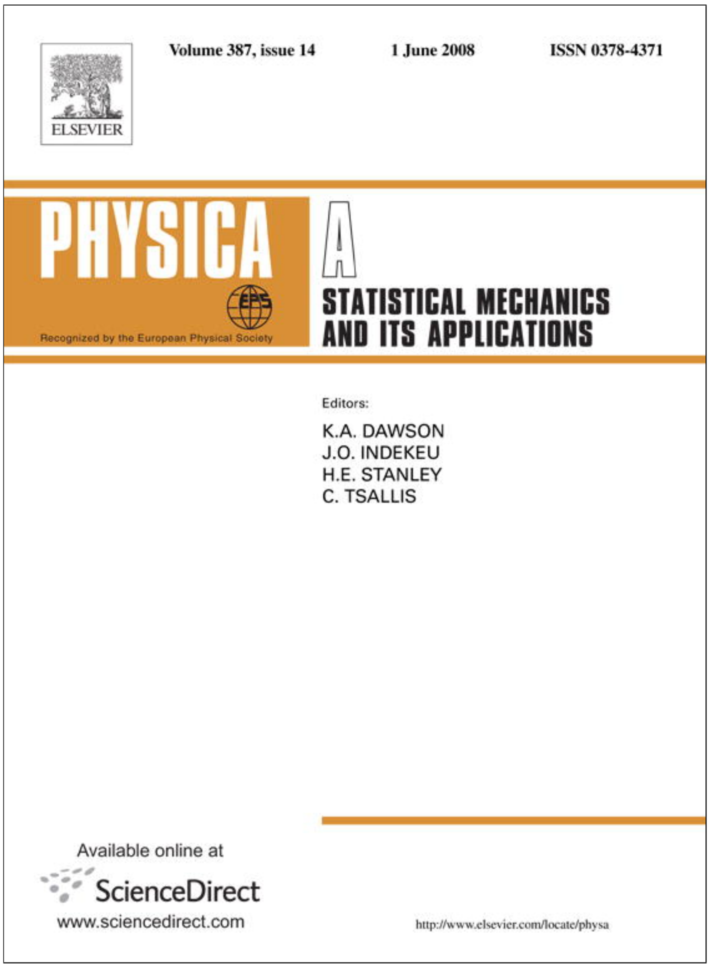

Volume 387, issue 14 1 June 2008 ISSN 0378-4371

ELSEVIER

PHYSICA A

STATISTICAL MECHANICS AND ITS APPLICATIONS

Recognized by the European Physical Society

Available online at
ScienceDirect
www.sciencedirect.com

http://www.elsevier.com/locate/physa

Available online at www.sciencedirect.com
ScienceDirect
Physica A 387 (2008) 3373–3383
PHYSICA A
www.elsevier.com/locate/physa

Randomizing nonlinear maps via symbolic dynamics

L. De Micco[a,e], C.M. González[a], H.A. Larrondo[a,e,*], M.T. Martin[b,e], A. Plastino[b,e], O.A. Rosso[c,d,e]

[a] *Facultad de Ingeniería, Universidad Nacional de Mar del Plata, Av. J.B. Justo 4302, 7600 Mar del Plata, Argentina*
[b] *Instituto de Física, Facultad de Ciencias Exactas, Universidad Nacional de La Plata (UNLP), C.C. 727, 1900 La Plata, Argentina*
[c] *Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine, School of Electrical Engineering and Computer Science, The University of Newcastle, University Drive, Callaghan NSW 2308, Australia*
[d] *Chaos & Biology Group, Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Pabellón II, Ciudad Universitaria. 1428 Ciudad Autonoma de Buenos Aires, Argentina*
[e] *CONICET, Argentina*

Received 16 October 2007; received in revised form 8 February 2008
Available online 21 February 2008

Abstract

Pseudo Random Number Generators (PRNG) have attracted intense attention due to their obvious importance for many branches of science and technology. A *randomizing technique* is a procedure designed to improve the PRNG randomness degree according the specific requirements. It is obviously important to quantify its effectiveness. In order to classify *randomizing techniques* based on a symbolic dynamics' approach, we advance a novel, physically motivated representation based on the statistical properties of chaotic systems. Recourse is made to a plane that has as coordinates (i) the Shannon entropy and (ii) a form of the statistical complexity measure. Each statistical quantifier incorporates a different probability distribution function, generating thus a representation that (i) sheds insight into just how each randomizing technique operates and also (ii) quantifies its effectiveness. Using the Logistic Map and the Three Way Bernoulli Map as typical examples of chaotic dynamics it is shown that our methodology allows for choosing the more convenient randomizing technique in each instance. Comparison with measures of complexity based on diagonal lines on the recurrence plots [N. Marwan, M.C. Romano, M. Thiel, J. Kurths, Phys. Rep. 438 (2007) 237] support the main conclusions of this paper.

PACS: 02.70.Rr; 05.40.-a; 05.45.-a; 89.70.+c

Keywords: Symbolic dynamics; Pseudo random number generators; Permutation entropy; Statistical complexity

1. Introduction

In spite of the fact that the existence (or not) of truly random number generators (RNG) remains an open question, Pseudo Random Number Generators (PRNG's) are widely used in science and technology. It is clear that a complex

* Corresponding author at: Facultad de Ingeniería, Universidad Nacional de Mar del Plata, Av. J.B. Justo 4302, 7600 Mar del Plata, Argentina. Tel.: +54 223 4816600; fax: +54 223 4816600.
E-mail addresses: lucianadm55@hotmail.com (L. De Micco), cmgonzal@fi.mdp.edu.ar (C.M. González), larrondo@fi.mdp.edu.ar, larrondo@ciudad.com.ar (H.A. Larrondo), mtmartin@venus.fisica.unlp.edu.ar (M.T. Martin), plastino@sinectis.com.ar (A. Plastino), oarosso@fibertel.com.ar (O.A. Rosso).

0378-4371/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2008.02.037

dynamics does not necessarily implies the workings of a complicated model, if nonlinearities are present. Low-dimensional, chaotic dynamic systems constitute paramount examples of such an assertion and have been employed as Pseudo Random Number Generators because they are able to generate stochastic-like signals out of underlying simple models that are easy to implement via appropriate software or hardware. Usually, a suitable manipulation of the time-series that these models generate is required to improve their statistical properties. Here we will be interested precisely in these manipulations or *randomizing techniques*, that is, procedures that increase the quality of a PRNG. Obviously, we want to be in a position to quantitatively assess a "degree of quality" of a given randomizing techniques, an issue that will be the leitmotif of the present effort, in the wake of many tests that have been proposed in the literature to characterize the Pseudo Random Number Generator's quality [1].

Our approach takes advantage of the statistical properties of chaotic systems, a theme that has been addressed by several authors. Excellent reviews are found in the works of Beck–Schögl [2], Lasota–Mackey [3], and Setti et al. [4]. We will show that Pseudo Random Number Generators based on very simple chaotic systems **may be greatly improved** by means of two types of symbolic dynamics randomizing techniques, namely, *Discretization* [5–7] and *Skipping* [4].

Note that symbolic dynamics are usually regarded as coarse-grain descriptions of a "real" dynamic system's continuous time-evolution [8,9]. Such a viewpoint stimulates research focused on generating suitable partitions of the relevant state space, thus producing symbolic dynamics descriptions without information loss. Contrariwise, *in this effort we look at symbolic dynamics as a tool* for randomizing a chaotic Pseudo Random Number Generator. That is, starting from a chaotic time series, our goal is to get a "symbolic" timeseries with more convenient statistical properties than those of the original one.

The expressions "more convenient", or "better" statistical properties are to be understood within an Information Theory framework. We employ a normalized-to-unity version of the Shannon entropy [10] to be denoted as H_S. One wishes for it to be maximal (i.e. equal to unity) if randomness is the desideratum. However, recourse to just H_S does not guarantee the random nature of a time series. Our proposal is *to combine the H_S-measure with another quantifier*, called a statistical complexity measure (SCM) whose original functional product form was proposed by Lopez-Ruiz, Mancini and Calbet in the seminal paper [11], and baptized as C_{LMC}.

Of course there exist many other complexity measures. For a comparison amongst them see the paper by Wackerbauer et al. [12]. Note that here we used a modified version of the SCM designed by Martin, Plastino and Rosso that overcomes some troublesome characteristics of the original measure (see, for example, Refs. [13–16]).

The intensive statistical complexity version (for short C_{MPR}) has been shown to be a convenient tool for different purposes [17–23]. We are, of course, taking advantage of an important statistical complexity measure property: it vanishes for completely random signals [11,15,16], thus guaranteeing that no "hidden" structures (or correlations) exist. In Refs. [11,24] a representation in a plane defined by the two quantifiers $H \times C$ is proposed, where the H-axis is considered *a time coordinate* of a dynamic system. In a similar way we use here $H_S \times C_{\mathrm{MPR}}$ in which the ideal situation is represented by the $(1, 0)$ point, will allow us to establish the main result of this endeavour: *Discretization and Skipping yield markedly different trajectories towards the ideal point.*

To exhibit that effect we will analyse chaotic time-series generated by two well-known maps: the Logistic Map (LOG) and the Three Way Bernoulli Map (TWBM). They have been selected, among other possibilities, because they are representative of two different classes of systems:

- The LOG-map represents continuous systems that may be approached with the Lorenz procedure via a $1D$-map. A *non uniform* natural invariant density is an important feature in this instance [2].
- On the other hand, TWBM is representative of many piecewise linear maps as, for example, the Four Way Tailed Shift Map, the Skew Tent Map, the Three Way Tailed Shift Map, etc., that share a *uniform* natural invariant density but have different mixing properties [2].

The present work makes two main contributions (a) two randomizing techniques that may be widely used to randomize chaotic time series and (b) a representation plane where the effectiveness of each one is clearly exhibited. The paper is organized as follows: we describe in Section 2 the symbolic dynamics' randomization processes (discretization and skipping). As a complement we revisit in Appendix the use of the Perron-Frobenius operator in connection with chaotic maps together with two central concepts for our present purposes, i.e. those of (i) invariant probability measure and (ii) mixing. Section 3 provides details concerning the evaluation of the two statistical

Table 1
Illustrating the **Discretization** procedure

$\mathcal{S}_{\mathrm{IN}}$ ($\mathbb{R}$)	$\mathcal{S}_1$ ($\mathbb{N}$)	$\mathcal{S}_2$ (*binary*)	$\mathcal{S}_3$ — MSB 4-dim embedding	$\mathcal{S}_4$ ($\mathbb{N}$)	$\mathcal{S}_{\mathrm{OUT}}$ ($\mathbb{R}$)
0.010559404	0	**0**000			
0.041791613	0	**0**000			
0.160180296	2	**0**010			
0.538090276	8	**1**000	**0001**	1	0.066666667
0.994196523	14	**1**110			
0.023079185	0	**0**000			
0.090186145	1	**0**001			
0.328210418	4	**0**100	**1000**	8	0.533333333
0.881953358	13	**1**101			
0.416446528	6	**0**110			
0.972075269	14	**1**110			
0.10857976	1	**0**001	**1010**	10	0.666666667
0.387160782	5	**0**101			
0.949069244	14	**1**110			
0.193347257	2	**0**010			
0.62385638	9	**1**001	**0101**	5	0.333333333

Table 2
Illustrating the **Skipping** procedure

$\mathcal{S}_{\mathrm{IN}}$	$\mathcal{S}_{\mathrm{OUT}} \equiv f^d$				
($\mathbb{R}$)	$d = 2$	$d = 3$	$d = 4$	$d = 5$	$d = 6$
0.010559404					
0.041791613	0.041791613				
0.160180296		0.160180296			
0.538090276	0.538090276		0.538090276		
0.994196523				0.994196523	
0.023079185	0.023079185	0.023079185			0.023079185
0.090186145					
0.328210418	0.328210418		0.328210418		
0.881953358		0.881953358			
0.416446528	0.416446528			0.416446528	
0.972075269					
0.10857976	0.10857976	0.10857976	0.10857976		0.10857976
0.387160782					
0.949069244	0.949069244				
0.193347257		0.193347257		0.193347257	
0.62385638	0.62385638		0.62385638		

quantifiers that we employ while results and conclusions are presented in Section 4. A brief summary is provided in Section 5.

2. Symbolic dynamics-based randomization processes

Let f be a chaotic map. Starting from a randomly chosen initial condition, the map is iterated generating the chaotic timeseries (CHTS) $\mathcal{S}_{\mathrm{IN}} = \{x_0, x_1, \ldots\}$. This CHTS is to be regarded as the input for the randomizing process (see Tables 1 and 2). Let us assume that x_i is a floating-point number (in the IEEE normalized representation). Without loss of generality, we consider values restricted to the interval [0, 1]. As previously stated, the central idea here is that of employing two kinds of well-known symbolic dynamics randomization processes (Discretization and Skipping). After several steps, the output of each of these processes will be a new timeseries (STS) $\mathcal{S}_{\mathrm{OUT}}$ obtained as described below (see Tables 1 and 2).

- **Discretization:** The discretization process is carried out according to the following steps:
  (1) Each value of $\mathcal{S}_{\rm IN}$ is first discretized using N-bits. Accordingly, the continuous (real) interval $[0, 1]$ is mapped onto the discrete interval $[0, 2^N - 1]$. The new series $\mathcal{S}_1$ ensues. In Table 1 the case $N = 4$ is illustrated (see columns $\mathcal{S}_1$, $\mathcal{S}_2$, etc.). Remark that this first step is unavoidable when the chaotic system is hardware-implemented, using for example, field-programmable gate arrays. After this first step the symbolic representation space contains 2^N different symbols. Note that in some *pathological* cases (the Tent map, for instance) this procedure renders a series for which the chaotic behaviour completely disappears. In these cases several approaches can be invoked so as to avoid this problem [25–27].
  (2) Translate now the preceding time-series $\mathcal{S}_1$ into a new $\mathcal{S}_2$-one (binary format). In column $\mathcal{S}_2$ the most significative digit (MSB) is displayed in bold typeface, as it is the one to be considered in the following conversion steps.
  (3) Regard now N consecutive $\mathcal{S}_2$-members as "coordinates" in an N-dimensional embedding space $\mathcal{E}$. "Points" in $\mathcal{E}$ are represented by binary numbers. We now form a new timeseries $\mathcal{S}_3$ by retaining only the most significative digit (MSB) of each coordinate of the embedding space. Note that this MSB-representation is equivalent to the commonly employed symbolic-dynamics process that consists of splitting the interval $[0, 1]$ into two subintervals $A = [0, 0.5)$ and $B = [0.5, 1]$, assigning a "0" if $x_i \in A$ or a "1" if $x_i \in B$.
  (4) Reconvert $\mathcal{S}_3$ into N-bits natural numbers $\mathcal{S}_4$ in the interval $[0, 2^N - 1]$.
  (5) Normalize $\mathcal{S}_4$ (dividing each member by $2^N - 1$).
  (6) The ensuing series is the output one $\mathcal{S}_{\rm OUT}$, again a series of floating-point numbers in the interval $[0, 1]$.

  Alternatively, the *least* significative digit (LSB) may be employed above, instead of the MSB one. The concomitant procedure is, of course, identical to the one summarized above, but now each symbol (number) in $\mathcal{S}_3$ represents the parity sequence of N consecutive members of the original CHTS. Note that information is lost by discretization because many points of the N-dimensional embedding space will share the same symbol ($\mathcal{S}_{\rm OUT}$) in the final timeseries.
- **Skipping:** This is a quite different approach (see Table 2). We deal with a two-stage process:
  (1) Partition the original CHTS $\mathcal{S}_{\rm IN}$ into groups of length d, without superposition, regarding each group as a point in a d-dimensional embedding space $\mathcal{E}$. This entails that in the embedding space each point represents now a d-length vector of IEEE-floating-point numbers.
  (2) Attach to each point in $\mathcal{E}$ a symbol consisting of only one coordinate (for example, the d-th one), generating thereby the output timeseries (STS) $S_{\rm OUT}$ of Table 2.

  Note that $d - 1$ values of $\mathcal{S}_{\rm IN}$ are "skipped" to get the STS $\mathcal{S}_{\rm OUT}$ which originates the name **Skipping** for this technique. In other words, we employ, instead of the original map f, its d-times iterated one f^d. This randomization technique is routinely (and successfully) used with piecewise linear maps in many applications [4]. In Table 2, an example with different values of d is displayed.

3. Normalized entropy and statistical complexity measures

As explained in Section 1, the second contribution of the present endeavour is a "two-probabilities" representation plane (see below), that uses H_S and $C_{\rm MPR} = Q \cdot H$ as coordinates. In Ref. [15] the disequilibrium Q was built using Wootters' statistical distance and taking H as a normalized Shannon entropy (see Ref. [15] and references therein for details). The ensuing SCM is neither an intensive nor extensive quantity in the thermodynamic sense, although it yields useful results. A natural SCM improvement is *to give it an intensive character*, as achieved in Ref. [16]. The concomitant SCM version is (i) able to grasp essential details of the dynamics, (ii) an intensive quantity and, (iii) capable of discerning among different degrees of periodicity and chaos [20]. The ensuing measure, to be referred to as the intensive statistical complexity, is a functional $C_{\rm MPR}[P]$ that characterizes the probability distribution function P associated with the time series of length M, generated by the dynamic system under study. It writes

$$C_{\rm MPR}[P] = Q_J[P, P_e] \cdot H_S[P], \tag{1}$$

where

$$H_S[P] = S[P]/S_{\max} = \left[-\sum_{j=1}^{N} p_j \ln(p_j) \right] \Big/ S_{\max}, \tag{2}$$

with $S_{\max} = S[P_e] = \ln N$, $(0 \leq H_S \leq 1)$, while N represent the total number of states of the system in phase space. We denote by $P_e = \{1/N, \ldots, 1/N\}$ the *uniform* distribution, while S stands for Shannon's entropy. Following the nomenclature introduced in Ref. [11] Q_J is the above referred to "disequilibrium", defined in terms of the extensive Jensen-Shannon divergence [16] (it induces a squared metric, in contrast to the Kullback–Leiber divergence) and writes

$$Q_J[P, P_e] = Q_0 \cdot \{S[(P + P_e)/2] - S[P]/2 - S[P_e]/2\}, \tag{3}$$

with Q_0 a normalization constant $(0 \leq Q_J \leq 1)$ that reads

$$Q_0 = -2\left\{\left(\frac{N+1}{N}\right)\ln(N+1) - 2\ln(2N) + \ln N\right\}^{-1}. \tag{4}$$

We see that the disequilibrium Q_J is an intensive quantity that reflects on the system's "architecture", being different from zero only if there exist "privileged", or "more likely" states among the accessible ones. $C_{\mathrm{MPR}}[P]$ quantifies the presence of correlational structures as well [15,16]. The opposite extremes of perfect order and maximal randomness possess no structure to speak of and, as a consequence, $C_{\mathrm{MPR}}[P] = 0$. In between these two special instances a wide range of possible degrees of physical structure exist, degrees that should be reflected in the features of the underlying probability distribution.

Both quantifiers $H_S[P]$ and $C_{\mathrm{MPR}}[P]$ can be calculated for any probability distribution function P. However P itself is not a uniquely defined object and several approaches have been employed in the literature so as to "extract" it from the given time series. Just to mention some frequently used extraction procedures: (a) amplitude-statistics [28], (b) binary symbolic-dynamics [29], (c) Fourier analysis [30], (d) wavelet transform [31,32], (e) partition entropies [8], (f) permutation entropy [33,34], (g) discrete entropies [35], etc. There is ample freedom to choose among them. An essential aspect of our work refers to this election. In fact we work with **two different PDF's**: one based on amplitudes statistics and the other devised via the attractor reconstruction procedure proposed by Bandt and Pompe [33], usually called in the literature the *permutation entropy*. The reason for this double PDF-choice is that the one based on amplitudes-statistics reflects on changes produced by each randomizing technique over the chaotic map's invariant-measure, while the Bandt–Pompe's procedure's PDF reflects on the mixing quality of the map under analysis (see the Appendix). As for the concomitant details:

- For extracting P via amplitude statistics, divide the interval $[0, 1]$ into a finite number n_{bin} of non overlapping subintervals A_i: $[0, 1] = \bigcup_{i=1}^{n_{\mathrm{bin}}} A_i$ and $A_i \bigcap A_j = \emptyset\ \forall i \neq j$. Note that N in Eq. (2) is equal to n_{bin}. We then employ the usual histogram method, based on counting the relative frequencies of the time series values within each subinterval. Of course, in this approach the temporal order of the time-series plays no role at all. The quantifiers obtained via the ensuing PDF are called in this paper $H_S^{(\mathrm{hist})}$ and $C_{\mathrm{MPR}}^{(\mathrm{hist})}$. Let us stress that for a timeseries with finite length M it is relevant to consider an optimal value of n_{bin}, as will be explained in Section 4. Finite-size effects for the estimation of Shannon's entropy have already been considered in the literature. Enlightening results can be consulted in Ref. [36].
- In extracting P by recourse to the Bandt–Pompe method (BPM) [33] the resulting probability distribution P is based on the details of the attractor reconstruction procedure. *Causal information* is, consequently, duly incorporated into the construction-process that yields P. The quantifiers obtained via the ensuing PDF are called in this paper $H_S^{(\mathrm{BP})}$ and $C_{\mathrm{MPR}}^{(\mathrm{BP})}$. A notable Bandt–Pompe result is getting a clear improvement of the Information Theory based quantifiers obtained by using their P-generating algorithm [6,7,19,20,22,23,37]. The extracting procedure is as follows: For the timeseries $\{x_t : t = 1, \ldots, M\}$ and an embedding dimension $D > 1$, we look for "ordinal patterns" of order D [9,33,34] generated by

$$(s) \mapsto \left(x_{s-(D-1)}, x_{s-(D-2)}, \ldots, x_{s-1}, x_s\right), \tag{5}$$

which assign to each "time s" a D-dimensional vector of values pertaining to the times $s, s-1, \ldots, s-(D-1)$. Clearly, the greater the D-value, the more information on "the past" is incorporated into these vectors. By the "ordinal pattern" related to the time (s) we mean the permutation $\pi = (r_0, r_1, \ldots, r_{D-1})$ of $(0, 1, \ldots, D-1)$ defined by

$$x_{s-r_{D-1}} \leq x_{s-r_{D-2}} \leq \cdots \leq x_{s-r_1} \leq x_{s-r_0}. \tag{6}$$

Table 3
r_{mix} as a function of the iteration-order j for the TWBM and LOG chaotic maps, respectively

j	LOG	TWBM
1	0.56789	0.333333333
2	0.31848	0.111111111
3	0.13290	0.037037037
4	0.05788	0.012345679
5	0.03646	0.004115226
6	0.01791	0.001371742
7	0.01152	0.000457247
8	0.00515	0.000152416

In order to get a unique result we consider that $r_i < r_{i-1}$ if $x_{s-r_i} = x_{s-r_{i-1}}$. Thus, for all the $D!$ possible permutations π of order D, the probability distribution $P = \{p(\pi)\}$ is defined by

$$p(\pi) = \frac{\sharp\{s|s \leq M - D + 1; (s) \texttt{ has type } \pi\}}{M - D + 1}. \tag{7}$$

In the last expression the symbol $\sharp$ stands for "number". We then evaluate the normalized entropy $H_S^{(\mathrm{BP})}$ and the intensive statistical complexity measure $C_{\mathrm{MPR}}^{(\mathrm{BP})}$ using *this* "permutation" probability distribution. The advantages of the Bandt–Pompe method reside in (a) its simplicity, (b) the associated extremely fast calculation-process, (c) its robustness in presence of observational and dynamic noise, and (d) its invariance with respect to nonlinear monotonous transformations. The Bandt–Pompe's methodology is not restricted to a timeseries representative of low dimensional dynamic systems but can be applied to any type of time series (regular, chaotic, noisy or reality based), with a very weak stationary assumption (for $k = D$, the probability for $x_t < x_{t+k}$ should not depend on t [33]). One also assumes that enough data are available for a correct attractor reconstruction. Of course, the embedding dimension D plays an important role in the evaluation of the appropriate probability distribution because D determines the number of accessible states $D!$. Also, it conditions the minimum acceptable length M of the timeseries that one needs in order to work with a reliable statistics. In relation to this last point Bandt and Pompe suggest, for practical purposes, to work with $3 \leq D \leq 7$ with a time lag $\tau = 1$. This is what we do here (in present work we use $D = 6$).

In this paper the representation plane has $H_S^{(\mathrm{hist})}$ as the x-coordinate and $C_{\mathrm{MPR}}^{(\mathrm{BP})}$ as the y-coordinate. Note that two probability distributions are thus being employed. This is an essential feature that allows us to obtain the results to be reported next.

4. Results and discussion

The examples below illustrate the preceding considerations.

- The Three Way Bernoulli Map (TWBM) given by

$$x_{n+1} = \begin{cases} 3x_n & \text{if } 0 \leq x_n \leq 1/3 \\ 3x_n - 1 & \text{if } 1/3 < x_n \leq 2/3 \\ 3x_n - 2 & \text{if } 2/3 < x_n \leq 1. \end{cases} \tag{8}$$

This map shares with many others (the fourway tailed shift, the threeway tailed shift, the skew tent, etc.) a uniform invariant density ρ_{inv} (see Appendix) over the interval [0, 1] while the mixing constant of the whole family of maps f^j is given by $r_{\mathrm{mix}}^j = (1/3)^j$ (see Table 3).

- Let us consider now the Logistic Map (LOG) given by

$$x_{n+1} = 4x_n(1 - x_n), \tag{9}$$

whose natural invariant density can be exactly determined, being expressed as

$$\rho_{\mathrm{inv}}(x) = \frac{1}{\pi\sqrt{x(1-x)}}. \tag{10}$$

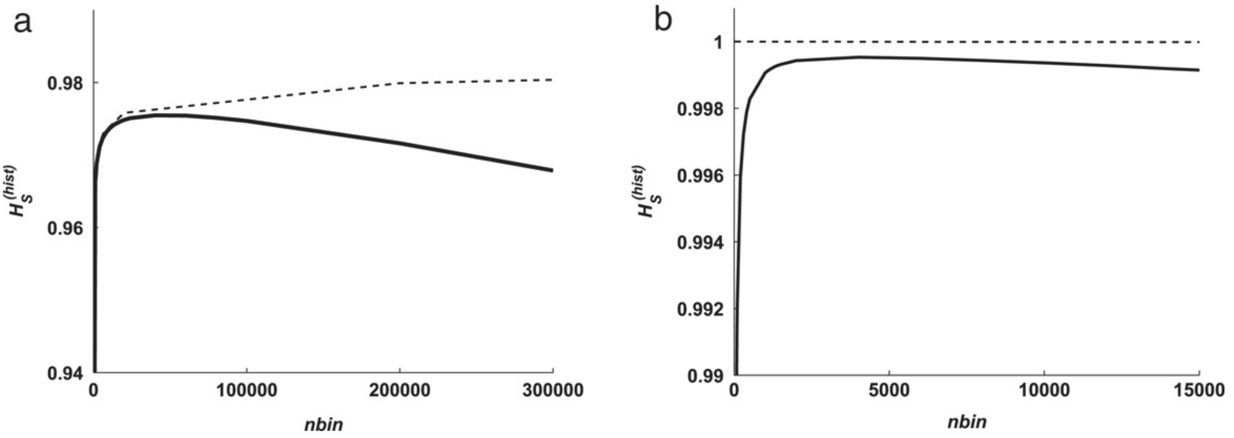

Fig. 1. Shannon Entropy $H_S^{(\text{hist})}$ as a function of the number of bins n_{bin} for different number of data (file sizes) $M = 10^6$ (solid line) and $M = 5 \times 10^7$ (dashed line). (a) LOG map, (b) TWB map.

The ensuing r_{mix}-values are also displayed in Table 3. They have been obtained by recourse to the Transfer Operator Method, as described in Ref. [2].

As pointed out in the previous section, some considerations regarding the number of subintervals in the histogram evaluation are in order. Fig. 1 displays $H_S^{(\text{hist})}$ as a function of $n_{\text{bin}} = N$ for LOG (Fig. 1(a)) and TWBM (Fig. 1(b)). Two different lengths M were considered: 1×10^6 and 5×10^7. In the case of these two M values the entropy $H_S^{(\text{hist})}$ first increases with $n_{\text{bin}} = N$, reaches a maximum, and finally decreases. It would thus seem plausible that n_{bin} must tend to ∞ so as to get the maximum entropy $H_S^{(\text{hist})}$. However, Fig. 1 shows that there exists an optimum value (it is 5×10^4 for $M = 1 \times 10^6$). The reason for this behaviour is the finite time-series' length $M < \infty$. Thus, the number of points within each subinterval decreases as n_{bin} increases, and, consequently, we cannot have a good statistics with n_{bin} values that are larger than the optimum one. This is because, in the case of a small data-file, a finer grid will not add new accessible states to the system. In consequence, the term $\sum p_j \log p_j$ does not change, although the normalization constant increases with n_{bin}, thus making H_s a decreasing function for these finer grids. The plateau-size for which $H_S^{(\text{hist})}$ remains almost constant does grow as the file-size increases. Another important issue is whether putative special features of the particular data set used can have some influence on the entropy computation. In order to confront the issue one generates, for a given length M, several "surrogates" iterating the map from a different initial condition. Then one evaluates the mean value (over the surrogates) of the quantifier (for example $\widehat{H_S} \equiv \langle H_S \rangle$). In this work convergence tests have been made employing 8 surrogates with $M = 5 \times 10^7$ data each and $\widehat{H_S^{(\text{hist})}}$ is taken as the entropy value in subsequent considerations. We have verified that neither a higher number of surrogates or a larger file-size affects the five most significant decimals of $\widehat{H_S^{(\text{hist})}}$, the x-coordinate of our representation plane. For notation simplicity the wide hat symbol is omitted.

The effects of the two randomization techniques discussed above on $H_S^{(\text{hist})}$ are illustrated in Fig. 2 for the LOG-instance as a function of n_{bin}. In Fig. 2(a) the dashed line is obtained with IEEE floating point numbers and the solid line with small circles shows the effect of a 16 bits-**Discretization**.

Applying the most significative digit (MSB) treatment in concocting the desired "symbolic" time-series (STS) (solid line with little "squares" in the graph), we ascertain that the ensuing entropic values considerably increase, as expected.

The maximum (valid) n_{bin}-value is $2^n - 1$ (65535 for 16-bits), and the maximum entropy is obtained precisely for this value. Furthermore, the n-bits representation has the same entropy than the floating point numbers representation only when the number of possible values 2^n is an exact multiple of n_{bin} (note in Fig. 2(a) the coincidence for values 65536, 65536/2, 65536/3, etc). This is an usual situation when the histogram of a discrete distribution is made. In fact the histogram grid is equivalent to collapse a subinterval of length $1/n_{\text{bin}}$ into a unique value. On the other hand a n-bits discretization collapses each subinterval of length $1/2^n$ into a unique natural value and consequently induces another grid. That is the reason the entropy of the discrete series is lower than that of the floating point representation if $n_{\text{bin}} \neq 2^n/k$, $(k = 1, 2, \ldots)$ as Fig. 2(a) shows.

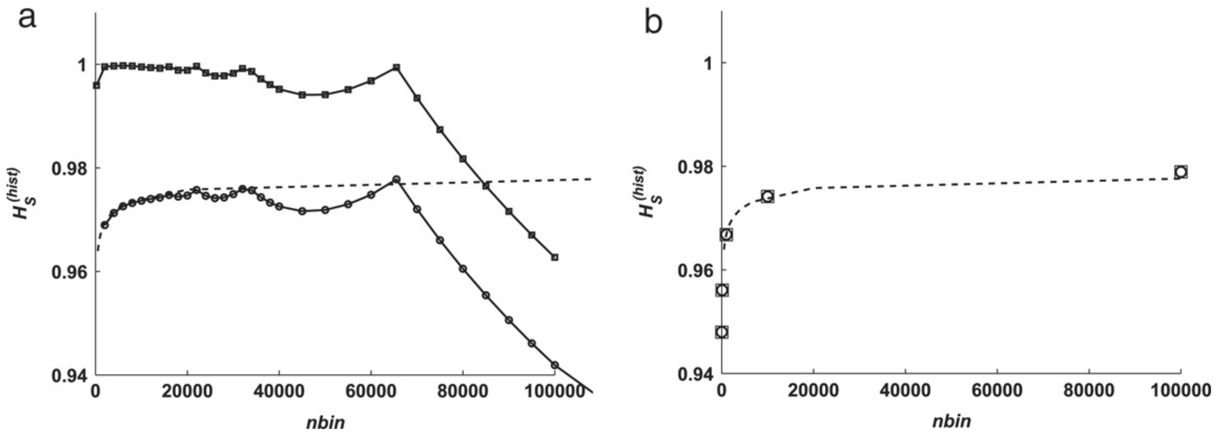

Fig. 2. Shannon Entropy $H_s^{(\text{hist})}$ as a function of the number of bins n_{bin} for both randomization process for the LOG map. (a)*Discretization:* IEEE floating point (dashed line); 16 bits (solid line with circles); MSB (solid line with squares). (b) *Skipping:* IEEE floating point — original map f (dashed line); second iterate map f^2 (circles); forth iterate map f^4 (squares).

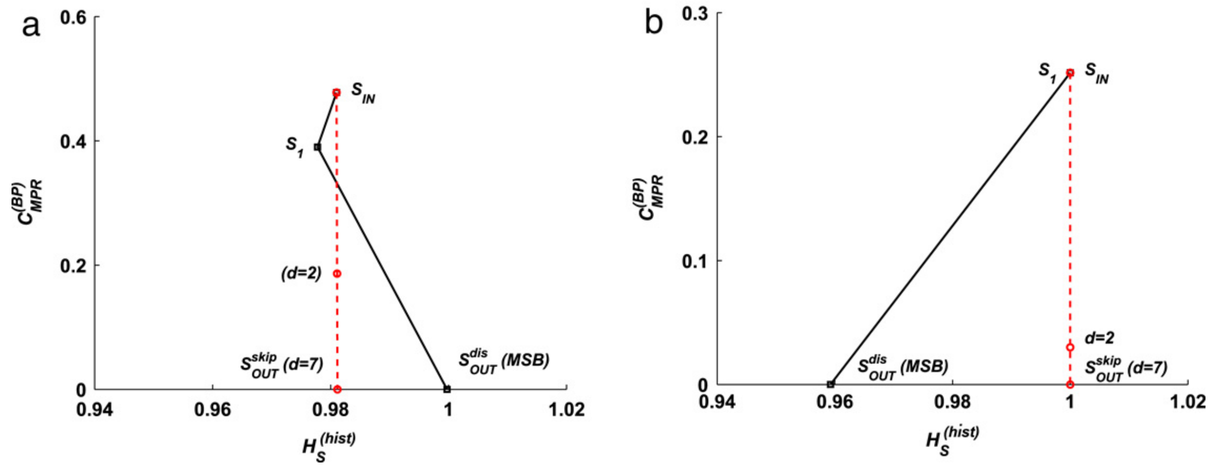

Fig. 3. The representation plane $H_S^{(\text{hist})}$–$C_{\text{MPR}}^{(\text{BP})}$ for both randomization processes. (a) For the LOG map **Discretization** produces an STS with the ideal coordinates $(1, 0)$ but **Skipping** is not capable to improve $H_S^{(\text{hist})}$ and the STS has coordinates $(0.98, 0)$; (b) For the TWBM map **Discretization** decreases the y-coordinate to the ideal value $C_{\text{MPR}}^{(\text{BP})} = 0$ but it also decreases the x-coordinate to $H_S^{(\text{hist})} = 0.96$ while **Skipping** improves the y coordinate and it does not change the entropy. Then the STS reaches the ideal point $(1, 0)$.

In Fig. 2(b) **Skipping** effects are illustrated. The dashed line is again that obtained with IEEE floating point numbers using the original map f. The small circles correspond to the second iterate of the map, f^2, and the squares to f^4. Note that the same entropy $H_S^{(\text{hist})}$ is obtained in all cases. This behaviour confirms that **Skipping** does not affect the P when amplitude statistics is used [4]. As a final point regarding this figure we note that the Shannon entropy of the STS obtained using the least significant digit (LSB) in the randomization process yields entropy values almost identical to the MSB-ones (they are not displayed in Fig. 2(a)).

Our most significative result is depicted in Fig. 3, for both randomization techniques and for the LOG (a) and TWBM (b) maps: the $H_S^{(\text{hist})} \times C_{\text{MPR}}^{(\text{BP})}$-plane. The goal of any randomization technique is to approach the ideal $(1, 0)$-point in this plane, as closely as possible, since there randomization is optimal. The planar representation clearly helps in ascertaining features of the randomization-approaches quality.

In the LOG map the initial CHTS point in the $H_S^{(\text{hist})} \times C_{\text{MPR}}^{(\text{BP})}$-plane is approximately $(0.98, 0.5)$. This point corresponds to the entropy of a nonuniform histogram with high complexity value indicating that it has a geometric structure. **Skipping** destroys this structure pushing the y-coordinate toward the ideal value 0, but the histogram is not modified by it and, consequently, the x-coordinate does not change. In spite of the fact that the STS (indicated as S_{OUT}^{skip}) has better statistical properties than the CHTS (indicated as S_{IN}), it does not reach the ideal point $(1, 0)$. On the other hand, **Discretization** decreases the y-coordinate and increases the x-coordinate so that the STS does reach the ideal coordinates $(1, 0)$. The reason is that the original CHTS follows the logistic equation with a nonuniform histogram. However, the number of values in $[0, 0.5)$ is almost equal to the corresponding number in $[0.5, 1]$, while the MSB technique produces a random sequence of $1's$ and $0's$ (a pseudo random bit generator). The same applies to the LSB alternative.

Note that the CHTS generated by the TWBM map "has" the ideal value $H_S^{(\text{hist})} = 1$ from the very beginning. For the TWBM map **Discretization** decreases the y-coordinate to the ideal value $C_{\text{MPR}}^{(\text{BP})} = 0$ and at the same time it also decreases the x-coordinate to $H_S^{(\text{hist})} = 0.96$. We conclude then that the best choice is **Skipping**, that improves the y-coordinate without any change of the x-coordinate, allowing for the STS to reach the ideal point $(1, 0)$.

5. Summary

After performing extensive simulation-runs with many other maps, characterized by uniform as well as by nonuniform natural invariant measures (see the Appendix), we are in a position to summarize our main conclusion: **Skipping** is better than **Discretization** for maps with a uniform natural invariant measure, and viceversa in the non uniform case. The reason is clear: **Skipping** decreases $C_{\text{MPR}}^{(\text{BP})}$ without changing $H_S^{(\text{hist})}$. Further, the endpoint in our representation plane of the symbolic timeseries lies in very close proximity to the ideal point $(1, 0)$ because maps with uniform natural invariant distributions have $H_S^{(\text{hist})}$ from the very beginning and only the mixing properties of the map can be improved. The number of iterations that must be used depends on the number of significant figures required. On the contrary, for maps with non-uniform invariant measures (the LOG one is a paradigmatic example) **Skipping** diminishes the complexity $C_{\text{MPR}}^{(\text{BP})}$ without changing $H_S^{(\text{hist})}$ and, consequently, the ideal point $(1, 0)$ is never reached, while **Discretization** decreases $C_{\text{MPR}}^{(\text{BP})}$ and increases $H_S^{(\text{hist})}$ as well, thus leading our representative point towards the ideal one. A different and very fruitful approach to study complex systems are *Recurrence Plots* (RP) (see the excellent review by Marwan et al. [38]). Such an analysis is very efficient, specially for nonsuperlong sequences. Then we have used Marwan et al.'s tools to analyse the series in Fig. 3 and found that *measures of complexity based on diagonal lines on the RP (see, for example, DET in the CRP toolbox for Matlab* [39]*) do support the main conclusions drawn in the above paragraph*. A more detailed comparison between both methods will be published elsewhere.

Acknowledgments

This work was partially supported by the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina (PIP 5569/04, PIP 5687/05, PIP 6036/05), ANPCyT, Argentina (PICT 11-21409/04) and Universidad Nacional de Mar del Plata. OAR gratefully acknowledge support from Australian Research Council (ARC) Centre of Excellence in Bioinformatics, Australia.

Appendix. Probabilistic description of a map evolution

Two central concepts are essential for our present purposes: those of (i) invariant probability measure and (ii) mixing. We revisit them next. Let a timeseries over $[0, 1]$ generated by a $1D$-chaotic map f by iterating the map starting from a single initial value x_0, so that the CHTS $S_{\text{IN}} = \{x_0, x_1, \ldots, x_\infty\}$ ensues. Another approach is also possible by considering the evolution of an ensemble of different initial values [2]. The relative frequency of initial values in a subset $A \subset [0, 1]$ can be interpreted as the probability $\mu_0(A)$ of having an initial value $x_0 \in A$ and is called the probability measure of A

$$\mu_0(A) = \int_A \rho_0(x)\,\mathrm{d}x. \tag{A.1}$$

Eq. (A.1) defines a probability density ρ_0 over the whole phase-space $[0, 1]$ (for more refined mathematical details see Ref. [2]). We now define $\mu_n(A)$ as the probability of finding an iterate x_n in the subset A. The appropriate density is now

$$\mu_n(A) = \int_A \rho_n(x)\,\mathrm{d}x. \tag{A.2}$$

Thus, instead of the evolution of a particular initial condition, we can consider that of ρ_0 in probability space. The concomitant process is generated by the so-called Perron-Frobenius operator $L_f : PDF \to PDF$ [2–4] in the form

$$\rho_{n+1} = L_f \rho_n, \tag{A.3}$$

whose two largest (in absolute value) eigenvalues $(\eta_0;\ \eta_1)$ acquire special relevance. Conservation of probability leads to the following (trivial) condition for arbitrary subsets A

$$\mu_{n+1}\ (A) = \mu_n\ (f^{-1}(A)), \tag{A.4}$$

where $f^{-1}(A)$ is the pre-image of A, i.e. the set of all points that are mapped onto A by one iteration step. Eq. (A.4) tells us that the relative frequency of iterates x_{n+1} in the subset A must be equal to the relative frequency of iterates x_n in the subset $f^{-1}(A)$. Particular interest is attached to invariant probability measures (also called invariant measures) that satisfy

$$\mu_{n+1}\ (A) = \mu_n\ (A), \tag{A.5}$$

and their corresponding invariant PDF's, also called invariant densities. A map is called ergodic if for any integrable test function $\mathcal{T}(x)$ the time average equals the ensemble average. For such maps the time average does not depend on the initial x_0 [2]. There may exist several invariant measures for an ergodic map, but only one of them is really important in the sense that, if we iterate a randomly chosen initial point, the iterates will be distributed according to this measure "almost surely". This measure is called the *natural invariant measure* μ_{inv} and its corresponding PDF is called the *natural invariant density* ρ_{inv}.

A map f is called *mixing* if an arbitrary, smooth initial probability density ρ_0 converges to ρ_{inv} [40–44]. Since each element of the initial set of initial conditions (with PDF ρ_0) evolves in time independently of the others, the operator L_f is linear in spite of the map's nonlinearity. ρ_{inv} is the L_f-eigenvector corresponding to the eigenvalue $\eta_0 = 1$. It can be analytically obtained only in a few cases but it may be approximated numerically. The eigenvalue with the second largest absolute value, η_1, has a special and quite important meaning: its absolute value gives the "speed" with which one approaches the natural invariant density, called the *mixing constant* r_{mix} of the chaotic map. The smaller r_{mix}, the faster the mixing process. It has been shown that maps $f^2, f^3, \ldots$ obtained by iterating a given f share a common ρ_{inv} but the pertinent r_{mix} decreases as the number of iterations increases (see Table 3). In other words, *the iterated maps are better mixing maps* [4].

References

[1] For PRNG's quality tests see for instance the following web pages: http://stat.fsu.edu/pub/diehard/, http://www.iro.umontreal.ca/simardr/random.html, http://csrc.nist.gov/rng/.

[2] C. Beck, F. Schlögl, Thermodynamics of chaotic systems: An introduction, in: Cambridge Nonlinear Science Series 4, Cambridge University Press, 1997.

[3] A. Lasota, M.C. Mackey, Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics, 2nd edition, in: Applied Mathematical Sciences, vol. 97, Springer Verlag, 1994.

[4] G. Setti, G. Mazzini, R. Rovatti, S. Callegari, Proc. IEEE 90 (2002) 662.

[5] C.M. González, H.A. Larrondo, O.A. Rosso, Physica A 354 (2005) 281.

[6] H.A. Larrondo, C.M. González, M.T. Martin, A. Plastino, O.A. Rosso, Physica A 356 (2005) 133.

[7] H.A. Larrondo, M.T. Martin, C.M. González, A. Plastino, O.A. Rosso, Phys. Lett. A 352 (2006) 421.

[8] W. Ebeling, R. Steuer, M.R. Titchener, Stoch. Dyn. 1 (2001) 1.

[9] K. Keller, H. Lauffer, Int. J. Bifurcation Chaos 13 (2003) 2657.

[10] C.E. Shannon, Bell Syst. Technical J. 27 (1948) 379 and 623.

[11] R. López-Ruiz, H.L. Mancini, X. Calbet, Phys. Lett. A 209 (1995) 321.

[12] R. Wackerbauer, A. Witt, H. Atmanspacher, J. Kurths, H. Scheingraber, Chaos Solitons Fractals 4 (1994) 133.

[13] D.P. Feldman, J.P. Crutchfield, Phys. Lett. A 238 (1998) 244.

[14] C. Anteneodo, A.R. Plastino, Phys. Lett. A 223 (1997) 348.

[15] M.T. Martin, A. Plastino, O.A. Rosso, Phys. Lett. A 311 (2003) 126.

[16] P.W. Lamberti, M.T. Martin, A. Plastino, O.A. Rosso, Physica A 334 (2004) 119.

[17] A.M. Kowalski, M.T. Martin, A. Plastino, A.N. Proto, O.A. Rosso, Phys. Lett. A 311 (2003) 180.

[18] A.M. Kowalski, M.T. Martin, A. Plastino, O.A. Rosso, Int. J. Mod. Phys. B 19 (2005) 2273.

[19] A.M. Kowalski, M.T. Martin, A. Plastino, O.A. Rosso, Physica D 233 (2007) 21.

[20] O.A. Rosso, H.A. Larrondo, M.T. Martin, A. Plastino, M.A. Fuentes, Phys. Rev. Lett. 99 (2007) 154102.

[21] O.A. Rosso, M.T. Martin, A. Figliola, K. Keller, A. Plastino, J. Neurosci. Meth. 153 (2006) 163.
[22] O.A. Rosso, R. Vicente, C.R. Mirasso, Phys. Lett. A 372 (2008) 1018.
[23] L. Zunino, D.G. Pérez, M.T. Martín, A. Plastino, M. Garavaglia, O.A. Rosso, Phys. Rev. E 75 (2007) 021115.
[24] X. Calbet, R. Lopez-Ruiz, Phys. Rev. E 63 (2001) 066116.
[25] S. Callegari, G. Setti, P.J. Langlois, International Symposium on Circuits and Systems, ISCAS97, 1997, p. 781.
[26] M. Jessa, IEEE Trans. Circuits and Syst. I 49 (2002) 84.
[27] C. Beck, G. Röpstorff, Physica D 25 (1987) 173.
[28] M.T. Martin, Ph.D. Thesis, Department of Mathematics, Faculty of Sciences, University of La Plata, La Plata, Argentina, 2004.
[29] K. Mischaikow, M. Mrozek, J. Reiss, A. Szymczak, Phys. Rev. Lett. 82 (1999) 1114.
[30] G.E. Powell, I.C. Percival, J. Phys. A 12 (1979) 2053.
[31] S. Blanco, A. Figliola, R. Quian Quiroga, O.A. Rosso, E. Serrano, Phys. Rev. E 57 (1998) 932.
[32] O.A. Rosso, S. Blanco, J. Jordanova, V. Kolev, A. Figliola, M. Schürmann, E. Başar, J. Neurosci. Meth. 105 (2001) 65.
[33] C. Bandt, B. Pompe, Phys. Rev. Lett. 88 (2002) 174102.
[34] K. Keller, M. Sinn, Physica A 356 (2005) 114.
[35] J.M. Amigó, L. Kocarev, I. Tomovski, Physica D 228 (2007) 77.
[36] U. Schwarz, A.O. Benz, J. Kurths, A. Witt, Astron. Astrophys. 277 (1993) 215.
[37] O.A. Rosso, L. Zunino, D.G. Pérez, A. Figliola, H.A. Larrondo, M. Garavaglia, M.T. Martín, A. Plastino, Phys. Rev. E 76 (2007) 061114.
[38] N. Marwan, M.C. Romano, M. Thiel, J. Kurths, Phys. Rep. 438 (2007) 237.
[39] Cross Reference Plot Toolbox for Matlab, provided by TOCSY: http://tocsy.agnld.uni-potsdam.de.
[40] M. Dellnitz, G. Froyland, S. Sertl, Nonlinearity 13 (2000) 1171.
[41] D. Pingel, P. Schmelcher, F.K. Diakonos, Chaos 9 (1999) 357.
[42] A. Rogers, R. Shorten, D.M. Heffernan, Phys. Lett. A 330 (2004) 435.
[43] A. Lasota, J.A. Yorke, Trans. Amer. Math. Soc. 186 (1973) 481.
[44] J. Ding, A. Zhou, Physica D 92 (1996) 61.