ORIGINAL ARTICLE

Confidence Regions for Information-Theoretic Descriptors of Time Series

Eduarda T. C. Chagas[1*] | Marcelo Queiroz[2†] | Osvaldo A. Rosso[3‡] | Heitor S. Ramos[1*] | Christopher G. S. Freitas[2†] | Leonardo V. Pereira[2†] | Alejandro C. Frery[4§]

[1]Departamento de Ciência da Computação, Universidade Federal de Minas, Brazil

[2]Laboratório de Computação Científica e Análise Numérica, Universidade Federal de Alagoas, Brazil

[3]Instituto de Física, Universidade Federal de Alagoas, Brazil

[4]School of Mathematics and Statistics, Victoria University of Wellington, New Zealand

Correspondence
Eduarda T. C. Chagas, Departamento de Ciência da Computação, Universidade Federal de Minas, Brazil
Email: eduarda.chagas@dcc.ufmg.br

Funding information
CNPq, Fapeal

The **?** methodology has been used successfully in the analysis of time series. It consists of computing information theory descriptors, using a histogram of ordinal patterns, which are found in a 2D variety: the Entropy-Complexity plane. So far, the analysis of the dynamics underlying the time series has been carried out from two reference points: those corresponding to a deterministic time series and a series of white noise. In this article, we provide a first proposal for the construction of empirical confidence regions in the Entropy-Complexity plane for white noise models and we use these regions to verify whether we can capture the randomness of PRNGs in short sequences. The proposed methodology showed consistency and coherence in its results, managing to discriminate sequences of true random samples, capturing the randomness of generators previously analyzed in the literature and proving to be robust the addition of correlation structures.

KEYWORDS
Random number generators, Bandt–Pompe approach, Entropy-complexity plane, Information Theory

[*] Equally contributing authors.

1 | INTRODUCTION

Time Series carry valuable information about the system which produces the data. Their analysis is usually based on two approaches **?**: in the (natural) time and transformed domains (for instance, frequency and wavelet). In the context of time-domain analysis, a new methodology was proposed by **?** PermutationEntropyBandtPompe. Its approach is non-parametric and based on descriptors of Information Theory. Through this, the time series is transformed into ordinal patterns, with which a histogram is formed. Using such patterns, the resulting distribution becomes less sensitive to outliers and, as it does not depend on any model, can be applied to a variety of situations.

The Bandt-Pompe methodology and its variants have been used successfully in the analysis of many types of dynamics, receiving so far more than 2500 citations, according to the Journal of Citation Records. We found works using such an approach in multiple areas of scientific knowledge such as, for example, the study of electroencephalography signals using wavelet decomposition **?**, characterization of household appliances through their energy consumption **?**, online signature classification and verification **?**.

Each time series is described by a point in the range of $\mathbb{R}^2$, the entropy complexity plane. Two points are well known in this plane: those of white noise and a completely deterministic sequence. Through these references, we can characterize the time series according to the dynamics of its generating process. Based on this premise, studies with different applications managed to obtain relevant results from time series through information on the nature of the data provided by the $H \times C$ plane. Examples include the analysis of **?** echegoyen2020permutation in magnetoencephalography recordings of individuals suffering from mild cognitive impairment and individuals diagnosed with Alzheimer's disease by trajectories in the $H \times C$ plane, **?** InformationTheoryPerspectiveNetworkRobustness verified the effect of attacks on complex networks by displacing their points in the $H \times C$ plane, **?** CharacterizationVehicleBehaviorInformationTheory described the behavior of vehicles depending on the topology of cities, and **?** Chagas2020Characterization succeeded in expanding the use of such techniques for analyzing SAR texture images, making characterization, and classification of them.

In the context of theoretical work, chaotic and random component analysis has been developed with the aim of improving the understanding of the plane's properties. We can cite as highlights: **?** GeneralizedStatisticalComplexityMeasuresGeometricalAnalyticalProperties analyzed the logistic chaotic map and discuss the boundaries of the $H \times C$ plane. **?** De$_{M}$icco$_{2}$009$studied chaotic components in pseudo-random number generators$.? $Distinguishing Noise From Chaos tackle scale approach to analyze the interplay between chaotic and stochastic dynamics. With the knowledge of the expected variabi Complexity plane$(H\times C)$ is a good indicator of the results of Diehard tests for pseudorandom number generators. **?** De$_{M}$icco$_{2}$008$assessed ways of improving pseudorandom sequences by their representation in this plane$.? $Libor Invisible H$ plane representation.

Motivated by such works, in this paper we advance the state-of-the-art providing confidence regions for white noise points in the $H \times C$ plane. In the proposed approach, the input is a sequence of true random observations generated by a physical procedure, and the confidence regions are obtained by performing an orthogonal projection of the data onto a space of principal components analysis (PCA), thus eliminating the restrictions imposed by the curvilinear space of the Entropy-Complexity Plane. Our contributions can be summarized as follows:

- Provide the first contribution in construction that confidence regions in Entropy-Complexity Plane.
- Evaluate the discrimination power of these regions in the analysis of small random sequences generated by physical procedures and pseudorandom generators (PRNGs).

The paper is structured as follows: Section 1 introduces the elements of the study (the Bandt and Pompe method-

ology, the random deviates, and model). The confidence regions are presented in Section 6 with some analysis of this application, and the conclusions are discussed in Section 7.

2 | ENTROPY-COMPLEXITY PLANE IN THE LITERATURE

The first works on the characterization of white noises with permutation entropy arose from the need to discriminate them in relation to chaotic maps rosso2013characterization, xiong2020complexity, olivares2012contrasting. However, it was found that the measure of statistical complexity was able to efficiently quantify the performance of pseudorandom number generators, expanding the possibilities of using information theory descriptors with the Bandt-Pompe symbolization larrondo2002statistical, gonzalez2005statistical.

Table 1 presents a summary of the main works in the literature that perform analysis of non-chaotic algorithmic generators, according to their features (h, c). For this, we also provide the length T and embedding dimension D applied in the analysis of the time series. The following algorithmic generators were analyzed:

- Mother RNG, available in Marsaglia website **?** (MOT);
- Multiple with carry RNG (MWC) **?**;
- Combo RNG (COM) **?**;
- Lehmer RNG (LEH) **?**;
- Fractional Gaussian noise with $\alpha = 0$ (fGn);
- Fractional Brownian motion with $\alpha = 1.2$ (fBm);
- f^{-k} noise with $k = 0$;
- Linear Congruential Generator (LCG) **?**.

3 | BANDT-POMPE SYMBOLIZATION: A BACKGROUND

In our work, we consider that ordinal patterns are formed from white noise sequences using Bandt-Pompe symbolization and mapped in the two-dimensional plane of information theory descriptors, permutation entropy and statistical complexity. Through the space formed by this tuple of features, we were able to obtain regions of confidence and a test statistic that can discriminating random series.

3.1 | The Bandt-Pompe Methodology

Let $\mathcal{X} \equiv \{x_t\}_{t=1}^{T}$ be a real valued time series of length T, without ties. As stated by **?** PermutationEntropyBandtPompe in their seminal work:

> *"If the $\{x_t\}_{t=1}^{T}$ attain infinitely many values, it is common to replace them by a symbol sequence $\Pi \equiv \{\pi_j\}$ with finitely many symbols, and calculate source entropy from it".*

Also, as stressed by these authors,

> *"The corresponding symbol sequence must come naturally from the $\{x_t\}_{t=1}^{T}$ without former model assump-*

tions".

Let $\mathbb{A}_D$ (with $D \geq 2$ and $D \in \mathbb{Z}$) be the symmetric group of order $D!$ formed by all possible permutation of order D, and the symbol component vector $\boldsymbol{\pi}^{(D)} = (\pi_1, \pi_2, \ldots, \pi_D)$ so every element $\boldsymbol{\pi}^{(D)}$ is unique ($\pi_j \neq \pi_k \ \forall \ j \neq k$). Consider for the time series $\mathcal{X} \equiv \{x_t\}_{t=1}^{T}$ its time delay embedding representation, with embedding dimension $D \geq 2$ ($D \in \mathbb{Z}$) and time delay $\tau \geq 1$ ($\tau \in \mathbb{Z}$, also called "embedding time"):

$$\mathbf{X}_t^{(D,\tau)} = (x_t, x_{t+\tau}, \ldots, x_{t+(D-1)\tau}) \,, \tag{1}$$

for $t = 1, 2, \ldots, N$ with $N = T - (D-1)\tau$. Then the vector $\mathbf{X}_t^{(D,\tau)}$ can be mapped to a symbol vector $\boldsymbol{\pi}^{(D)} \in \mathbb{A}_D$. This mapping should be defined in a way that preserves the desired relation between the elements $x_t \in \mathbf{X}_t^{(D,\tau)}$, and all $t \in T$ that share this pattern (also called motif) have to mapped to the same $\boldsymbol{\pi}^{(D)}$. The two most frequent ways to define the mapping $\mathbf{X}^{(D,\tau)} \mapsto \boldsymbol{\pi}^{(D)}$ are:

1. ordering the ranks of the $x_t \in \mathbf{X}^{(D,\tau)}$ in chronological order (*Rank Permutation*) or,
2. ordering the time indexes according to the ranks of $x_t \in \mathbf{X}^{(D,\tau)}$ (*Chronological Index Permutation*);

see details in **?** BPRepeatedValuesChaos. Without loss of generality, in the following we will only use the latter.

Consider, for instance, the time series $\mathcal{X} = (1.8, 1.2, 3.2, 4.8, 4.2, 4.5, 2.3, 3.7, 1.2, .5)$ depicted in Fig. 1. Assume we are using patterns of length $D = 5$ with unitary time lag $\tau = 1$. The code associated to $\mathbf{X}_3^{(5,1)} = (x_3, \ldots, x_7) = (3.2, 4.8, 4.2, 4.5, 2.3)$, shown in black, is formed by the indexes in $\boldsymbol{\pi}^{(5)} = (1, 2, 3, 4, 5)$ which sort the elements of $\mathbf{X}_3^{(5,1)}$ in increasing order: 51342. With this, $\widetilde{\pi}^{(5)} = 51342$, and we increase the counting related to this motif in the histogram of all possible patterns of size $D = 5$.

The dash-dot line in Fig. 1 illustrates $\mathbf{X}_1^{(5,2)}$, i.e. the sequence of length $D = 5$ starting at x_1 with lag $\tau = 2$. In this case, $\mathbf{X}_1^{(5,2)} = (1.8, 3.2, 4.2, 2.3, 1.2)$, and the corresponding motif is $\widetilde{\pi}^{(5)} = 51423$.

Once all symbols have been computed, one obtains the histogram of proportions $\boldsymbol{h} = (h(j))_{1 \leq j \leq D!}$. This is an estimate of the (unknown, in general) probability distribution function of these patterns. The next step into the characterization of the time series is computing descriptors from this histogram.

The first descriptor is a measure of the disorder of the system. The most frequently used feature for this is the Normalized Shannon entropy, defined as

$$H(\boldsymbol{h}) = -\frac{1}{\log D!} \sum_{j=1}^{D!} h(j) \log h(j), \tag{2}$$

with the convention that terms in the summation for which $h(j) = 0$ are null. This quantity is bounded in the unit interval, and is zero when $h(j) = 1$ for some j (and, thus, all other bins are zero), and one when $h(j) = 1/D!$ for every j (the uniform probability distribution function).

Although very expressive, the Normalized Shannon Entropy is not able to describe all possible underlying dynamics. In particular, for intermediate values of H, there is a wide variety of situations worth characterizing. To this aim, **?** LopezRuiz1995 proposed using Q, the disequilibrium, a measure of how far $\boldsymbol{h}$ is from an equilibrium or non-informative distribution. They employed the Euclidean distance between $\boldsymbol{h}$ and the uniform probability distribution function.

With this, they proposed $C = HQ$ as a measure of the Statistical Complexity of the underlying dynamics. A time

TABLE 1 Result of the main works of white noise sequences analysis in the $H \times C$ plane.

Reference	PRNG	T	D	H	C	Is white noise?	p-value
? (?)	MOT	NA	6	$\cong 0.9969$	$\cong 0$	no	NA
? (?)	MWC	65536	NA	$\cong 1$	0.3	yes	NA
	MOT	65536	NA	$\cong 1$	0.3	yes	NA
	COM	65536	NA	$\cong 1$	0.05	yes	NA
? (?)	LEH	5×10^6	5	NA	10^{-4}	yes	NA
	MOT	5×10^6	5	NA	10^{-4}	yes	NA
	MWC	5×10^6	5	NA	10^{-4}	yes	NA
? (?)	fGn	2×10^{15}	6	$\cong 0.998$	NA	yes	NA
	fBm	2×10^{15}	6	$\cong 0.993$	NA	yes	NA
	f^{-k}	2×10^{15}	6	$\cong 0.997$	NA	yes	NA
? (?)	LCG	1×10^7	6	0.997871	0.005101	no	NA
? (?)	fGn	2×10^{17}	6	$\cong 1$	$\cong 0$	yes	NA
	f^{-k}	2×10^{17}	6	$\cong 1$	$\cong 0$	yes	NA

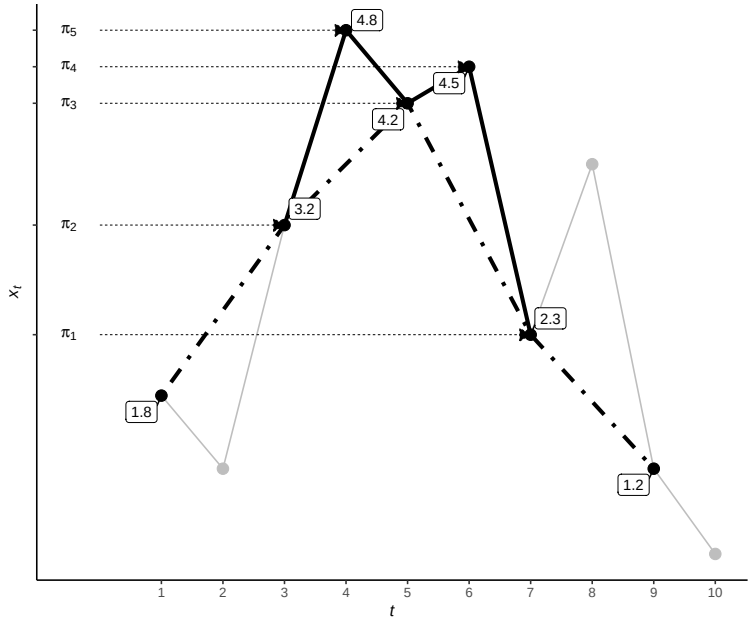

FIGURE 1 Illustration of the Bandt and Pompe coding

series can then be mapped into a point in the $H \times C$ plane.

3.2 | The Entropy-Complexity Plane

We illustrate the use of the Entropy-Complexity ($H \times C$) with the following time series:

- Colored k-noise: white ($k = 0$), $k = -1/2$, pink ($k = 1$), $k = 3/2$, red ($k = 2$), $= 5/2$, and $k = 3$;
- Chaotic logistic series $x_t = r x_{t-1}(1 - x_{t-1})$, with $r = 3.6$ and 4;
- Deterministic series: monotonic increasing ($\log(x_t + 0.1)$, $x_t = \{1, 2, \ldots, 10^4\}$) and periodic ($\sin(2x_t)\cos(2x_t)$, with $0 \leq x_t \leq 2\pi$ over ten thousand equally spaced points).

In all cases, we used $D = 6$ and $\tau = 1$. Fig. 2 shows nine of the histograms produced by these series using the Mersenne-Twister pseudorandom number generator; we omitted those corresponding to the deterministic series, as they produce one and two nonzero bins.

Fig. 3 shows the $H \times C$ plane with the bounds for $D = 6$, the time series and the points they were mapped onto. The points due to f^{-k} noises appear joined by dotted segments. It is noticeable that deterministic patterns have more complexity than random ones. Also, points related to f^{-k} noises tend to clutter for $k < 1$, having the highest entropy values, as can be seen in Fig. 4.

Fig. 4 shows the rightmost lower corner of the $H \times C$ plane, emphasizing the location of the white ($k = 0$), $k = -1/2$, and pink ($k = 1$) noises.

The focus of our study is to assess pure randomness by analyzing the empirical distribution of the points produced by true random sequences, providing regions of confidence in the $H \times C$ plane.

4 | PROPOSED METHOD

In this section, we formalize the task of building a confidence region in the Entropy-Complexity manifold. Then, we present our proposal to change space through the algorithm of the principal components analysis. Our goal is to find a latent space representative of the data, without the restrictions of a curvilinear space. Through this new representation of the data, we calculate empirical regions with different levels of confidence. Finally, after calculating these regions, we build a test statistic that determines the probability that a given sequence belongs to the distribution of the points provided.

4.1 | Overall Framework

The structure of our proposal is shown in Fig. 5. It consists of two steps:

- **Empirical confidence region**: With the data present in a Euclidean plane, we can easily calculate empirical regions that involve the data with a certain level of confidence.
- **Construction of a test statistic**: To measure the similarity of new data sequences with the empirical points, a test statistic was proposed. By acquiring a p-value less than 0.5, we can reject the null hypothesis, which states that such data belong to the empirical probability distribution used for the construction of the confidence region.

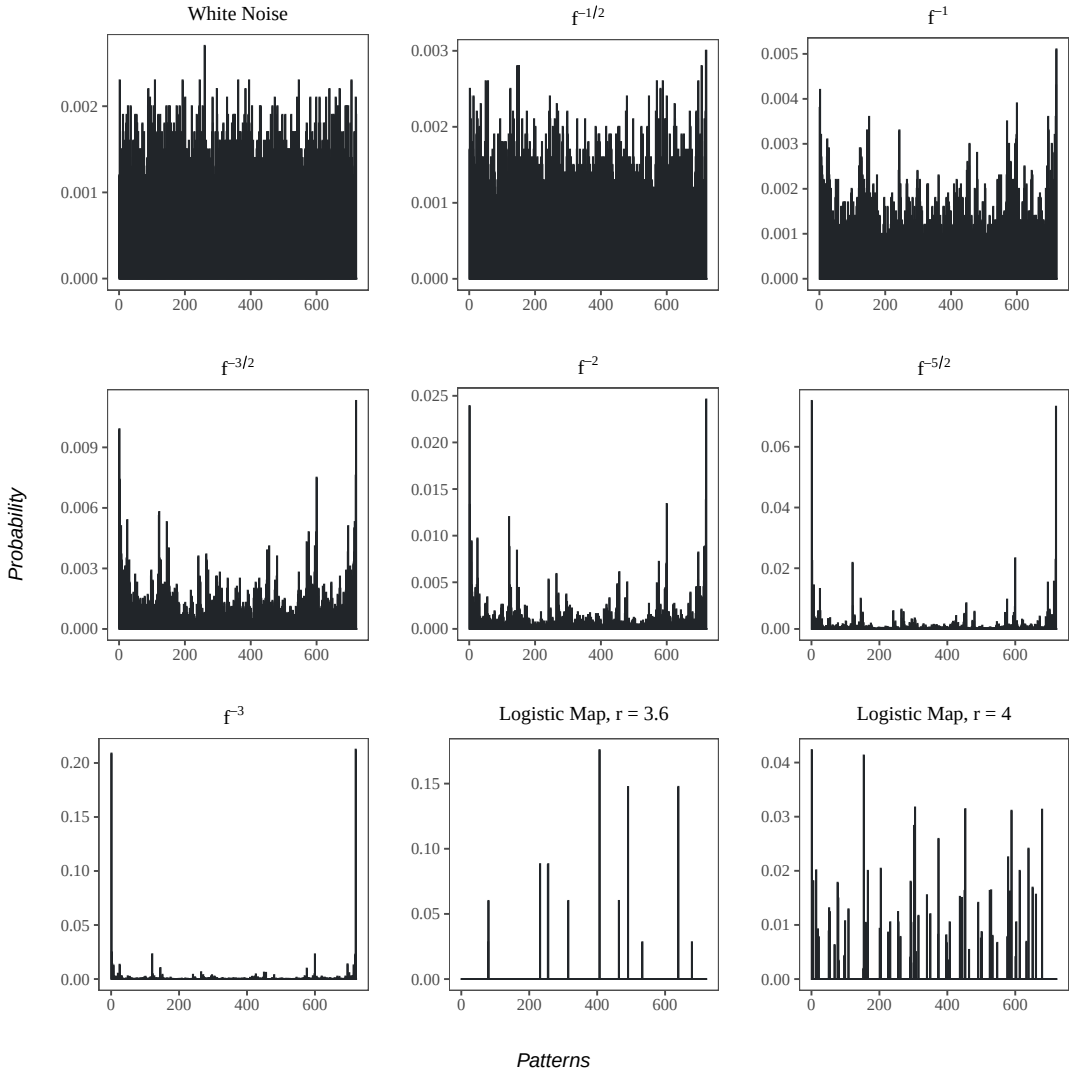

FIGURE 2 Patterns histograms of selected time series, with $D = 6$ and $\tau = 1$

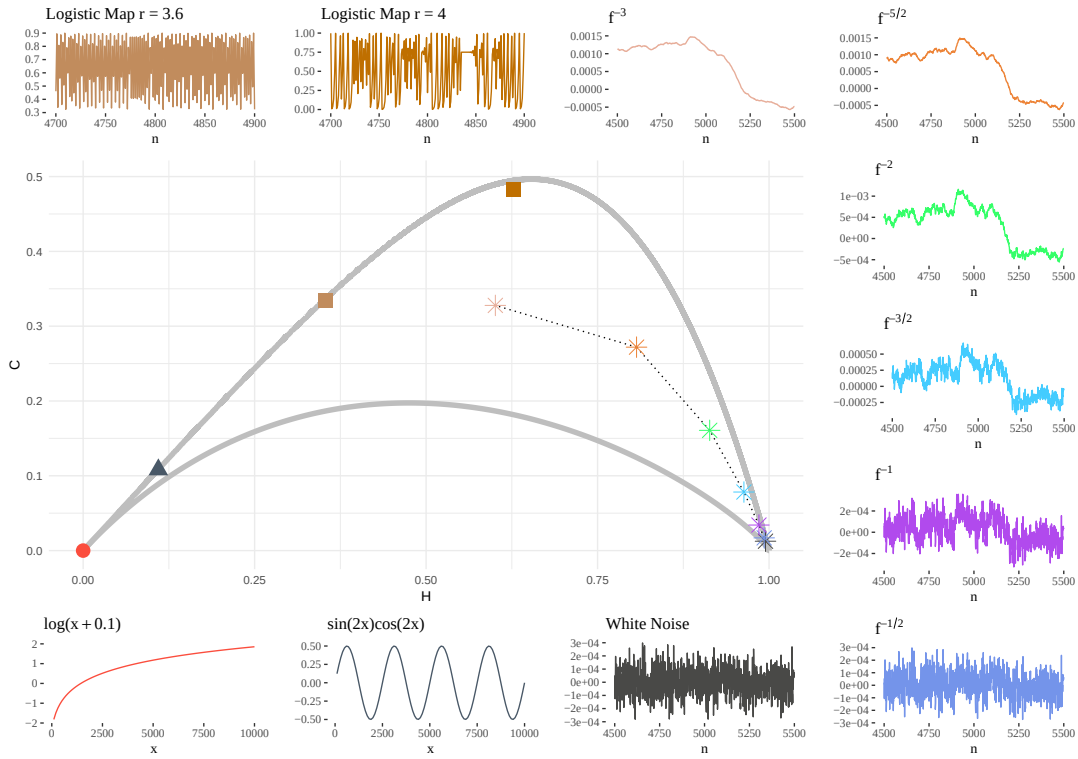

FIGURE 3 Eleven systems and their points in the $H \times C$ plane

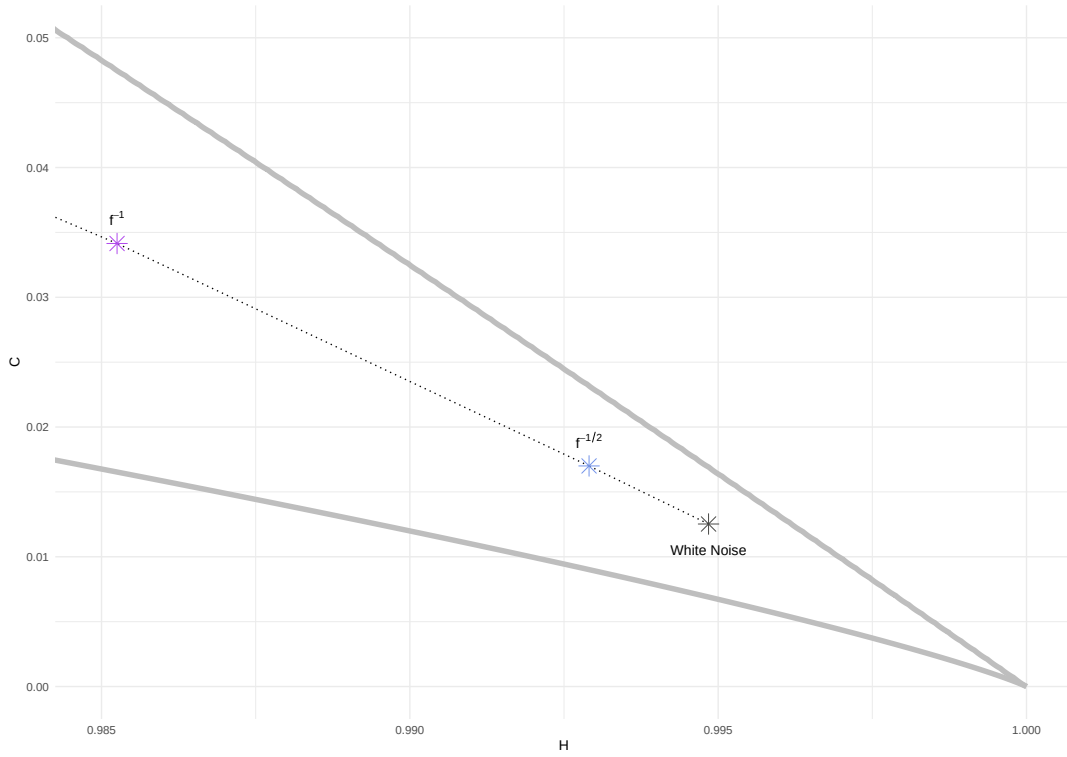

FIGURE 4 White Noise, $f^{-1/2}$ and f^{-1} noise points

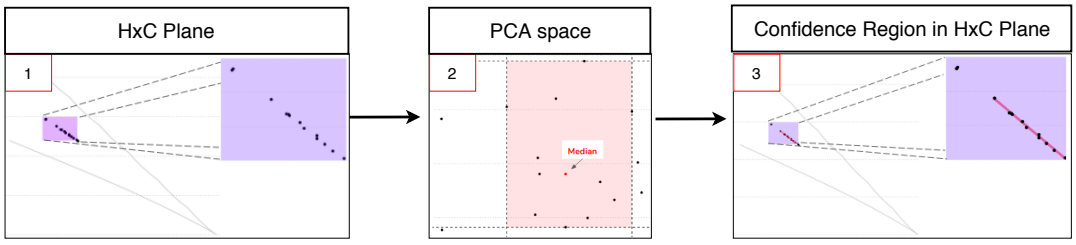

FIGURE 5 Outline of the methodology used for the construction of confidence regions.

4.2 | Empirical Confidence Regions

As we do not know the joint probability distribution of the pair (h, c) for a sequence of random variables collectively independent and identically distributed according to a uniform law, studies involving classical bi-variate analysis, linear regression, and generalized linear models become unfeasible. Therefore, for the construction of our proposal, we adopted a non-parametric approach, making an empirical analysis of data obtained from physical sources and using them as our reference in the search for confidence regions.

The set of all feasible pairs in the $H \times C$ plane is found in a compact subset of $\mathbb{R}^2$, which has limits with explicit expressions for the boundaries of this closed manifold, dependent only on the dimension of the probability space considered, that is D! in the traditional Bandt-Pompe method **?**. Due to such quotas, some limitations are generated, such as the absence of a representative distance metric and the difficulty of proposing confidence regions. In view of this, it is necessary to apply an orthogonal projection in the data for a new two-dimensional coordinate system to solve these restrictions. A classic proposal in these categories of problems is the principal component analysis algorithm **?**.

Let $\mathcal{P} = \{p_n\}_{n=1}^{N}$ be a set of N observations in the $H \times C$ plane, where $p_n = (h_n, c_n)$ corresponds to a time series. We propose in this work the use of a latent space $\Omega_1 \times \Omega_2$ obtained by the PCA for the construction of empirical confidence regions, which hereinafter referred to we call HC-PCA regions. Let $\{(u_n, v_n)\}_{n=1}^{N}$ uncorrelated observations obtained by the representation of $\mathcal{P}$ in the latent space $\Omega_1 \times \Omega_2$. For simplicity, and without loss of generality, assume N odd; in the proposed methodology, we find a parallelepiped that contains $100\,(1-\alpha)\%$ of the points in the $H \times C$ plane with the following steps:

1. Find the ranks that sort the values of the first principal component $\boldsymbol{u} = (u_1, u_2, \ldots, u_N)$ in ascending order: $\boldsymbol{r} = (r_1, r_2, \ldots, r_N)$, i.e., u_{r_1} is the minimum value, and u_{r_N} is the maximum value.
2. Find point (u, v) whose first principal component is the median: $(u_{r_{(N+1)/2}}, \cdot)$.
3. Find the point (u, v) whose first principal component is the quantile $\alpha/2$: $(u_{r_{[N\alpha/2]}}, \cdot)$.
4. Find the point (u, v) whose first principal component is the quantile $1-\alpha/2$: $(u_{r_{[N(1-\alpha/2)]}}, \cdot)$.
5. The values $u_{r_{[N\alpha/2]}}$ and $u_{r_{[N(1-\alpha/2)]}}$ are the rightmost and leftmost bounds of the box, respectively.
6. The top and bottom bounds of the box are the minimum and maximum values of the second principal component of the points whose first principal component is at least $u_{r_{[N\alpha/2]}}$ and at most $u_{r_{[N(1-\alpha/2)]}}$; denote these values $v_{\min}$ and $v_{\max}$, respectively.
7. The corners of the box are $(u_{r_{[N\alpha/2]}}, v_{\min})$, $(u_{r_{[N\alpha/2]}}, v_{\max})$, $(u_{r_{[N(1-\alpha/2)]}}, v_{\min})$ and $(u_{r_{[N(1-\alpha/2)]}}, v_{\max})$.
8. Apply the inverse PCA transform to these corners obtaining (h_{v_1}, c_{v_1}), (h_{v_2}, h_{v_2}), (h_{v_3}, c_{v_3}) and (h_{v_4}, c_{v_4}).

The visual representation of the proposed technique can be seen in Fig. 5.

4.3 | Construction of a test statistic

Using the method of construction of confidence regions described in subsection 4.2 in M reference points in the $H \times C$ plane produced by sequences obtained from true-random generators of length T and embedding dimension D, we formulate a hypothesis test that we hereinafter referred the HC-PCA test. Whether $\boldsymbol{x}'$ is a sequence of length T and embedding dimension D, the main idea of the HC-PCA test is based on the following null hypothesis:

$\mathcal{H}_0$: $\boldsymbol{x}'$ consists of a sequence of independent uniform random variables.

The empirical p-value of $\boldsymbol{x}'$ with respect to the null hypothesis is given by the percentage of points outside the lowest confidence region obtained by the reference points to which $\boldsymbol{x}'$ belongs. Therefore, it can be achieved by the following steps:

1. Compute (h', c') the point in the $H \times C$ plane produced by $\boldsymbol{x}'$;
2. Compute $b_{T,D}(\boldsymbol{x}')$: the smallest box which contains (h', c') using the M reference points;
3. The empirical p-value of $\boldsymbol{x}'$ is obtained by the percentage of reference points outside $b_{T,D}(\boldsymbol{x}')$.

Thus, we can verify that:

- the smallest $b_{T,D}(\boldsymbol{x}')$ is observed when (h', c') "coincides" with the the point produced by the emblematic sequence, that is, point whose first principal component is the median of the M reference points, and
- the largest box is any box associated to (h', c') "outside" the largest box produced by the M reference points.

Assuming the critical value $\alpha = 0.05$, we obtain:

$$p - \text{value} \geq \alpha, \text{ the null hypothesis not can be reject.}$$

5 | EXPERIMENTAL SETTINGS

We evaluated the performance of the proposed method in relation to a large set of random sequences provided by state-of-the-art pseudo-random number generators. In this section, we present the settings of the parameters that we use as a reference, the true random physical generators used to calculate the empirical distribution, and descriptive analysis of representative points in relation to the confidence regions.

5.1 | Parameters Settings and Dataset

In the calculation of the ordinal symbol histogram, we employed the following factors in this study:

- Sequence length $T \in \mathcal{T} = \{10^3, 5 \times 10^4\}$,
- Embedding dimension $D \in \mathcal{D} = \{3, 4, 5, 6\}$,and
- Time delay $\tau = \{1, 10, 30, 50\}$.

For the construction of the confidence regions presented, we use:

- Set of 104596 points in the $H \times C$ plane, referring to sequences of length $T = 1000$, for each combination of the factors $\mathcal{T} \times \mathcal{D} \times \tau$, and
- Another set of 2093 points in the $H \times C$ plane, referring to sequences of length $T = 50000$, for each combination of factors $\mathcal{T} \times \mathcal{D} \times \tau$.

Since the results involving the variation of the time delay parameter did not show a significant difference in repeated experiments, we don't consider it as a determinating factor during the execution of the algorithm. On the other hand,

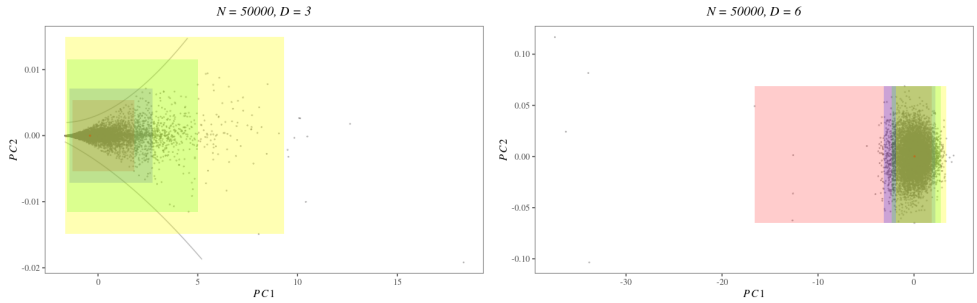

FIGURE 6 Representation of truly random sequences with length $T = 50.000$ in the PCA space for $D = 3$ and $D = 6$, and the quantiles of 90 %, 95 %, 99 %, and 99.9 %.

we consider two determining variables during the generation of such sub-spaces: the embedding dimension and the length of the sequence.

The data generation and analyses were performed using the `R` platform **?** v. 3.6.3. We used the `ggplot2` library **?** for generating the plots.

5.2 | True Random Numbers

Gerador - Marcelo

Random numbers are used in many fields, from gambling to criptography, aiming to guarantee a secure, realistic or unpredictable behavihor. Pseudo randomic results can be achieved by software in a deterministic way. But, some applications need actual random numbers (despite the somewhat elusive nature of actual randomness). Randomness can be observed in unpredictable real world phenomena like cathodic radiation or atmospheric noise. In this study we used two sources of real random numbers. The first is based on vacuum states to generate random quantum numbers described by **?** (**?**), the second one is based on atmospheric noise captured by a cheap radio receiver presented at `www.random.org`.

6 | RESULTS

6.1 | Descriptive analysis of empirical confidence regions

The regions used as a reference in this work are obtained through true random sequences, where we extract the empirical distribution of white noises in the Entropy-Complexity plane. In Fig. 6 we show the results obtained for $T = 50000$ in the scenarios of $D = 3$ and $D = 6$ in the new space defined by the PCA, together with the quantiles of 90 %, 95 %, 99 %, and 99.9 %. We also show the projection of the $H \times C$ plane limits in this new representation space, as well as identifying the median of each data set, the latter being represented as the red dots present in the graphs.

As we can see in Fig. 7 in the new representation space produced by the PCA, the data are not evenly distributed among the axes of the first main component, maintaining the character presented in the $H \times C$ plane, since such points tend to be concentrated close to the point $(1, 0)$.

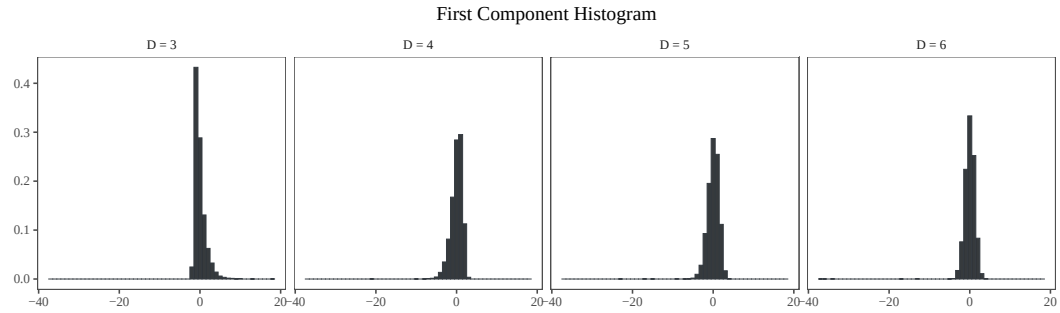

FIGURE 7 Histogram of the PCA first component

6.2 | Testing White Noise in the confidence regions

To analyze the efficiency of the confidence region calculated, we tested its applicability on a set of true random data generated physically not used by the algorithm during its construction. The results can be seen in Fig. 8.

For small series, $T = 1000$, and $D = 3$ we managed to maintain exactly 99 % of the data in the confidence region of the same value, and as the dimension increased reach 96 % of the points. On the other hand, there was a very large loss of points located in the region with 95 % confidence as the dimension increased. A reasonable explanation for this event is given in the choice of the parameter itself. It is known by definition that $D! << N$, which does not happen for such a sample size, thus presenting many missing patterns that lead to an unrepresentative probability distribution. For larger series, $T = 50000$, although we observed a small drop in the percentage of data present in the region with 99 % confidence, there was a significant increase in points in the region with 95 % confidence, showing between 90 % and 88 % of the points when we vary the embedding dimension.

6.3 | Analyzing Robustness to Correlation Structures

Fig. 3 shows the behavior of random time series with different levels of correlation (by means of the f^{-k} model) in the (h, c) plane. Knowing that such plane can discriminate between different system dynamics, several studies in the literature have used this approach in the investigation of methods of identification and characterization of randomness. Although this same strategy can be used to characterize different levels of correlation structures, in our case, we want to analyze the impact of injecting such dynamics into noise under the aspect of confidence regions.

For carrying out the experiment, we used an ”emblematic” time series as a basis. This series consists of the sample corresponding to the median of the (h, c) points used to build the confidence regions, thus expressing a representative sample of the dataset. Fig. 9a. illustrates, respectively, the effect of a white noise time series when adding correlation structures related to the f^{-k} series for $k \in \{0, 1, 2, 3\}$. As we can observe in the plane as the correlation between the observations increases, that is, $k > 0$, the randomness decreases and the entropy presented decreases, informing the loss of its stochastic characteristic.

Fig. 9b. illustrates the degree of limit correlation structure that can be added in white noise to eliminate it from the regions of confidence, where the red dot represents the original "emblematic" series. When we have $k = 0$, the features of the sequence have a small variation, corroborating the premise that series of noise $f^{-k} = 0$ have a minimum correlation measure, not significantly changing the dynamics of the system.

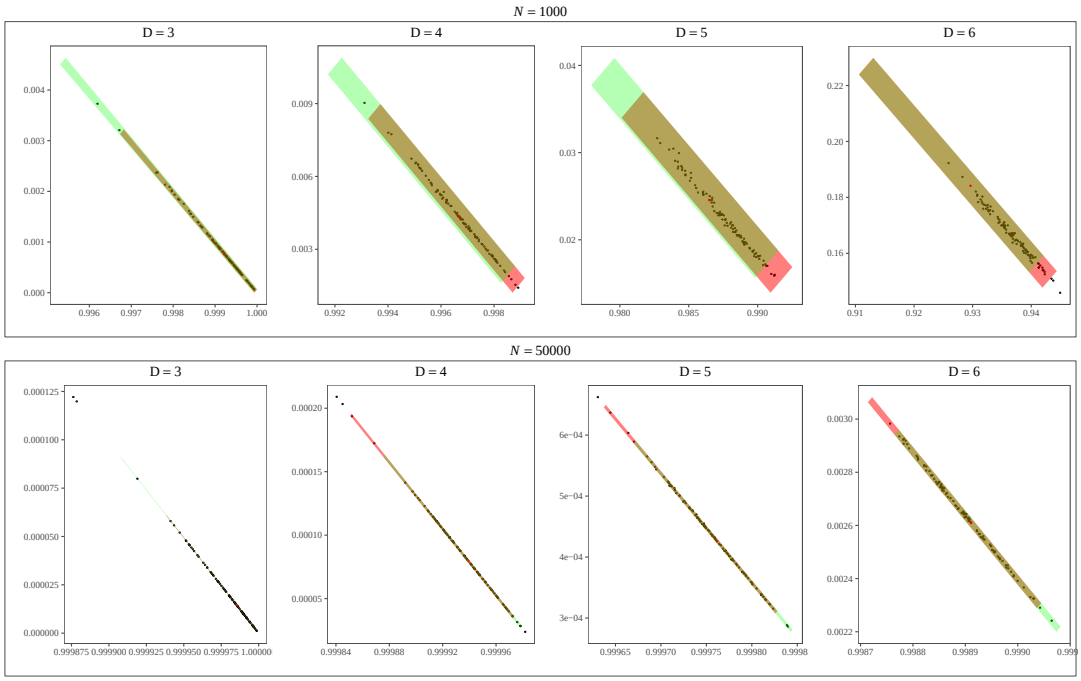

FIGURE 8 Results of the analysis behavior of true random noises in the regions of confidence built.

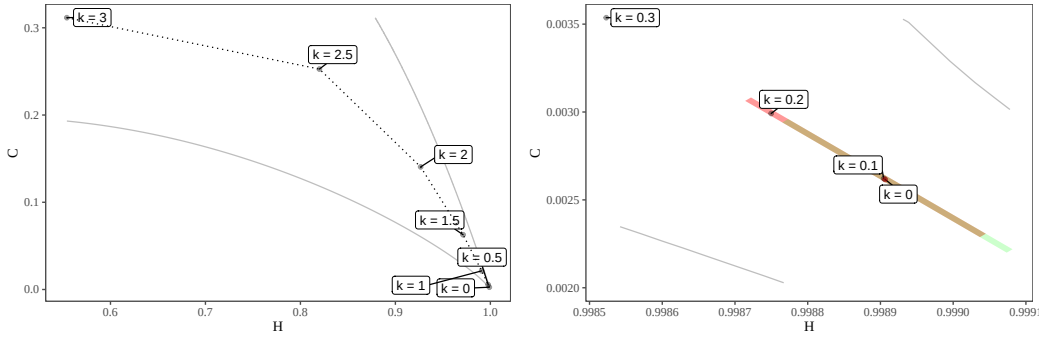

FIGURE 9 Correlation Structure Analysis

6.4 | Revisiting the White Noise Hypothesis in the Literature

In this section, we evaluate the quality of the confidence regions preceded by HC-PCA with sequences of previously analyze generators in the literature with the Entropy-Complexity plane. For this, we produced 100 sequences of length $T = 5 \times 10^4$ for each generator present in table 1 and calculated the respective p-values for each configuration of $D = \{3, 4, 5, 6\}$. The results obtained can be seen in table 2, where we categorize the sets of sequences among those that did not reject (NR) the null hypothesis presented in the HC-PCA test and those that rejected (R).

Performing a comparative analysis of the results produced by tables 1 and 2, we can see that the proposed methodology of confidence regions can capture the random dynamics of the sequences produced by most of the analyzed generators well, although in our experiments we consider only short sequences.

However, two results deserve special attention: the sequences produced by the generators fBm and Combo RNG. The first one, although it was not rejected by the analysis of **?**, presented low p-values, due to the characteristic point in the $H \times C$ plane where its sequences belong. For the case analyzed in **?** with $D = 6$, we verified that the produced sequences present an average of $H = 0.997$, therefore they do not belong to the empirical 95 % confidence region produced by HC-PCA. Analyzing the results produced by the Combo RNG, we can verify that only for $D = 3$ the proposed method can characterize the sequences produced as random. This is due to the fact that by presenting higher dimension values, we were able to analyze a greater number of ordinal patterns, which may not be presented in their entirety in short sequences.

7 | CONCLUSIONS

We present and evaluate a new method of building empirical confidence regions in the Entropy-Complexity plane for the analysis of white noise. The following proposal consists of two stages: (1) the construction of empirical confidence regions obtained through the mapping of points in the latent space formed by the application of the principal components analysis under the reference points, (2) the application of a hypothesis test to measure the similarity of new sequences to the reference points. Sequences of true random samples were used to calculate theses empirical regions.

Experiments with true random samples showed that the presented methodology can represent them with good performance. Although the present work focuses on the study of short sequences, we were able to capture the random behavior of PRNGs already analyzed in the literature and we verified the robustness of our technique and the correlation structure present in the sequences.

8 | SOURCE CODE AVAILABILITY

The text, source code, and data used in this study are available at the *Confidence-Regions* repository `https://github.com/EduardaChagas/ConfidenceRegions`.

9 | ACKNOWLEDGEMENTS

This work was partially funded by the Coordination for the Improvement of Higher Education Personnel (CAPES) and National Council for Scientific and Technological Development (CNPq).

TABLE 2 Results of the sequences generated by the main PRNGs in the literature. The analyzed sequences have length $T = 5 \times 10^4$.

Algorithm	D	p-value	HC-PCA	Algorithm	D	p-value	HC-PCA
MOT	3	0.305	NR	fGn	3	0.521	NR
	4	0.572	NR		4	0.519	NR
	5	0.455	NR		5	0.498	NR
	6	0.508	NR		6	0.470	NR
MWC	3	0.501	NR	fBm	3	1.11×10^{-16}	R
	4	0.477	NR		4	1.11×10^{-16}	R
	5	0.496	NR		5	1.11×10^{-16}	R
	6	0.496	NR		6	1.11×10^{-16}	R
COM	3	0.123	NR	f^{-k}	3	0.482	NR
	4	0.002	R		4	0.520	NR
	5	1.11×10^{-16}	R		5	0.513	NR
	6	1.11×10^{-16}	R		6	0.508	NR
LEH	3	0.531	NR	LCG	3	0.009	R
	4	0.515	NR		4	1.11×10^{-16}	R
	5	0.495	NR		5	1.11×10^{-16}	R
	6	0.501	NR		6	1.11×10^{-16}	R