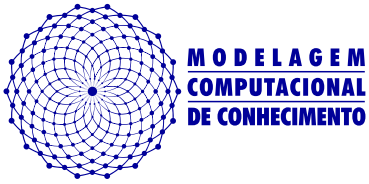


Dissertação de Mestrado

Teste para Verificação da Hipótese de Ruído Branco utilizando Teoria da Informação

Marcelo Queiroz de Assis Oliveira
marceloqao@gmail.com

Orientadores:

Dr. Alejandro C. Frery
Dr. Heitor Soares Ramos Filho

Maceió, Novembro de 2017

Marcelo Queiroz de Assis Oliveira

Teste para Verificação da Hipótese de Ruído Branco utilizando Teoria da Informação

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Modelagem Computacional de Conhecimento do Instituto de Computação da Universidade Federal de Alagoas.

Orientadores:
Dr. Alejandro C. Frery

Dr. Heitor Soares Ramos Filho

Maceió, Novembro de 2017

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Bibliotecária Responsável: Helena Cristina Pimentel do Vale

O48t Oliveira, Marcelo Queiroz de Assis.

Teste para Verificação da Hipótese de Ruído Branco utilizando Teoria da Informação/ Marcelo Queiroz de Assis Oliveira - 2017
52 f. : il.

Orientadores: Alejandro C. Frery, Heitor Soares Ramos Filho.
Dissertação (Mestrado em Modelagem Computacional de Conhecimento) – Universidade Federal de Alagoas. Instituto de Computação. Programa de Pós-Graduação em Modelagem Computacional de Conhecimento. Maceió, 2017.

Bibliografia: f. 50–52.

1. Teoria da Informação. 2. Geradores de números aleatórios. 3. Processamento de dados - Testes teóricos. 4. Matemática computacional – Testes estatísticos.

CDU: 004.94:519.724

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Modelagem Computacional de Conhecimento
Avenida Lourival Melo Mota, Km 14, Bloco 12, Cidade Universitária
CEP 57.072-900 – Maceió – AL – Brasil
Telefone: (082) 3214-1364/1825

Membros da Comissão Julgadora da Dissertação de Mestrado de Marcelo Queiroz de Assis Oliveira, intitulada: "Teste para verificação da hipótese de ruído branco utilizando teoria da informação", apresentada ao Programa de Pós-Graduação em Modelagem Computacional de Conhecimento da Universidade Federal de Alagoas, em 9 de novembro de 2017, às 9h30min, no laboratório 15 do Centro de Pesquisas em Matemática Computacional da Ufal.

COMISSÃO JULGADORA

Prof. Dr. Alejandro Cesar Frery
Ufal – Instituto de Computação
Orientador

Prof. Dr. Heitor Soares Ramos Filho
Ufal – Instituto de Computação
Orientador

Prof. Dr. Osvaldo Anibal Rosso
Ufal – Instituto de Física
Examinador

Profa. Dra. Juliana Gambini
Itba – Departamento de Ingeniería Informática
Examinadora

Maceió, 09 de novembro de 2017.

À minha amada esposa Élida, aos meus *filhos* Duda,
Gabriel, João Marcelo e Melissa, a minha querida Mãe,
Francisca *(In Memorian)* dedico este trabalho.
Por vocês, faria tudo de novo.

RESUMO

O nosso ponto de partida é o desejo de analisar se é viável verificar no plano ($H \times C$), dentro de uma abordagem estatística, se sequências de observações são ruído branco.

Na literatura encontramos diversos trabalhos que fazem isso de forma "ad hoc", verificando se o ponto característico de uma sequência nesse plano é próximo ao ponto (1,0). Contudo, tal como afirma Bandt (2017), não encontramos análises detalhadas que permitam atribuir significância estatística a tais afirmações.

Para elucidar essa questão, e diante da impossibilidade de contar com sequências infinitamente longas e que garantidamente sejam ruído branco, coletamos sequências de três fontes diferentes: duas físicas e uma algorítmica considerada de qualidade. Verificamos se é possível considerá-las ideais para os nossos fins.

Analisamos a dispersão dos pontos característicos dessas sequências no plano ($H \times C$) utilizando quatro fatores: o tamanho da sequência (N), o tamanho da palavra (D), o *delay* (τ) e a fonte geradora, observando a distância dos pontos característicos ao ponto de referência. Sugiram então evidências de que a fonte geradora seria um fator irrelevante para a análise.

Com o intuito de consolidar essa possibilidade, aplicamos o teste de Kolmogorov-Smirnov a pares de sequências comparáveis, porém verificamos que apenas duas das fontes geradoras são realmente irrelevantes, as duas fontes físicas.

Agrupamos, então os dados das fontes físicas e passamos a tratá-los como nossa referência, em seguida procuramos por regiões de confiança. Adotamos uma abordagem não-paramétrica por não termos nenhuma evidência teórica acerca da distribuição que segue a distância do ponto característico ao de referência quando é analisada uma sequência finita de ruído branco.

Calculamos então os quantis, respeitados os fatores tamanho da sequência (N), tamanho da palavra (D) e *delay* (τ), que servem como regiões de confiança para o teste que deu origem a este trabalho.

Concluímos a dissertação verificando que sequências produzidas por geradores aceitos pela comunidade geram pontos característicos dentro de regiões de confiança, enquanto que um gerador que foi descartado pelas estruturas que as suas sequências apresentam leva a pontos fora dessas mesmas regiões. Analisamos também sequências estacionárias e não estacionárias, e para as primeiras fazemos uma análise preliminar do poder do teste.

Palavras-chave: Geradores de Números Aleatórios. Testes Teóricos. Testes Estatísticos. Teoria da Informação.

ABSTRACT

We want to verify if sequences of observations are white noise in the plane ($H \times C$).

In the literature, we find several works that do this in an "ad hoc" way, checking if the characteristic point of a sequence in that plane is close to the point (1,0).

However, as Bandt (2017) states, we do not found detailed analyzes to assign statistical significance to such statements.

To elucidate this question, and in the face of the impossibility of counting infinitely long sequences that are guaranteed to be white noise, we gather candidate sequences from three different sources: two physical and one algorithmic considered as a good source. We checked, if we may consider them ideals for our purposes.

We analyze the dispersion of the characteristic points of these sequences in the plane ($H \times C$) using four factors: the size of the sequence (N), the size of the word (D), the delay (τ) and the generating source, observing the distance from the characteristic points to the reference point. We observe that the generating source would be an irrelevant factor to the analyses.

To verify this hipothesis, we applied the Kolmogorov-Smirnov test to pairs of comparable sequences, however we verify that only two sources are equivalents, the both physical.

Therefore, we grouped the data from physical sources, and it became our groundtruth, then in sequence, we searched for trust regions. We adopted a non-parametric approach because we did not have no theoretical evidence about the distribution that follows the distance from the characteristic point to our groundtruth when a finite sequence of white noise candidate is analyzed.

We then calculated the quantiles, respecting the factors sequence size (N), size of the word (D) and delay (τ), which serve as confidence regions for the test that gave rise to this work.

We conclude the dissertation by verifying that sequences produced by community accepted generators generate characteristic points within confidence regions, while a generator that has been discarded by the structures that its sequences present leads to points outside these regions. We also applied the test to stationary and nonstationary sequences and, for the former, we make a preliminary assessment of the test power.

Keywords: Random Number Generators. Theoretical tests. Statistical Tests. Information theory.

AGRADECIMENTOS

A Deus pelo dom da vida e por manter-me são durante o desenvolvimento deste trabalho.

A meus Pais pelos ensinamentos que não temos dentro da escola e pelo incentivo a buscar o aperfeiçoamento intelectual.

A minha esposa Elida, pelo companheirismo em tudo!

A meus filhos, Maria Eduarda, Gabriel, João Marcelo e Melissa, por sua enorme alegria de viver e seu sorriso.

Aos colegas do mestrado, com os quais vivi frutíferos e divertidos momentos de estudo, valeu Rodrigo, Domarques, Davy, Felipe, Brião, Pedro, Torres e todos os colegas do curso e do LaCCAN.

Ao colega Joao Borges pela grande contribuição na implementação.

Ao Professor Christoph Marquardt do *Max Planck Institute for the Science of Light* e ao Professor Mads Haahr do *Trinity College Dublin* e criador do *www.random.org* pela gentileza em fornecer dados realmente aleatórios, sem os quais este trabalho não teria sido realizado.

A todos os colegas da UFAL que direta ou indiretamente contribuiram para este trabalho, no LCCV, NTI e na gesão central da universidade, em especial ao companheiro Kleymerson por segurar a barra durante os momentos finais e também à Magnífica Reitora da UFAL, Valéria Correia e José Vieira, Vice Reitor. E ainda a todos que sempre inquiriam acerca do desenvolvimento deste trabalho.

Aos servidores do Instito de Computação da UFAL, em especial Vitor Torres pela presteza e atenção em atender os alunos do curso de Mestrado em Modelagem Computacional de Conhecimento, da mesma forma aos professores do curso.

Ao professor Osvaldo Rosso pela imensa contribuição relativa à teoria da informação e pelas ricas conversas acerca do tema

E finalmente aos meus orientadores, Alejandro Frery e Heitor Ramos sem os quais este trabalho não teria se concretizado, obrigado pelo incentivo, pela disponibilidade e pela paciência.

Muito obrigado!!

Marcelo Queiroz de A. Oliveira

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1

Introdução e Delimitação do Problema

ESTE capítulo tem como objetivo apresentar uma fundamentação teórica necessária para embasar os conceitos aplicados no trabalho. Delimitaremos aqui o problema a ser tratado nesta tese.

Em relação a fundamentação teórica, utilizou-se como principal fonte de pesquisa a área de indexação de periódicos científicos ISI *Web of Knowledge*, onde foram obtidas a grande maioria das referências, usando como parâmetros o fator de impacto dos periódicos pesquisados, a quantidade de citações de cada publicação, o grau de relevância para o tema pesquisado e o nível de produtividade (fator-H) dos autores envolvidos. O apoio em livros, *surveys, lecture notes* e ferramentas complementares de busca, como o *Google Acadêmico* foram utilizadas para complementar esta pesquisa.

1.1 Números Aleatórios

A noção de aleatoriedade é fundamental em diversas áreas, entretanto uma definição precisa, até mesmo do ponto de vista matemático rigoroso, é bastante difícil. Algumas questões emergem naturalmente como, o que vem a ser aleatoriedade? Existem eventos aleatórios na natureza? Faz algum sentido buscar leis da aleatoriedade? É possível simular a aleatoriedade? Estas são questões muito difíceis, envolvendo, inclusive, os primórdios da investigação filosófica como discute (Volchan, 2002). Essa dificuldade é bem ilustrada de forma humorística na Figura 1.1.

Números aleatórios perfazem uma das partes mais importantes em aplicações computacionais nos vários campos do conhecimento, como aborda (Knuth, 1998):

- *Simulação* - Quando um computador é usado para simular fenômenos naturais, números aleatórios são necessários para fazer as coisas de forma realística. Simulação

Figura 1.1: Visão de Dilbert de um gerador de números aleatórios

abrange diversas áreas, desde o estudo de física nuclear (onde particulas são submetidas a colisões aleatórias) até pesquisa operacional (como, por exemplo, a taxa de pessoas que entram num aeroporto em intervalos aleatórios).

- *Amostragem* - É praticamente impossível examinar todos os possíveis casos, porém uma amostra aleatória provê um palpite sobre como é o comportamento típico do fenômeno em questão.

- *Análise Numérica* - Técnicas elaboradas para a solução de complexos problemas numéricos foram desenvolvidas utilizando-se números aleatórios.

- *Programação de Computadores* - Valores aleatórios são uma ótima fonte de dados para testar a eficácia de algoritimos computacionais.

- *Tomada de Decisão* - Existem relatos de que diversos executivos tomam suas decisões lançando moedas ou atirando dardos. Também há rumores de que professores universitários lançam suas notas de forma similar. Em alguns momentos é importante tomar decisões de forma não influenciada por qualquer agente externo.

- *Criptografia* - Uma fonte de bits não viesada é essencial para diversos tipos de comunicações seguras, quando os dados precisam ser mantidos em sigilo.

- *Estética* - Um pouco de aleatoriedade faz com que gráficos e músicas geradas por computador aparentem ser menos artificiais.

- *Diversão* - Rolar dados, embaralhar cartas, girar rodas de roletas, etc., são passatempos facinantes para muitos. Estes usos tradicionais de números aleatórios sugeriram o nome “Método de Monte Carlo”.

Existem dois tipos básicos de geradores utilizados para produzir sequências aleatórias: GNAs – Geradores de Números Aleatórios, do inglês (*RNGs – Random Number Generators*) e GNPA – Geradores de Números Pseudoaleatórios, do inglês (*PRNGs* – Pseudorandom Number Generators).

1.1.1 Geradores de Números Aleatórios – GNA

Geradores de Números Aleatórios utilizam uma fonte não determinística juntamente com algumas funções de processamento para produzir aleatoriedade. As saídas deste tipo de gerador podem ser usadas diretamente como números aleatórios, desde que satisfaçam critérios de aleatoriedade, ou ainda podem servir como parâmetro de entrada para geradores de números pseudoaleatórios, vistos com mais detalhes na sequência.

A maior parte dos geradores utiliza-se de fenômenos físicos naturais como, decaimento radioativo, ruidos termais em semiconcutores, amostras de som num local ruidoso, ruído no espectro eletromagnético, dentre outros que, por óbvia dedução, carecem de algum hardware específico para serem capturados.

Na literatura é possível encontrar trabalhos relatando detalhadamente a criação de geradores de números aleatórios utilizando fontes de aleatoriedade apropriadas. Fairfield et al. (1985) descrevem a geração de um fluxo de bits aleatórios baseado na instabilidade da frequência de um oscilador. Evidentemente a construção de um gerador como o descrito anteriormente demanda o emprego de técnicas apuradas e um grande conhecimento teórico, além de necessitar, em sua grande maioria, de hardware especializado, o que gera uma barreira para os usuários que demandam tais dados aleatórios. Desta sorte, frequentemente os usuários precisam se valer de técnicas alternativas para obter aleatoriedade.

1.1.2 Geradores de Números Pseudoaleatórios – GNPA

Dadas as dificuldades descritas anteriormente, atualmente a maneira mais conveniente e confiável de se gerar números aleatórios para diversas aplicações é através de algoritmos com um sólido embasamento matemático. Tais algoritmos produzem sequências de números sabidamente não aleatórios ao todo, mas que aparentam comportar-se como números aleatórios independentes, isto é, tomada uma sequência de variáveis aleatórias independentes e identicamente distribuídas sobre o Intervalo (0,1). Tal sequência pode ser chamada de “Pseudoaleatória” e o programa utilizado em sua produção de “Gerador de Números Pseudoaleatórios” como define L’Ecuyer (2007).

1.2 Principais Testes Clássicos

Existem duas abordagens para testar-se a capacidade de geradores aleatórios ou pseudoaleatórios produzirem sequências ditas aleatórias. Segundo L’Ecuyer (1992) são elencados em teóricos e empíricos.

Os testes teóricos são bastante específicos para cada tipo de GNPA, pois analisam o as propriedades das sequências a partir da definição do gerador. Já os testes empíricos valem-se de técnicas estatísticas objetivando avaliar o quão boas são as sequências produzidas por um

determinado gerador. Estes últimos podem ser aplicados tanto a GNAs quanto a GNPAs.

Neste trabalho propomos um teste empírico não paramétrico baseado em ferramentas da teoria da informação. A seguir daremos uma breve introdução aos testes disponíveis na literatura e ao estado da arte.

1.2.1 NIST

Fundado em 1991, o NIST (*National Institute of Standards and Technology*) é uma agência não regulatória do Departamento de Comércio dos Estados Unidos da América que tem por missão promover a inovação e a competitividade nos EUA através da ciência de medidas, padrões e tecnologia de forma a alavancar a segurança econômica e melhorar a qualidade de vida do povo americano.

A sua Divisão de Segurança de Computadores (CSD) e o Centro de Pesquisa em Segurança Computacional (CSRC) facilitam a ampla disseminação de práticas e ferramentas de segurança da informação, provendo recursos para a definição de padrões além de identificar recursos de segurança na Web para suportar usuários na industria, governo e academia.

CSRC é o portal de acesso primário para se ter acesso às publicações de segurança de computadores, padrões e instruções, além de outras informações relacionadas a segurança.

Desde 1997, o Grupo de Trabalho Técnico em Geração de Números Aleatórios (RNG-TWG) tem trabalhado no desenvolvimento de uma bateria de testes estatísticos apropriados para a avaliação de geradores de números aleatórios e pseudoaleatórios utilizados em aplicações criptográficas.

Os principais objetivos do grupo são:

- Desenvolvimento de uma bateria de testes estatísticos para detectar não aleatoriedade em sequencias binárias construidas através de geradores de números aleatórios e pseudoaleatórios utilizados em aplicações criptográficas;
- Produzir documentação e uma implementação em software destes testes;
- Prover auxílio no uso e aplicação destes testes.

1.2.2 Diehard e Dieharder

George Marsaglia desenvolveu a bateria de testes Diehard em 1995, e os disponibilizou em CD-ROM. Robert Brown identificou limitações nessa bateria de testes, os implementou novamente na linguagem de programação C, acrescentou testes da bateria NIST e disponibilizou um conjunto ampliado de testes denominado Dieharder. A página Web `http://webhome.phy.duke.edu/~rgb/General/dieharder.php` é o portal de acesso a esses testes bem como a resultados de aplicá-los a diversas fontes de dados.

1.2.3 ENT

ENT Walker (2017) realiza uma variedade de testes no fluxo de bytes de um arquivo (ou na entrada padrão se nenhum arquivo for especificado) e produz uma saída como esta:

```
Entropy = 7.980627 bits per character.

Optimum compression would reduce the size
of this 51768 character file by 0 percent.

Chi square distribution for 51768 samples is 1542.26, and randomly
would exceed this value less than 0.01 percent of the times.

Arithmetic mean value of data bytes is 125.93 (127.5 = random).
Monte Carlo value for Pi is 3.169834647 (error 0.90 percent).
Serial correlation coefficient is 0.004249 (totally uncorrelated = 0.0).
```

ENT é distribuido em forma de binário para a plataforma Win32 e pode ter descarregado o seu código fonte para construção sob ambiente *Unix-Like*

1.2.4 TestU01

Considerado como o estado da arte dos testes para geradores de números aleatórios L'Ecuyer & Simard (2007), o $TestU01$ se apresenta como uma biblioteca de software escrita em *ANSI C* que oferece uma coleção de utilitários para testagem, ele provê implemetações generalistas dos testes estatísticos clássicos para geradores de números aleatórios, bem como de vários outros propostos na literatura, além de propor alguns originais. Pode ser aplicado a números aleatórios produzidos por qualquer tipo de dispositivo ou armazenados em arquivos.

O principal problema com o uso dos testes NIST, Dieharder, Test$U01$ e outros similares é que estão direcionados a usuários especializados. Uma parcela significativa de pesquisadores prefere ferramentas visuais, como a descrita a seguir.

1.2.5 Visualização

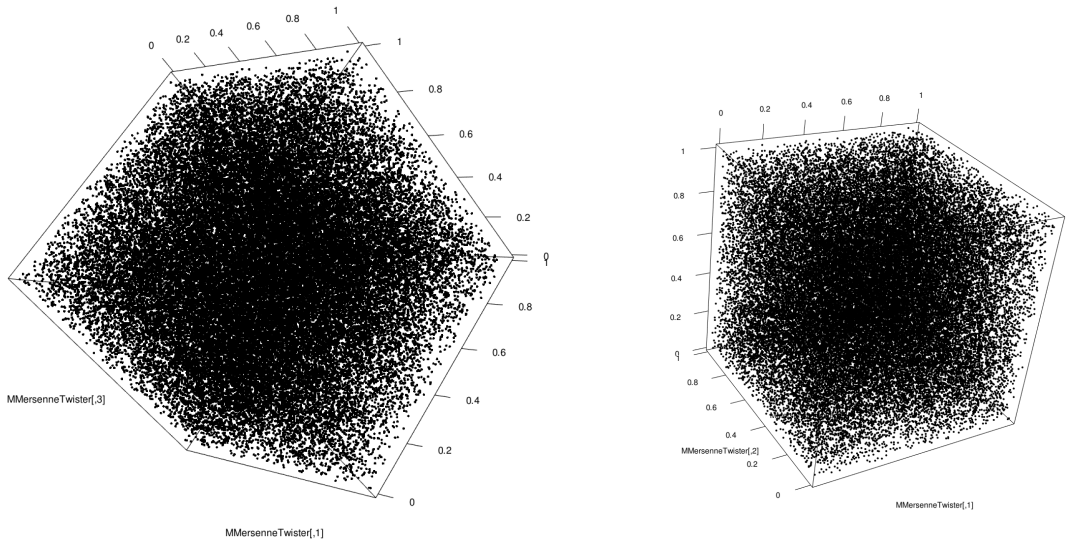
O intuito desses testes é a verificação de que os dados produzidos por um gerador, seja ele algoritmico ou físico, não se afastam significativamente da hipótese de serem eventos de variáveis independentes e identicamente distribuídas segundo uma lei Uniforme no intervalo (0,1].

Verificar que os dados seguem uma lei Uniforme é relativamente simples, e há para eles uma bateria de testes que observam diversas propriedades.

A componente mais difícil é a de verificar a independência, por se tratar de independência coletiva e não apenas aos pares.

A falta de independência de uma sequência de N pontos pode se manifestar de várias maneiras. Uma delas é quando os pontos jazem em subespaços de dimensão menor a N, ao invés de preencher o espaço por completo.

As figuras 1.2 e 1.3 mostram sequências de três pontos não sobrepostos desenhadas no cubo unitário. A primeira perspectiva (figs. 1.2(a) e 1.3(a)) mostram uma perspectiva dos pontos que não levanta nenhuma suspeita. Já as figuras 1.2(b) e 1.3(b) mostram que as sequências produzidas pelo gerador RANDU não preenchem o espaço, mas ficam confinadas a alguns planos, enquanto que as produzidas por Mersenne-Twister não apresentam essa deficiência.



(a) Primeira perspectiva

(b) Segunda perspectiva

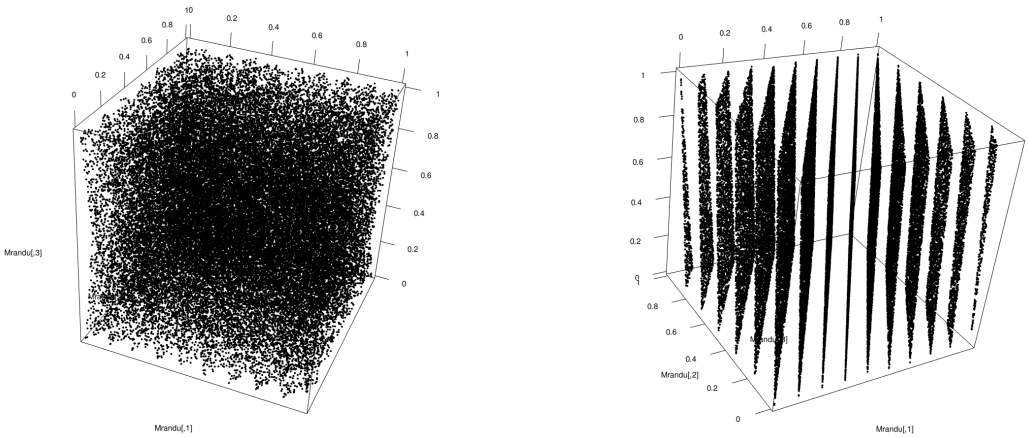
Figura 1.2: Visualização 3D de sequências disjuntas produzidas pelo gerador Mersenne-Twister

Embora o exemplo mostre claramente que o gerador RANDU apresenta problemas, nem sempre é fácil identificar esse tipo de estrutura. Isso se deve a que ela pode ocorrer em espaços de dimensão maior, ou seguindo estruturas mais complexas.

1.2.6 Descritores baseados em Teoria da Informação

O trabalho pioneiro de Bandt & Pompe (2002) representa uma mudança de paradigma na análise de séries temporais, e será o nosso marco referencial teórico.

Eles propõem uma técnica não-paramétrica de análise de sequências que consiste em transformar palavras de D dados não necessariamente subsequentes em símbolos ordinais. Esses símbolos codificam a ordem que as D observações têm na sequência e, portanto, são bem menos sucetíveis a contaminação dos que os próprios valores. Forma-se então um histograma de proporções dos símbolos observados, e calculam-se duas quantidades: a Entropia e a Divergência de Jensen-Shannon a uma distribuição de referência (usualmente a



(a) Primeira perspectiva (b) Segunda perspectiva

Figura 1.3: Visualização 3D de sequências disjuntas produzidas pelo gerador RANDU

uniforme). A série, finalmente, é representada pelo par de valores Entropia-Complexidade Estatística, sendo que esta última é o produto da Entropia e a Divergência de Jensen-Shannon.

O conjunto de valores possíveis dos pontos característicos de qualquer série não varre $\mathbb{R}^2$, mas constitui-se em um subconjunto compacto do plano: o plano Entropia-Complexidade, visto na Figura 1.4. Uma grande quantidade de trabalhos mostra que a posição dos pontos característicos no plano Entropia-Complexidade é capaz de caracterizar diversos tipos de dinâmicas, sendo as duas extremas o ponto (0,0), que corresponde a sequências determinísticas, e o ponto (1,0), típico de ruído branco.

Alguns dos trabalhos emblemáticos que empregam essa abordagem são:

Larrondo et al. (2006) mostram que o plano Entropia-Complexidade permite prever o resultado dos testes Diehard de qualidade de GNPA.

Martin et al. (2006) analisam o mapa caótico logístico e discutem cotas no plano Entropia-Complexidade.

Rosso et al. (2006) analisam dados de eletroencefalogramas de pacientes com epilepsia utilizando decomposição wavelet e o plano Entropia-Complexidade.

De Micco et al. (2008) avaliam melhorias da qualidade de sequências pseudoaleatórias.

De Micco et al. (2009) estudam as componentes caóticas de GNPAs.

Carpi et al. (2011) analisam a evolução de redes dinâmicas com descritores baseados em Teoria da Informação.

Zunino et al. (2012) analisam a relação entre dinâmicas caóticas e estocásticas com uma abordagem multiescala.

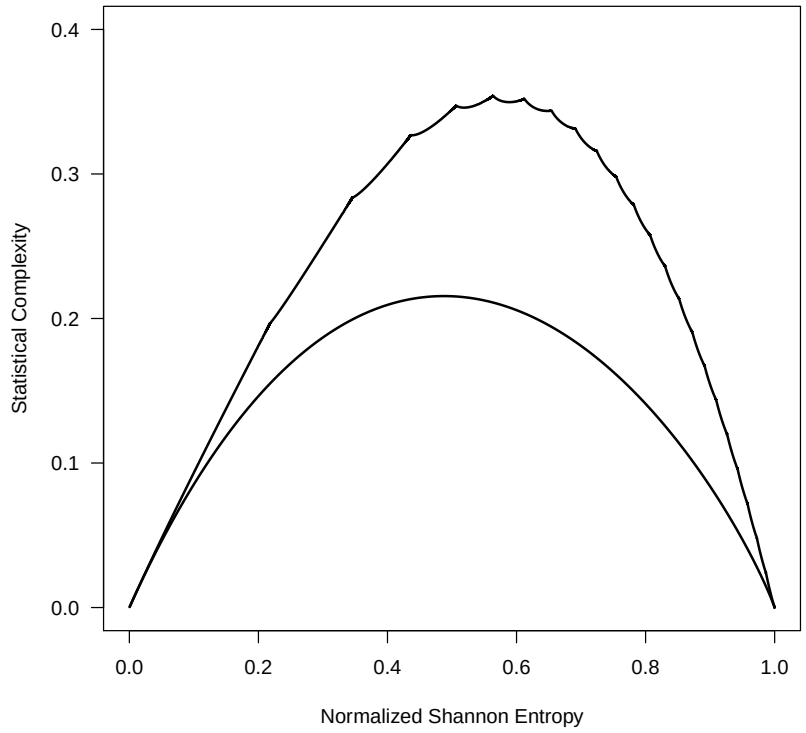


Figura 1.4: Plano Entropia-Complexidade.

Cabral et al. (2013) utilizam descritores de Teoria da Informação para descrever a dinâmica de redes de sensores sem fios.

Ravetti et al. (2014) caracterizam sequências com componentes caóticas e estocásticas no plano Entropia-Complexidade.

Aquino et al. (2015) analisam o comportamento de veículos em larga escala em função da topologia de diversas cidades.

Sippel et al. (2016) analisam séries temporais de fenômenos ambientais.

Schieber et al. (2016) verificam o efeito de ataques em redes complexas através de descritores no plano Entropia-Complexidade.

Rosso et al. (2016) mostram a expressividade dos descritores de Teoria da Informação para o problema de classificação e verificação de assinaturas.

Aquino et al. (2017) conseguem caracterizar o tipo de dispositivo elétrico observando o ponto no plano Entropia-Complexidade em que o histórico do seu consumo é mapeado.

Uma limitação de todos esses trabalhos é que os pontos no plano são atribuídos a padrões prototípicos, isto é, a modelos, de forma *ad hoc*. O autor desta dissertação só conhece um

trabalho em que é feita uma análise da significância da posição de pontos característicos no plano Entropia-Complexidade: o artigo de Bandt (2017).

Nesse contexto, esta dissertação fornece regiões de confiança para uma boa diversidade de situações que permitem verificar a significância da hipótese de uma sequência aleatória ou pseudoaleatória é aderente à hipótese dela ser ruído branco.

1.3 Delimitação do problema

A motivação deste trabalho é o desenvolvimento de um teste baseado em Teoria da Informação para verificar a hipótese de que uma sequência é ruído branco, isto é, formada por observações de variáveis aleatórias independentes e identicamente distribuídas. Para tanto, trabalharemos com atributos derivados da simbolização de Bandt & Pompe. Cada sequência sob análise será transformada em um ponto no plano Entropia-Complexidade, e será medida a sua distância ao ponto característico de sequências ideais. Analisaremos, então, a distribuição empírica de uma variedade de situações de interesse para, finalmente, propor regiões de confiança da hipótese nula. Por fim, analisaremos sequências com o ferramental aqui proposto.

2

Proposta

ESTE capítulo tem como objetivo apresentar a proposta bem como os materiais e métodos utilizados na realização do trabalho.

2.1 Fundamentação Teórica

Trataremos a sequência de saída de um gerador de números pseudoaleatórios como uma série temporal. Procuraremos por estruturas indesejadas em um gerador ideal utilizando ferramentas de Teoria da Informação.

Há diversas abordagens para a análise de séries temporais, sendo que uma das mais importantes surge a partir do trabalho de Bandt & Pompe (2002).

Essa abordagem consiste em transformar a série de entrada em uma sequência de símbolos. Feito isso, computa-se o histograma de proporções dos símbolos, e calculam-se descritores a partir dele.

A hipótese subjacente a toda análise de séries temporais é que os dados observados são o resultado da operação de um sistema causal, possivelmente sujeito a ruído observacional. Esse sistema, ou dinâmica, é responsável pela criação de padrões através de cuja observação deseja-se inferir a respeito da dinâmica.

A análise de séries temporais é um ramo clássico da Estatística Brockwell & Davis (1991) que se divide, tipicamente, na análise nos domínios do tempo e da frequência. Ambas abordagens empregam diretamente os valores observados e, portanto, são suscetíveis ao efeito danoso de diversos tipos de contaminação. Uma forma de tornar as análises mais imunes a contaminação é através de técnicas robustas Bustos & Fraiman (1984). Outra, mais moderna, é pelo uso de métodos não-paramétricos.

Há diversas ferramentas que auxiliam na análise clássica de séries temporais; na data de redação deste trabalho havia 238 bibliotecas para essa finalidade (ver https://cran.

r-project.org/web/views/TimeSeries.html). Para essa mesma plataforma, apenas três delas trabalham exclusivamente com técnicas não paramétricas.

Seja a série temporal $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. Ao invés de analisarmos os valores, transformaremos grupos de N valores (não necessariamente adjacentes) em padrões ordinais, e analisaremos a sua distribuição de frequência. Por exemplo e sem perda de generalidade, com $N = 3$ e para qualquer i viável, se $x_i < x_{i+1} < x_{i+2}$ atribuiremos a esta tripla o padrão π_0; caso $x_i > x_{i+1} > x_{i+2}$ o padrão será π_1 e assim por diante. Com isso, há $N!$ possíveis padrões. Esta é conhecida como *simbolização de Bandt & Pompe* Bandt & Pompe (2002). Forma-se, então, um histograma e, a partir dele, extraem-se quantificadores como, por exemplo, entropia, distância estocástica a uma distribuição de equilíbrio, e complexidade estatística.

Esta simbolização é muito resistente a vários tipos de contaminação, por exemplo, o padrão π_0 não será alterado para qualquer $k > 1$ que afete de forma multiplicativa x_{i+2}. Ainda que o padrão seja alterado, por exemplo se $k = -1$, a mudança será local e afetará, no máximo, N padrões.

A análise da dinâmica subjacente a uma série temporal utilizando a simbolização de Bandt & Pompe tem sido usada com sucesso em diversas áreas como, por exemplo, a discriminação entre fenómenos estocásticos e caóticos Ravetti et al. (2014), a identificação de padrões de comportamento em redes veiculares Aquino et al. (2015), a classificação e verificação de assinaturas *online* Rosso et al. (2016), na análise da robustez de redes Schieber et al. (2016), a classificação de padrões de consumo de energia elétrica Aquino et al. (2017).

O processo de simbolização, também conhecido como particionamento de dados, representa o procedimento de distribuição dos elementos em conjuntos de símbolos capazes de fornecer a sua informação causal. De acordo com a abordagem de Bandt e Pompe, substituímos a série por sequências de postos, obtidos pela análise desta ao longo do tempo.

Dada uma série temporal a tempo discreto $X = x_t : 1 \leq t \leq M$ de comprimento N, uma dimensão D e um tempo de atraso (*delay*) τ, o particionamento é efetuado por meio da reorganização do sistema em conjuntos seguindo os passos:

- **Composição dos grupos:** Os conjuntos, ou palavras, de comprimento D e *delay* τ são definidas por um segmento da série:

$$(x_{t+1}, x_{t+\tau}, \ldots, x_{t+D\tau}).$$

- **Formação dos padrões:** Cada palavra é então relacionada a um padrão ordinal π_j de ordem D, isto é, um elemento indexado univocamente por

$$j \in \{1, 2, \ldots, D-1, D!\}.$$

Há várias formas de atribuir palavras a símbolos; neste trabalho utilizaremos a atribuição

lexicográfica, isto é, se os valores da palavra $(x_{t+1}, x_{t+\tau}, \ldots, x_{t+D\tau})$ são tais que, ordenados, eles têm índices crescentes $b_1, b_2, \ldots, b_D$, então o padrão correspondente será $\pi = b_1 b_2 \ldots b_{D!}$.

Calcula-se, então, o histograma de proporções $\mathcal{H} = (p_1, \ldots, p_{D!})$ dos padrões observados:

$$p_j = \frac{1}{D-N+1} \#\{\text{padrões } \pi_j \text{ observados}\}.$$

O seguinte passo consiste em calcular descritores a partir desse histograma.

Os trabalhos já citados utilizam dois descritores: a Entropia de Shannon e a Complexidade Estatística da série.

A Entropia de Shannon é definida como

$$E(\mathcal{H}) = -\sum_{j=1}^{D!} p_j \log p_j,$$

em que adotamos a convenção $0(-\infty) = 0$. Esta é uma medida da desordem ou imprevisibilidade da lei subjacente a $\mathcal{H}$. Quando todas as proporções são iguais, isto é, quando se tem uma lei uniforme $p_1 = \cdots = p_{D!}$, então a entropia é máxima e igual a 1. Chamaremos a lei uniforme "medida de referência", e a denotaremos $\mathcal{H}_R$. Quando toda a probabilidade está concentrada em um único símbolo, ou seja há um símbolo k tal que $p_k = 1$, então a entropia é nula.

A entropia apenas não consegue caracterizar de forma plena a dinâmica que produz a série. Torna-se interessante, então, o uso de um outro descritor baseado em quão diferente o histograma $\mathcal{H}$ é de uma lei de probabilidade de referência. A nossa referência será a lei uniforme, e a medida de distância entre elas será a distância de Jensen-Shannon:

$$JS(\mathcal{H}, \mathcal{H}_R) = E((\mathcal{H} + \mathcal{H}_R)/2) - \frac{1}{2}(E(\mathcal{H}) + E(\mathcal{H}_R)).$$

2.2 Regiões de Confiança no Plano Entropia-Complexidade

Seja $\boldsymbol{X} = (X_1, \ldots, X_n)$ uma amostra a respeito da qual temos uma conjectura que queremos verificar. Essa conjectura é a respeito dos parâmetros que caracterizam a distribuição da amostra, ou a respeito de parâmetros que caracterizam a distribuição de atributos relacionados à distribuição da amostra. Chamaremos "hipótese nula" àquela que convimos em não rejeitar a não ser que obtenhamos suficiente evidência para isso; a denotaremos H_0. Por vezes precisaremos da "hipótese alternativa", que denotaremos H_1.

Classicamente, um teste de hipótese se baseia em uma estatística de teste T que depende exclusivamente da amostra $\boldsymbol{X}$, isto é, $T(\boldsymbol{X})$, e é construída de tal forma que adota valores "pequenos" sob H_0 e "cresce" conforme "se afasta" de H_0. Idealmente, conhecemos a distribuição de T sob a hipótese nula e, com isso, somos capazes de aferir a probabilidade de observarmos valores "grandes" mesmo sob H_0. Definimos, assim, o p-valor do teste baseado

em $T(\boldsymbol{X})$ para o valor observado η como $\Pr_{H_0}(T(\boldsymbol{X}) \geq \eta)$. O procedimento básico consiste em rejeitar a hipótese nula ao nível de significância $100(1-\alpha)\%$ se o p-valor for inferior a α, e em não rejeitá-la caso contrário. Mais modernamente, não se fala em "rejeição", reporta-se o p-valor, deixando a decisão para o leitor. Desta forma, chamaremos de "valor crítico" o conjunto de valores, definidos pelo leitor tais que, quando assumidos pela estatística de teste T levam à rejeição da hipótese, nula H_0.

Um teste pode incorrer em dois tipos de erro, ou rejeitar a hipótese nula quando, na verdade, os dados provêm dela, ou não rejeitá-la quando os dados provêm de outro modelo. O primeiro chama-se "Erro Tipo I" ou "Falso Positivo", e o segundo "Erro Tipo II" ou "Falso Negativo". A probabilidade de incorrer no Erro Tipo I chama-se "Tamanho do Teste", enquanto a probabilidade do complementar do evento que leva ao Erro Tipo II é o "Poder do Teste".

Diversos testes foram desenvolvidos ao longo do tempo com o objetivo de testar a aleatoriedade de sequências de números. Como "aleatoriedade" é uma noção vaga, cada teste procura identificar uma ou algumas falhas dos dados. Por esse motivo é que há classes ou conjuntos de testes que, aplicados de forma criteriosa, permitem aferir o quanto uma sequência se afasta da hipótese de aleatoriedade. A partir dessa análise, pode se ter uma ideia do comportamento global do gerador que a produziu.

Sob a conjectura de aleatoriedade, cada estatística de teste T tem uma distribuição, que pode ser determinada ou de forma exata ou de forma aproximada, mas sempre precisa ser conhecida. Com ela é possível informar o p-valor de uma determinada amostra.

Não conhecemos, contudo, a distribuição conjunta do par (E,JS) para uma sequência de variáveis aleatórias coletivamente independente e identicamente distribuídas segundo uma lei uniforme. Como também não conhecemos a distribuição do par (H,C) decidimos fazer a análise empírica de dados obtidos de fontes "confiáveis".

Para isso, utilizamos três fontes de dados: duas físicas e uma algorítmica. As fontes físicas foram dados de medidas de estados quânticos (Gabriel et al., 2010) (que denominaremos *sequências quânticas*) e de sinais de rádio (Stamey, 2016) (que chamaremos *sequências de rádio*). A fonte algorítmica é o gerador Mersenne-Twister (Matsumoto & Nishimura, 1998), considerado um padrão de qualidade de algoritmos de geração de números pseudoaleatórios.

Obtivemos com os autores 54 000 000 de observações de cada gerador físico, e as submetemos à análise dos padrões de Bandt & Pompe.

Obtivemos os valores de entropia e complexidade estatística para todas as sequências possíveis de tamanho 18 000, palavras de tamanho $D \in \{3,4,5,6\}$ e *lag* $\tau \in \{1,10,30,50\}$.

Seguindo a proposta de Bandt (2017), nossa medida de qualidade é a distância do ponto padrão observado (E,C) no plano Entropia-Complexidade ao ponto ideal (1,0). Empregamos, como esse autor, a distância euclidiana.

2.3 Artefatos

Para organizar, catalogar e facilitar a consulta a todo material obtido, as referências foram gerenciadas com a ferramenta *JabRef*. Um gerenciador de referências bibliográficas multiplataforma gratuito que permite organizar de maneira centralizada vários vínculos entre as referências utilizadas, bem como visualizar e anotar as mesmas dentro da própria ferramenta. Quanto à editoração eletrônica do trabalho, fez-se uso da plataforma LaTeX, com editor de textos *TeXstudio* de código aberto. Este trabalho foi desenvolvido num equipamento com as configurações constantes na tabela 2.1.

Adicionalmente, para o processamento dos dados das sequências, foi utilizado a infraestrutura computacional do Laboratório de Computação Científica e Análise Numérica da Universidade Federal de Alagoas *LaCCAN-UFAL*, descrito na tabela 2.2.

Tabela 2.1: Ambiente local utilizado no desenvolvimento do trabalho

Arquitetura	Intel core $i5$ 2,4 GHz 64 bits - 8GB RAM
S.O.	MacOS High Sierra 10.13
Ambiente R	R v3.4.2 e RStudio v1.0.153
LaTeX	MacTeX v2017 e TeXstudio v2.1.3

Tabela 2.2: Ambiente remoto virtual utilizado no desenvolvimento do trabalho

Arquitetura	Intel Xeon $E312xx$ 16 Cores 32GB RAM
S.O.	Debian 8 GNU/Linux Kernel v3.16.0 – 4 – $amd64$
Ambiente R	R v3.4.1 - “Single Candle”

Do ponto de vista técnico deste trabalho, foi utilizada a plataforma de análise estatística R. Esta plataforma foi desenvolvida originalmente por Ross Ihaka e Robert Gentleman, com o intuito de ser uma linguagem de código aberto voltada para a análise estatística e, consequentemente, a precisão numérica com fortes características funcionais (R Core Team, 2017). Por este motivo, ela será utilizada para a geração e manipulação das sequências pseudoaleatórias, análise dos dados e geração dos gráficos desse trabalho. A precisão numérica desta ferramenta, sendo aferida por Almiron et al. (2009), é adequada para essa abordagem.

As ferramentas utilizadas no desenvolvimento deste trabalho, são preferencialmente multiplataforma e código aberto com licença de uso *GNU General Public License* (GPL).

Todos esses aplicativos, métodos e informações obtidas, forneceram grandes contribuições para o traçado da linha mestra deste trabalho, indicando que o mesmo está na fronteira do conhecimento produzindo um estado da arte fidedigno aos temas e ferramentas adotadas para norteá-lo.

3

Resultados

Trabalhamos com sequências de observações provindas de três geradores: dois físicos (considerados "verdadeiramente aleatórios") e um algorítmico (o gerador Mersenne-Twister, que é reputado um dos melhores geradores pseudoaleatórios).

Para cada gerador considerado coletamos sequências disjuntas de tamanho 1000 e 50 000, de cada tamanho de palavra D e a cada *lag* τ. Temos, assim, quatro fatores a serem analisados.

Cada sequência passou pelo processo de simbolização, e foi calculado o histograma dos símbolos. Foram então calculados os valores de Entropia e de Complexidade de cada histograma, bem como a distância euclidiana desses valores ao ponto de referência (1,0).

O primeiro passo consistiu em fazer uma análise visual das distâncias dentro do plano (H,C). Dessa análise visual concluímos que é plausível eliminar o fator gerador e, para tanto, aplicamos o teste de Kolmogorov-Smirnov aos pares de distâncias oriundas de distâncias comparáveis, isto é, provindas dos mesmo fatores N, D e τ.

Os testes de Kolmogorov-Smirnov não nos fazem rejeitar a hipótese de que os diferentes geradores produzem sequências com idênticas propriedades no que diz respeito à distância ao ponto de referência. Assim, nos passos seguintes analisamos o agregado das distâncias provindas dos três geradores considerados.

Acompanhando a análise realizada por Bandt (2017), calculamos os quantis de ordem 999/1000, 1/100, 5/100 e 10/100 das distâncias agrupadas pelos fatores relevantes (N, D e τ). Com isso, produzimos intervalos de confiança para testar a hipótese de que a distância do ponto característico de uma sequência é compatível com a de uma sequência de ruído branco. No repositório associado a esta dissertação deixamos disponíveis as funções de distribuição acumuladas dessas distâncias, com as quais é possível calcular o p-valor, e não apenas a decisão binária "rejeita" ou "não rejeita".

3.1 Análise global das sequências

Em todos os casos os pontos foram desenhados com 1 % de transparência, para evidenciar as regiões mais e menos densas.

As figuras 3.1 e 3.2 mostram os planos Entropia-Complexidade com as respectivas curvas de complexidade mínima e máxima para cada par $D \in \{3,4,5,6\}$ (colunas) e $\tau \in \{1,10,30,50\}$ (linhas), com os 52 429 pontos observados a partir das sequências obtidas pelo gerador quântico tomando 1.000 e 50.000 sequências respectivamente.

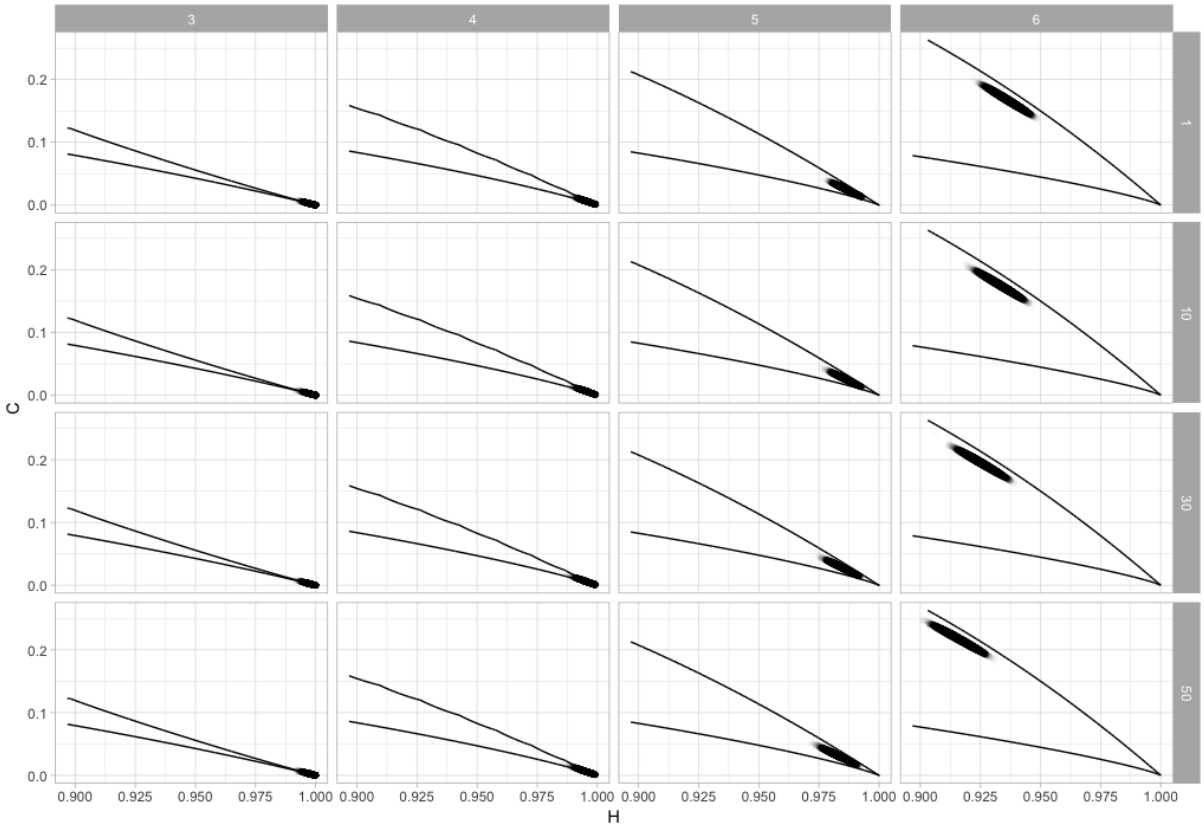


Figura 3.1: Diagramas de dispersão das sequências quânticas com 1.000 observações para $D \in \{3,4,5,6\}$ (colunas) e $\tau \in \{1,10,30,50\}$ (linhas), com curvas de complexidade mínima e máxima no plano Entropia-Complexidade.

As figuras 3.3 e 3.4 mostram os planos Entropia-Complexidade com as respectivas curvas de complexidade mínima e máxima para cada par $D \in \{3,4,5,6\}$ (colunas) e $\tau \in \{1,10,30,50\}$ (linhas), com os 52 429 pontos observados a partir das sequências obtidas pelo gerador de rádio tomando 1.000 e 50.000 sequências respectivamente.

As figuras 3.5 e 3.6 mostram os planos Entropia-Complexidade com as respectivas curvas de complexidade mínima e máxima para cada par $D \in \{3,4,5,6\}$ (colunas) e $\tau \in \{1,10,30,50\}$ (linhas), com os 52 429 pontos observados a partir das sequências obtidas pelo gerador Mersenne-Twister tomando 1.000 e 50.000 sequências respectivamente.

A olho nu, D é um fator relevante pois os diagramas de dispersão mostram comportamen-

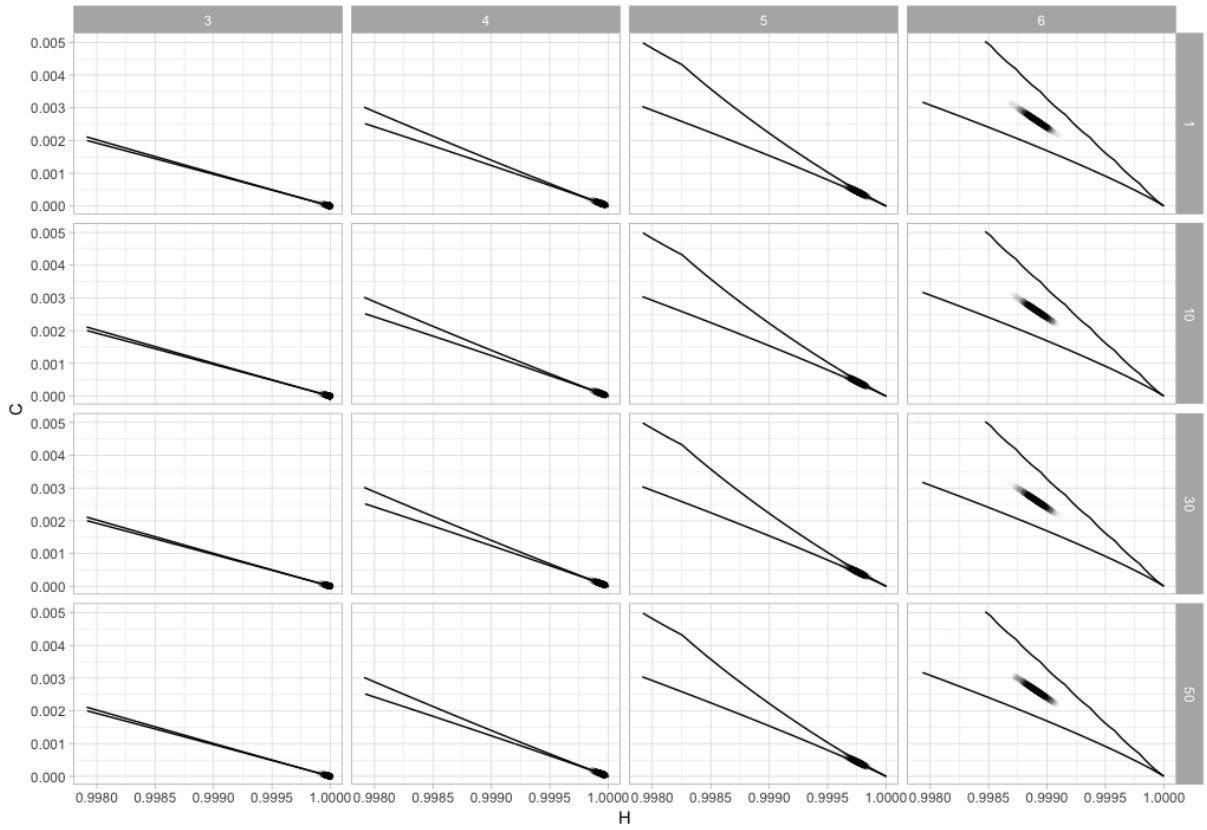


Figura 3.2: Diagramas de dispersão das sequências quânticas com 50.000 observações para $D \in \{3,4,5,6\}$ (colunas) e $\tau \in \{1,10,30,50\}$ (linhas), com curvas de complexidade mínima e máxima no plano Entropia-Complexidade.

tos que merecem uma análise mais aprofundada, por outro lado não temos certeza de como τ e o gerador influenciam os resultados.

As figuras 3.7 e 3.8 mostram os histogramas suavizados das distâncias para sequências de tamanho ($N = 1.000, 50.000$) respectivamente, com palavras de tamanho $D = 6$ e valores de *lag* ($\tau = 1,50$), sobrepondo-se os três geradores. Os gráficos sugerem que o gerador é um fator irrelevante para a distribuição da distância. Mais adiante veremos que esta impressão não se confirma totalmente.

A figura 3.9 mostra os histogramas suavizados das distâncias dos três geradores para palavras de tamanho $D = 6$ e dois valores de *lag* ($\tau = 1, 50$), sobrepondo os resultados obtidos com os dois tamanhos de sequências ($N = 1.000, 50.000$). Os gráficos sugerem fortemente que o tamanho das sequências é um fator relevante para a distribuição da distância.

A figura 3.10 mostra os histogramas suavizados das distâncias dos três geradores para dois tamanhos de sequências ($N = 1.000, 50.000$) e dois valores de *lag* ($\tau = 1, 50$), sobrepondo os resultados obtidos com os diferentes tamanhos de palavras ($D = 3$, $D = 4$, $D = 5$, $D = 6$). Os gráficos sugerem fortemente que o tamanho das palavras é um fator relevante para a distribuição da distância.

Para consolidar esta análise realizamos a seguir testes Kolmogorov-Smirnov afim de

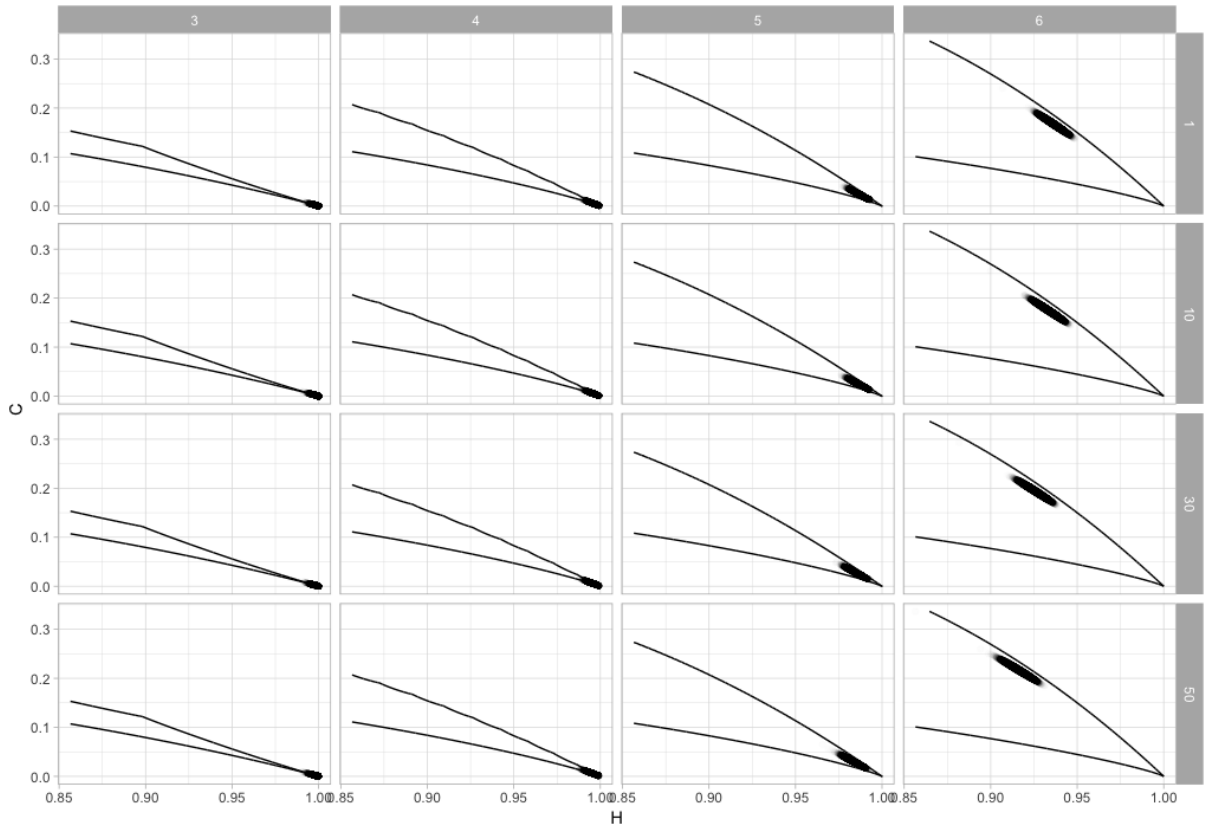


Figura 3.3: Diagramas de dispersão das sequências de rádio com 1.000 observações para $D \in \{3,4,5,6\}$ (colunas) e $\tau \in \{1,10,30,50\}$ (linhas), com curvas de complexidade mínima e máxima no plano Entropia-Complexidade.

analisar a influência dos fatores envolvidos.

A Tabela 3.1 mostra os p-valores dos testes de Kolmogorov-Smirnov aplicados a pares de distâncias calculadas sobre sequências de tamanho 1000, variando D e τ. Verificamos que há excelente aderência entre os pares de distâncias de sequências Quânticas e de Rádio. Já quando a comparação é feita com distâncias de sequências de Mersenne-Twister (M-T), a aderência diminui um pouco.

Os p-valores reportados para distâncias obtidas com sequências de tamanho $N = 1000$ não permitem concluir que haja diferenças significativas. Essa constatação será revertida ao analisar distâncias entre sequências de tamanho $N = 50000$.

A Tabela 3.2 mostra os p-valores dos testes de Kolmogorov-Smirnov aplicados a pares de distâncias calculadas sobre sequências de tamanho 50000, variando D e τ. Verificamos que há excelente aderência entre os pares de distâncias de sequências Quânticas e de Rádio. Já quando a comparação é feita com distâncias de sequências de Mersenne-Twister (M-T), a aderência diminui de forma sistemática e significativa para $\tau = 1$.

Os p-valores observados na Tabela 3.2 nos levam a concluir que não é possível desconsiderar a fonte de dados como um fator relevante quando se trata do gerador de Mersenne-Twister. Já as distâncias das sequências produzidas pelos geradores Quântico e de Rádio são indistin-

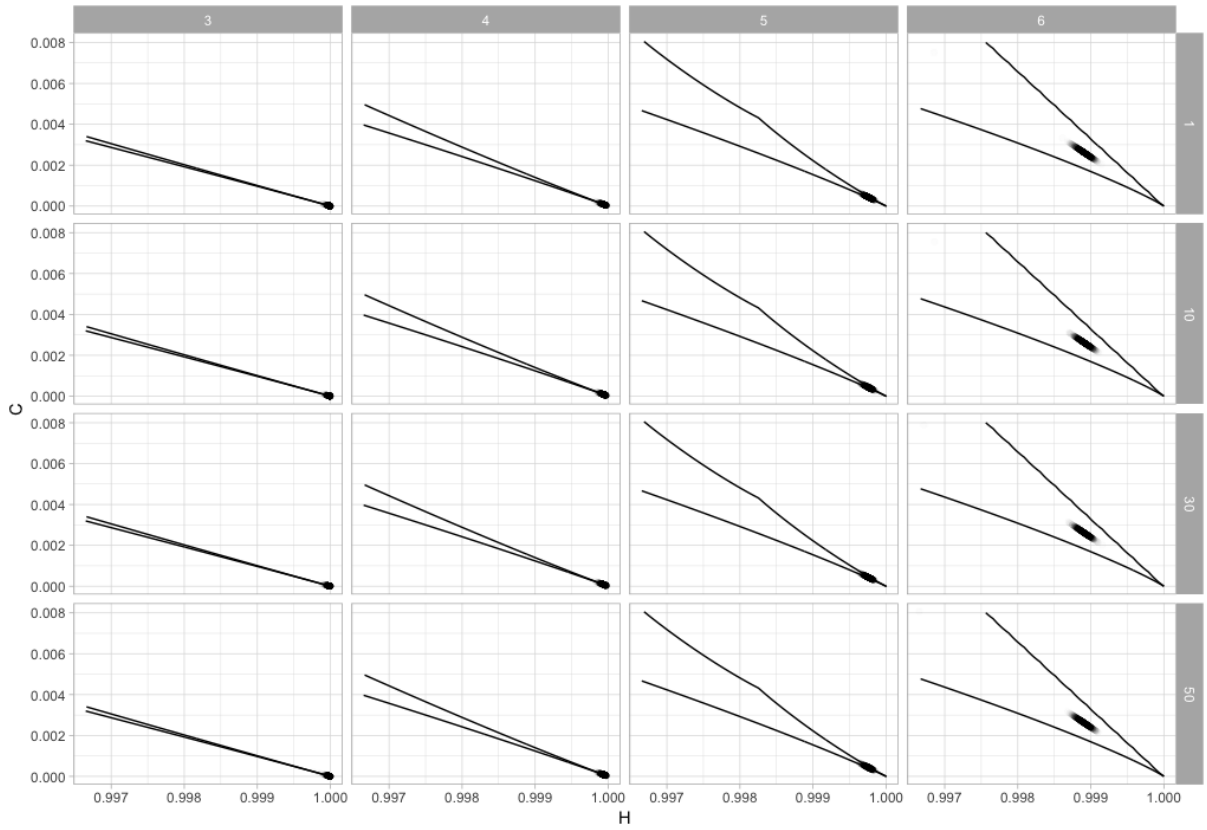


Figura 3.4: Diagramas de dispersão das sequências de rádio com 50.000 observações para $D \in \{3,4,5,6\}$ (colunas) e $\tau \in \{1,10,30,50\}$ (linhas), com curvas de complexidade mínima e máxima no plano Entropia-Complexidade.

guíveis e, portanto, não podemos descartar a hipótese dessas fontes serem idênticas para a medida considerada.

A figura 3.12 sugere que o comportamento das distâncias euclidianas ao ponto de referência muda conforme o tamanho do padrão D varia. Além disso, também são perceptíveis mudanças de comportamento em função do *lag* τ quando $D = 5{,}6$.

Da análise aqui apresentada concluímos que, à luz da distância do ponto característico de uma sequência ao ponto de referência, os dois geradores físicos produzem sequências indistinguíveis. Diante disso, nos cálculos subsequentes faremos a junção desses conjuntos de dados, que denominaremos simplesmente “aleatórios” no que segue.

Já o gerador de Mersenne-Twister apresenta diferenças em relação aos geradores de origem física. Por se tratar de um gerador algorítmico e, portanto, pseudoaleatório, ele será tratado como objeto de análise e não como padrão para estabelecer critérios de qualidade.

No que segue analisaremos o comportamento dessas distâncias em detalhes.

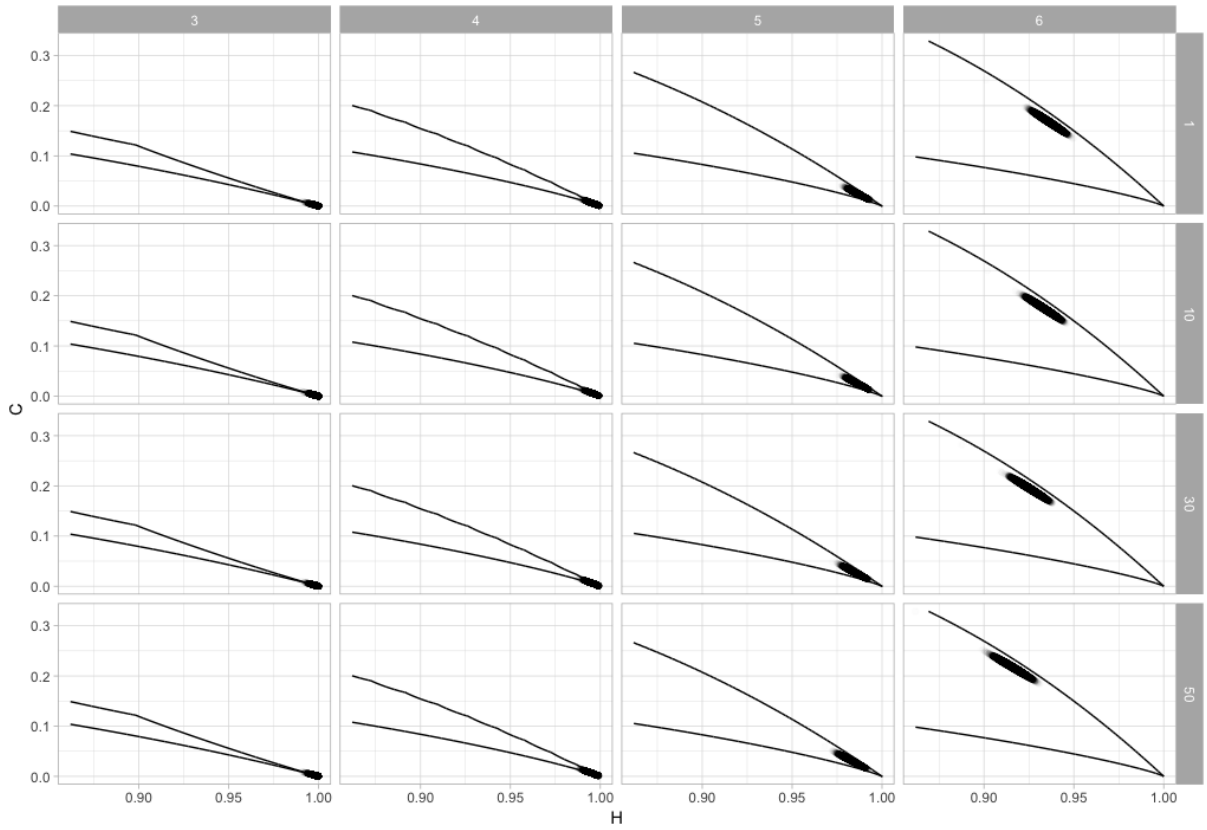


Figura 3.5: Diagramas de dispersão das sequências de Mersenne-Twister com 1.000 observações para $D \in \{3,4,5,6\}$ (colunas) e $\tau \in \{1,10,30,50\}$ (linhas), com curvas de complexidade mínima e máxima no plano Entropia-Complexidade.

3.2 Análise das regiões de confiança

Feita a junção das distâncias dos pontos característicos ao ponto de referência das sequências produzidas pelos geradores quântico e de rádio (sequências aleatórias), para cada situação de $N = 1.000, 50.000$, $D = 3,4,5,6$ e $\tau = 1,10,30,50$, o próximo passo consiste em calcular os quantis relevantes.

Inicialmente ilustraremos apenas duas situações. A figura 3.13 mostra os padrões das sequências aleatórias para o caso $N = 1.000$, $D = 3$ e $\tau = 1$ no plano Entropia-Complexidade, junto com os quantis de 90 %, 95 %, 99 % e 99.9 % em escala linear (figura 3.13(a)) e em escala logarítmica (figura 3.13(b)).

A figura 3.14 mostra os padrões das sequências aleatórias para o caso $N = 50.000$, $D = 6$ e $\tau = 50$ no plano Entropia-Complexidade, junto com os quantis de 90 %, 95 %, 99 % e 99.9 % em escala linear (figura 3.14(a)) e em escala logarítmica (figura 3.14(b)).

Nestas duas figuras, os pontos característicos foram desenhados com um gradiente de cores que vai do amarelo ao preto em função da distância ao ponto de referência. Identificamos também os valores de Entropia e de Complexidade Estatística correspondentes a cada quantil de interesse, estes últimos plotados em vermelho.

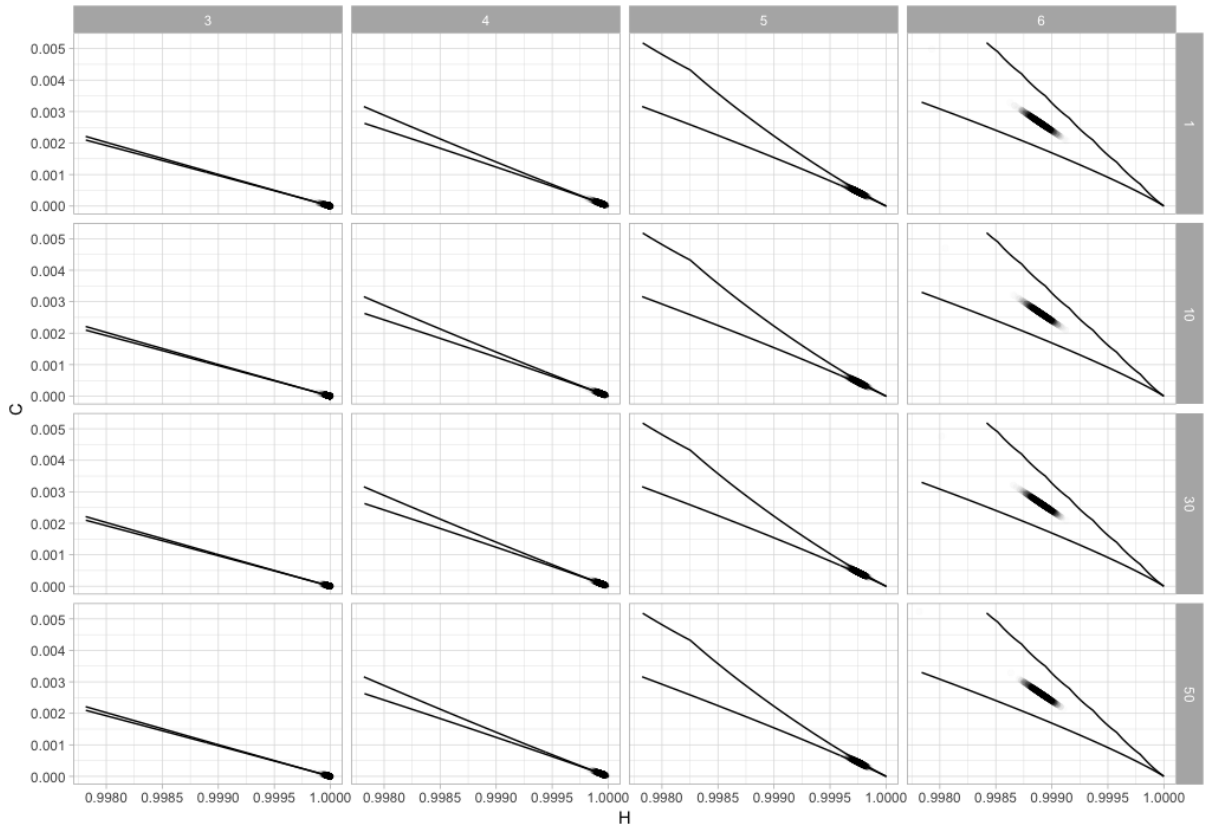


Figura 3.6: Diagramas de dispersão das sequências de Mersenne-Twister com 50.000 observações para $D \in \{3,4,5,6\}$ (colunas) e $\tau \in \{1,10,30,50\}$ (linhas), com curvas de complexidade mínima e máxima no plano Entropia-Complexidade.

Os resultados centrais dessa dissertação são exibidos nas figuras 3.15, 3.16, 3.17 e 3.18 (os quantis de interesse para amostras de tamanho 1.000 e $\tau = 1,10,30,50$), e nas figuras 3.19, 3.20, 3.21 e 3.22 (os quantis de interesse para amostras de tamanho 50.000 e $\tau = 1,10,30,50$). Cada uma destas figuras inclui as quatro situações de $D = 3,4,5,6$.

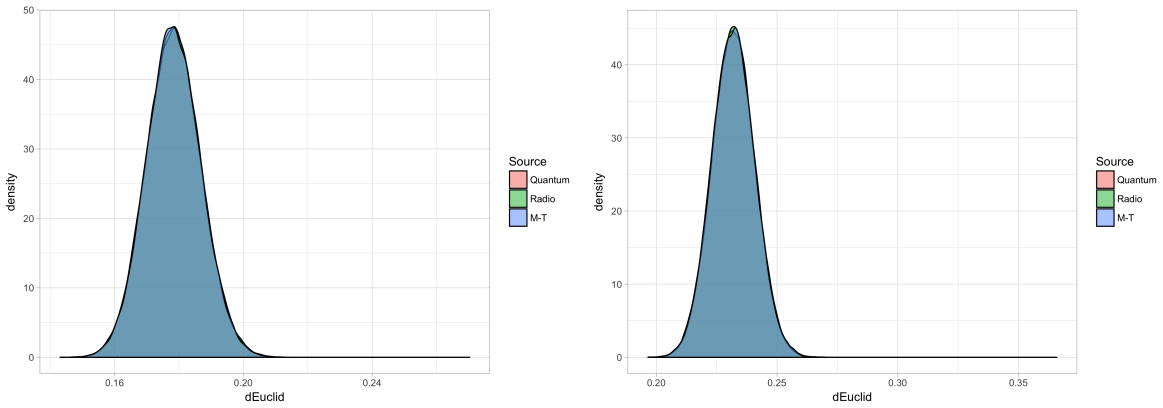
Adicionalmente, aos gráficos são apresentadas tabelas 3.3 e 3.4 mostrando os valores da distância euclidiana dos pontos de interesse dos quantis ao ponto de referência.

3.3 Aplicações

Nesta seção mostramos a aplicação da nossa proposta a sequências de tamanho 1000.

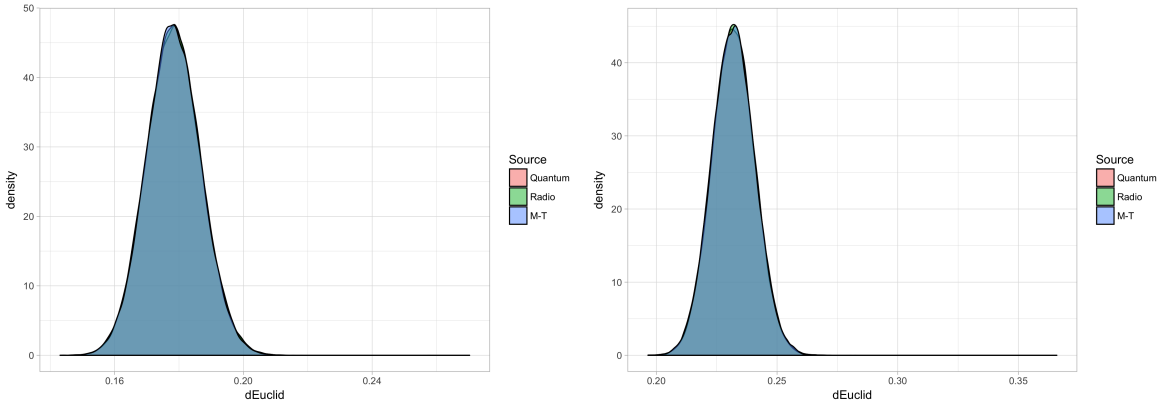
Utilizando a metodologia descrita, aplicamos o teste a uma seqência de 1000 observações produzidas pelos geradores Mersenne-Twister e Randu, além de séries não estacionárias, estacionárias e mapas logísticos, todos com a mesma dimensão.

Os resultados de aplicar a nossa abordagem às sequências oriundas dos geradores de números pseudoaleatórios podem ser observados na figura 3.23: Mersenne-Twister na figura 3.23(a) Randu na figura 3.23(b). Para obter as sequências de Mersenne-Twister utilizou-se a implementação interna do *R* (ver código A.1), já para as sequências de Randu foi implemen-



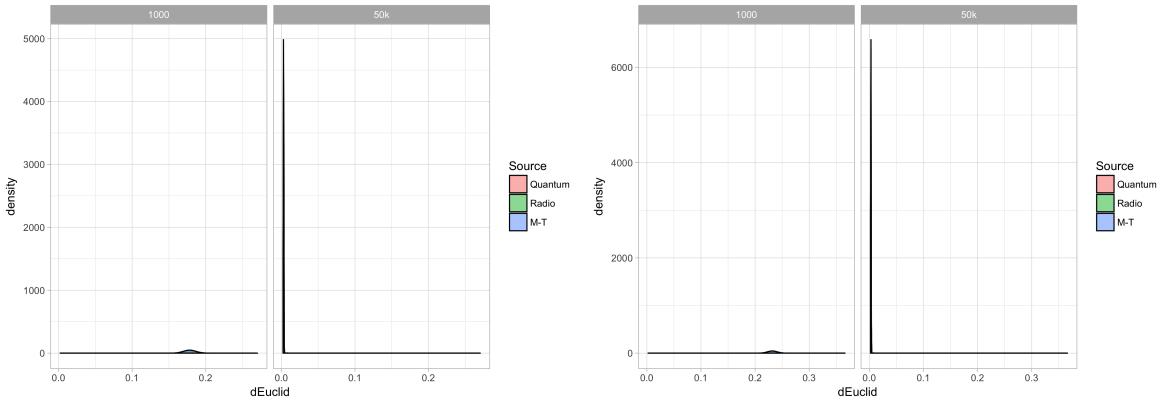
(a) $D = 6, \tau = 1$ (b) $D = 6, \tau = 50$

Figura 3.7: Histogramas suavizados de situações que sugerem que o gerador é um fator irrelevante para $N = 1.000$



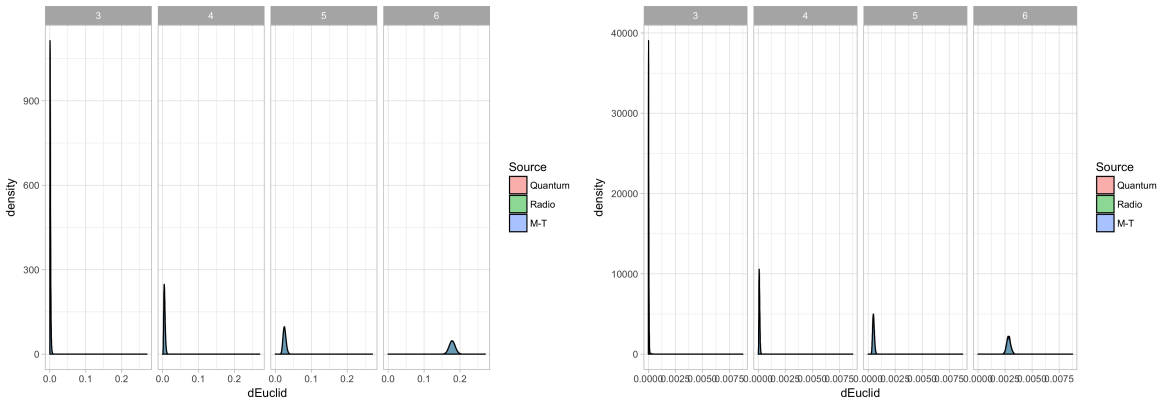
(a) $D = 6, \tau = 1$ (b) $D = 6, \tau = 50$

Figura 3.8: Histogramas suavizados de situações que sugerem que o gerador é um fator irrelevante também para $N = 50.000$



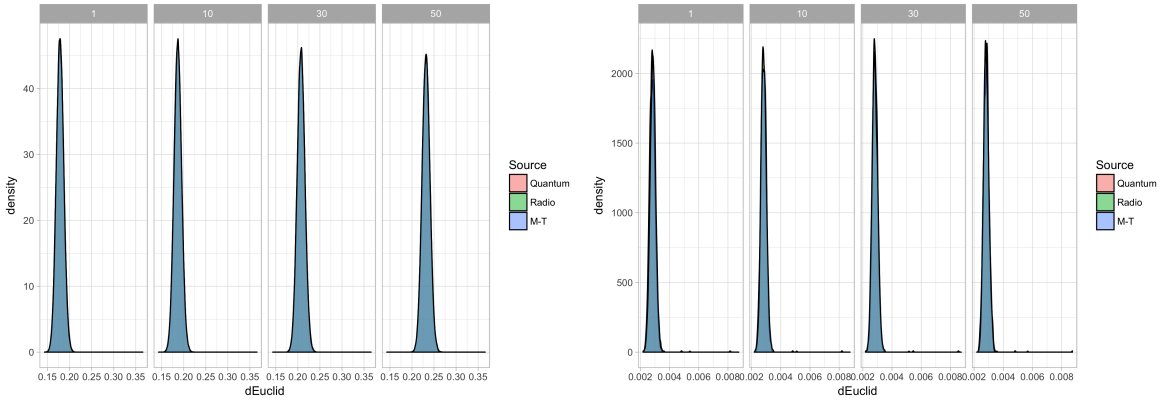
(a) $D = 6, \tau = 1$ (b) $D = 6, \tau = 50$

Figura 3.9: Histogramas suavizados de situações que sugerem que o N é um fator relevante



(a) $N = 1.000$, $\tau = 1$ (b) $N = 50.000$, $\tau = 50$

Figura 3.10: Histogramas suavizados de situações que sugerem que o D é um fator relevante



(a) $N = 1.000$, $D = 6$ (b) $N = 50.000$, $D = 6$

Figura 3.11: Histogramas suavizados de situações que sugerem que o τ é um fator relevante

Tabela 3.1: Teste de Kolmogorov-Smirnov aplicado a pares de sequências para 1.000 observações.

Par	D	$\tau = 1$	$\tau = 10$	$\tau = 30$	$\tau = 50$
Quântica vs. Rádio	$D = 3$	0.18034676	0.08582490	0.58096350	0.32626542
	$D = 4$	0.21708776	0.60690204	0.08116764	0.46372312
	$D = 5$	0.32388371	0.53394280	0.46970138	0.02768674
	$D = 6$	0.61501858	0.63403661	0.54795745	0.15353799
Quântica vs. M-T	$D = 3$	0.09400120	0.22995096	0.36766759	0.03706359
	$D = 4$	0.25188769	0.35844686	0.16768952	0.18237754
	$D = 5$	0.97552039	0.79878301	0.12852918	0.08764347
	$D = 6$	0.47615384	0.42420007	0.55290011	0.79669144
Rádio vs. M-T	$D = 3$	0.008560614	0.496450214	0.982419336	0.390237891
	$D = 4$	0.003804157	0.229503619	0.629158543	0.651783589
	$D = 5$	0.254216237	0.179697451	0.824743440	0.071709252
	$D = 6$	0.846441994	0.033726860	0.493286733	0.184856861

tada função A.2.

As figuras 3.23(a) e 3.23(b) mostram que a nossa técnica não é capaz de identificar desvios significativos da hipótese dessas duas fontes de dados não produzirem ocorrências de variáveis independentes e identicamente distribuídas. Embora já tenhamos mostrado evidências nesse sentido, através da figura 1.3 para Randu e dos testes de Kolmogorov-Smirnov para o gerador Mersenne-Twister (tabelas 3.1 e 3.2), as distâncias dos pontos característicos ao ponto de referência é inferior aos limites calculados para a rejeição da hipótese. Isso se deve ao poder limitado do teste, característica que será investigada em futuros trabalhos.

A seguir, geramos sequências estocásticas com estrutura de autocorrelação: uma estacionária (ruído gaussiano filtrado) e uma não estacionária (uma trajetória de movimento browniano).

A aplicação do teste à sequência obtida pelo método de séries não estacionárias pode ser observada na figura 3.24. O teste leva à rejeição da sequência, pois para qualquer tamanho de palavra $D = 3,4,5,6$ os descritores levam a pontos cuja distância euclidiana é maior do que quaisquer dos limites estabelecidos. O código A.3 foi utilizado para gerar a sequência; ver apêndice A.

A uma série estacionária em que aplicamos nosso teste foi obtida pela convolução de uma sequência de variáveis aleatórias independentes e identicamente distribuídas segundo uma lei gaussiana padrão, convolucionada com uma máscara de tamanho 3 com valores não negativos: $(\beta,1,\beta)$, $0 \leq \beta \leq 1$. Quando $\beta = 0$ temos a sequência original, e para valores crescentes de β temos sequências com cada vez maior estrutura de correlação. Calculamos o ponto característico da série assim obtida, e o contrastamos com os quantis empíricos obtidos anteriormente.

A aplicação do teste à sequência obtida pelo método de séries estacionárias pode ser

Tabela 3.2: Teste de Kolmogorov-Smirnov aplicado a pares de sequências para 50.000 observações.

Par	D	$\tau = 1$	$\tau = 10$	$\tau = 30$	$\tau = 50$
Quântica vs. Rádio	$D = 3$	0.13862662	0.93677447	0.07714702	0.46405291
	$D = 4$	0.68079537	0.90035466	0.60801914	0.77908261
	$D = 5$	0.14371256	0.76662067	0.64456996	0.91315843
	$D = 6$	0.02268670	0.49307044	0.53135926	0.30074267
Quântica vs. M-T	$D = 3$	6.074571e-10	1.388898e-01	2.682058e-01	4.822849e-01
	$D = 4$	6.592620e-09	2.987721e-02	8.438134e-01	7.923220e-01
	$D = 5$	9.424114e-08	3.289299e-02	7.681676e-01	8.405168e-01
	$D = 6$	9.058821e-04	7.075225e-01	2.982731e-01	3.614763e-01
Rádio vs. M-T	$D = 3$	1.226271e-09	1.609665e-01	1.465970e-01	1.914937e-01
	$D = 4$	2.190462e-10	1.277762e-02	2.069821e-01	9.963221e-01
	$D = 5$	1.438372e-11	1.267622e-02	4.364935e-01	9.999998e-01
	$D = 6$	3.749001e-06	2.263466e-01	7.419248e-01	4.014641e-01

observada na figura 3.25. Novamente, o teste leva à rejeição da sequência pois para qualquer $D = 3,4,5,6$ os descritores produzem pontos a distâncias euclidianas maiores do que as estipuladas pelos intervalos de confiança estabelecidos. O código A.4 foi utilizado para gerar a sequência; ver apêndice A.

As abscissas da figura 3.25 descrevem a variação de β, e para cada valor calculamos a distância euclidiana ao ponto de referência (1,0) para palavras de tamanho $D = 6$ e $\tau = 1$ No gráfico desenhamos também os quantis que correspondem a essa situação, e verificamos quais situações não são rejeitadas a cada nível de significância:

90 %**:** máscaras com $\beta \leq 0{,}15$.

95 %**:** máscaras com $\beta \leq 0{,}17$.

99 %**:** máscaras com $\beta \leq 0{,}17$.

99.9 %**:** máscaras com $\beta \leq 0{,}20$.

Estes resultados sugerem a necessidade de se fazer uma análise exaustiva do poder do teste em função de uma diversidade de parâmetros. Esse estudo é objeto de trabalhos futuros.

A figura 3.26 mostra os pontos característicos da sequência de valores estacionários obtida com $\beta = 1$, para $D = 3,4,5,6$ e $\tau = 1$. Percebe-se que ao aumentar o *lag* perde-se o poder discriminatório do teste, pois nenhuma delas é considerada excessivamente afastada do ponto de referência.

A seguir analisamos uma série determinística com comportamento caótico: o mapa logístico. O mapa logístico é a sequência obtida pela recursão

$$x_{n+1} = r\,x_n(1 - x_n), \tag{3.1}$$

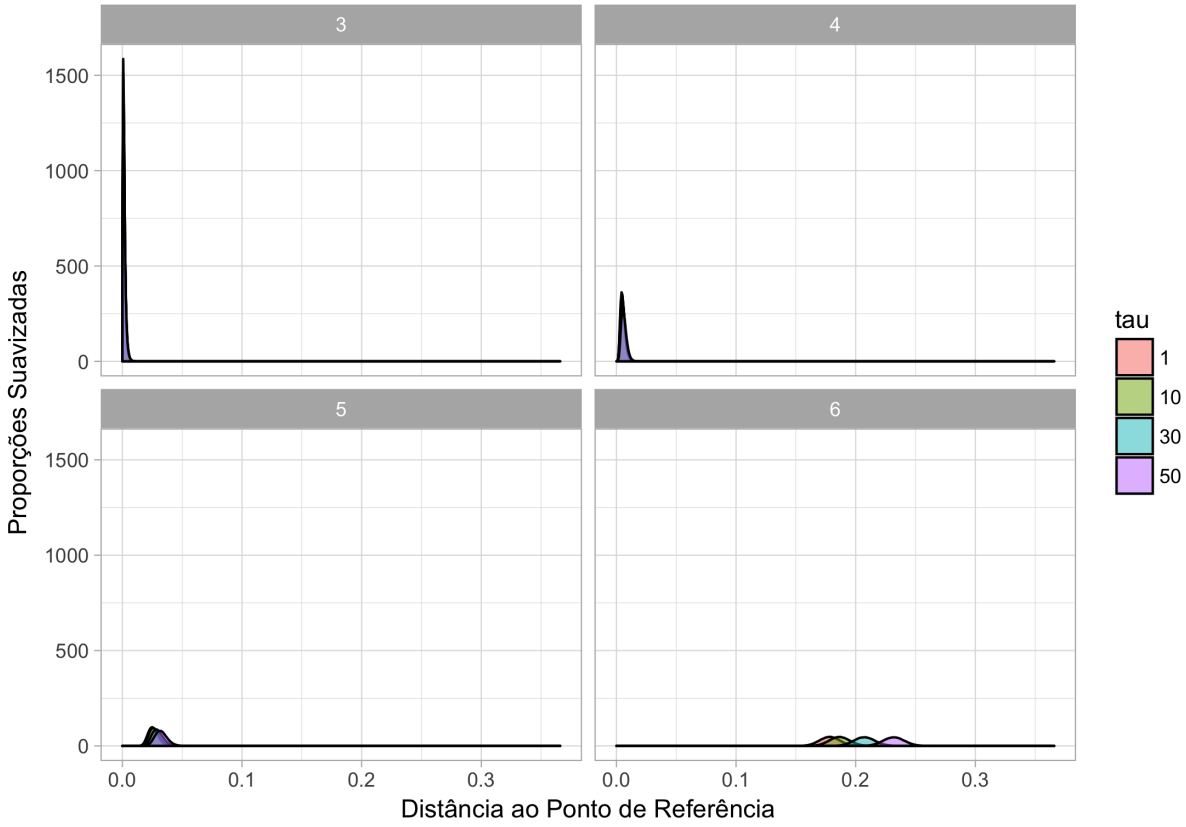
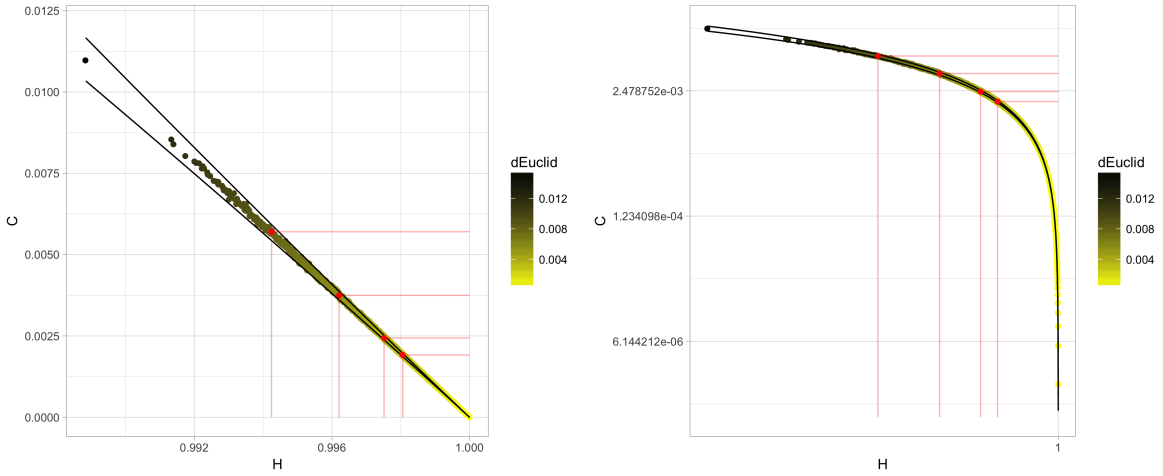


Figura 3.12: Histogramas suavizados das distâncias euclidianas dos padrões ao ponto de referência, para $D \in \{3,4,5,6\}$ e $\tau \in \{1,10,30,50\}$.

com $0 < x_1 < 1$ e $0 < r \leq 4$. Utilizamos $r = 4$, $x_0 = 0.01$. Iteramos o mapa 10000 vezes para alcançar estabilidade, e só coletamos a partir de $n = 10001$.

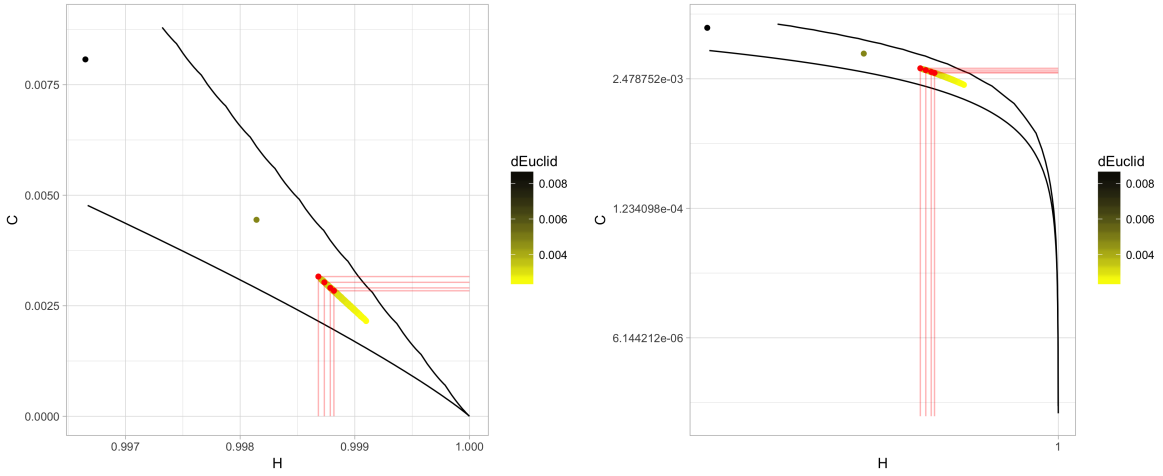
A aplicação do teste à sequência obtida pelo mapa logístico pode ser observada na figura 3.27. Mais uma vez, o teste leva à rejeição da sequência pois para qualquer $D = 3,4,5,6$ os descritores produzem pontos a distâncias euclidianas maiores do que as estipuladas pelos intervalos de confiança estabelecidos. O código A.5 foi utilizado para gerar a sequência; ver apêndice A.

Convém notar que o mapa logístico já foi usado como fonte de dados pseudoaleatórios, e este resultado mostra que ele não é adequado (com os parâmetros aqui escolhidos) para esse propósito.



(a) Escala linear (b) Escala logarítmica

Figura 3.13: Diagramas de dispersão das sequências aleatórias para o caso $N = 1.000$, $D = 3$ e $\tau = 1$.



(a) Escala linear (b) Escala logarítmica

Figura 3.14: Diagramas de dispersão das sequências aleatórias para o caso $N = 50.000$, $D = 6$ e $\tau = 50$.

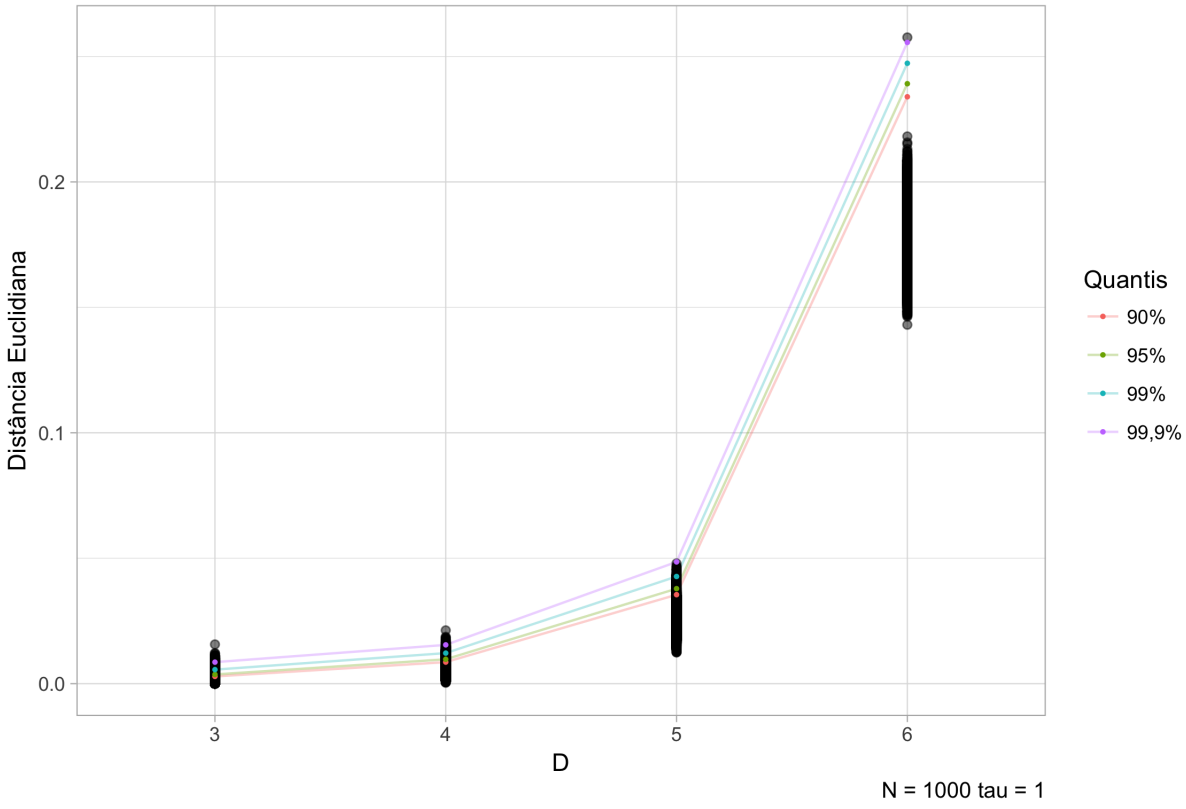


Figura 3.15: Intervalos de confiança para o caso $N = 1.000$ e $\tau = 1$.

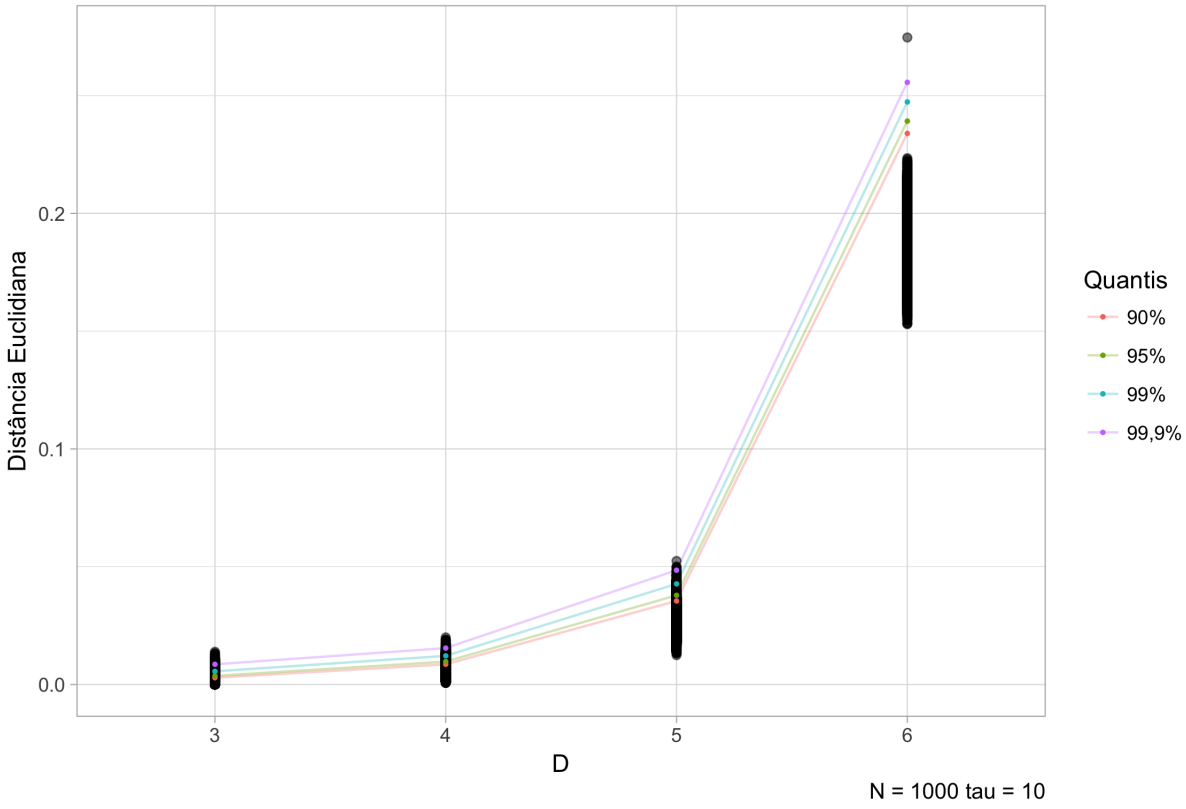


Figura 3.16: Intervalos de confiança para o caso $N = 1.000$ e $\tau = 10$.

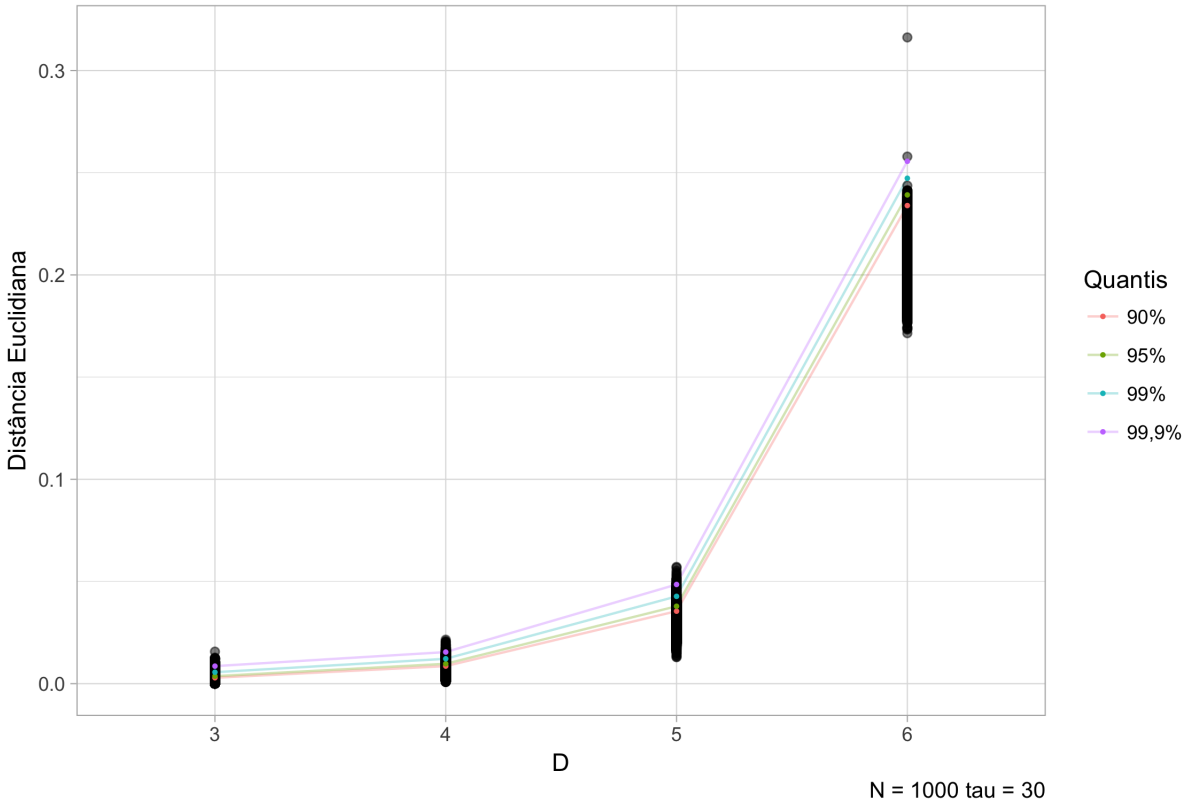


Figura 3.17: Intervalos de confiança para o caso $N = 1.000$ e $\tau = 30$.

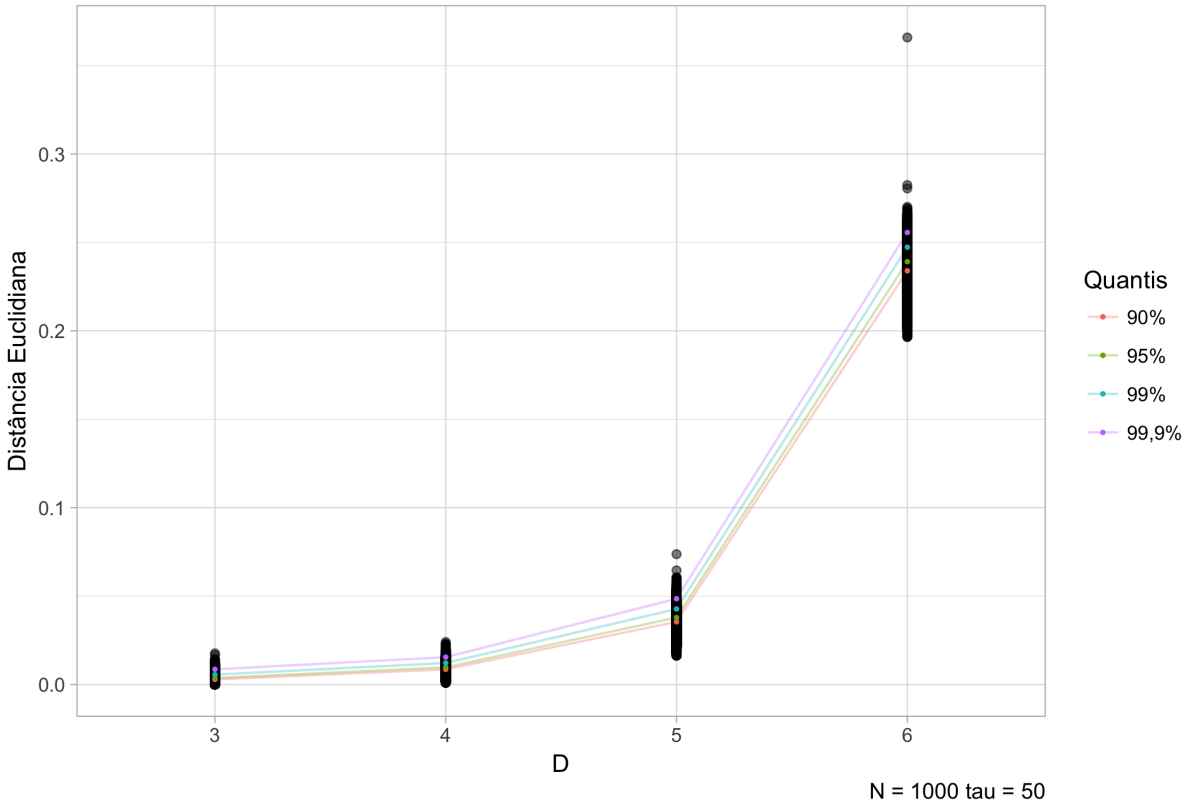


Figura 3.18: Intervalos de confiança para o caso $N = 1.000$ e $\tau = 50$.

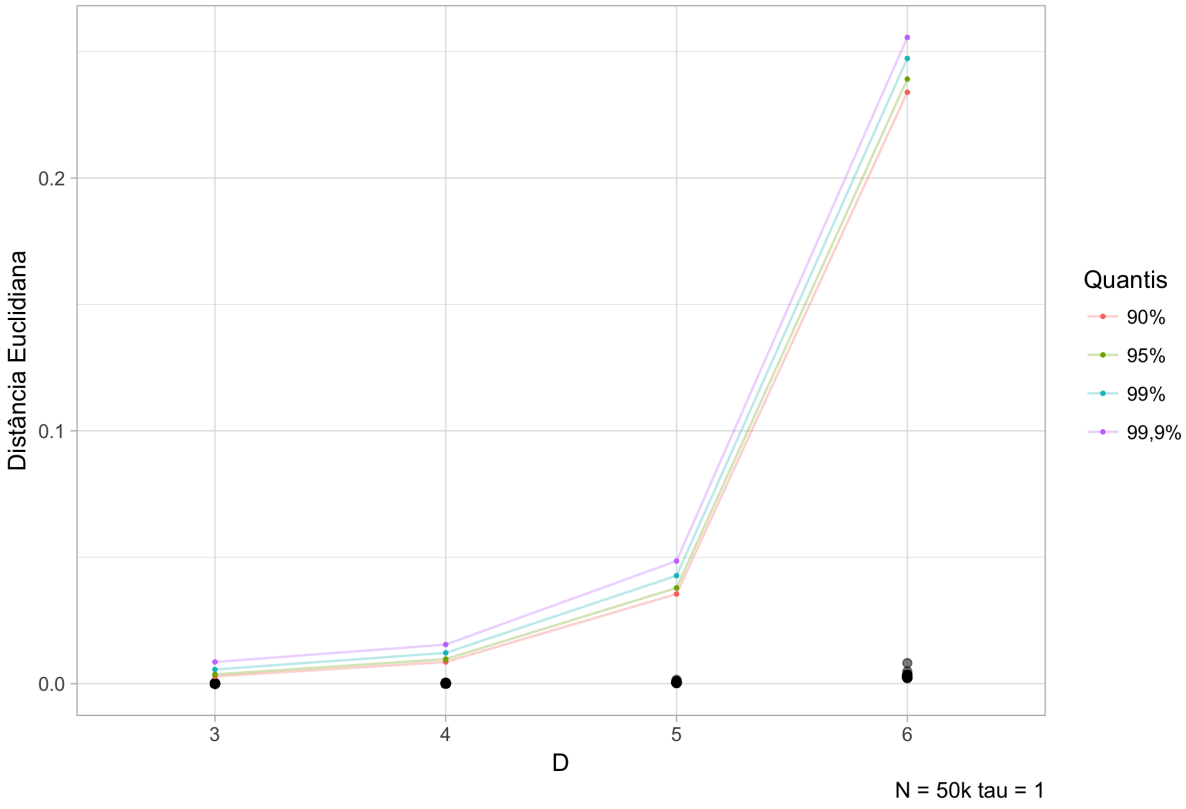


Figura 3.19: Intervalos de confiança para o caso $N = 50.000$ e $\tau = 1$.

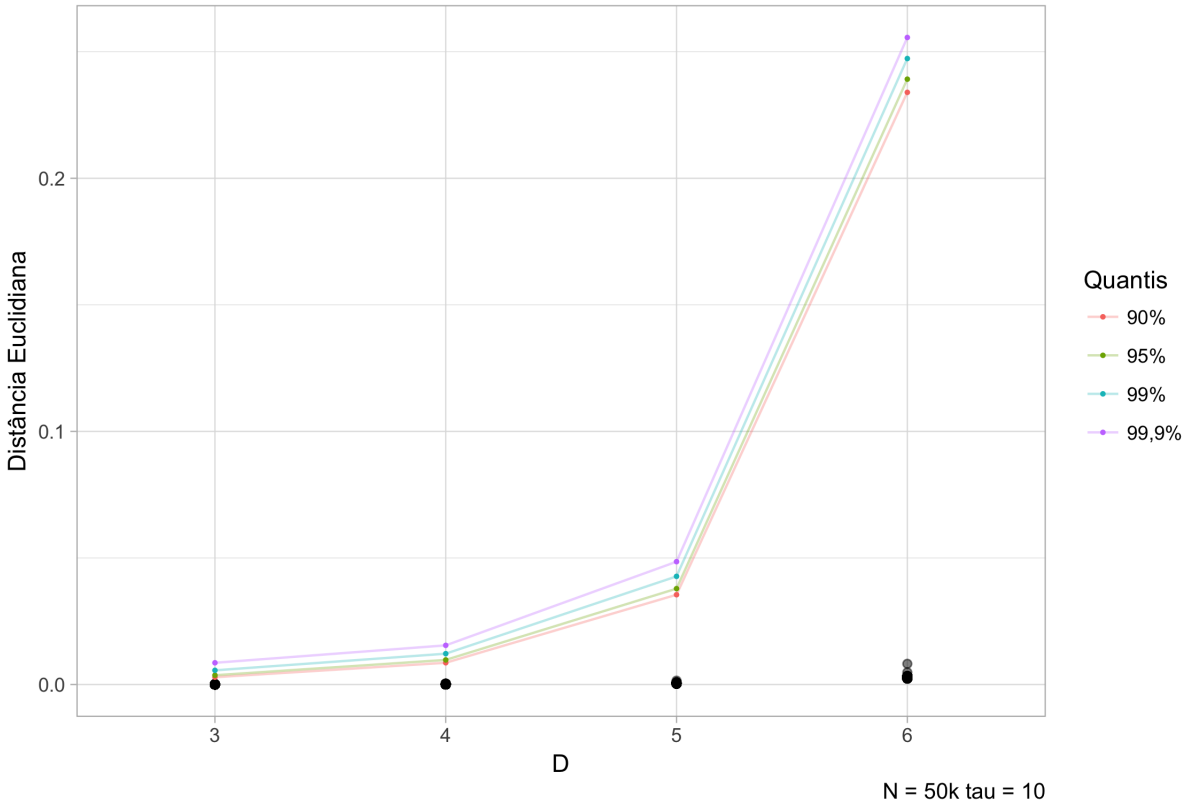


Figura 3.20: Intervalos de confiança para o caso $N = 50.000$ e $\tau = 10$.

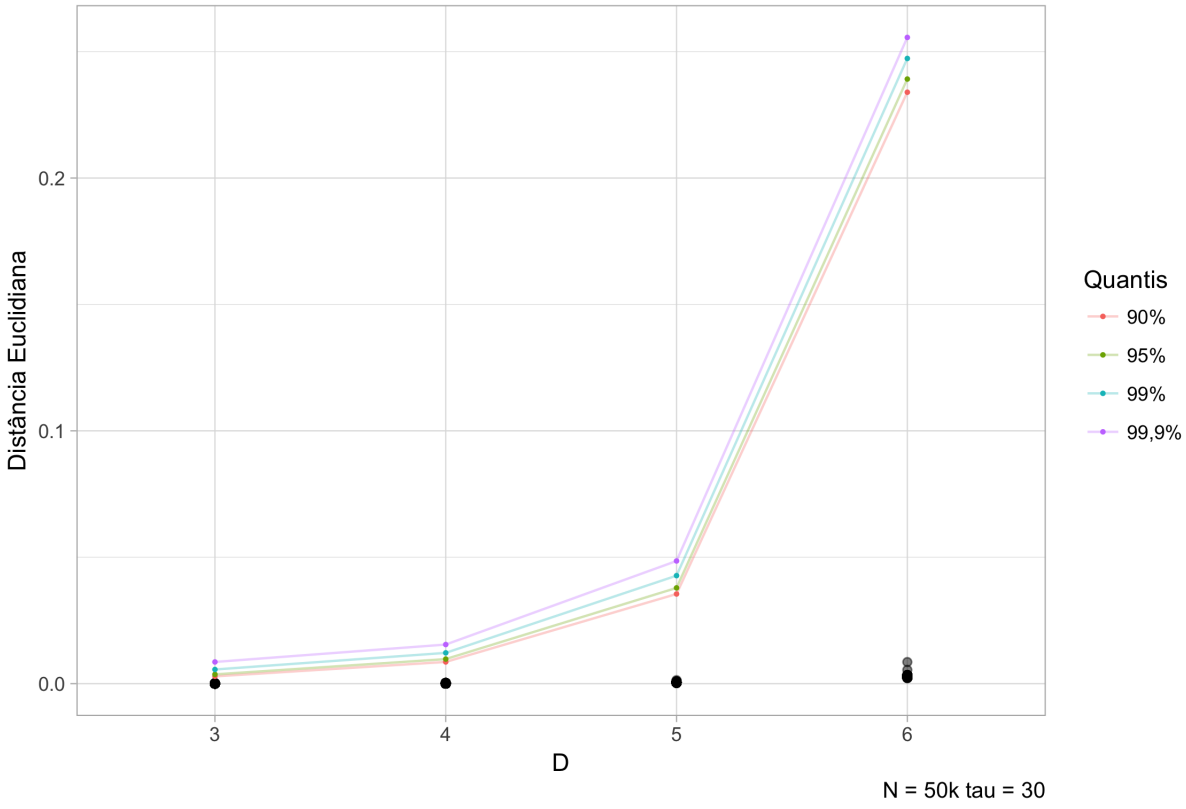


Figura 3.21: Intervalos de confiança para o caso $N = 50.000$ e $\tau = 30$.

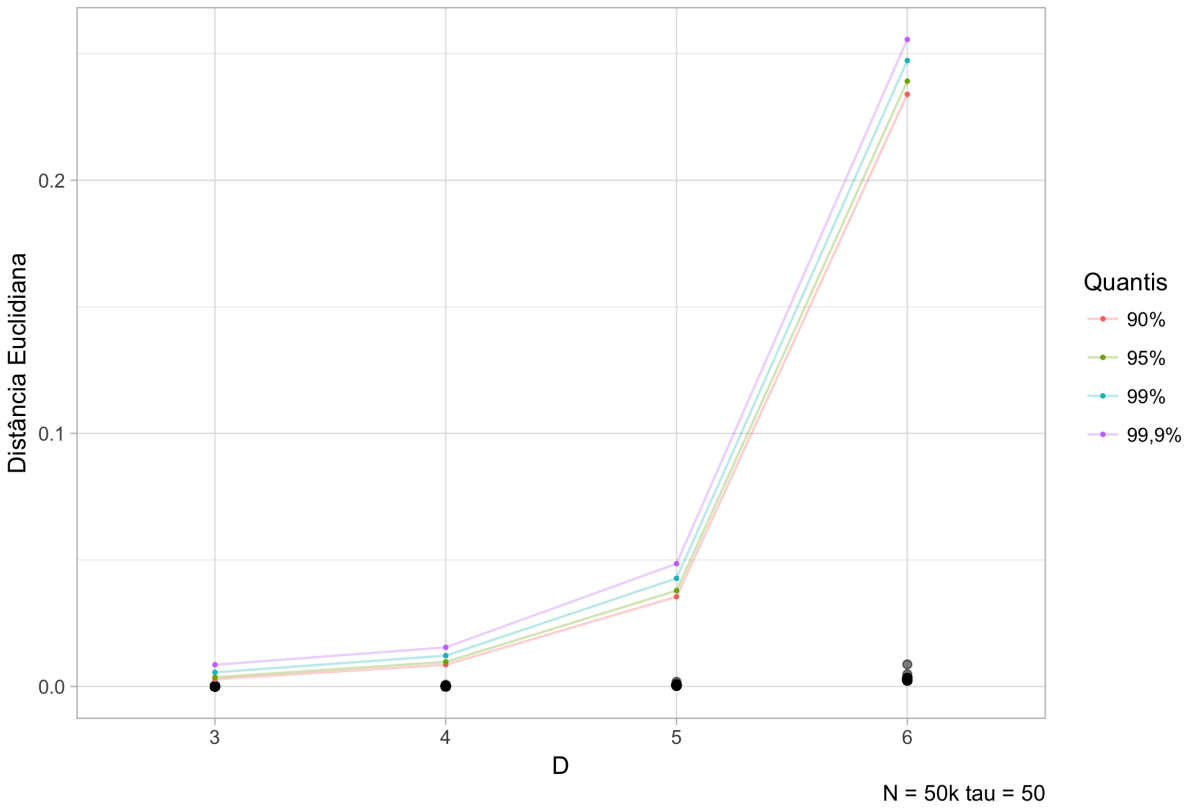
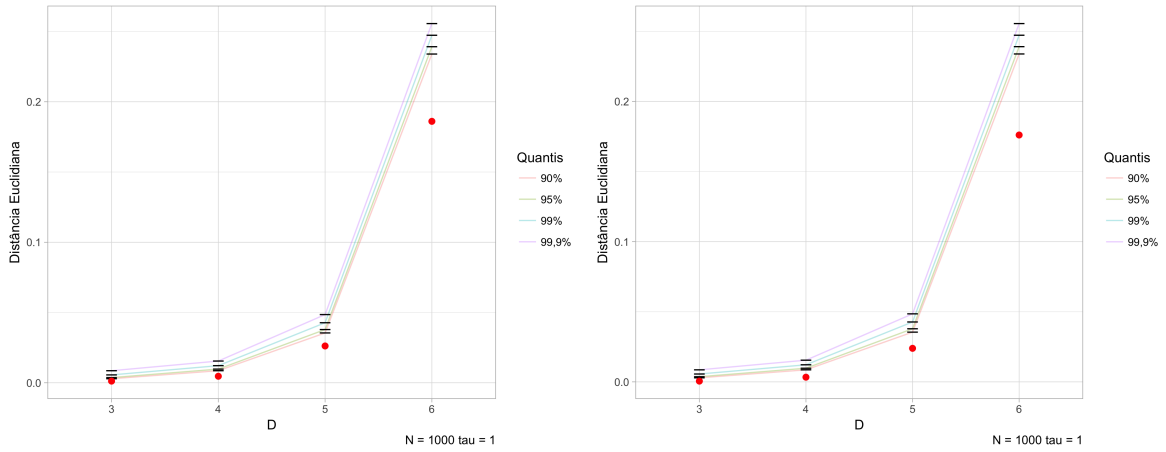


Figura 3.22: Intervalos de confiança para o caso $N = 50.000$ e $\tau = 50$.



(a) Mersenne-Twister. (b) Randu.

Figura 3.23: Aplicação do teste aos pontos de Mersenne-Twister e Randu

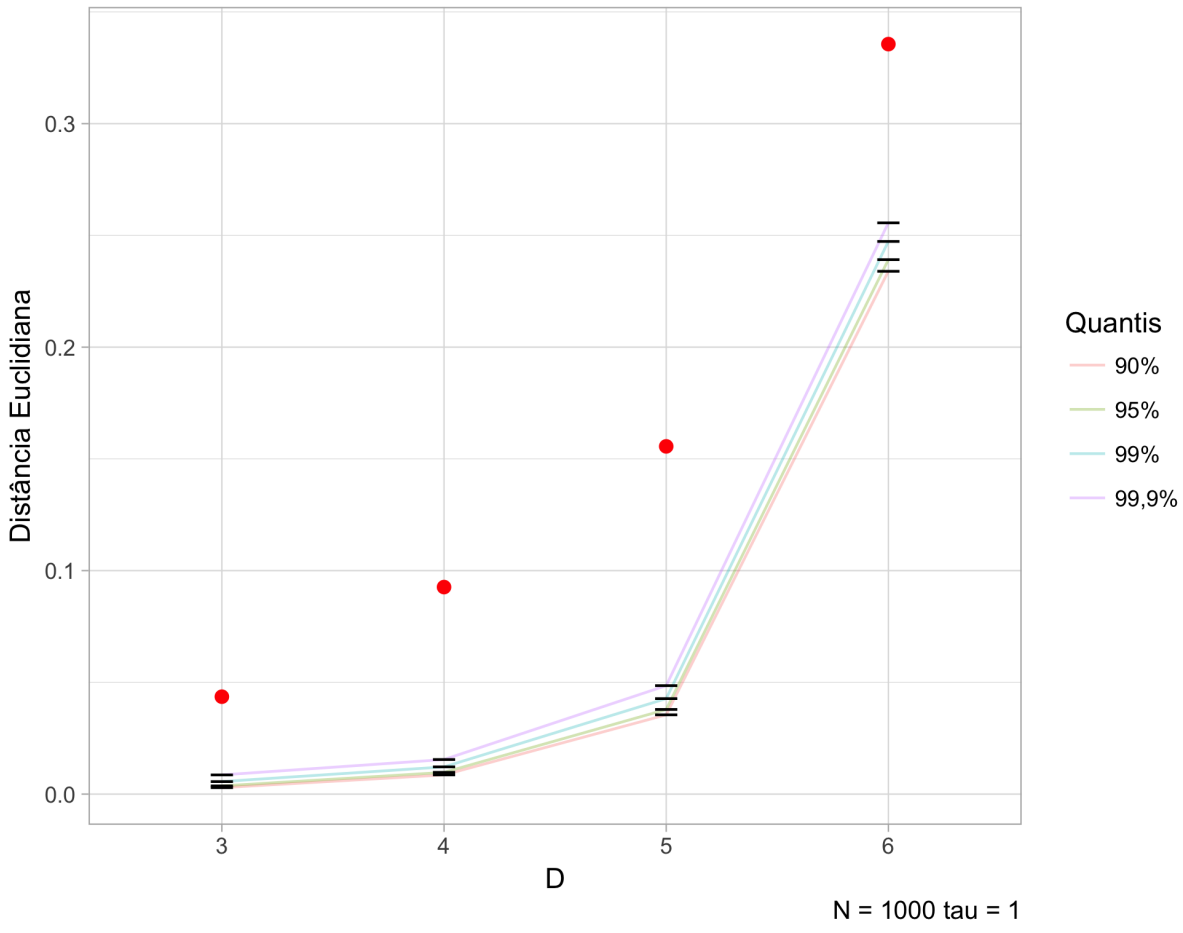


Figura 3.24: Pontos característicos das séries não estacionárias e intervalos de confiança.

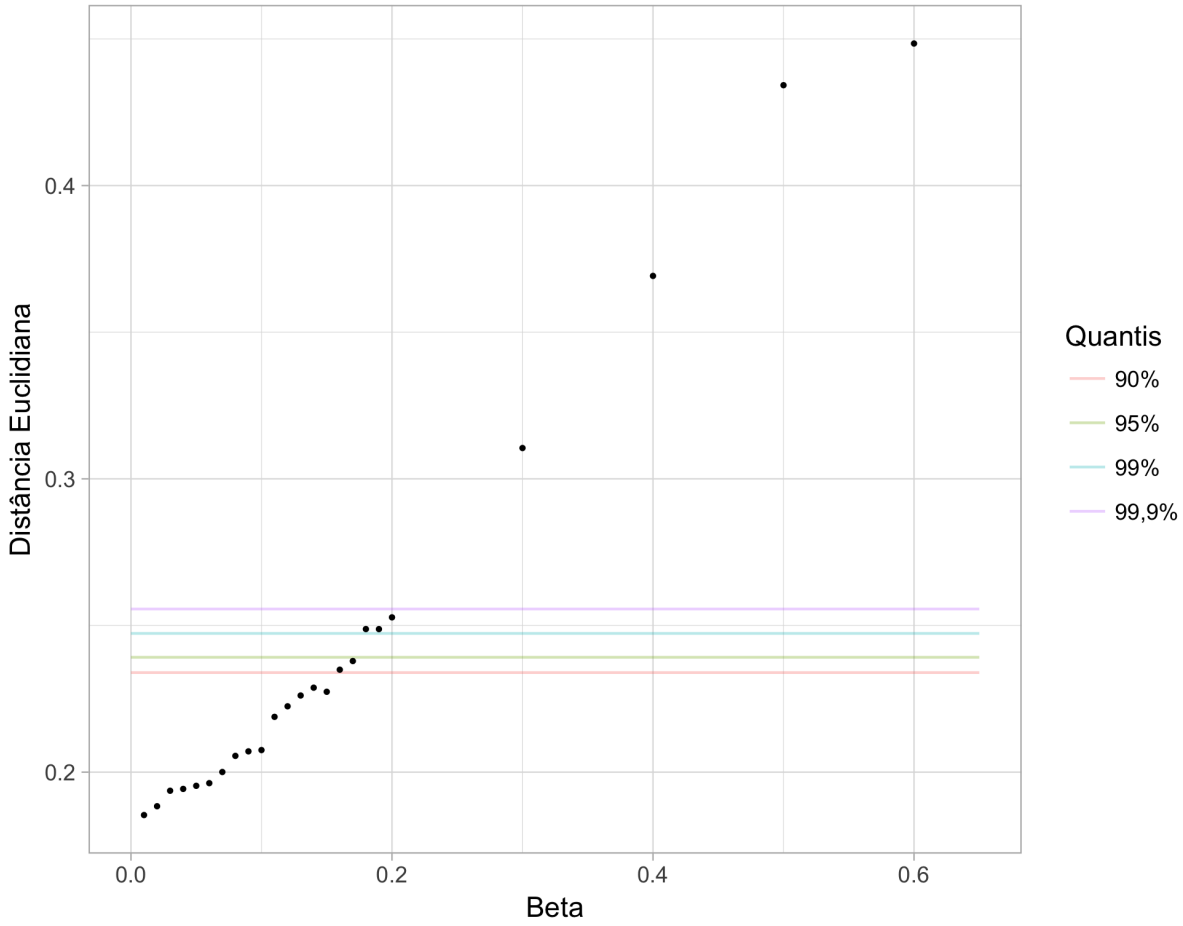


Figura 3.25: Poder do teste aplicado a uma sequência de séries estacionárias no caso particular $N = 1.000$, $\tau = 1$, variando-se a máscara de convolução $(\beta,1,\beta)$.

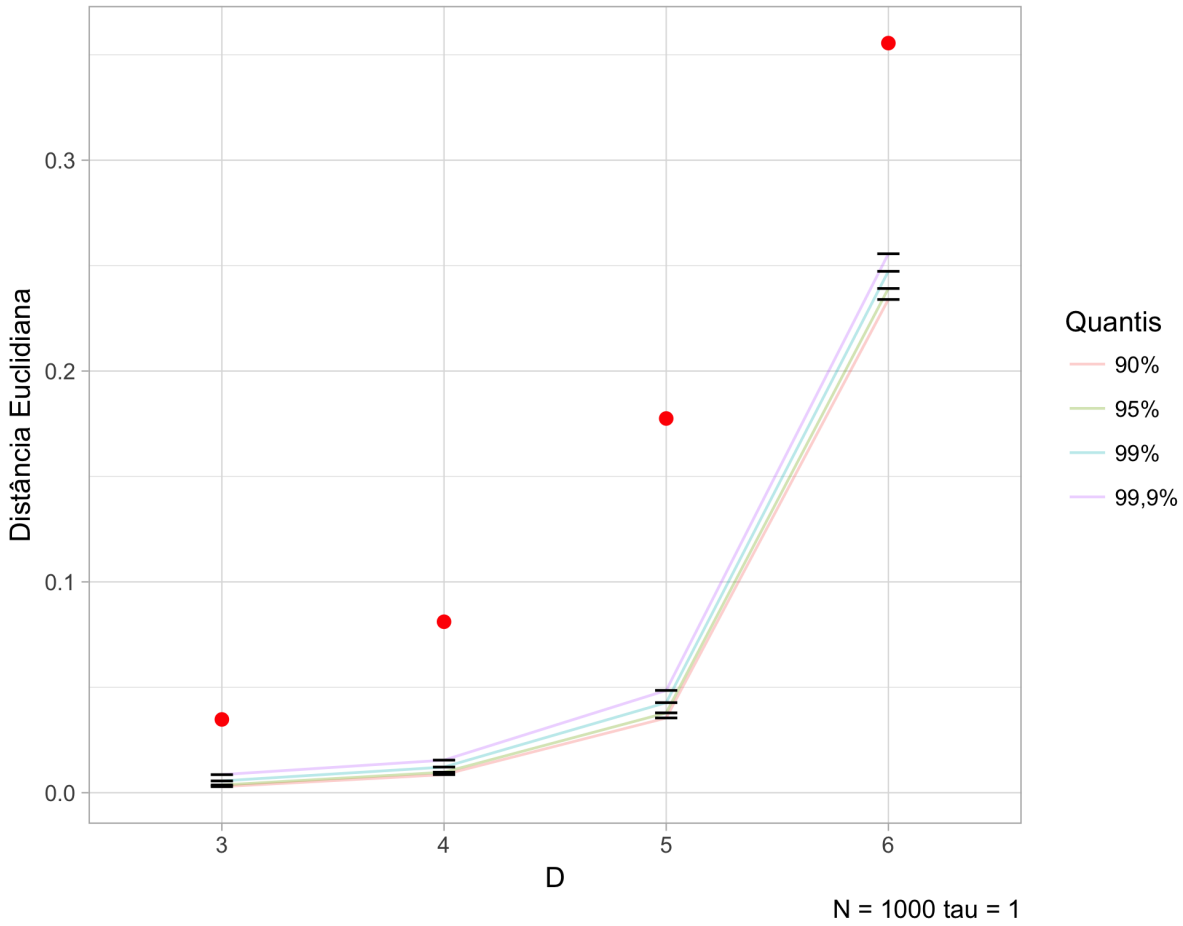


Figura 3.26: Pontos característicos das séries estacionárias e intervalos de confiança.

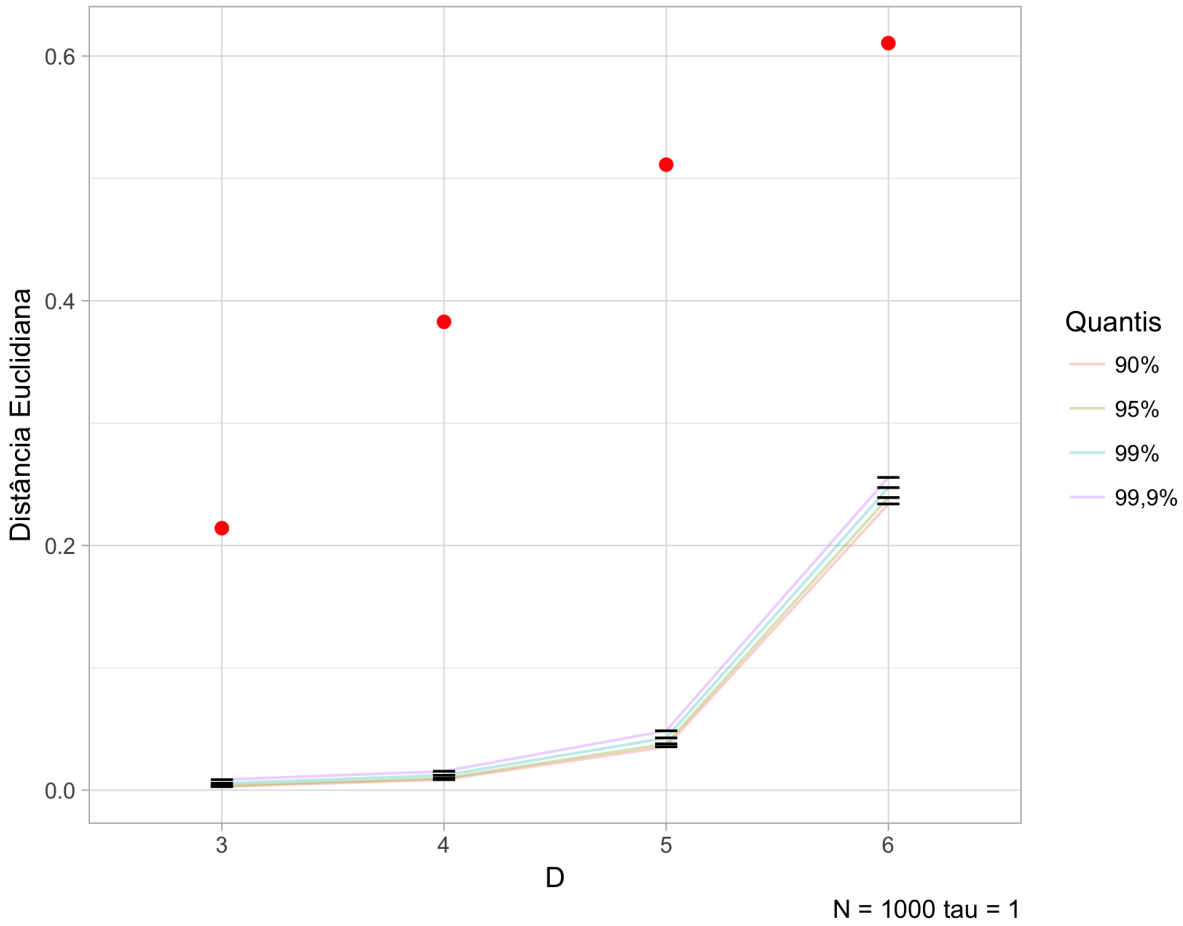


Figura 3.27: Pontos característicos do mapa logístico e intervalos de confiança.

Tabela 3.3: Quantis das distâncias euclidianas para os valores de $D = 3,4,5,6$ e $\tau = 1,10,30,50$ para sequências de 1.000 observações.

$N = 1.000$	D	τ	90 %	95 %	99 %	99.9 %
	3	1	2.728065e-03	3.478919e-03	0.0053313857	0.0081048900
	3	10	2.802528e-03	3.577539e-03	0.0054960059	0.0083091257
	3	30	2.961344e-03	3.749702e-03	0.0056871429	0.0087472165
	3	50	3.120298e-03	3.950138e-03	0.0059008109	0.0090272065
	4	1	7.964076e-03	9.015899e-03	0.0112777244	0.0143372255
	4	10	8.199472e-03	9.295153e-03	0.0116255590	0.0147762539
	4	30	8.738506e-03	9.883617e-03	0.0123589751	0.0156235388
	4	50	9.368242e-03	1.054840e-02	0.0131188349	0.0166817556
	5	1	3.117067e-02	3.304803e-02	0.0366895915	0.0413215927
	5	10	3.235895e-02	3.425898e-02	0.0380480169	0.0427998033
	5	30	3.545788e-02	3.752600e-02	0.0417425086	0.0467667352
	5	50	3.914194e-02	4.142425e-02	0.0459567507	0.0514563584
	6	1	1.891794e-01	1.923893e-01	0.1984529319	0.2050583463
	6	10	1.975446e-01	2.007669e-01	0.2069269144	0.2139321164
	6	30	2.185870e-01	2.218804e-01	0.2280312709	0.2345272440
	6	50	2.431686e-01	2.464034e-01	0.2526103765	0.2596196960

Tabela 3.4: Quantis das distâncias euclidianas para os valores de $D = 3,4,5,6$ e $\tau = 1,10,30,50$ para sequências de 50.000 observações.

$N = 50.000$	D	τ	90 %	95 %	99 %	99.9 %
	3	1	5.407679e-05	7.015910e-05	0.0001178519	0.0001917689
	3	10	5.636160e-05	7.130762e-05	0.0001000853	0.0001585019
	3	30	5.731458e-05	7.245769e-05	0.0001101931	0.0001580179
	3	50	5.951595e-05	7.541980e-05	0.0001093136	0.0001983728
	4	1	1.589081e-04	1.818144e-04	0.0002250318	0.0003038889
	4	10	1.588585e-04	1.790301e-04	0.0002264259	0.0002903681
	4	30	1.631504e-04	1.863802e-04	0.0002330694	0.0002893557
	4	50	1.619028e-04	1.809200e-04	0.0002311957	0.0003017229
	5	1	6.062508e-04	6.389830e-04	0.0007179730	0.0008832568
	5	10	5.985032e-04	6.289105e-04	0.0007041024	0.0007855387
	5	30	6.040569e-04	6.400727e-04	0.0007268196	0.0008213228
	5	50	6.055769e-04	6.381391e-04	0.0007134216	0.0008042788
	6	1	3.071590e-03	3.150152e-03	0.0033104814	0.0035923810
	6	10	3.050841e-03	3.117779e-03	0.0032553996	0.0033734229
	6	30	3.066360e-03	3.129649e-03	0.0032669327	0.0034581195
	6	50	3.074748e-03	3.147889e-03	0.0032828968	0.0034178859

4

Conclusão

Neste trabalho analisamos a possibilidade de a distância euclidiana de pontos no plano ($H \times C$) de sequências ao ponto (1,0), referência teórica de ruído branco, poderem ser usadas como uma estatística de teste para a hipótese de a sequência ser ruído branco. Verificamos que essa possibilidade existe, e que essa estatística é capaz de identificar, com limitações, o mapa logístico (que já foi usado como gerador de números pseudoaleatórios), movimento browniano e ruído com autocorrelação. Para este último, fizemos uma análise preliminar do poder do teste em função da intensidade da correlação.

Verificamos, também, que os geradores Mersenne-Twister e Randu são considerados ruído branco, mesmo sendo eles técnicas algorítmicas de geração de observações pseudoaleatórias.

Sendo a máscara parametrizada, fizemos uma análise preliminar do poder do teste.

Uma limitação deste trabalho é que apenas verificamos a qualidade do gerador em relação a um de estrutura ideal. Com isso, limitamos a aplicabilidade do nosso trabalho à análise de séries que, potencialmente, são ocorrências de variáveis aleatórias independentes e identicamente distribuídas.

Há farta literatura que caracteriza diferentes tipos de estruturas como, por exemplo, processos estocásticos do tipo f^{-k}. A nossa metodologia pode, em princípio, ser aplicada a quaisquer processos mas, para isso, é necessário o conhecimento da distribuição dos padrões ordinais do processo de referência. No nosso caso, trata-se da lei uniforme sobre os padrões, que é característica de ruído branco. Não conhecemos resultados que caracterizem de forma teórica as leis de outros processos.

Há, contudo, uma solução para esse problema: estimar a lei característica do padrão de interesse. Isso pode ser feito através de estudos Monte Carlo, mas tal extensão foge ao objetivo deste trabalho.

Apêndice A

Apêndice 1 - Agoritmos

Algoritmo A.1: Mersenne-Twister

```
set.seed(1234567890, kind = "Mersenne-Twister")

MT1k <- runif(1000)
MT50k <- runif(50000)
```

Algoritmo A.2: Randu

```
RANDU <- function() {
seed <<- ((1103515245 * seed) + 12345 ) %% (2^31)
}

Randu1k <- vector()
for(i in 1:1000) {

Randu1k[i] <- c(RANDU())
}
```

Algoritmo A.3: Não Estacionária

```
NoEst1k <- diffinv(rnorm(1000))
NoEst1k <- abs(NoEst1k/max(NoEst1k))

NoEst50k <- diffinv(rnorm(50000))
NoEst50k <- abs(Randu50k/max(NoEst50k))
```

Algoritmo A.4: Estacionária

```
Est1k <- filter(rnorm(1000), filter=rep(1,3), circular=TRUE)

Est50k <- filter(rnorm(50000), filter=rep(1,3), circular=TRUE)
```

Algoritmo A.5: Mapa Logístico

```
logisticmap <- function(N, x0) {

saida <- vector(mode="double", length=10000)
saida[1] <- x0

for(i in 2:10000)
saida[i] <- 4 * saida[i-1] * (1 - saida[i-1])

x0 <- saida[10000]
saida <- vector(mode="double", length=N)
saida[1] <- x0

for(i in 2:N)
saida[i] <- 4 * saida[i-1] * (1 - saida[i-1])

return(saida)
}

LogMap1k <- logisticmap(1000, .01)
LogMap50k <- logisticmap(50000, .01)
```

REFERÊNCIAS BIBLIOGRÁFICAS

Almiron, M. G., Almeida, E. S. & Miranda, M. N. (2009), 'The reliability of statistical functions in four software packages freely used in numerical computation', *Brazilian Journal of Probability and Statistics* **23**(2), 107–119.

Aquino, A. L. L., Cavalcante, T. S. G., Almeida, E. S., Frery, A. C. & Rosso, O. A. (2015), 'Characterization of vehicle behavior with information theory', *The European Physical Journal B: Condensed Matter and Complex Systems* **88**(10), 257–269. URL http://dx.doi.org/10.1140/epjb/e2015-60384-x.

Aquino, A. L. L., Ramos, H. S., Frery, A. C., Viana, L. P., Cavalcante, T. S. G. & Rosso, O. A. (2017), 'Characterization of electric load with information theory quantifiers', *Physica A* **465**, 277–284.

Bandt, C. (2017), 'A new kind of permutation entropy used to classify sleep stages from invisible EEG microstructure', *Entropy* **19**(5), 197.

Bandt, C. & Pompe, B. (2002), 'Permutation entropy: A natural complexity measure for time series', *Physical Review Letters* **88**, 174102–1–174102–4.

Brockwell, P. J. & Davis, R. A. (1991), *Time Series: Theory and Methods*, 2 ed., Springer-Verlag, Berlin.

Bustos, O. & Fraiman, R. (1984), Robust and nonlinear time series, Vol. 29 of *Lectures Notes in Statistics*, Springer, New York, chapter Asymptotic behavior of the estimates based on residual autocovariances for ARMA models, pp. 26–49.

Cabral, R. S., Aquino, A. L. L., Frery, A. C., Rosso, O. A. & Ramírez, J. A. (2013), 'Structural changes in data communication in wireless sensor networks', *Central European Journal of Physics* **11**(12), 1645–1652.

Carpi, L. C., Rosso, O. A., Saco, P. M. & Gómez Ravetti, M. (2011), 'Analyzing complex networks evolution through Information Theory quantifiers', *Physics Letters A* **375**, 801–804.

De Micco, L., González, C. M., Larrondo, H. A., Martin, M. T., Plastino, A. & Rosso, O. A. (2008), 'Randomizing nonlinear maps via symbolic dynamics', *Physica A: Statistical Mechanics and its Applications* **387**(14), 3373–3383. URL http://dx.doi.org/10.1016/j.physa.2008.02.037.

De Micco, L., Larrondo, H. A., Plastino, A. & Rosso, O. A. (2009), 'Quantifiers for randomness of chaotic pseudo-random number generators', *Philosophical Transactions of the Royal*

Society A: Mathematical, Physical and Engineering Sciences **367**(1901), 3281–3296. URL http://dx.doi.org/10.1098/rsta.2009.0075.

Fairfield, R. C., Mortenson, R. L. & Coulthart, K. B. (1985), An LSI Random Number Generator (RNG), *in* 'Proceedings of CRYPTO 84 on Advances in Cryptology', Springer-Verlag New York, Inc., New York, NY, USA, pp. 203–230. URL http://dl.acm.org/citation.cfm?id=19478.19496.

Gabriel, C., Wittmann, C., Sych, D., Dong, R., Mauerer, W., Andersen, U. L., Marquardt, C. & Leuchs, G. (2010), 'A generator for unique quantum random numbers based on vacuum states', *Nature Photonics* **4**(10), 711–715. URL http://dx.doi.org/10.1038/NPHOTON.2010.197.

Knuth, D. E. (1998), *The Art of Computer Programming, Volume 2: (2Nd Ed.) Seminumerical Algorithms*, Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA.

Larrondo, H. A., Martín, M. T., González, C. M., Plastino, A. & Rosso, O. A. (2006), 'Random number generators and causality', *Physics Letters A* **352**(4–5), 421–425. URL http://www.sciencedirect.com/science/article/pii/S0375960105018232.

L'Ecuyer, P. (1992), Testing Random Number Generators, *in* 'Proceedings of the 1992 Winter Simulation Conference', {IEEE} Press, pp. 305–313.

L'Ecuyer, P. (2007), *Random Number Generation*, John Wiley & Sons, Inc., pp. 93–137. URL http://dx.doi.org/10.1002/9780470172445.ch4.

L'Ecuyer, P. & Simard, R. (2007), 'Testu01: A c library for empirical testing of random number generators', *ACM Trans. Math. Softw.* **33**(4), 22:1–22:40. URL http://doi.acm.org/10.1145/1268776.1268777.

Martin, M. T., Plastino, A. & Rosso, O. A. (2006), 'Generalized statistical complexity measures: Geometrical and analytical properties', *Physica A* **369**, 439–462.

Matsumoto, M. & Nishimura, T. (1998), 'Mersenne Twister: a 623-dimensionally equidistributed uniform pseudo-random number generator', *ACM Transactions Model. Comput. Simul.* **8**(1), 3–30. URL https://doi.org/10.1145/272991.272995.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ravetti, M. G., Carpi, L. C., Gonçalves, B. A., Frery, A. C. & Rosso, O. A. (2014), 'Distinguishing noise from chaos: objective versus subjective criteria using Horizontal Visibility Graph', *PLOS One* **9**(9), 1–15.

Rosso, O. A., Martin, M. T., Figliola, A., Keller, K. & Plastino, A. (2006), 'EEG analysis using wavelet-based information tools', *Journal of Neuroscience Methods* **153**, 163–182.

Rosso, O. A., Ospina, R. & Frery, A. C. (2016), 'Classification and verification of handwritten signatures with time causal information theory quantifiers', *PLoS ONE* **11**(12), e0166868.

Schieber, T. A., Carpi, L., Frery, A. C., Rosso, O. A., Pardalos, P. M. & Ravetti, M. G. (2016), 'Information theory perspective on network robustness', *Physics Letters A* **380**, 359–364.

Sippel, S., Lange, H., Mahecha, M. D., Hauhs, M., Bodesheim, P., Kaminski, T., Gans, F. & Rosso, O. A. (2016), 'Diagnosing the dynamics of observed and simulated ecosystem gross primary productivity with time causal information theory quantifiers', *PLoS One* **11**(10), e0164960.

Stamey, L. (2016), 'How RANDOM.ORG's journey from radio static to true randomness generates reliable results for games, security, and clinical trials'. URL http://www.hostingadvice.com/blog/random-dot-org-true-randomness-reliable-results/.

Volchan, S. B. (2002), 'What is a random sequence?', *AMERICAN MATHEMATICAL MONTHLY* **109**(1), 46–63.

Walker, J. (2017), 'Ent. a pseudorandom number sequence test program'. URL http://www.fourmilab.ch/random/.

Zunino, L., Soriano, M. C. & Rosso, O. A. (2012), 'Distinguishing chaotic and stochastic dynamics from time series by using a multiscale symbolic approach', *Phys. Rev. E* **86**, 046210. URL http://link.aps.org/doi/10.1103/PhysRevE.86.046210.

Este trabalho foi redigido em LaTeX utilizando uma modificação do estilo IC-UFAL. As referências bibliográficas foram preparadas no JabRef e administradas pelo BibTeX com o estilo LaCCAN. O texto utiliza fonte Fourier-GUT*enberg* e os elementos matemáticos a família tipográfica Euler Virtual Math, ambas em corpo de 12 pontos.