UNIVERSIDADE FEDERAL DO PARANÁ

ADRIANA CAMILA BRAGA

CARACTERIZAÇÃO EM LARGA ESCALA DAS FLUTUAÇÕES DAS VAZÕES EM RIOS VIA MÉTODOS DE FÍSICA ESTATÍSTICA

CURITIBA
2016

ADRIANA CAMILA BRAGA

CARACTERIZAÇÃO EM LARGA ESCALA DAS FLUTUAÇÕES DAS VAZÕES EM RIOS VIA MÉTODOS DE FÍSICA ESTATÍSTICA

Tese apresentada ao Programa de Pós-Graduação em Métodos Numéricos em Engenharia na área de concentração em Programação Matemática e na linha de pesquisa em Métodos Estatísticos Aplicados à Engenharia, setores de Tecnologia e Ciências Exatas da Universidade Federal do Paraná, como requisito parcial à obtenção do grau de Doutor.

Orientador: Prof. Dr. Ademir Alves Ribeiro

Coorientador: Prof. Dr. Manoel Messias Alvino de Jesus

CURITIBA
2016

Catalogação na Publicação elaborada pela Biblioteca UTFPR – Câmpus Apucarana – PR., Brasil.

B813c

Braga, Adriana Camila

Caracterização em larga escala das flutuações das vazões em rios via métodos de física estatística / Adriana Camila Braga.- - Curitiba, 2016.

72 f. : il. color. ; 30cm.

Orientador: Prof. Dr. Ademir Alves Ribeiro – Co-orientador: Prof. Dr. Manoel Messias Alvino de Jesus.

Tese – Universidade Federal do Paraná, Programa de Pós-Graduação em Métodos Numéricos em Engenharia, 2016.

Bibliografia: p. 64-69.

1. Sistemas complexos. 2. Séries temporais. 3. Física estatística. 4. Redes complexas. 5. Descargas fluviais. I. Ribeiro, Ademir Alves. II. Jesus, Manoel Messias Alvino de. III. Título.

CDD 22.ed. 530.23

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
Setor Tecnologia
Programa de Pós Graduação em MÉTODOS NUMÉRICOS EM ENGENHARIA
Código CAPES: 40001016030P0

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em MÉTODOS NUMÉRICOS EM ENGENHARIA da Universidade Federal do Paraná foram convocados para realizar a arguição da Tese de Doutorado de **ADRIANA CAMILA BRAGA**, intitulada: "**CARACTERIZAÇÃO EM LARGA ESCALA DAS FLUTUAÇÕES DAS VAZÕES EM RIOS VIA MÉTODOS DE FÍSICA ESTATÍSTICA**", após terem inquirido a aluna e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO.

Curitiba, 02 de Setembro de 2016.

ADEMIR ALVES RIBEIRO
Presidente da Banca Examinadora (UFPR)

ANSELMO CHAVES NETO
Avaliador Interno (UFPR)

ERVIN KAMINSKI LENZI
Avaliador Externo (UEPG)

MARCELO KAMINSKI LENZI
Avaliador Externo (UFPR)

HAROLDO VALENTIN RIBEIRO
Avaliador Externo (UEM)

À minha filha amada, Ana Beatriz Braga Moraes.
Melhor justificativa de minha existência.

AGRADECIMENTOS

À Deus por me amparar nos momentos difíceis, me dar força interior para superar as dificuldades, mostrar o caminho nas horas incertas e me suprir em todas as minhas necessidades.

À Ana Beatriz, minha filha amada, que com seus seis anos cheios de vida e de graça, ensina-me a ser mãe. Filha, perdão pelos momentos de ausência exigidos para minha formação no doutorado. Agora a tese chegou ao fim. Prometo ser muito mais sua. Vamos poder dançar juntas, cantar juntas, ler juntas, brincar juntas, passear juntas e viver muito mais vezes juntas. Amo-a muito e sempre!

Ao Arnaldo, meu marido, pela força que nos une e faz do nosso amor o mais intenso e o maior. Obrigada pela sua força, por sua dedicação, pela espera paciente nos momentos de ausência, por toda a sua capacidade de compreensão, por sua confiança em mim, enfim, pela sua presença em minha vida. Esta vitória é nossa!

À minha mãe Otilia, pela sabedoria em me educar, por seus gestos solidários, pelo amor e carinho de mãe que soube me proteger e me ensinar os limites da vida, por ter investido e acreditado sempre na educação e me incentivado a trilhar os caminhos do conhecimento capaz de transformar as pessoas sempre para melhor. Mãe, você é presença marcante em minha vida. Obrigada por me ensinar a não desistir dos meus sonhos, por acreditar em mim e por compartilhar de muitas das minhas angústias e conquistas. Mãe, você é exemplo de força-guerreira e renascer constante. Amo-a muito e sempre!

Agradeço ao meu estimado orientador, Prof. Dr. Ademir Alves Ribeiro, pela disponibilidade e por acreditar em minha capacidade. Obrigada pela confiança!

Ao Prof. Dr. Manoel Messias Alvino de Jesus, pela coorientação deste trabalho, pelo apoio e pelo incentivo.

Ao Prof. Dr. Haroldo Valentin Ribeiro, pelos ensinamentos transmitidos e pelas valiosas sugestões. Sua participação foi fundamental para realização deste trabalho. Obrigado pelas constantes demosntrações de sabedoria e humildade.

Aos meus colegas “dinterandos”, pela amizade, convívio e apoio.

Aos amigos Suellen, Dione, Marcos e Rodnny pelas boas risadas nas viagens e pela disposição em ajudar quando necessário.

À Universidade Tecnológica Federal do Paraná (UTFPR), por me concederem afastamento total das minhas atribuições.

A todos, que de alguma forma contribuíram para a realização deste trabalho.

Ora, a fé é o firme fundamento das coisas que se esperam e a prova das coisas que não se vêem. Porque por ela os antigos alcançaram bom testemunho. Pela fé entendemos que os mundos foram criados pela palavra de Deus; de modo que o visível não foi feito daquilo que se vê.
Carta aos Hebreus 11:1-3, A Bíblia Sagrada

RESUMO

Este trabalho apresenta caracterizações em larga escala das flutuações das vazões em rios, tendo como principais técnicas os métodos empregados na Física de Sistemas Complexos. Inicialmente, foi utilizada uma estrutura de grafos de visibilidade horizontal, resultante da análise e do mapeamento em redes complexas, das séries temporais diárias de 141 estações diferentes, localizadas em 53 rios brasileiros entre os anos de 1931 e 2012. Verificou-se que as distribuições de graus dessas redes são bem descritas por distribuições exponenciais, nas quais os expoentes característicos são, em sua maior parte, numericamente maiores que aqueles obtidos para séries temporais aleatórias. O decaimento mais rápido do ajuste da exponencial, quando comparado ao ajuste de modelo da distribuição aleatória, evidência que a dinâmica de flutuações subjacentes às vazões dos rios têm uma natureza correlacionada de longo alcance. A investigação da evolução das descargas fluviais, acompanhando os valores dos expoentes característicos e os coeficientes de aglomeração globais das redes ao longo dos anos mostrou que as vazões dos rios em várias estações evoluíram tornando-se correlacionadas. Por fim, o uso de outros dois métodos de Física Estatística, típicos no estudo de sistemas complexos que são: o *detrended fluctuation analysis* (DFA) e a entropia e complexidade de permutação foram aplicados, mostrando, sobretudo, que o uso do espectro de permutação permite encontrar o período associado à sazonalidade natural da vazão dos rios e que, após padronizadas, as vazões tornam-se aproximadamente bem descritas por uma mesma distribuição. Fica claro, também, que as séries temporais são correlacionadas de longo alcance pela análise DFA. Por outro lado, revisitou-se três dos principais métodos existentes na literatura, que conseguem identificar correlações de longo alcance. Nominalmente: a análise de flutuação DFA, as transformações Wavelet e a análise entrópica (DEA - *diffusion entropy analysis*). Fez-se uma comparação entre os três métodos quanto a sua convergência para o verdadeiro valor do expoente h de Hurst em função do tamanho das séries geradas. Nessa comparação, observou-se algumas peculiaridades de cada método; por exemplo: o DFA converge para valores superiores de h, enquanto as transformações Wavelet e o DEA o fazem por valores inferiores. Com base nessa observação empírica, se propõe aplicar simultaneamente DFA e Wavelet. Isso fez com que a convergência para o valor verdadeiro de h fosse alcançada para séries razoavelmente pequenas.

Palavras-chaves: sistemas complexos, séries temporais, física estatística, correlações de longo alcance, entropia e complexidade de permutação, redes complexas, descargas fluviais.

ABSTRACT

This paper presents broad-scale characterizations of fluctuation in river flows, principally using the methods employed in complex systems physics. Initially, we used a horizontal visibility graph structure produced by analysis and mapping of complex networks from daily temporal series of 141 different stations in 53 Brazilian rivers between 1931 and 2012. The degree distributions of these networks were found to be well-described by exponential distributions, in which most of the characteristic exponents are numerically greater than those obtained for random time series. The faster decay of exponential distribution in comparison with randomized distribution shows that the dynamics underlying the fluctuations in river flow have a long-range correlation. Investigation of the changes in river discharge, by following characteristic exponent values and global clustering coefficients of the networks over this period, showed that the river flow in several stations evolved and became correlated. Finally, we applied *detrended fluctuation analysis* (DFA) and entropy and complexity permutation, two methods from statistical physics which are typically used to study complex systems. These showed in particular that spectrum permutation permits the period associated with the natural seasonality of river flows to be found; after this value is normalized, flow rates are described approximately using the same distribution. The time series was also clearly seen to be correlated in the long term using DFA analysis. However, we also returned to three main methods available in the literature which can identify long-range correlations: DFA fluctuation analysis, wavelet transformation, and entropic analysis (*diffusion entropy analysis*, or DEA). The three methods were compared to assess their convergence for the true value for Hurst's h exponent depending on the size of the generated series. This comparison showed some peculiarities of each method; for example: DFA converges for higher h values while wavelet and DEA converge for lower values. Based on this empirical finding, we resolved to apply DFA and wavelet simultaneously, which caused convergence for the true h value to be achieved for relatively small series.

Key-words: complex systems, time series, statistical physics, long-range correlations, entropy and complexity of permutation, complex networks, river discharges.

LISTA DE FIGURAS

LISTA DE ABREVIATURA E SIGLAS

DFA	Detrend Fluctuation Analysis
DEA	Diffusion Entropy Analysis
HVG	Horizontal Visibility Graph
BHP	Modelo BramwellL-Holdsworth-Pinton

SUMÁRIO

1 INTRODUÇÃO

1.1 PROBLEMA DE ESTUDO

A Física de Sistemas Complexos tem como base fundamental a aplicação de métodos numéricos e estatísticos acompanhados por minuciosas observações empíricas e considerações teóricas, resultando em respostas altamente confiáveis do ponto de vista estatístico. Os métodos empregados nessa área não estão limitados a problemas tradicionais da Física e têm sido utilizados com grande funcionalidade em uma série de problemas interdisciplinares.

De acordo com Ribeiro (2012), não existe uma definição precisa do que venha a ser um sistema complexo, no entanto, em seu trabalho de doutoramento, descreve um conjunto de propriedades comuns a esses sistemas citando Boccara (2004), de acordo com o qual, um sistema complexo normalmente apresenta as seguintes características:

1. São constituídos de um grande número de agentes interagentes;
2. As interações são, em geral, não-locais e/ou não-lineares;
3. Existe um comportamento coletivo, auto-organizado, o qual é difícil de antecipar a partir do conhecimento da dinâmica individual dos agentes;
4. Esse comportamento coletivo não resulta da existência de um controle central.

Esses comportamentos se enquadram perfeitamente no estudo das vazões dos rios brasileiros, por conseguinte, em todos os rios que possuem as mesmas características dos rios estudados, em um cenário no qual físicos e engenheiros têm se empenhado na investigação de sistemas relacionados ao meio ambiente como um todo.

Estudos que têm como ciência fundamental a Física de Sistemas Complexos, podem ser evidenciados na investigação da dinâmica de terremotos (MENDES et al., 2010), comportamento das atividades geomagnéticas (TURNER et al., 2012), estudo do clima (BOETTLE; RYBSKI; P., 2013) e outros sistemas interdependentes relacionados ao clima (RYBSKI; HOLSTEN; KROPP, 2011; RIBEIRO et al., 2013), este último, por sua vez, tem como componente importante, os rios e suas vazões que impactam e são impactados em um cenário de interação homem/ambiente.

De maneira geral, este trabalho compila e associa os resultados publicados recentemente por Braga et al. (2016) decorrentes da investigação em larga escala das flutuações das vazões em rios brasileiros por meio de técnicas da Física de Sistemas Complexos, pelas quais foram estudadas as vazões em 141 estações diferentes (nas proximidades de usinas hidrelétricas) de 53 rios brasileiros via séries temporais diárias obtidas ao longo

de um período de 80 anos, compreendidos entre 1931 à 2012. Especificamente, foi utilizada a estrutura de grafo de visibilidade horizontal (LUQUE et al,2009; LACASA et al, 2008; LACASA e TORAL, 2010) para o mapeamento dessas séries temporais relacionadas as vazões de rios brasileiros em redes complexas. O acompanhamento da evolução de propriedades topológicas dessas redes revela que as vazões em várias estações estão se tornando mais ou menos correlacionadas (e exibindo estruturas de rede interna mais ou menos complexas) ao longo dos anos. Tal comportamento poderia estar relacionado às mudanças no sistema climático e outros fenômenos provocados pelo homem. Finalmente, o trabalho conclui que outros métodos de Física Estatística, típicos no estudo de sistemas complexos, tais como o *detrended fluctuation analysis* (DFA) e a entropia e complexidade de permutação podem ser aplicados a esses sistemas, mostrando, sobretudo, que o uso do espectro de permutação permite encontrar o período associado à sazonalidade natural da vazão dos rios e que, após padronizadas, as vazões são aproximadamente descritas por uma mesma distribuição. Fica claro, também, que as séries temporais são correlacionadas de longo alcance.

Assim como as séries temporais das vazões dos rios, muitas outras podem ser investigadas no contexto de sistemas complexos. Exemplos de séries temporais incluem resultados de jogos (RIBEIRO et al., 2010), som de papel amassando (MENDES et al., 2010), amplitude sonora de músicas (MENDES et al., 2011), som de pessoas aglomeradas (RIBEIRO et al., 2011), intensidade de um laser atravessando uma amostra de água fervente (RIBEIRO et al., 2011b), número de casos em uma epidemia (PICOLI JR. et al., 2011), sequências de símbolos geradas numericamente (RIBEIRO et al., 2009; RIBEIRO et al., 2011a) e muitas outras.

Nesses trabalhos com séries temporais, tal como no apresentado aqui, uma análise muito importante é a da correlação. Medidas de correlação indicam o quão associados estão os termos de uma série temporal. Com medidas como essa, pode-se intuir muitos aspectos diretamente ligados à dinâmica do sistema sob investigação, como se o sistema é Markoviano ou não Markoviano e se existe invariância por escala ou alguma outra propriedade fractal. Visando detectar correlação, alguns métodos foram propostos ao longo das últimas décadas. Entre eles destacam-se a análise de flutuação DFA (Detrend Fluctuation Analysis) (PENG et al., 1994; VJUSHIN et al., 2001; HU et al., 2001; KANTELHARDT et al., 2001; CHEN et al., 2002), as transformações wavelet (MUZY; BACRY; ARNEODO, 1991; ARNEODO et al., 1995; TORRENCE; COMPO, 1997; SIMONSEN; HANSEN, 1998; MANIMARAN; PANIGRAHI; PARIKH, 2005) e a análise de entropia difusiva DEA (Diffusion Entropy Analysis) (SCAFETTA; HAMILTON; GRIGOLINI, 2001; GRIGOLINI; PALATELLA; RAFFAELLI, 2001; SCAFETTA; GRIGOLINI, 2002). Além de empregar alguns desses métodos de detecção de correlação de longo alcance nas séries da vazão de rios (como mencionado anteriormente), revisitou-se esses três métodos visando compará-los de uma maneira objetiva.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Objetivo geral desse trabalho é propor a investigação e modelagem de um sistema específico, as vazões naturais de vários rios brasileiros, visando a extrair padrões, regularidades ou leis que estejam governando a dinâmica do sistema. No contexto de detecção de correlações de longo alcance, pretende-se revisar aspectos relacionados aos principais métodos, bem como compará-los de uma maneira objetiva.

1.2.2 Objetivos Específicos

- Revisar aspectos relacionados a alguns métodos propostos para detecção de correlações em séries temporais;
- Comparar, de uma maneira objetiva, três métodos propostos para detecção de correlações de longo alcance;
- Investigar dados empíricos de um sistema específico: as vazões naturais de vários rios brasileiros;
- Comparar os resultados obtidos com resultados de séries temporais das vazões naturais dos rios (descargas fluviais) existentes na literatura;
- Lançar novas possibilidades para investigar essas séries que podem encontrar implicações para modelagem e previsão da vazão de rios.

1.2.3 JUSTIFICATIVA

O estudo dos sistemas relacionados à Terra tornou-se ainda mais importante com as preocupações crescentes sobre mudanças ambientais e da consciência do desenvolvimento sustentável. Como um paradigma de sistemas complexos, esse tema de pesquisa baseia-se em esforços multidisciplinares e também tem sido abordada por físicos via métodos da Física Estatística. Terremotos (MENDES et al., 2010; RIBEIRO et al., 2015), atividades geomagnéticas (TURNER et al., 2012) , clima (BOETTLE; RYBSKI; P., 2013) e relacionados com o clima (RYBSKI; HOLSTEN; KROPP, 2011; RIBEIRO et al., 2013) são apenas alguns exemplos de sistemas que os pesquisadores tem abordado com esses métodos.

Em particular, como apontado por Dove e Kammen (2015), um dos desafios ambientais globais mais importantes é a mudança climática. Num sentido mais amplo, a complexidade dos sistemas climáticos está relacionada com as complexas interações entre a atmosfera, biosfera, criosfera, hidrosfera e litosfera. A última é a parte do sistema climático que compreende oceanos, rios e lagos, isto é, a água no estado líquido na superfície

terrestre e subterrâneas (WMO, 2015) e é bem conhecido o papel extremamente importante da água em mudanças ambientais globais (OLIVER; OLIVER, 1995; MACHIWAL; JHA, 2012).

A vazão dos rios é resultante de inúmeras interações complexas entre diversos fatores físicos tais como: sistemas climáticos (índice pluviométrico; temperatura e pressão atmosférica local - por consequência, taxa de evaporação), configuração do relevo (altitude, área da bacia e relevo) e, em muitos casos, atividades humanas (poluição, utilização na indústria, utilização na agricultura, transposições para regiões de escassez e geração de energia). Essa variedade de processos faz com que as vazões dos rios, que podem ser definidas como descargas fluviais, se enquadrem em um processo bastante complexo que têm estimulado seus estudos nos últimos sessenta anos.

O trabalho pioneiro de Hurst (1951) inaugurou uma série de discussões relacionadas às propriedades fractais e/ou multifractais da evolução temporal das vazões dos rios por meio do estudo da dependência de longa duração de escoamento de um grande número de rios (TESSIER et al., 1996; PORPORATO; RIDOLFI, 1997; JÁNOSI; GALLAS, 1999; BORDIGNON; LISI, 2000; KANTELHARDT et al., 2001; BRAMWELL; HOLDSWORTH; PORTELLI, 2002; LIVINA et al., 2003; KANTELHARDT et al., 2003; DAHLSTEDT; JENSEN, 2005; MOVAHED; HERMANIS, 2008; DOLGONOSOV; KORCHAGIN; KIRPICHNIKOVA, 2008; ZHANG; XU; YANG, 2009; ZHANG; WANG; W., 2015; HAJIAN; MOVAHED, 2010; DOMENICO; LATORA, 2011; BIGACHEV; BUND, 2012; YU et al., 2014; MIHAILOVIC et al., 2014; RABASSA; BECK, 2015).

Embora o número de trabalhos associados ao estudo da vazão de rios seja relativamente grande e crescente, sua maioria têm como base o estudo de um pequeno conjunto de rios, de maneira que a caracterização em larga escala de séries temporais desses sistemas pode ser considerada incipiente. Por outro lado, o Brasil tem a maior bacia hidrográfica do mundo (Amazonas) bem como um dos sistemas de rios mais complexos e extensos do globo, instigando fortemente um processo de incrementação do conhecimento nessa área a partir da investigação em larga escala sobre os rios brasileiros, até então, não relatada. Com base no que foi posto, este trabalho tem como objetivo a produção de conhecimento destinado a preencher um pouco mais essa lacuna.

1.2.4 ESTRUTURA DO TEXTO

Esse trabalho está organizado em quatro capítulos, sendo o primeiro capítulo destinado a essa introdução.

No segundo capítulo será abordada uma breve revisão do estado da arte à que se refere o estudo, apresentando, de forma breve, alguns conceitos e ferramentas básicas de Mecânica Estatística, além de uma pequena revisão sobre a função de correlação, invariância de escala e movimento browniano fracionário, os três métodos para detecção de correlações de longo alcance e, finalmente a entropia e complexidade de permutação.

O terceiro capítulo é dedicado à análise dos resultados, no qual são realizadas a descrição e discussão dos resultados obtidos, será exposto um estudo (usando alguns dos métodos anteriores) sobre a vazão natural de diversos rios brasileiros e uma comparação objetiva entre os três métodos mais comuns para detecção de correlações.

Finalmente, no quarto capítulo, apresentam-se as conclusões e sugestões para trabalhos futuros.

2 FUNDAMENTOS TEÓRICOS

2.1 FÍSICA ESTATÍSTICA E SISTEMAS COMPLEXOS

A seguir, são abordados, de forma breve, alguns conceitos e técnicas básicas de Mecânica Estatística utilizados nas investigações discutidas nesse trabalho.

2.1.1 Sistemas Complexos e Leis de Potências

É cada vez maior o interesse em estudar os chamados sistemas complexos. A prova disso são os recentes avanços neste campo, materializados sobre uma crescente quantidade de artigos e livros publicados (MANTEGNA; STANLEY, 1999; AUYANG, 1998; JENSEN, 1998; ALBERT; BARABÁSI, 2002; BOCCARA, 2004; HAKEN, 2006). Trata-se de um campo de pesquisa vasto e, em geral, multidisciplinar. Contudo, é interessante ressaltar (como dito anteriormente) que não existe uma definição muito precisa de um sistema complexo. O que se tem são propriedades gerais, as quais, muitas vezes, estão presentes nos sistemas complexos.

Neste contexto, destaca-se o conceito de *emergence*. É bem aceito que a maioria dos fenômenos macroscópicos - da supercondutividade à economia e à neurociência - resultam da dinâmica coletiva dos componentes microscópicos do sistema. O termo *emergence* refere-se a um padrão ou comportamento coletivo espaço-temporal apresentado pelo sistema complexo, o qual não é esperado. Nessa direção, "não esperado" mostra nossa dificuldade (seja analítica ou numérica) de prever o comportamento emergente a partir das interações que descrevem a dinâmica das partes microscópicas do sistema. Além disso, muitas vezes não temos o conhecimento da forma dessas interações ou mesmo de detalhes das estruturas microscópicas do sistema.

De fato, na maioria das vezes, o estado de arte das investigações em sistemas complexos baseia-se em uma pequena quantidade de informação relacionada ao sistema. Essa quantidade limitada de informação torna ainda mais impressionante as implicações dos estudos em sistemas complexos.

Apesar de não haver uma definição precisa de sistemas complexos, em geral eles apresentam propriedades emergentes que decorrem em grande parte da interação não-linear entre suas partes. Assim, suas propriedades apenas se tornam notáveis quando vistos como um todo. Uma das propriedades marcantes de muitos desses sistemas é a presença de leis de escala ou leis de potência (SORNETTE, 2006).

Uma distribuição tipo lei de potência é usualmente escrita como

$$P(x) \propto x^{-\alpha} \tag{2.1}$$

com $x_0 \leq x \leq \infty$ e $\alpha > 1$. Comumente, essa distribuição é usada para descrever o

comportamento assintótico de uma distribuição de probabilidade quando $x > x_0$. Um dos atributos da lei de potência é sua invariância de escala. Dada uma relação $f(x) = ax^{\alpha}$, ao escalar o argumento x por um fator constante c, apenas faz-se com que a função se torne proporcional a ela mesma, ou seja,

$$f(cx) = a(cx)^{\alpha} = c^{\alpha} f(x) \propto f(x). \tag{2.2}$$

Nota-se, ainda, que a equação 2.1 produz uma relação linear quando o logaritmo é tomado, o que pode ser visto na figura 1 (SILVA, 2015).

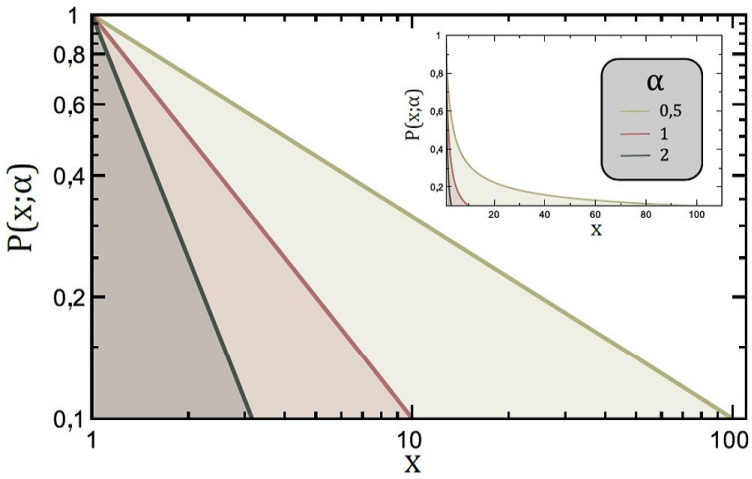

Figura 1 – Distribuições tipo lei de potência em escala log-log para $\alpha = 0, 5, 1\, e\, 2$. No detalhe, as mesmas distribuições em escala normal.

2.1.2 Distribuição Gaussiana e o Teorema Central do Limite

É usual, na investigação de distribuições empíricas simétricas em relação à média, compará-las com uma distribuição de probabilidades chamada Gaussiana (ou normal). Dois parâmetros identificam a distribuição de probabilidades gaussiana de uma variável aleatória: a média μ e o desvio padrão σ. Essa distribuição é da forma

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \; x \in \mathbb{R}. \tag{2.3}$$

Geralmente, utiliza-se a distribuição Gaussiana padronizada, em que a variável aleatória é transformada, ou seja, considera-se uma nova variável $z = (x - \mu)/\sigma$, com média zero e desvio-padrão unitário e, consequentemente, sua distribuição pode ser escrita como segue

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right), \; z \in \mathbb{R} \tag{2.4}$$

que é livre dos parâmetros.

Basicamente, quando um grande número de eventos aleatórios e independentes conduzem a determinado evento (o que aproxima muitos fenômenos da natureza), este poderá ter uma distribuição de probabilidades Gaussiana (2.3). Formalmente, se x_i são variáveis aleatórias independentes de média zero e desvio padrão finito σ, a função densidade de probabilidade $z(n) = \sum_{i=1}^{n} x_i/\sigma\sqrt{n}$ tende à distribuição Gaussiana padronizada à medida que n cresce. Esse resultado é chamado Teorema Central do Limite.

Supondo que um processo x possa ser decomposto no produto de n subprocessos x_i, ou seja, $x = x_1, x_2, \ldots, x_n$ em que cada x_i se comporta como uma variável aleatória não negativa e independente, pelo Teorema Central do Limite, para n suficientemente grande, a distribuição de probabilidade da variável aleatória $\log x = \log x_1 + \log x_2 + \cdots + \log x_n$ é normal. Assim, para variável x tem-se a distribuição

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2\right], \quad x > 0 \tag{2.5}$$

chamada logonormal e diz-se que x segue um processo multiplicativo.

2.1.3 Distribuição Tipo I de Fisher - Tippett ou Gumbel

Em 1958, Fisher e Tippett, tomando de vários conjuntos de muitas amostras o maior valor de cada conjunto, mostraram que a distribuição dos valores extremos é independente da distribuição original e se comporta como função limite. Gumbel, em 1945, sugeriu que essa distribuição de valores extremos seria apropriada para a análise de frequência das cheias, desde que a série fosse anual, isto é, cada vazão da série de valores extremos fosse a maior vazão de uma amostra de 365 possibilidades (maior vazão do ano).

A distribuição de Gumbel também conhecida como distribuição de valores extremos do tipo I, ou distribuição do tipo I de Fisher-Tippett, tem a função de probabilidade acumulada

$$F(x) = P\{X < x\} = e^{-e^{-y}}, \tag{2.6}$$

sendo x a vazão e y a variável reduzida de Gumbel dada por

$$y = \frac{x - \beta}{\alpha}, \tag{2.7}$$

em que α e β são os parâmetros característicos da distribuição de Gumbel; α representa o parâmetro de escala e β o parâmetro de posição.

A função densidade da distribuição de Gumbel é dada por:

$$f(x) = \frac{1}{\alpha} exp\left[-\frac{x - \beta}{\alpha} - exp\left(\frac{x - \beta}{\alpha}\right)\right], \quad x \in \mathbb{R}. \tag{2.8}$$

em que $\beta \in \mathbb{R}$ e $\alpha > 0$.

2.1.4 Redes Complexas

Uma rede complexa pode ser vista como uma abstração matemática para representar as interações entre partes de um sistema. Numa primeira aproximação, pode-se considerar uma condição binária para representar essa interação, isto é, as partes interagem ou não. Assim, para um conjunto de elementos/partes, sabe-se com quais outras partes cada elemento interage. Logo, uma rede complexada pode ser representada por um grafo.

Um grafo $G(\mathcal{N}, \mathcal{L})$ consiste de dois conjuntos: um conjunto $\mathcal{N}$ de N nós (ou sítios, ou vértices) e um conjunto $\mathcal{L}$ de L ligações (ou arestas) entre os nós. Uma ligação entre dois nós i e j é denotada por (i, j) ou l_{ij}, dois nós unidos por uma ligação são ditos adjacentes ou vizinhos próximos. Um grafo é dito direcionado quando a direção da ligação é relevante, ou seja, l_{ij} não é igual a l_{ji} para qualquer (i, j). O número máximo de ligações em um grafo $G(\mathcal{N}, \mathcal{L})$ é $N(N-1)/2$.

De acordo com Gabardo (2015), em uma rede social, pode-se exemplificar as pessoas como vértices e as ligações entre elas como arestas de um grafo. Se tomar como exemplo um mapa com cidades e rodovias, pode-se representar cada cidade por um vértice do grafo e as rodovias por arestas que as conectam. Como está claro, redes complexas podem representar dados de diversas áreas.

2.1.5 Distribuição de Grau

O grau de um nó i, denotado por k_i, é o número de ligações partindo desse nó que pode ser calculado usando a matriz adjacência $A = [a_{ij}]$, em que $a_{ij} = 1$ se há ligação entre i e j e $a_{ij} = 0$ se não há ligação entre i e j. Nesse caso, $k_i = \sum_{j=1}^{N} a_{ij}$. Se o grafo for direcionado, o grau terá duas componentes: $k_i^{out} = \sum_{j=1}^{N} a_{ij}$ para as ligações partindo de i e $k_i^{in} = \sum_{j=1}^{N} a_{ij}$ para as ligações chegando em i. A distribuição de grau é a característica mais simples de uma rede aleatória e, geralmente, é só o primeiro passo para a descrição da rede. A estrutura topológica de um grafo está totalmente relacionada com a distribuição de graus. Portanto, a escolha da distribuição é fundamental ao criar um modelo de rede, pois esta determina a classe do grafo. Em redes reais, o cálculo da distribuição de conectividade (ou de grau) $P(k)$ é importante na classificação dos diferentes tipos de redes quanto à sua topologia. Notavelmente, em muitas situações o conhecimento da distribuição de grau é suficiente para o conhecimento da rede e o que está acontecendo nela.

A distribuição de conectividade determina como estão distribuídas as ligações entre os sítios da rede. Matematicamente, $P(k)$ é definida como sendo a probabilidade de um sítio escolhido ao acaso, ter exatamente k ligações, de forma equivalente, mostra a fração de sítios de uma rede que apresentam conectividade k. Em redes direcionadas é necessário considerar $P_{in}(k)$ e $P_{out}(k)$ representando a conectividade de entrada e saída do nó, respectivamente. Os nós de uma rede, com exceção das redes regulares, possuem

um número distinto de ligações. A forma como a conectividade está distribuída na rede é analisada pelo gráfico $P(k)$, ou pelo cálculo de seus momentos de ordem n

$$\langle k^n \rangle = \sum_k k^n P(k). \tag{2.9}$$

O primeiro momento caracteriza a conectividade média da rede $\langle k \rangle$. O segundo momento é uma medida de flutuação da distribuição de conectividade (AGUIAR, 2012).

A distribuição de grau nas redes aleatórias segue a distribuição de Poisson. No entanto, em muitas redes reais a distribuição de grau segue a lei de potência, em que $P(k) \propto k^{-\alpha}$ para uma constante α qualquer, geralmente $1 < \alpha < 3$ (METZ et al., 2007).

2.1.6 Coeficiente de Aglomeração

Coeficiente de aglomeração, ou *clustering coefficient*, é uma métrica utilizada para avaliar a presença de *clusters* em uma rede complexa. *Cluster* é um grupo de vértices fortemente conectados. Os *clusters* em redes sociais são análogos a comunidades e, frequentemente, esses grupos são compostos de indivíduos que compartilham interesses ou características.

Existem dois coeficientes de aglomeração: o coeficiente de aglomeração local (que mede a aglomeração para um vértice específico da rede) e o coeficiente de aglomeração global (que mede a propensão de uma rede em apresentar grupos ou comunidades).

2.1.6.1 Coeficiente de Aglomeração Local

O coeficiente de aglomeração local mede a propensão que um vértice específico e seus vizinhos tem para formar um clique. Clique é um subgrafo conexo, ou seja, todos os vértices estão conectados entre si. O clique mais comum em redes é um trio. É possível calcular o coeficiente de aglomeração local por meio da seguinte equação:

$$C_i = \frac{2n_i}{k_i(k_i - 1)}, \tag{2.10}$$

em que C_i é o coeficiente de aglomeração local para o vértice i, n_i representa o número de conexões, arestas que estão conectadas ligando os k_i vértices vizinhos de i até esse vértice.

2.1.6.2 Coeficiente de Aglomeração Global

O coeficiente de aglomeração global mede a propensão de uma rede em apresentar grupos ou comunidades. Seu cálculo é baseado no número de trios conectados presentes na rede. Por um trio, entende-se o conjunto de três vértices conectados entre si, formando um triângulo. Este valor pode ser medido utilizando a seguinte equação:

$$C = \frac{\text{3x número de triângulos no grafo}}{\text{número de trios conectados no grafo}}, \tag{2.11}$$

em que “trio conectado” é um trio de vértices constituídos por um nó central, ligado aos outros dois; os vértices de acompanhamento são não ordenados. A equação conta a fração de trios conectados que, na verdade, formam triângulos; o fator de três indica que cada triângulo corresponde a três trios.

2.1.6.3 Coeficiente de Aglomeração Médio

É a média do coeficiente de aglomeração local de todos os vértices da rede. Essa métrica de rede ajuda a avaliar a qual modelo de rede um grafo empírico se assemelha. Redes de mundo pequeno devem apresentar um coeficiente de aglomeração médio maior que um grafo aleatório com o mesmo número de vértices. Este coeficiente pode ser calculado por meio da seguinte equação:

$$C = \frac{1}{N} \sum_i C_i, \tag{2.12}$$

em que N representa o número de vértices na rede e C_i representa o coeficiente de aglomeração local de cada um dos vértices do grafo.

2.1.7 Grafo de Visibilidade Horizontal (Horizontal Visibility Graph - HVG)

O método recentemente proposto, chamado de grafo de visibilidade horizontal (horizontal visibility graph - HVG) (LUQUE et al., 2009), tem como função transformar uma série temporal em um grafo e é definido da seguinte forma. Seja $\{x_i\}_{i=1,\ldots,N}$ uma série temporal de N dados reais. O algoritmo associa cada dado da série temporal a um nó/vértice no grafo de visibilidade horizontal. Dois nós, i e j, serão conectados sempre que for possível a construção de uma linha horizontal no espaço da série temporal que juntam os pontos x_i e x_j sem interseção em qualquer altura do ponto intermediário. Assim, i e j são dois nós conectados se o seguinte critério geométrico for satisfeito dentro da série temporal:

$$x_i,\, x_j \,>\, x_n \;\; \forall\, n \,|(i < n < j)\,. \tag{2.13}$$

Esse método vem sendo utilizado para estudar diversos sistemas dinâmicos, como uma ferramenta de distinção entre sistemas caóticos e estocásticos (LACASA; TORAL, 2010). Especificamente nesse trabalho, os autores propõem que a distribuição de grau pelo mapeamento utilizando HVG segue uma função exponencial, $P(k) \sim exp(-\lambda k)$, em que k é o grau do nó/vértice e λ é um parâmetro positivo para fazer a distinção entre dinâmicos estocásticos e dinâmicos caóticos, isto é, $\lambda < ln(3/2)$ caracteriza processos caóticos, enquanto $\lambda > ln(3/2)$ caracteriza processos estocásticos correlacionados. A fronteira $\lambda_{rand} = ln(3/2)$ corresponde à situações não correlacionadas.

2.2 CORRELAÇÕES EM SÉRIES TEMPORAIS

Nesta seção apresenta-se alguns aspectos relacionados à função de correlação com ênfase em séries temporais.

2.2.1 A Função de Autocorrelação

Medidas de correlação são bastante comuns em estatística e uma das mais populares é o coeficiente de correlação de Pearson (HOGG; CRAIG, 1995). Esse diz o quão é linear a relação entre dois conjuntos de variáveis. Para definir esse coeficiente, considere dois conjuntos x_i e y_i ($i \in \{1, 2, \ldots, n\}$). O coeficiente de Pearson pode ser definido como:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{2.14}$$

em que $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Esse coeficiente varia de -1 a 1, sendo que $r = 1$ implica numa relação linear crescente perfeita entre os conjuntos x_i e y_i; e $r = -1$ implica numa relação linear decrescente perfeita. Para $r = 0$ não há uma relação linear entre esses dois conjuntos. A figura 2 mostra alguns possíveis cenários relacionados ao coeficiente de Pearson. Os conjuntos x_i e y_i foram gerados numericamente de modo que $x_i = y_i + \varepsilon$, sendo $\varepsilon \sim N(0, \sigma^2)$ um número aleatório Gaussiano de média zero e desvio padrão σ. Os valores de r e σ são mostrados na figura.

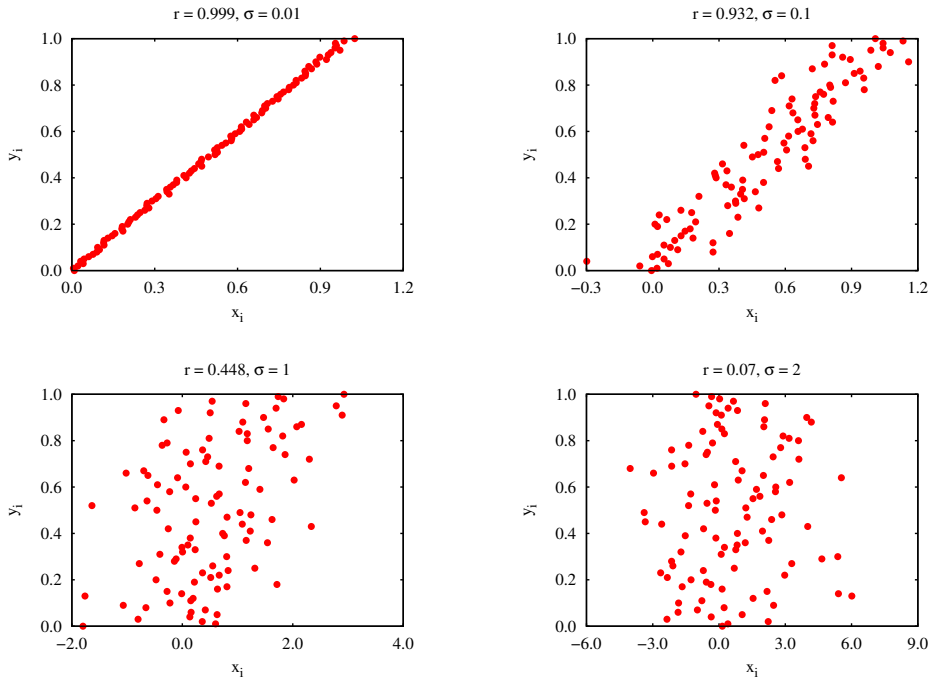

Figura 2 – Diferentes cenários para o coeficiente de correlação de Pearson.

A idéia de Pearson conduz naturalmente à função de autocorrelação[1]. Basta imaginar que, ao invés de investigar a correlação entre dois conjuntos, deseja-se saber como os elementos de uma série temporal estão relacionados com seus próprios elementos defasados por uma ou mais unidades de tempo. Suponha que z_i ($i \in \{1, 2, \ldots, n\}$) representa uma série temporal, para a qual deseja-se saber como é a correlação entre os elementos imediatamente espaçados. Para isso, pode-se usar o coeficiente de Pearson aplicado aos conjuntos $x_i = \{z_1, z_2, \ldots, z_{n-1}\}$ e $y_i = \{z_2, z_3, \ldots, z_n\}$, ou seja,

$$r_1 = \frac{\sum_{i=1}^{n-1}(z_i - \bar{x})(z_{i+1} - \bar{y})}{\sqrt{\sum_{i=1}^{n-1}(z_i - \bar{x})^2 \sum_{i=1}^{n-1}(z_{i+1} - \bar{y})^2}}\,. \tag{2.15}$$

Se a série temporal z_i é estacionária[2] e n é razoavelmente grande, pode-se considerar que $\bar{x} = \bar{y} = \bar{z}$ e $\sum_{i=1}^{n-1}(z_i - \bar{x})^2 = \sum_{i=1}^{n-1}(z_{i+1} - \bar{y})^2$, conduzindo a

$$r_1 = \frac{\sum_{i=1}^{n-1}(z_i - \bar{z})(z_{i+1} - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2}\,, \tag{2.16}$$

sendo $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$. Naturalmente, essa ideia pode ser facilmente generalizada para elementos espaçados por k unidades temporais. Matematicamente tem-se

$$r_k = \frac{\sum_{i=1}^{n-k}(z_i - \bar{z})(z_{i+k} - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2}\,. \tag{2.17}$$

Essa última quantidade é conhecida como função de autocorrelação defasada em k unidades de tempo (*lag-k auto-correlation*).

Após essa definição, é interessante observar que o numerador da expressão anterior é uma generalização da variância. De fato, a expressão

$$c_k = \frac{1}{n-k}\sum_{i=1}^{n-k}(z_i - \bar{z})(z_{i+k} - \bar{z}) \tag{2.18}$$

é conhecida como covariância, sendo que

$$c_0 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2 \tag{2.19}$$

é precisamente[3] a variância da série temporal z_t. Assim, pode-se reescrever a função de autocorrelação (2.17) da seguinte maneira alternativa

$$r_k = \frac{c_k}{c_0}\,. \tag{2.20}$$

---

[1] Muitas vezes referida apenas por função de correlação. Trata-se de um abuso de linguagem muito comum na literatura em geral e este texto não é uma exceção.

[2] Um processo estocástico $x(t)$ é dito ser estacionário se sua distribuição de probabilidade é invariante por translação temporal. Essa definição é muitas vezes considerada bastante restritiva, dando origem a outras definições mais amplas. Veja, por exemplo, em (MANTEGNA; STANLEY, 1999).

[3] Alguns autores usam uma definição ligeiramente diferente, sendo $\frac{1}{n-k}$ substituído por $\frac{1}{n-k-1}$ na equação (2.18), ou seja,

$$c_k = \frac{1}{n-k-1}\sum_{i=1}^{n-k}(z_i - \bar{z})(z_{i+k} - \bar{z})\,.$$

Essa definição tem um erro estatístico menor para amostras pequenas. Contudo, as duas definições são praticamente indistinguíveis para séries de tamanho razoável (HOGG; CRAIG, 1995).

Obviamente, esses resultados podem ser facilmente estendidos para o caso contínuo. Com relação a isso, é interessante notar que alguns autores usam diferentes denominações para o caso discreto e contínuo, sendo a função de autocorrelação no caso discreto também chamada de coeficiente de autocorrelação. Para o caso contínuo, a série temporal z_i deve ser substituída pela variável estocástica $x(t)$ com $t \in \mathbb{R}$. Assim, a covariância fica determinada por

$$C(\tau) = \langle x(t)x(t+\tau) \rangle \,, \tag{2.21}$$

sendo que $\langle \dots \rangle$ representa o valor médio sobre um conjunto de realizações do processo estocástico que gera $x(t)$. Uma vez definida a covariância, pode-se definir também a autocorrelação

$$R(\tau) = \frac{C(\tau)}{C(0)}. \tag{2.22}$$

A definição na forma contínua é bastante conveniente para descrever o tipo de memória existente no processo estocástico $x(t)$. Uma questão fundamental diz respeito à existência ou não de uma escala típica para a função de autocorrelação $R(\tau)$. Para investigar essa questão, considere a integral (MANTEGNA; STANLEY, 1999)

$$\int_0^\infty R(\tau)d\tau \tag{2.23}$$

e os três cenários descritos abaixo.

- Se a integral $\int_0^\infty R(\tau)d\tau$ é finita, então existe uma escala de tempo de memória típica, o qual é chamado de tempo de correlação do processo $x(t)$. Por exemplo:
  - $R(\tau) \sim \exp(-\frac{\tau}{\tau_c})$ implica em $\int_0^\infty R(\tau)d\tau \sim \tau_c$;
  - $R(\tau) \sim \exp(-\frac{\tau^\nu}{\tau_c})$ implica em $\int_0^\infty R(\tau)d\tau \sim \tau_c^{1/\nu}$.
- Se a integral $\int_0^\infty R(\tau)d\tau$ é indeterminada, então é possível que a função de correlação seja oscilante. Por exemplo:
  - $R(\tau) \sim \sin(a\tau)$ conduz à uma integral oscilante que não pode ser determinada. Nesses casos, a correlação é harmônica.
- Se a integral $\int_0^\infty R(\tau)d\tau$ não é finita, então não existe uma escala de tempo de memória típica. Por exemplo:
  - $R(\tau) \sim \tau^{\eta-1}$ com $0 < \eta \leq 1$ implica em $\int_0^\infty \tau^{\eta-1}d\tau \to \infty$.

Variáveis aleatórias caracterizadas por uma função de autocorrelação do tipo

$$R(\tau) \sim \tau^{\eta-1} \qquad \text{ou} \qquad R(\tau) \sim \tau^{-\gamma}, \tag{2.24}$$

com $\gamma = 1 - \eta$, são ditas serem correlacionadas de longo alcance (*long-range correlated*). Por outro lado, variáveis aleatórias caracterizadas por funções de autocorrelação

do tipo $R(\tau) \sim \exp(-\frac{\tau}{\tau_c})$ são ditas serem correlacionadas de curto alcance (*short-range correlated*).

Outra maneira de fazer a distinção entre os tipos de memória presentes em uma variável estocástica, é considerar a transformada de Fourier da função de autocorrelação (MANTEGNA; STANLEY, 1999)

$$S(f) = \int_{-\infty}^{+\infty} R(\tau) \exp(-if\tau) d\tau \,. \tag{2.25}$$

Essa função é chamada de espectro de potência (*power spectrum*). Note que para $R(\tau) \sim \exp(-\tau/\tau_c)$, obtem-se $S(f) \sim 1/f^2$ e, para $R(\tau) \sim \tau^{\eta-1}$, segue que $S(f) \sim 1/f^\eta$ com $0 < \eta \leq 1$. Além disso, para um caso em que não existe nenhuma memória, ou seja, $R(\tau) \sim \delta(\tau)$ a função $S(f)$ é uma constante. Esses três resultados permitem uma classificação mais elegante:

$$S(f) \sim 1/f^\eta \tag{2.26}$$

com $0 \leq \eta \leq 2$. Assim, quando $\eta = 0$ não há nenhuma memória na variável $x(t)$, o que é conhecido como ruído branco. Para $\eta = 2$ existe uma escala típica de memória e esse processo estocástico é chamado de processo Wiener. Quando $0 < \eta \leq 1$ o processo possui correlações de longo alcance. De uma maneira geral, um processo caracterizado por um espectro de potência $S(f) \sim 1/f^\eta$ com $0 < \eta < 2$ é denominado ruído $1/f$.

Após essa breve revisão relacionada à função de correlação, é interessante observar que o cálculo da função de correlação aplicando diretamente a definição (2.20) requer uma série de tamanho considerável. Isso ocorre porque quanto maior for a defasagem k, menor serão os termos da somatória $\sum_{i=1}^{n-k}(z_i - \bar{z})(z_{i+k} - \bar{z})$, conduzindo a um erro estatístico crescente em k. Esse problema será ilustrado na próxima seção.

2.2.2 Invariância de Escala e Movimento Browniano Fracionário

Uma outra propriedade que está diretamente relacionada aos aspectos de correlação descritos na seção anterior é a invariância de escala. De uma maneira matemática, uma função $\Phi(x_1, x_2, \dots)$ é invariante por escala (ou também auto-similar) sempre que

$$\Phi(x_1, x_2, \dots) = \gamma^a \Phi(\gamma^b x_1, \gamma^c x_2, \dots) \,. \tag{2.27}$$

No caso de uma série temporal estacionária, z_i ($i \in \{1, 2, \dots, n\}$), uma maneira de investigar invariância de escala, é definir a série das diferenças ou incrementos

$$\Delta z_i(k) = x_{i+k} - x_i \,. \tag{2.28}$$

Essa nova série é dita ser invariante por escala ou auto-similar ou fractal, se a distribuição de probabilidade de $\Delta z_i(k)$ possuir a mesma forma funcional para diferentes valores de k, isto é,

$$p(\Delta z_i, k) = \frac{1}{k^\delta} \Psi \left(\frac{\Delta z_i}{k^\delta} \right) , \tag{2.29}$$

sendo δ o expoente de escala e $\Psi(x)$ a função escala. Por exemplo, se z_i representar uma série temporal oriunda de um movimento Browniano usual, $\Psi(x)$ será uma Gaussiana e $\delta = 1/2$. Outra propriedade interessante das flutuações de uma série fractal é a dependência em k do segundo momento

$$\langle \Delta z_i(k)^2 \rangle = \frac{1}{n-k} \sum_{i=1}^{n-k} \Delta z_i(k)^2 \sim k^{2\delta} \,, \tag{2.30}$$

em que δ é novamente o expoente de escala.

Uma maneira de fazer a conexão entre invariância de escala e correlações de longo alcance é por meio do movimento Browniano fracionário. Assim como o movimento Browniano usual pode ser visto como a integração de um ruído branco, o movimento Browniano fracionário é usualmente definido por meio da seguinte integração estocástica (MANDELBROT; NESS, 1968):

$$\begin{aligned} B_h(t) &= \frac{1}{\Gamma(h+1/2)} \left(\int_{-\infty}^{0} [(t-s)^{h-1/2} - (-s)^{h-1/2}] dB(s) \right. \\ &+ \left. \int_0^t (t-s)^{h-1/2} dB(s) \right) , \end{aligned} \tag{2.31}$$

em que $\Gamma(y) = \int_0^\infty x^{\alpha-1} \exp(-x) dx$, $B(s)$ representa um movimento Browniano usual e $0 < h < 1$ é um parâmetro chamado de expoente de Hurst. Note que para $h = 1/2$ recupera-se o movimento Browniano usual

$$B_{1/2}(t) = \int_0^t dB(s) \,. \tag{2.32}$$

Esse processo possui algumas propriedades, limitando apenas a enunciá-las[4]. Trata-se de um processo estocástico estacionário e de valor médio nulo. Além disso, sua variância é dada por $\langle B_h(t)^2 \rangle \sim t^{2h}$ e a covariância é $\langle B_h(t) B_h(\tau) \rangle \sim t^{2h} + \tau^{2h} - |t-\tau|^{2h}$.

Para investigar a invariância de escala no processo Browniano fracionário, deve-se considerar sua série dos incrementos

$$\begin{aligned} \Delta B_h(\tau) &= B_h(t+\tau) - B_h(t) = \frac{1}{\Gamma(h+1/2)} \\ &\times \left[\int_{-\infty}^{t+\tau} (t+\tau-s)^{h-1/2} dB(s) - \int_{-\infty}^{t} (t-s)^{h-1/2} dB(s) \right] . \end{aligned} \tag{2.33}$$

Usando essa definição é possível mostrar que $\Delta B_h(\tau)$ é invariante por escala, caracterizado por uma função escala $\Psi(x)$ Gaussiana e

$$\langle \Delta B_h(\tau)^2 \rangle \sim \tau^{2h} \,. \tag{2.34}$$

Além disso, sua covariância é dada por

$$\langle \Delta B_h(\tau) \Delta B_h(\tau+k) \rangle \sim h(2h-1) k^{2h-2} \,. \tag{2.35}$$

[4] Para uma descrição detalhada veja (MANDELBROT; NESS, 1968; KOUTSOYIANNIS, 2002)

Por essas duas expressões, pode-se observar que para o caso dos incrementos do movimento Browniano fracionário, também chamado de ruído Gaussiano fracionário, o expoente de escala δ e o expoente da função de correlação de longo-alcance γ estão diretamente relacionados com o expoente h de Hurst. De fato, por comparação direta das equações (2.34) e (2.35) com as equações (2.24) e (2.30), obtemos $\delta = h$ e $\gamma = 2(1-h)$.

Essa relação entre invariância de escala e correlações de longo alcance é a base para os métodos que será apresentado mais adiante. Sabe-se ainda que estes resultados valem em um cenário mais geral, por exemplo, uma característica necessária (mas não suficiente) para a validade da igualdade $\delta = h$ é que função de escala $\Psi(x)$ tenha o segundo momento finito.

Do ponto de vista formal, a definição do movimento Browniano fracionário dada pela equação (2.31) é bastante útil. Contudo, do ponto de vista da geração de séries temporais com propriedades fractais seu uso é bastante limitado. De fato, para gerar séries temporais com as propriedades de escala do movimento Browniano fracionário, é comum recorrer a métodos numéricos específicos. Um desses é particularmente simples[5] e pode ser definido pela seguinte expressão:

$$\begin{aligned} \xi_i &= \frac{m^{-h}}{\Gamma(h+1/2)} \left\{ \sum_{j=1}^{m} j^{h-1/2} \theta_{1+m(M+i)-j} \right. \\ &+ \left. \sum_{j=1}^{m(M-1)} \left[(m+j)^{h-1/2} - j^{h-1/2} \right] \theta_{1+m(M-1+i)-j} \right\}, \end{aligned} \tag{2.36}$$

em que M e m são inteiros e $\{\theta_i\}$ é um conjunto de variáveis aleatórias Gaussianas de média zero e desvio padrão unitário. Neste caso, ξ_i são os incrementos do movimento Browniano fracionário. Para obtermos a trajetória fracionária, devemos integrar essa série, isto é,

$$B_h(t) = \sum_{i=1}^{t} \xi_i \,. \tag{2.37}$$

Com relação aos parâmetros M e m, observa-se empiricamente que bons resultados são obtidos para M e m moderadamente grandes, especificamente bons resultados foram obtidos para $M = 1000$ e $m = 10$. Existem muitos outros métodos mais modernos, como o algoritmo de Hosking (DIEKER; MANDJES, 2003). As figuras 3 e 4, mostra alguns exemplos de séries geradas pelo método anterior.

Agora tendo um método para gerar séries com propriedades fractais bem definidas, pode-se tentar aplicar diretamente a definição da função de correlação (2.22) visando a obter (numericamente) o expoente de correlação $\gamma = 1 - \eta$, como mostra a figura 5. As séries temporais foram obtidas usando o método de Mandelbrot, definido na equação (2.36) para diferentes valores de h, com $M = 1000$ e $m = 10$. Note que para $h < 0.5$ o valor absoluto de r_k foi utilizado para possibilitar a construção dos gráficos em escala

[5] Este é o algoritmo original proposto por Mandelbrot e pode ser encontrado em (FEDERS, 1988).

logarítmica. A linha tracejada representa a lei de potência $r_k \propto h(2h-1)k^{2h-2}$. Como deve estar visível, calcular a função de correlação diretamente de sua definição requer séries temporais com muitos termos, o que raramente ocorre com séries oriundas de sistemas complexos naturais.

Observe que, mesmo para séries temporais com um grande número de termos, os resultados não podem ser considerados bons. A seguir, será apresentado alguns métodos que podem contornar esse problema.

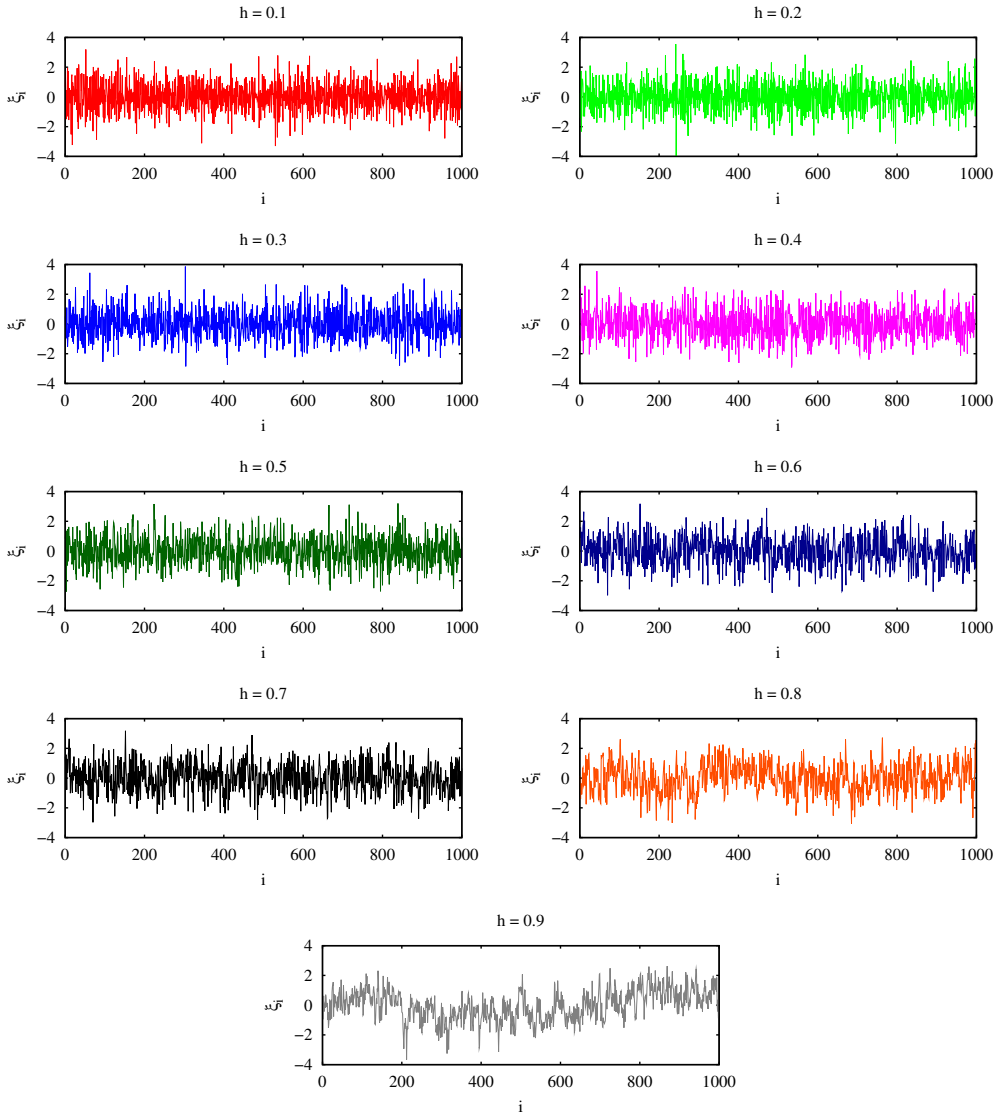

Figura 3 – Possíveis realizações para os incrementos do movimento Browniano gerados pela equação (2.36) com $M = 1000$ e $m = 10$, considerando diferentes valores de h.

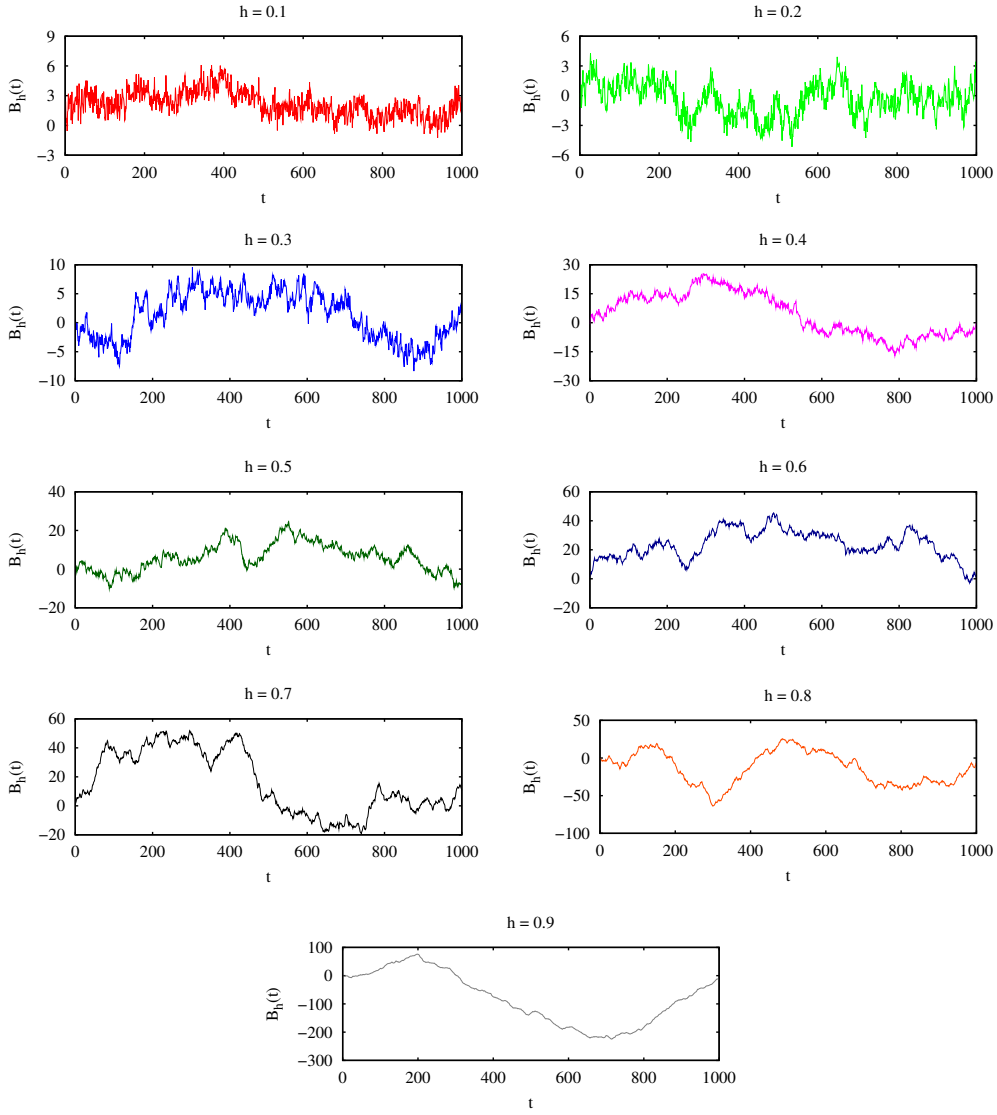

Figura 4 – Possíveis realizações para as trajetórias do movimento Browniano gerado pela equação (2.37) com $M = 1000$ e $m = 10$, considerando diferentes valores de h.

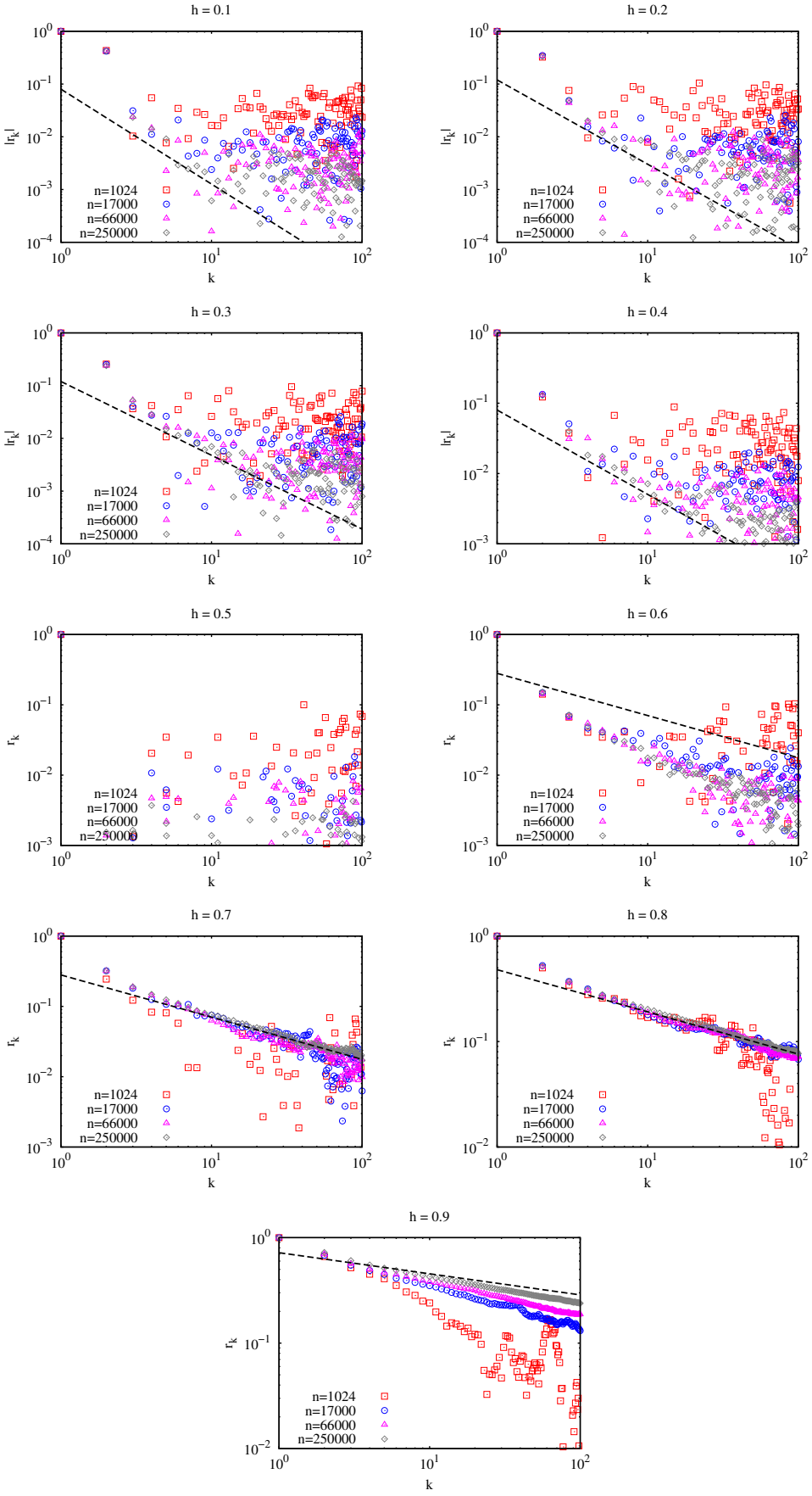

Figura 5 – Coeficiente de autocorrelação dado pela equação (2.20) versus tempo de defasagem k.

2.2.3 Análise de flutuações e o método DFA

Um dos métodos mais utilizados para detectar correlações de longo alcance em séries temporais é o chamado DFA, sigla inglesa para *detrended fluctuation analysis*. Ele foi proposto por Peng et al. (1994) (veja também (VJUSHIN et al., 2001; HU et al., 2001; KANTELHARDT et al., 2001; CHEN et al., 2002)) para o estudo de correlações em séries temporais construídas a partir do DNA. Como indica o nome, o método DFA analisa as flutuações de séries temporais removendo uma possível tendência local. Trata-se de um método de fácil implementação e que produz excelentes resultados, mesmo para séries temporais moderadamente pequenas (empiricamente, da ordem de mil termos). Para apresentar o método, considere novamente uma série temporal x_i ($i \in \{1, 2, \ldots, n\}$). O primeiro passo do DFA consiste em obter a série integral de x_i, isto é,

$$y_i = \sum_{j=1}^{i} x_j \,. \tag{2.38}$$

Logo após, divide-se a série integrada em $s = n/l$ partições não superpostas, de maneira que cada partição tenha l elementos, ou seja,

$$\underbrace{\{y_1, y_2, \ldots, y_l\}}_{w_j^{(1,l)}}, \underbrace{\{y_{l+1}, y_{l+2}, \ldots, y_{2l}\}}_{w_j^{(2,l)}}, \ldots, \underbrace{\{y_{(s-1)l+1}, y_{(s-1)l+2}, \ldots, y_{sl}\}}_{w_j^{(s,l)}}, \tag{2.39}$$

em que $w_j^{(i,l)}$ ($j \in \{1, 2, \ldots, l\}$) representa o conjunto dos elementos contidos na i-ésima partição. Para cada conjunto $w_j^{(i,l)}$, ajusta-se um polinômio de grau v e obtém-se a flutuação em torno desse ajuste

$$\chi_2^{(i,l)} = \frac{1}{l-1} \sum_{j=1}^{l} [w_j^{(i,l)} - f_v(j)]^2 \,, \tag{2.40}$$

em que $f_v(j)$ representa o polinômio ajustado ao conjunto $w_j^{(i,l)}$. Em seguida, calcula-se o valor médio dessa flutuação sobre todas as s partições,

$$F(l) = \left(\frac{1}{s} \sum_{i=1}^{s} \chi_2^{(i,l)} \right)^{1/2} \,. \tag{2.41}$$

Naturalmente, essa flutuação média será uma função de l, que está diretamente relacionada com o expoente de escala δ da seguinte maneira:

$$F(l) \sim l^{\delta} \,. \tag{2.42}$$

Para o caso em que a função escala $\Psi(x)$ possui o segundo momento finito, tem-se a igualdade $\delta = h$. Nos casos em que $\Psi(x)$ não possui o segundo momento, o DFA pode conduzir a falsas correlações. Na prática, costuma-se aplicar o método na série embaralhada de maneira aleatória para verificar a validade da igualdade $\delta = h$. Se $h \neq 0.5$ for obtido para a série embaralhada, nada pode-se dizer sobre o expoente h da série não embaralhada.

Por outro lado, se $h \approx 0.5$ para a série embaralhada, o expoente h da série original será verdadeiramente o expoente de Hurst para aquela série.

A derivação do resultado (2.42) pode ser encontrada em (TAQQU; TEVEROVSKY; WILLINGER, 1995) para o caso em que remove-se uma tendência linear ($v = 1$). Para evitar as longas (embora simples) manipulações algébricas relacionadas a remoção das tendências, aqui será apresentado uma dedução considerando o caso em que a série x_i não apresenta tendências. Sob essa hipótese, calcular a flutuação $\chi_2^{(i,l)}$ é o mesmo que calcular $\langle y_i^2 \rangle$, isto é,

$$\begin{aligned}
\langle y_i^2 \rangle &= \left\langle \left(\sum_{j=1}^{i} x_j \right)^2 \right\rangle = \left\langle \sum_{j=1}^{i} x_j^2 \right\rangle + \left\langle \sum_{j=1}^{i} \sum_{k \neq j}^{i} x_j x_k \right\rangle \\
&= \sum_{j=1}^{i} \langle x_j^2 \rangle + \sum_{j=1}^{i} \sum_{k \neq j}^{i} \langle x_j x_k \rangle \\
&= i\langle x_j^2 \rangle + \sum_{j=1}^{i} \sum_{k \neq j}^{i} C(|k-j|) \\
&= i\langle x_j^2 \rangle + 2 \sum_{j=1}^{i-1} (i-j) C(j) \,, \qquad (2.43)
\end{aligned}$$

em que $C(|k-j|) = \langle x_j x_k \rangle$. Agora, supondo $i \gg 1$, pode-se aproximar a soma do segundo termo de (2.43) pela integral

$$2 \sum_{j=1}^{i-1} (i-j) C(j) \sim \int_1^i (i-j) C(j) dj \sim \int_1^i (i-j) j^{-\gamma} dj \,, \qquad (2.44)$$

considerando que a função correlação seja uma lei de potência, isto é, $C(j) \sim j^{-\gamma}$ com $0 < \gamma < 1$. Assim, obtém-se

$$\langle y_i^2 \rangle \sim i^{-\gamma+2} + i[\langle x_j^2 \rangle + (\gamma - 1)^{-1}] - i^{-\gamma} \,, \qquad (2.45)$$

a qual tem como termo dominante

$$\langle y_i^2 \rangle \sim i^{-\gamma+2} \,, \qquad (2.46)$$

considerando i muito grande. Note que esse deslocamento quadrático médio cresce mais rápido que uma função linear, correspondendo a um processo superdifusivo. Finalmente, substituindo esse resultado em (2.41) e imaginando que i faça o papel de l, encontra-e

$$F(l) \sim l^{1-\gamma/2} \qquad \text{ou} \qquad F(l) \sim l^h \qquad \text{ou} \qquad F(l) \sim l^\delta \,, \qquad (2.47)$$

sendo que $h = 1 - \gamma/2$ e $\delta = h$. Se a série for correlacionada de curto alcance, $C(j)$ decai exponencialmente e o primeiro termo em (2.43) será dominante, conduzindo a $\langle y_i^2 \rangle \sim i$ (difusão usual) e $F(l) \sim l^{1/2}$, de modo que $h = 0.5$ indica ausência de correlações de longo alcance, como já discutido. Observe que a validade desses resultados está condicionada à existência do segundo momento $\langle x_j^2 \rangle$.

A aplicação da equação (2.47) é bastante simples. Basta calcular essa função de flutuação para um conjunto de valores de l e construir um gráfico log-log. A inclinação dessa reta será numericamente igual ao expoente δ. Observe que a grande engenhosidade desse método é o fato dele construir diversas trajetórias aproximadamente independentes a partir de uma única série temporal. Isto faz com que, mesmo para séries moderadamente pequenas, os erros envolvendo a determinação do expoente h sejam pequenos, como ilustra a figura 6. A figura 6 mostra a aplicação do método DFA para séries temporais, obtidas usando o método de Mandelbrot definido na equação (2.36), para diferentes valores de h (mostrados na figura), $M = 1000$ e $m = 10$. Note que em gráficos log-log como esses, a relação lei de potência torna-se uma função linear, sendo a inclinação da reta numericamente igual ao expoente h de Hurst. Além disso, observe que mesmo para séries com poucos termos a função de flutuação empírica é bastante próxima do seu valor teórico dado por $F(l) \sim l^h$ (linha tracejada).

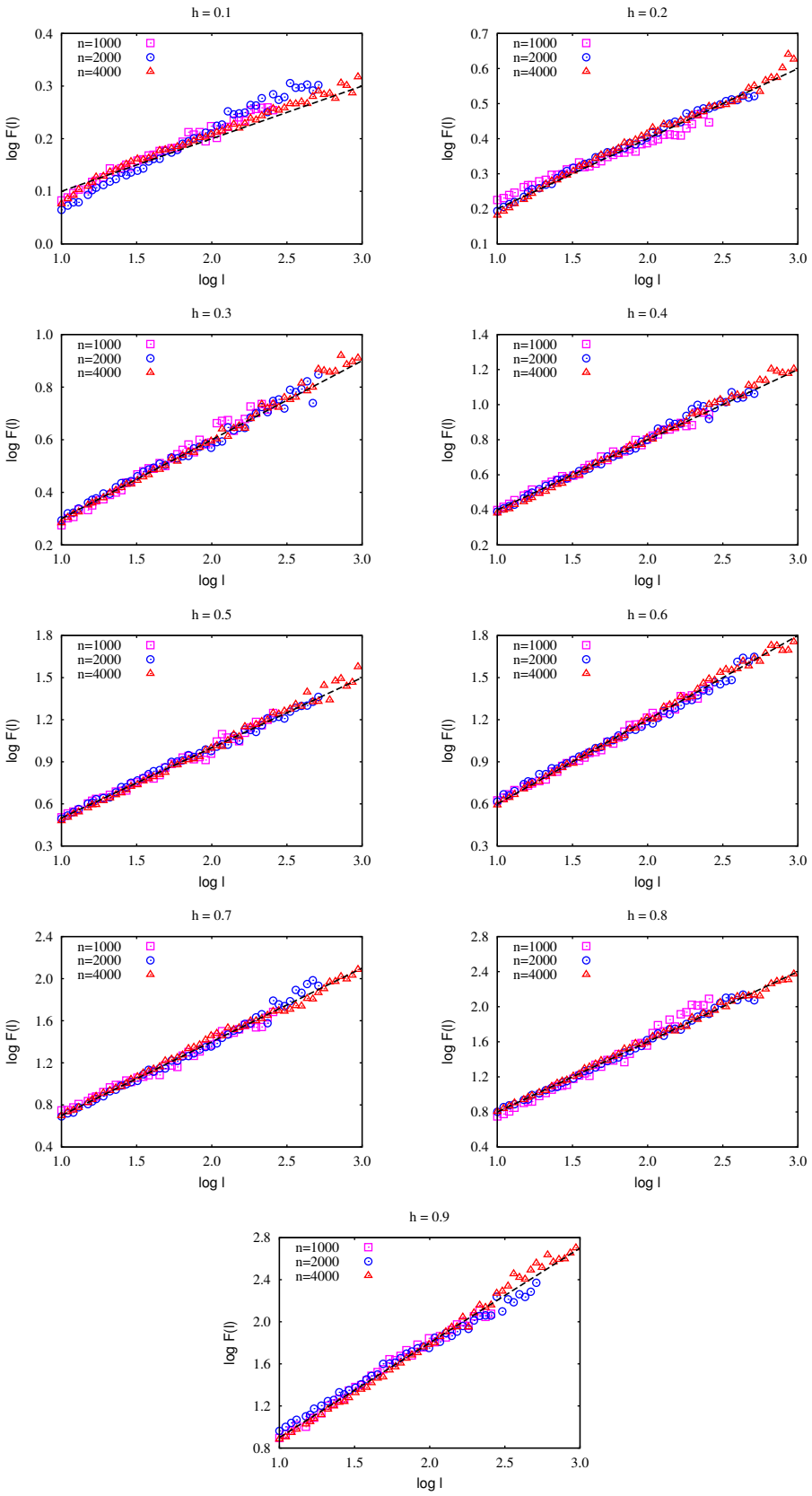

Figura 6 – Aplicação do método DFA para séries temporais.

2.2.4 Transformações Wavelet

A literatura de análise de séries temporais parece estar dividida (essencialmente) em duas vertentes concorrentes no que diz respeito à análise de correlações de longo alcance. Como foi dito, DFA é um dos métodos mais utilizados para determinar o expoente de Hurst e talvez o seu maior "concorrente" seja as transformações Wavelet (MUZY; BACRY; ARNEODO, 1991; ARNEODO et al., 1995; TORRENCE; COMPO, 1997; SIMONSEN; HANSEN, 1998; MANIMARAN; PANIGRAHI; PARIKH, 2005).

As transformações Wavelet são parametrizadas por um parâmetro de escala $s > 0$ e por um parâmetro de translação u. Além disso, deve-se escolher uma função do tipo

$$\psi_{s,u}(x) = \psi\left(\frac{x-u}{s}\right), \tag{2.48}$$

a qual é denominada de função mãe ou *analyzing Wavelet*. Esta função deve ter média nula (quando $u = 0$), ser localizada no espaço usual e também no espaço de Fourier. Assim, dada uma função $g(x)$, define-se sua transformada Wavelet como

$$\mathcal{W}[g(x)](s,u) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} \psi^*_{s,u}(x)\, g(x)\, dx\,, \tag{2.49}$$

em que $\psi^*(x)$ denota o complexo conjugado de $\psi(x)$. Para o caso de um conjunto discreto, isto é, uma série temporal x_i ($i \in \{1, 2, \ldots, n\}$), a definição acima toma a seguinte forma

$$\mathcal{W}[x_i](s,u) = \frac{1}{\sqrt{s}} \sum_{i=1}^{n} \psi^*\left(\frac{i-u}{s}\right) x_i\,. \tag{2.50}$$

No caso de uma função $g(x)$ invariante por escala[6], isto é,

$$g(x) \doteq \lambda^{-h} g(\lambda x)\,, \tag{2.51}$$

em que o símbolo $\doteq$ representa uma igualdade estatística, ou seja, ambos os lados possuem a mesma distribuição (função escala), deve-se obter

$$\begin{aligned}
\mathcal{W}[g(x)](s,u) &\doteq \mathcal{W}[\lambda^{-h} g(\lambda x)](s,u) \\
&= \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} \psi^*\left(\frac{x-u}{s}\right) \lambda^{-h} g(\lambda x)\, dx \\
&= \lambda^{-(1/2)-h} \frac{1}{\sqrt{\lambda s}} \int_{-\infty}^{\infty} \psi^*\left(\frac{x'-\lambda u}{\lambda s}\right) \lambda^{-h} g(x')\, dx' \\
&= \lambda^{-(1/2)-h} \mathcal{W}[g(x)](\lambda s, \lambda u)\,.
\end{aligned}$$

Note que fizemos a mudança de variável $x' = \lambda x$. Portanto,

$$\mathcal{W}[g(x)](\lambda s, \lambda u) \doteq \lambda^{(1/2)+h}\, \mathcal{W}[g(x)](s,u)\,. \tag{2.52}$$

[6] Essa definição permite uma outra interpretação para o expoente h de Hurst (SIMONSEN; HANSEN, 1998): ele expressa a tendência para $dg = [dg(x)/dx]dx$ mudar de sinal. Para $1/2 < h \leq 1$ o sinal tende a não mudar, enquanto para $0 \leq h < 1/2$ existe uma tendência para mudança de sinal. Em ambos os casos, como já sabe-se, existem correlações de longo alcance. Normalmente, quando $h > 1/2$ ($h < 1/2$), diz-se que $g(x)$ é persistente (anti-persistente). Quando tem-se $h = 1/2$, dg muda de sinal aleatoriamente e $g(x)$ não possui correlações de longo alcance.

Como deve estar claro, o domínio de uma transformação Wavelet de uma função unidimensional é bidimensional: uma dimensão correspondente ao parâmetro de escala s e outra dimensão correspondente ao parâmetro de translação u. Assim, fixando-se o parâmetro s, tem-se um número infinito de amplitudes relacionadas ao parâmetro de translação u.

Uma maneira de contornar essa dupla dependência é tomar o valor médio sobre o parâmetro de translação u para cada valor de escala s. Essa manipulação conduz a

$$\mathcal{W}[g(x)](s) = \langle |W[g(x)](s,u)| \rangle_u \,, \tag{2.53}$$

em que $\langle \dots \rangle_u$ denota o valor médio com respeito à variável u. Outros tipos de média podem ser considerados, como por exemplo, a média harmônica. A inclusão do valor absoluto também é opcional. Porém, seu uso é bastante comum e acredita-se que resultados numéricos melhores são obtidos usando essa forma.

Após considerar esse valor médio, a relação de escala (2.52) leva a

$$\mathcal{W}[g(x)](\lambda s) \doteq \lambda^{1/2+h}\, \mathcal{W}[g(x)](s) \,. \tag{2.54}$$

Assim, em um gráfico log-log do valor médio $\mathcal{W}[g(x)](s)$ versus s, temos uma reta cuja inclinação é numericamente igual a $1/2 + h$. Quando se lida com séries temporais, a expressão (2.53) pode ser escrita como

$$\mathcal{W}(s) = \frac{1}{n} \sum_{u=1}^{n} \left[\frac{1}{\sqrt{s}} \sum_{i=1}^{n} \psi^* \left(\frac{i-u}{s} \right) x_i \right] , \tag{2.55}$$

sendo que se omiti a notação de funcional por questão de simplicidade.

Resta escolher qual função mãe $\psi_{s,u}(x)$ a usar. Algumas são bastante comuns na literatura, como é o caso da derivada de ordem $\beta \in \mathbb{N}$ da Gaussiana

$$\psi_{s,u}(x) = \frac{(-1)^{\beta+1}}{\sqrt{\Gamma(\beta + 1/2)}} \frac{d^\beta}{dx^\beta} \exp(-x^2/2) \tag{2.56}$$

e também da Wavelet de Paul de ordem $\beta \in \mathbb{N}$

$$\psi_{s,u}(x) = \frac{(2i)^\beta \beta!}{\sqrt{\pi(2\beta)!}} (1 - ix)^{-\beta-1} \,. \tag{2.57}$$

Empiricamente, observa-se que a escolha da função mãe não é crucial para o desempenho do método. Por conta disso, usa-se aqui somente as derivadas de ordem β da gaussiana. Na figura (7), mostra-se o valor médio do coeficiente Wavelet $\mathcal{W}(s)$ para algumas trajetórias do movimento Browniano fracionário geradas usando a equação (2.37). Novamente, foram considerados diferentes valores de h (mostrados na figura), $M = 1000$ e $m = 10$. Além disso, utilizou-se como função mãe a derivada de ordem 3 da Gaussiana, equação (2.56) com $\beta = 3$. A linha tracejada representa o valor teórico da lei de potência, $\mathcal{W}(s) \propto s^{1/2+h}$.

É interessante observar que o DFA utiliza a série integrada para calcular a função de flutuação, enquanto as transformações Wavelet usam a série original. Assim, para uma comparação correta entre os dois métodos é necessário o uso da série integrada do movimento Browniano fracionário.

Vale notar também que, se por um lado o DFA gera muitas trajetórias aproximadamente independentes, as transformações Wavelet levam uma série temporal para um espaço bidimensional. Dessa maneira, bons resultados podem ser obtidos para séries razoavelmente pequenas, como mostra a figura 7.

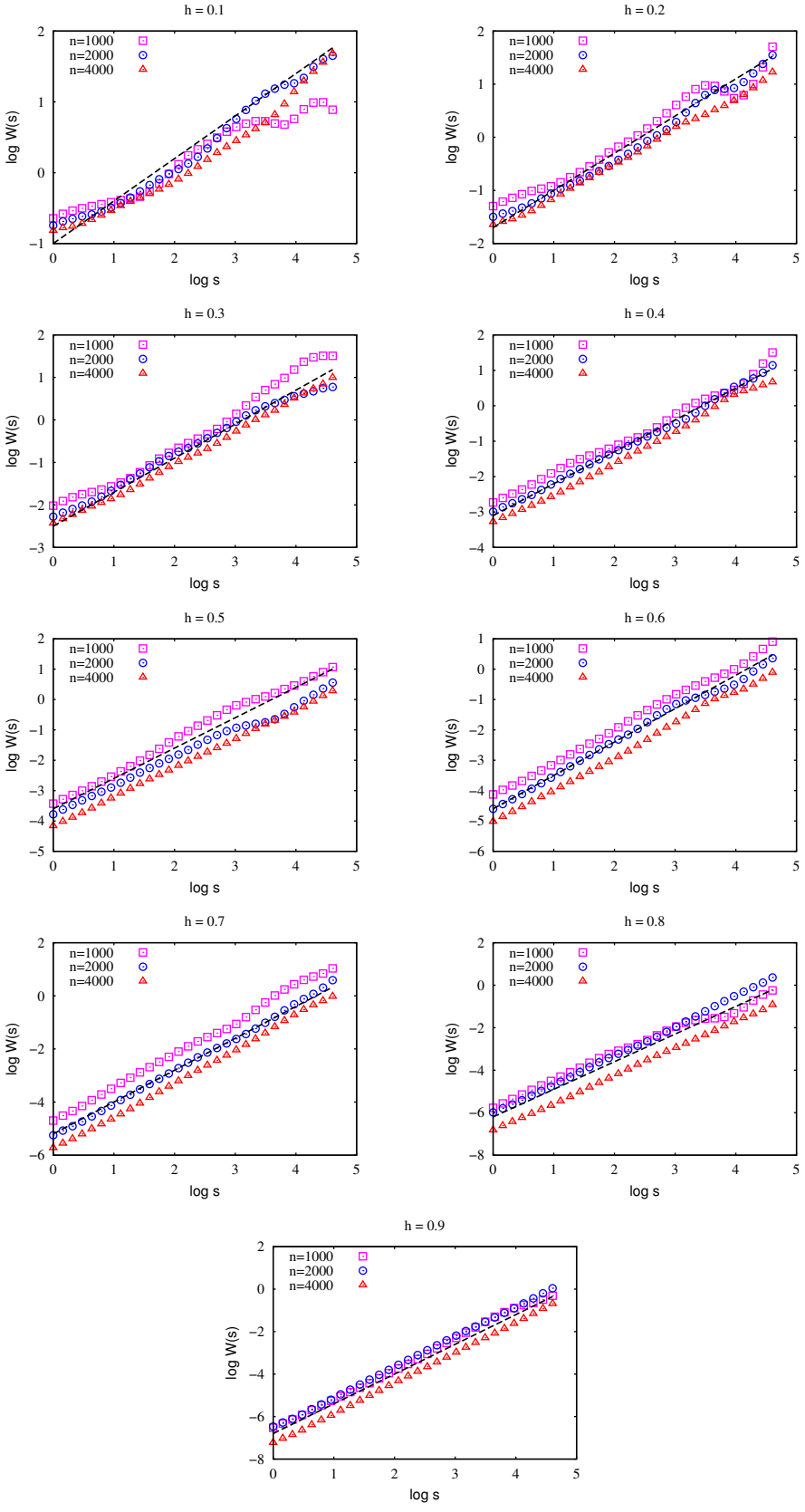

Figura 7 – Valor médio do coeficiente Wavelet $\mathcal{W}(s)$ versus s.

2.2.5 Análise entrópica e o método DEA

Outro método para detecção de correlações em séries temporais é o chamado DEA, sigla inglesa para *diffusion entropy analysis*. Trata-se de um método mais recente, proposto pelo grupo liderado pelo professor Grigolini e seus colaboradores (SCAFETTA; HAMILTON; GRIGOLINI, 2001; GRIGOLINI; PALATELLA; RAFFAELLI, 2001; SCAFETTA; GRIGOLINI, 2002). Sua implementação também é bastante simples e, como sugere o nome, este método está baseado no cálculo da entropia de trajetórias criadas a partir da série temporal.

Analogamente aos outros dois métodos, o DEA está baseado no cálculo do expoente de escala δ. Para apresentar o algoritmo, considere a série temporal z_i ($i \in \{1, 2, \ldots, n\}$). Usando essa série, constroem-se as seguintes trajetórias

$$x^{(s)}(t) = \sum_{i=1}^{t} x_{i+s} \qquad \forall t \in \{1, 2, \ldots, n\}\,, \tag{2.58}$$

em que $s \in \{0, 1, \ldots, n-t\}$. Note que essas trajetórias nada mais são do que a série original integrada de diferentes pontos iniciais. Dessa maneira, para cada t, tem-se um conjunto de $n - t + 1$ trajetórias denotadas por $x^{(s)}(t)$. Pode-se imaginar que essas trajetórias representam partículas em um movimento difusivo, para o qual pode-se definir a probabilidade $\rho(x, t)$ de encontrar uma partícula localizada entre x e $x + dx$, em um intervalo de tempo entre t e $t + dt$. Usando essa distribuição, é possível calcular a entropia do processo difusivo

$$S(t) = -\int_{-\infty}^{\infty} \rho(x,t) \ln[\rho(x,t)] dx\,. \tag{2.59}$$

Nos casos em que existir invariância de escala, isto é, $\rho(x, t) = t^{-\delta}\, \Psi(x\, t^{-\delta})$, tem-se

$$\begin{aligned} S(t) &= -\int_{-\infty}^{\infty} t^{-\delta}\, \Psi(x\, t^{-\delta}) \ln[t^{-\delta}\, \Psi(x\, t^{-\delta})] dx \\ &= -\int_{-\infty}^{\infty} \Psi(y) \ln[t^{-\delta}\, \Psi(y)] dy \\ &= -\int_{-\infty}^{\infty} \Psi(y) \ln[\Psi(y)] dy + \delta \ln(t)\,. \end{aligned} \tag{2.60}$$

Note que foi feito a mudança de variável $y = x\, t^{-\delta}$ e considerou-se $\int_{-\infty}^{\infty} \Psi(y) dy = 1$. Podendo escrever ainda

$$S(t) = A + \delta \ln(t)\,, \tag{2.61}$$

visto que o primeiro termo em (2.60) não depende de t. Assim, em um gráfico de $S(t)$ versus $\ln(t)$, deve-se obter uma reta cuja inclinação é numericamente igual ao expoente δ, o qual coincide com o expoente h de Hurst quando a função escala $\Psi(x)$ possuir o segundo momento finito. Naturalmente, o cálculo da distribuição de probabilidade $\rho(x, t)$ deve ser feito utilizando histogramas. Uma possível implementação consiste em construir M partições de tamanho ϵ no espaço das trajetórias $x^{(s)}(t)$ e contar o número de partículas

existentes em cada partição para um dado t. Denotando esse número por $N_j(t)$, pode-se definir a versão discreta de $\rho(x,t)$ como

$$p_j(t) = \frac{N_j(t)}{n-t+1}\,, \tag{2.62}$$

sendo que j representa o índice da partição. Além disso, deve-se utilizar a versão discreta da fórmula (2.59), ou seja,

$$S(t) = -\sum_{j=1}^{M} p_j(t)\ln[p_j(t)]\,. \tag{2.63}$$

Empiricamente, uma maneira eficiente de escolher o tamanho ϵ das partições, é assumi-lo independente de t e determiná-lo por uma fração conveniente do desvio padrão da série z_i. Na figura 8, as séries temporais são oriundas dos incrementos do movimento browniano fracionário, geradas usando a expressão (2.36) para diferentes valores de h, $M = 1000$ e $m = 10$. Além disso, utilizou-se partições ϵ de tamanho igual ao desvio padrão da série original, construídas do menor ao maior valor de $x^{(s)}(t)$. Novamente, bons resultados são obtidos para séries moderadamente pequenas.

Mais adiante, apresenta-se uma comparação objetiva entre os três métodos apresentados anteriormente. Mais especificamente, investiga-se como é a convergência desses métodos com relação ao tamanho (número de termos) das séries.

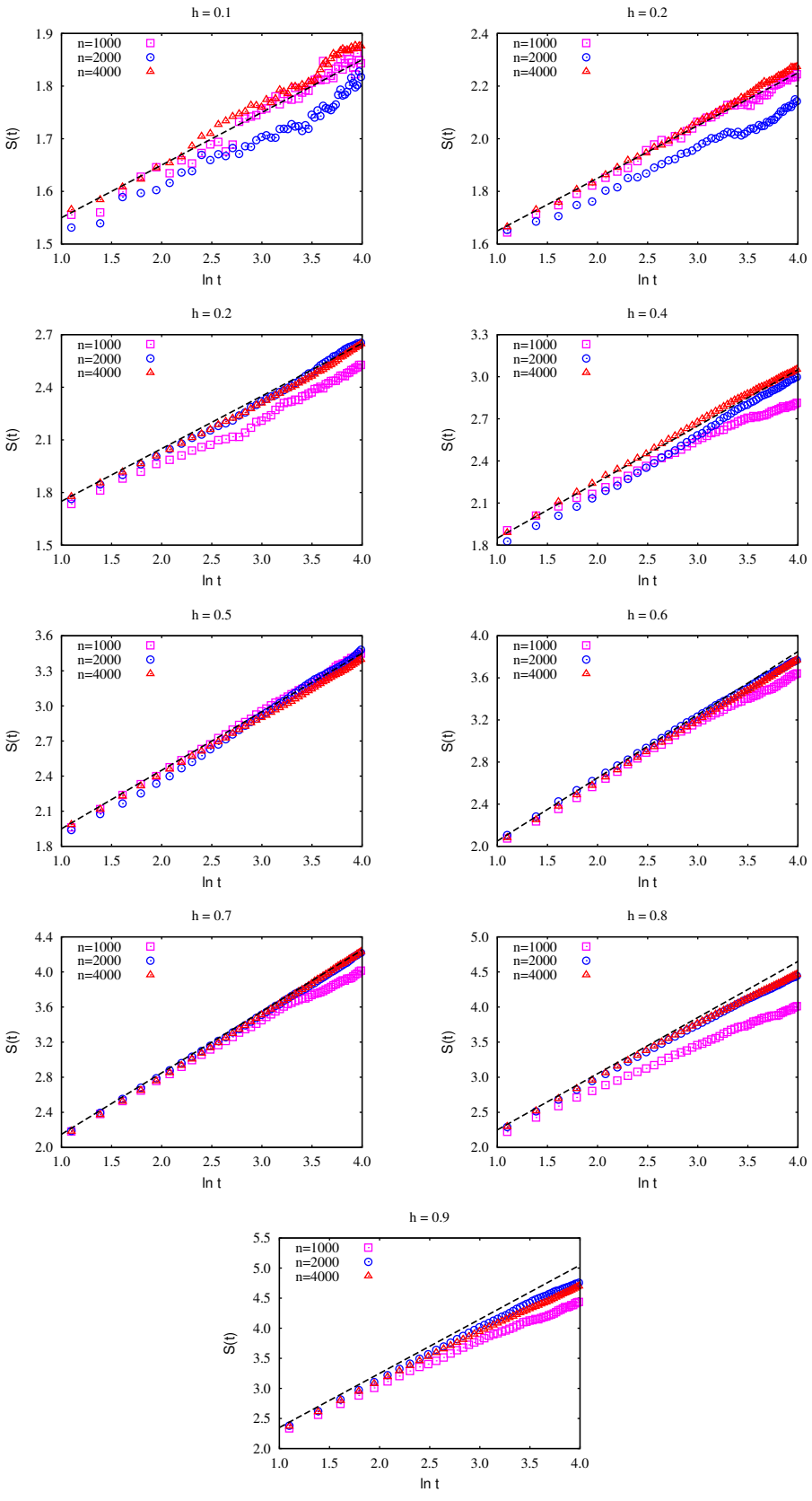

Figura 8 – Determinação do expoente de Hurst usando DEA.

2.3 ENTROPIA E COMPLEXIDADE DE PERMUTAÇÃO

Além da função de correlação, ou mais especificamente o expoente de Hurst, existem outras medidas que fornecem informações, de uma maneira geral, sobre as relações entre os elementos de uma série temporal

Nessa direção um método recente para definir uma medida de complexidade mais natural para séries temporais foi proposto por Bandt e Pompe (2002). O método chamado entropia de permutação baseia-se em associar uma sequência simbólica a segmentos da série temporal por meio de um ordenamento local. Esses "símbolos" ou estados definirão probabilidades que serão usadas no cálculo de índices entrópicos, como a entropia de Shannon ou outras medidas.

Para apresentar o método, considera-se uma série temporal $\{x(t)\}_{t=1,\ldots,N}$. E também a partição representada pelo vetor d-dimensional ($d > 1$, $d \in \mathbb{N}$)

$$s \mapsto \{x(s), x(s-\tau), \ldots, x(s-(d-2)\tau), x(s-(d-1)\tau)\}, \tag{2.64}$$

com $s = 1, 2, \ldots, N-(d-1)\tau$. O parâmetro d é chamado *embedding dimension* e τ é o *embedding delay*. Para cada um desses $[(N-(d-1)\tau]$ vetores, investiga-se as permutações $\pi = (r_0, r_1, \ldots, r_{d-1})$ dos símbolos $(0, 1, \ldots, d-1)$, definidas pelo ordenamento

$$x(s-r_0\tau) \geq x(s-r_1\tau) \geq \cdots \geq x(s-r_{d-2}\tau) \geq x(s-r_{d-1}\tau). \tag{2.65}$$

As $d!$ possíveis permutações de π definem os estados acessíveis ao sistema, para os quais calcula-se o conjunto de probabilidades $P = \{p(\pi_i)\}$, usando a frequência relativa de suas ocorrências ao longo da série temporal, dada por

$$p(\pi_i) = \frac{\#\{s \mid s \leq N-(d-1)\tau; \quad s \quad \text{do tipo} \quad \pi_i\}}{N-(d-1)\tau}, \tag{2.66}$$

no qual o símbolo # representa o número de ocorrências da permutação π_i. Além disso, define-se a distribuição de padrões ordinais $P = \{p(\pi_i)\}$ (com $i = 1, 2, \ldots, d!$) como o conjunto de probabilidades associadas a cada uma das possíveis $d!$ permutações.

A fim de deixar mais claro o procedimento anterior, considera-se um exemplo em que $x(t) = \{8, 7, 4, 9, 2, 5, 1\}$ ($N = 7$) e suponha-se que queira avaliar as permutações de π considerando o *embedding dimension* $d = 3$ e o *embbeding delay* $\tau = 2$. Assim, tem-se $(s = 1) \mapsto \{8, 4, 2\}$, $(s = 2) \mapsto \{7, 9, 5\}$ e $(s = 3) \mapsto \{4, 2, 1\}$ ordenados (respectivamente) por $\pi_{(s=1)} = (2, 1, 0)$, $\pi_{(s=2)} = (2, 0, 1)$, $\pi_{(s=3)} = (2, 1, 0)$. Desta maneira, $p(\pi = (2, 1, 0)) = 2/3$, $p(\pi = (2, 0, 1)) = 1/3$ e todas as outras probabilidades são zero ($p(\pi = (0, 1, 2)) = 0$, $p(\pi = (0, 2, 1)) = 0$, $p(\pi = (1, 0, 2)) = 0$, e $p(\pi = (1, 2, 0)) = 0$ produzindo $P = \{2/3, 1/3, 0, 0, 0, 0\}$.

A distribuição de padrões ordinais $P = \{p(\pi_i\}$ retém informações sobre a dinâmica de ordenamento dos elementos da série temporal $x(t)$; claramente, quanto maior o valor de d mais informações sobre essa dinâmica estará disponível. Contudo, a escolha

de d está intimamente ligada ao tamanho N da série, de tal modo que, para uma boa estatística a condição $d \ll N$ deve ser satisfeita. Na maioria dos casos práticos, empregar $d = 3, \ldots, 7$, como recomendado por Bandt e Pompe (2002) é suficiente. O *embedding delay* τ tem outra proposta: para $\tau > 1$, os valores de $x(t)$ não são tomados sucessivamente, e assim, o mapeamento da dinâmica de ordenamento da série temporal ocorre em diferentes resoluções temporais, permitindo uma análise mais aprofundada dos processos relacionados à série $x(t)$. Uma vez definido o processo para a obtenção da distribuição de padrões ordinais $P = \{p(\pi_i\}$, pode-se calcular a entropia (normalizada) de permutação de Shannon (1948) definida por

$$H[P] = -\frac{1}{S_{max}} \sum_{i=1}^{d!} p(\pi_i) \log p(\pi_i) \tag{2.67}$$

onde $S_{max} = \log d!$ é obtido quando todas as permutações π_i são equiprováveis, ou seja, $P = P_e = \{1/d!\}$. Por definição, $0 < H < 1$, sendo que o limite superior ocorrerá para séries completamente aleatórias e espera-se valores de $H < 1$ para séries que possuam algum tipo de dinâmica de ordenamento correlacionadas.

A outra quantidade que foi calculada à partir de $P = \{p(\pi_i\}$ é a medida de complexidade estatística de López-Ruiz et al. (1995) dada por

$$C[P] = Q[P, P_e]H[P] \tag{2.68}$$

no qual $Q[P, P_e]$ é uma métrica entrópica relativa entre a distribuição ordinal empírica $P = \{p(\pi)\}$ e o estado equiprovável $P_e = \{1/d!\}$. A quantidade $Q[P, P_e]$ é muitas vezes denominada desequilíbrio e é definida em termos da divergência de Jensen-Shannon (GROSSE et al., 2002) (ou ainda em termos da divergência de Kullback-Leibler simetrizada (LIN, 1991)):

$$Q[P, P_e] = \frac{S[(P + P_e)/2] - S[P]/2 - S[P_e]/2}{Q_{max}} \tag{2.69}$$

em que

$$Q_{max} = -\frac{1}{2} \left\{ \frac{d! + 1}{d!} \log(d! + 1) - 2 \log(2d!) + \log(d!) \right\}$$

é o valor máximo possível de $Q[P, P_e]$, obtido quando um dos componentes de P é igual a um e todos os outros são nulos. O desequilíbrio C quantifica o grau de estruturas correlacionadas, fornecendo informações adicionais importantes que podem não ser obtidas somente pela entropia de permutação. Além disso, para um dado valor de $H[P]$ existe um intervalo de possíveis valores para $C[P]$ (MARTIN; PLASTINO; ROSSO, 2006). Esse comportamento foi a principal razão pela qual Rosso et al. (2007) propuseram empregar um diagrama de $C[P]$ versus $H[P]$ como uma ferramenta de diagnóstico, construindo o que vem sendo chamado de *complexity-entropy causality plane*. Além disso, pode-se calcular $H[P]$ e $C[P]$ em função do *embedding delay* τ, definindo os chamados espectros de permutação. Essa última análise pode revelar, por exemplo, estruturas periódicas da série $x(t)$, como pode ser visto no estudo das vazões de rios brasileiros (apresentada a seguir).

3 APRESENTAÇÃO DOS DADOS E ANÁLISE DE RESULTADOS

3.1 UMA APLICAÇÃO AO ESTUDO DAS VAZÕES DE RIOS

Os dados analisados são provenientes de séries temporais diárias das vazões naturais dos rios (descargas fluviais) medidas em 141 estações diferentes nas proximidades das usinas hidrelétricas e abrangem 53 rios brasileiros no período de 1931 à 2012. Essas informações são de domínio público, disponibilizados gratuitamente pelo Operador Nacional do Sistema Elétrico - ONS - (órgão federal do governo brasileiro que controla o sistema de energia no Brasil) disponível em <http://www.ons.org.br/operacao/vazoes_naturais.aspx>. As figuras 9(a) e 9(b), ilustram, respectivamente, a evolução do número de estações ao longo dos anos e o histograma do comprimento de registro (em anos) para todas as estações analisadas. O conjunto de dados começa com 45 estações em 1931 e desde 1995 há 141 estações que cobrem 53 rios brasileiros (BRAGA et al., 2016).

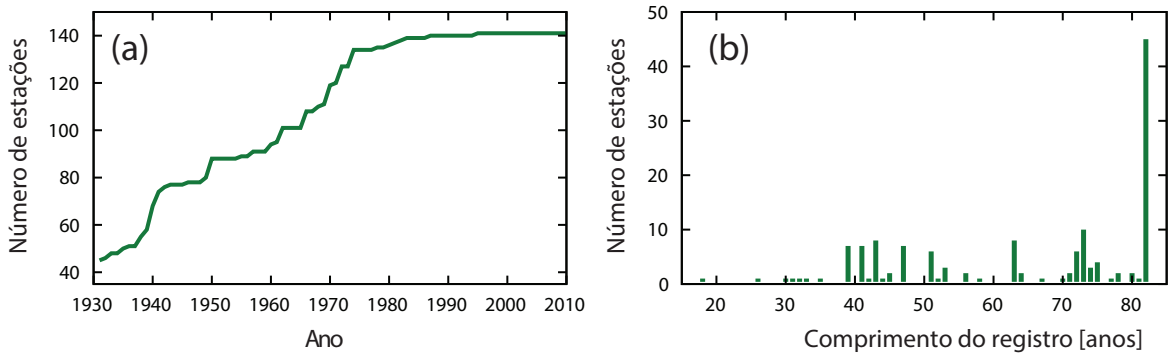

Figura 9 – Descrição esquemática do conjunto de dados.

A vazão será denotada por $F_t(i)$, em que $i = 1, 2, \ldots, 365$, é uma variável discreta indexada aos dias do ano e t representa o ano associado à vazão; dessa forma, $F_{1986}(10)$ representa a vazão em 10 de janeiro de 1986 em uma dada estação. Por questão de conveniência, os ponto dos dados associado a 29 de fevereiro de todas as séries temporais de anos bissextos foram removidos, garantindo o mesmo comprimento para todas as séries temporais analisadas.

3.1.1 CARACTERIZAÇÃO EM LARGA ESCALA DAS FLUTUAÇÕES DAS VAZÕES DE RIOS POR MEIO DE GRAFO DE VISIBILIDADE HORIZONTAL

O método de visibilidade horizontal tem sido aplicado em vários contextos associados a sistemas complexos (LACASA et al., 2009; YANG et al., 2009; ELSNER; JAGGER; FOGARTY, 2009; AHMADLOU; ADELI; ADELI, 2010; MURKS; PERC, 2011; TELESCA; LOVALLO, 2012; GAO; JIN, 2012; JIANG et al., 2013; ZHUANG; SMALL; FENG, 2014; TELESCA; LOVALLO; TOTH M., 2014; ZOU Y.; KURTHS,

2014; ZHANG; WANG; W., 2015) permitindo que pesquisadores façam uso recorrente da análise de séries temporais como ferramentas de análises de redes complexas.

O procedimento utilizado nesse trabalho de caracterização foi proposto por Luque et al. (2009) como uma versão mais restritiva do método do grafo de visibilidade apresentado por Lacasa et al. (2008), com a vantagem de proporcionar expressões analíticas de séries temporais totalmente aleatórias (LUQUE et al., 2009; LACASA; TORAL, 2010).

Nessa análise foi considerada uma versão padronizada das vazões, definida pela equação (3.1)

$$f_t(i) = \frac{F_t(i) - \mu(i)}{\sigma(i)}\,, \tag{3.1}$$

na qual

$$\mu(i) = \frac{1}{n}\sum_{t=1}^{n} F_t(i) \quad \text{e} \quad \sigma^2(i) = \frac{1}{n-1}\sum_{t=1}^{n}[F_t(i) - \mu(i)]^2 \tag{3.2}$$

são, respectivamente, a média e a variância do perfil da vazão ao longo dos dias do ano para uma dada estação e n é o número de anos disponíveis para aquela estação.

O painel superior da figura 10(a) ilustra a definição das vazões normalizadas $f_t(i)$ ao longo dos dias ($i = 1, 2, \ldots, 365$) para um determinado ano e estação, mostrando um exemplo concreto da sua construção. O procedimento consiste em subtrair o fluxo médio $\mu(i)$ do fluxo $F_t(i)$ em um determinado ano e dividir o resultado pelo desvio-padrão do fluxo diário $\sigma(i)$ (todos em unidades de $10^4 \times \mathrm{m}^3/\mathrm{s}$). Ao fazer este procedimento, tem-se a garantia que, pelo menos, a principal tendência sazonal é removida das vazões originais. Uma vez tendo a vazão normalizada $f_t(i)$, aplicou-se o método do grafo de visibilidade horizontal para o mapeamento de cada ano da série temporal em uma rede complexa correspondente.

Conforme descrito anteriormente o algoritmo de visibilidade horizontal é um mapa que atribui a cada ponto de uma série temporal um nó/vértice em uma rede complexa. Dois nós, i e j, serão conectados sempre que for possível a construção de uma linha horizontal no espaço de séries temporais que juntam os pontos $f_t(i)$ e $f_t(j)$ sem interseção em qualquer altura do ponto intermediário, ou seja:

$$\textcircled{i} \leftrightarrow \textcircled{j} \quad \text{quando} \quad [f_t(i), f_t(j)] > f_t(l)\ \forall\, l\,|(i < l < j)\,. \tag{3.3}$$

Observe que, nesse trabalho, o algoritmo de visibilidade horizontal produzirá uma rede por ano da série temporal de uma determinada estação. O painel inferior da figura 10(a) mostra um exemplo de uma rede construída a partir do procedimento anterior, considerando dados do rio Paraná, coletadas na estação de Itaipu no ano $t = 1931$, na qual, cada nó representa um dia do ano (mostrado na rede). O tamanho dos nós são proporcionais a seus graus e as cores são ilustrativas. Por outro lado, a figura 10(b) ilustra o método do grafo de visibilidade horizontal aplicado a uma série temporal simples. O painel esquerdo mostra uma série temporal representada por barras verticais. Aqui, as

linhas tracejadas horizontais indicam as conexões de rede criado de acordo com o critério geométrico da equação (3.3). O painel da direita mostra a rede que emerge dessa série temporal.

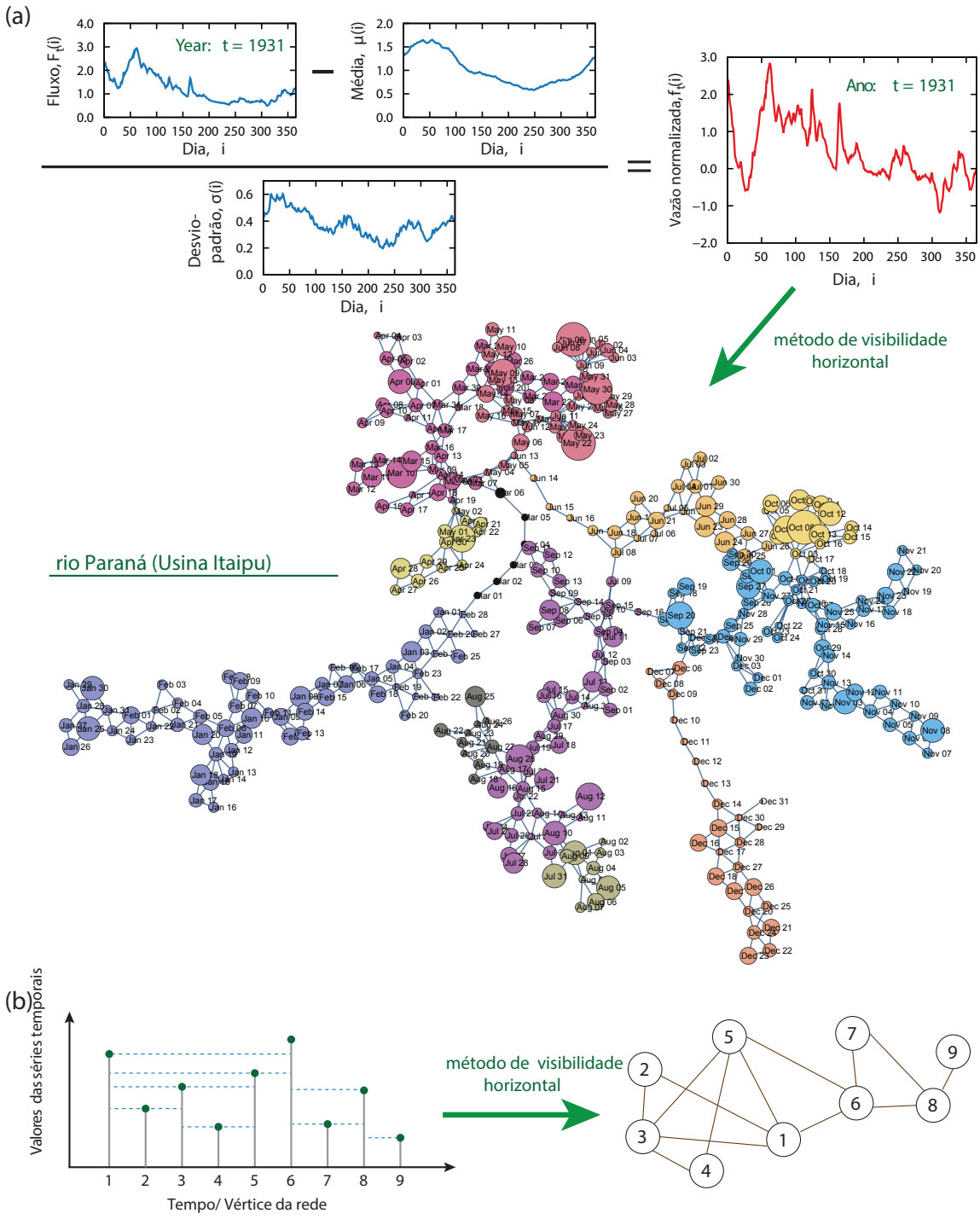

Figura 10 – Ilustração esquemática da construção da rede das descargas fluviais.

Foram avaliados pela primeira vez a distribuição de graus da rede $P(k)$, na qual k representa o grau do nó ou seu número de conexões (há que se notar que as redes são não-direcionadas). Ainda com relação aos trabalhos de Luque et al. (2009) e Lacasa e

Toral (2010), tem-se que a distribuição de grau assume a forma:

$$P(k) \sim \exp(-\lambda\, k) \;\; \text{com} \;\; \lambda = \lambda_{\text{rand}} = \ln(3/2) \tag{3.4}$$

para as séries temporais totalmente aleatórias, independentemente da distribuição de probabilidade subjacente aos valores da série temporal. Além disso, séries temporais decorrentes de dinâmicas mais complexas também são geralmente descritas por distribuições exponenciais assintótica. No entanto, o valor de λ é geralmente diferente do λ_{rand}. Casos em que $\lambda < \lambda_{\text{rand}}$ estão associados a processos caóticos (quanto menor o valor numérico de λ, menor a dimensionalidade do sistema) e casos em que $\lambda > \lambda_{\text{rand}}$ estão associados a processos estocásticos (quanto maior o valor numérico de λ, maior será o tempo de correlações do sistema) (LACASA; TORAL, 2010).

Para o caso específico desse trabalho, a figura 11(a) mostra as distribuições cumulativas de grau para todas as 82 redes (um para cada ano) construídas com os dados do rio Paraná na posição em que situa-se a Usina de Itaipu. Há que se destacar, que a construção concreta da hidrelétrica ocorreu entre os anos de 1975 a 1982, quando foram fechadas as comportas da represa. Na figura 11(a) verifica-se as distribuições cumulativas de grau $P(k)$ (em escala log-lin) para cada rede construída a partir de séries temporais disponível para o rio Paraná (Estação de Itaipu). As curvas em cinza são as distribuições para cada ano e a linha preta sólida é um ajuste exponencial à média por janela destas distribuições, nas quais o expoente característico médio é $\langle\lambda\rangle = 0.74 \pm 0.01$. A linha tracejada vermelha ilustra o decaimento exponencial esperado para uma série temporal aleatória, isto é, $P(k) \sim \exp(-k/\lambda_{\text{rand}})$, com $\lambda_{\text{rand}} = \ln(3/2) \approx 0.41$.

Observa-se que todas as distribuições são assintoticamente bem descrita por decaimentos exponenciais, que na escala log-lin são representados por linhas retas cujas inclinações coincidem com os valores de λ. Ao comparar o comportamento assintótico dessas distribuições ao esperado para uma série temporal aleatória $[P(k) \sim \exp(-\lambda_{\text{rand}}\, k)]$ é possível verificar que as distribuições empíricas mostram um decaimento mais rápido.

A fim de estimar numericamente o valor empírico de λ para cada distribuição, foi montado um modelo linear para cada distribuição cumulativa (em escala log-lin) após a remoção do comportamento não-exponencial inicial ($k > 4$). Foi obtido ainda, o valor médio de λ depois de embaralhar as séries temporais 100 vezes.

A figura 11(b) mostra os valores de λ para cada ano da série original, bem como a média (e intervalos de confiança) de λ para as versões embaralhadas dessas séries, isto é, verifica-se a evolução do expoente característico λ para cada uma das distribuições anteriores. A linha preta mostra os valores empíricos de λ após embaralhar a série temporal (média sob 100 realizações) e as áreas sombreadas claras (escuras) representa seus 95% (99%) intervalos de confiança *bootstrap*. Observa-se que os valores empíricos de λ (para essa estação) são sempre maiores que λ_{rand} e que eles estão fora do limite de confiança relativo as versões aleatórias dessas séries.

As figuras 11(c) e (d) fornecem outro exemplo representativo desta análise (para o rio Tocantins na estação Tucuruí), em que, novamente, foram encontrados valores de λ maiores do que λ_{rand}.

A fim de caracterizar plenamente a distribuição de graus de todo conjunto de dados, foi realizado um procedimento idêntico ao dos dois exemplos anteriormente discutidos, com o objetivo de estimar os valores de λ para cada série temporal. Em seguida, foi calculada a distribuição de probabilidade $P(\lambda)$ para os valores de λ obtidos a partir da série temporal original, bem como para uma versão embaralhada de cada série temporal.

A figura 11(e) mostra a distribuição $P(\lambda)$ para os valores de λ obtidos a partir da série temporal original (linha preta) e para as embaralhadas aleatoriamente (linha vermelha). Note que os valores de λ estão concentrados em torno de λ_{rand} para as séries aleatórias; considerando que os valores de λ das séries originais são quase sempre maior do que λ_{rand} (99.7% das series), com um valor médio (avaliado ao longo de todos os rios e anos) de 0.65. A inserção mostra essas distribuições para uma maior parcela do gráfico. Portanto, a figura 11 mostra ambas as distribuições e deixa evidente que os valores de λ para a série original são maiores do que λ_{rand} para praticamente todas as séries temporais no conjunto de dados estudados, de maneira que apenas 0.3% das séries temporais assumem valores ligeiramente pequenos (mais próximos de λ_{rand}).

Vale a pena notar que o valor médio de λ é 0.65, um valor que pode estar relacionado com correlações de longo alcance das vazões dos rios. Em particular, se levado em consideração que as séries temporais analisadas são descritas por uma função de autocorrelação ajustada por uma lei de potência $R(\tau) \sim \langle f_t(i+\tau) f_t(i) \rangle \sim \tau^{-\gamma}$, a média $\langle \lambda \rangle = 0.65$ corresponde a $\gamma \approx 0.5$ (figura 3 do trabalho de Lacasa e Toral (2010)) e, consequentemente, a um expoente de Hurst em torno de 0.75, obtém-se um resultado que é compatível com a existência de correlações de longo alcance persistente nas vazões.

Outra questão que os dados do trabalho permitem abordar está relacionada a possíveis características evolutivas na dinâmica das vazões dos rios. Para resolver essa questão, foi investigado se os valores de λ de uma determinada estação apresentam uma dependência temporal ao longo do ano t. Especificamente, foi testada a hipótese da relação λ em função de t apresentar uma tendência linear, obtida pelo ajuste do modelo linear (via método dos mínimos quadrados ordinários)

$$\lambda = a + b\,t + \varepsilon, \qquad\qquad \varepsilon \sim N(0, \sigma^2) \tag{3.5}$$

sendo a e b os parâmetros do modelo. Dessa forma, o valor de b e sua significância estatística fornece pistas sobre se λ está mudando ao longo do ano t ou não.

A figura 12 mostra os valores de b para cada estação. O painel principal mostra os valores dos coeficientes lineares b obtidos pelo método dos mínimos quadrados ajustados pelo modelo para as relações entre o expoente característico λ e o tempo t (o ano associado com a série temporal) para todos os rios. Os círculos em cinza mostram os valores de b que não são estatisticamente significativos (isto é, rios em que a relação entre λ e t

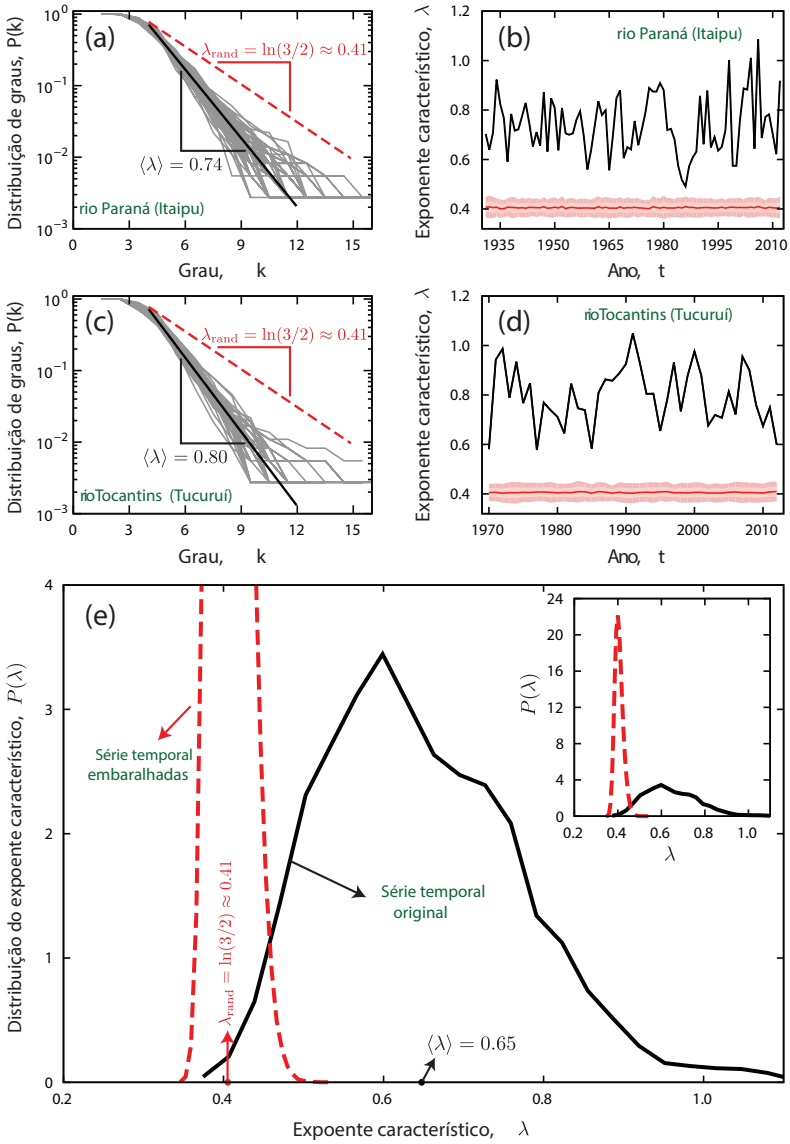

Figura 11 – Distribuição de grau e a natureza correlacionada das vazões normalizadas dos rios.

são aproximado por uma função constante) e os asteriscos vermelhos mostram os valores significativamente diferentes de zero. Os painéis secundários, indicados pelas setas fornecem casos representativos para a relação λ em função de t (círculos vermelhos), em que as linhas tracejadas representam os modelos lineares ajustados.

Para um nível de confiança de 95% (ou seja, p-valor menor que $0,05$), foi descoberto que a hipótese nula do coeficiente b ser diferente de zero não pode ser rejeitada em 46 entre 141 estações de medição (figura 12). Além disso, entre as estações onde b é estatisticamente significativo, foi revelado que a maioria das estações deste grupo (72%) mostra uma tendência crescente. Esses resultados indicam que algumas estações estão apresentando uma dinâmica mais correlacionada ao longo dos anos. É importante mencionar que o *t-teste* assume que as distribuições residuais são normais e que podem ser consideradas como não ideal neste contexto (KOUTSOYIANNIS, 2002). No entanto, aplicando ainda mais o método de regressão *bootstrap* (EFRON; TIBSHIRANI, 1993) os

resultados obtidos são bastante semelhantes, ou seja, as tendências lineares consideradas significantes pelo *t-teste*, também foram consideradas estatisticamente significativas pela regressão *bootstrap*.

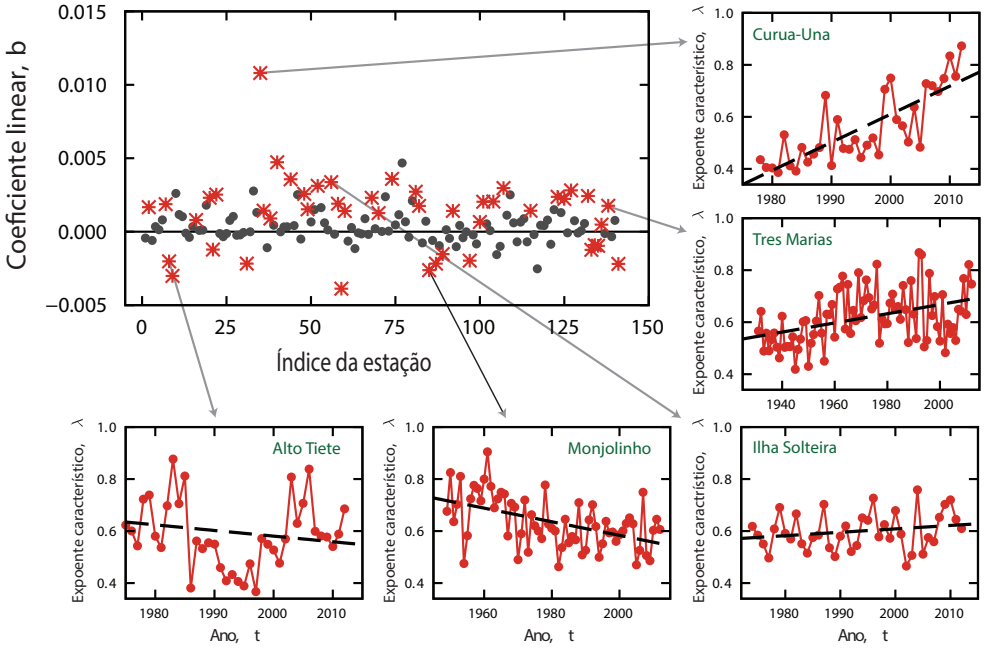

Figura 12 – Tendências evolutivas na distribuição de graus.

Em adição a distribuição de grau, uma outra propriedade comum no estudo de redes é o coeficiente de aglomeração (WATTS; STROGATZ, 1998; NEWMAN; STROGATZ; WATTS, 2011; NEWMAN, 2003). Lembrando que essa quantidade mede a probabilidade dos nós criarem grupos muito unidos com uma densidade relativamente alta de conexões. Especificamente, será calculado o coeficiente de aglomeração global C.

Similarmente à análise da distribuição de grau, foram comparados os valores de C obtidos a partir das redes emergentes da série temporal original com as redes obtidas a partir de versões embaralhadas dessas séries (média sob 100 realizações).

As figuras 13(a) e 13(b) mostram, respectivamente, os casos representativos para os valores de C avaliados a partir de duas estações de medição: o rio Paraná (Estação de Itaipu) e o rio Tocantins (Estação de Tucuruí). Para ambas as estações, nota-se que os valores de C avaliados para a série temporal original (linhas pretas) são consideravelmente maiores do que os valores obtidos a partir da série aleatória (linhas vermelhas). As linhas pretas mostram os valores de C para cada t. As linhas vermelhas mostram os valores de C após embaralhar a série temporal (média sob 100 realizações) e as áreas sombreadas claras (escuras) representa seus 95% (99%) intervalos de confiança *bootstrap*. Figura 13(c) mostra a distribuição do coeficiente de aglomeração $P(C)$ para os valores de C obtidos a partir da série temporal original (linha preta) e para as embaralhadas aleatoriamente (linha vermelha). Observe que os valores de C estão concentrados em torno de $\langle C_{\text{rand}} \rangle = 0.346$ para as séries embaralhadas; considerando que os valores de C da série temporal original

são quase sempre maior do que $\langle C_{\text{rand}} \rangle$ (99.1% das séries), com um valor médio (avaliado ao longo de todos os rios e anos) de 0.399. A inserção mostra essas distribuições para uma maior parcela do gráfico.

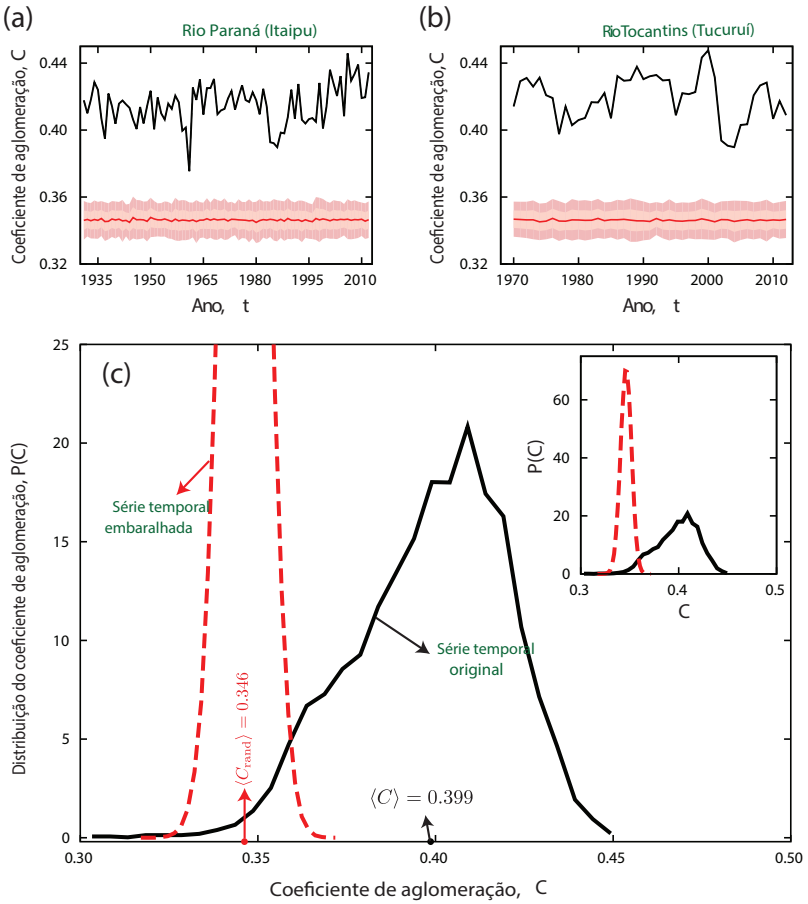

Figura 13 – Coeficientes de aglomeração global das redes.

De fato, pode-se constatar a partir da análise da figura 13(c), que os valores de C, das séries originais, são distribuídos em torno de $\langle C \rangle = 0.399$, enquanto que os valores de C relacionados às séries embaralhadas apresentam uma distribuição mais aguda em torno $\langle C_{\text{rand}} \rangle = 0.346$. Esse resultado indica, portanto, que as flutuações subjacentes às vazões dos rios produzem redes com estruturas internas mais complexas, os quais são observados visualmente na figura 10, observando a formação de grupos entre os dias mais próximos.

As características evolutivas dos valores de C ao longo dos anos t para todas as estações também podem ser investigadas. Como feito anteriormente para a distribuição de grau, foi verificada a hipótese de C em função de t exibir uma tendência linear por meio do modelo linear

$$C = a' + b'\, t\, + \varepsilon, \qquad\qquad \varepsilon \sim N(0, \sigma^2) \tag{3.6}$$

onde a' e b' agora são os parâmetros do modelo. De maneira equivalente à equação (3.5), os valores de b' na equação (3.6) e a sua significância fornecem indícios que informam se o valor de C está mudando ao longo do ano t ou não.

A figura 14 mostra os valores de b' para cada estação. O painel principal mostra os valores dos coeficientes lineares b' obtidas por meio do método dos mínimos quadrados para todos os rios. Os círculos em cinza mostram os valores de b' que não são estatisticamente significativos (isto é, rios para os quais a relação entre C e t são bem aproximados por uma função constante) e os asteriscos azuis mostram os significativos. Os painéis indicados pelas setas fornecem casos representativos para a relação C em função de t (círculos azuis), em que as linhas tracejadas representam os modelos lineares ajustados. Note que a maior parte das estações que apresentam evoluções significativas para os valores de C também apresentam tendências evolutivas nos valores λ (ver Fig. 12). Na figura 14, os asteriscos indicam os valores que são estatisticamente significativos (nível de confiança de 95%). Os resultados mostram que entre as 141 estações de medição, 68 estações apresentam uma tendência linear estatisticamente significativa e que a maioria destas estações (72%) apresentam uma tendência crescente em C (resultados similares são obtidos com o método da regressão *bootstrap*). A figura 14 ilustra, ainda, alguns casos representativos dessas tendências evolutivas.

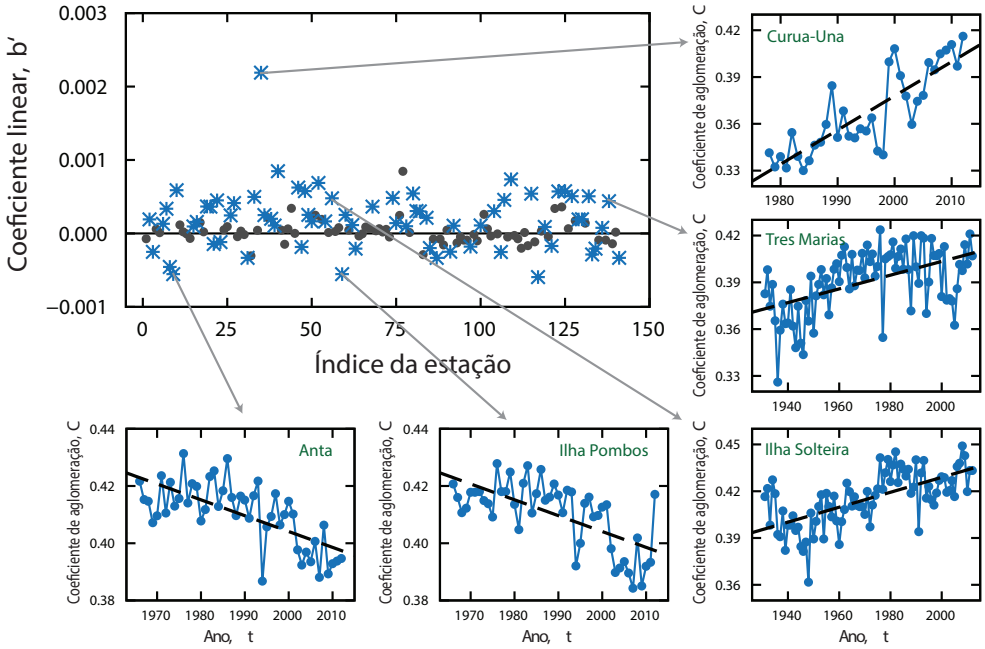

Figura 14 – As tendências evolutivas no coeficiente de aglomeração.

Outro aspecto intrigante verificado na figura 14 é que a maioria das estações que apresentam características evolutivas em C também apresentaram uma dinâmica semelhante em λ. Esse comportamento sugere que os valores de C e λ podem ser de alguma forma acoplados.

Para verificar essa possibilidade, foi construído o gráfico dos valores de C em função dos valores de λ, avaliados para cada estação e ano da série temporal. A figura 15 mostra essa relação na qual (apesar da dispersão) verifica-se que o aumento de λ é (em média) seguido de um aumento de C. Os pontos em cinza são os valores do coeficiente de

aglomeração C em função do expoente característico λ para cada série temporal em nosso conjunto de dados. Para contornar a dispersão dos dados e concentrar-se na principal tendência, foram calculados os valores médios por janela dessa relação. Os pontos vermelhos são valores médios por janela da relação anterior, as barras de erro representam um desvio padrão e a linha contínua é um ajuste exponencial aos valores médios (parâmetros são mostrados nos gráficos). Os resultados, sugerem que uma função exponencial descreve bem a relação média, o que confirma um acoplamento (em média) entre os valores de C e λ.

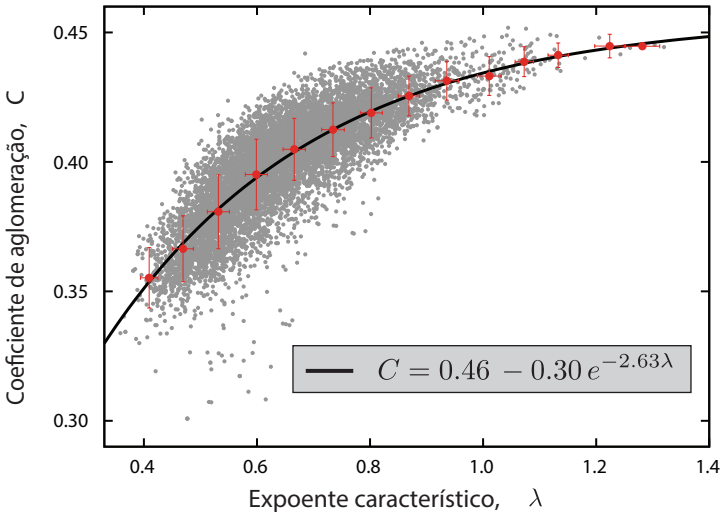

Figura 15 – A relação entre o coeficiente de aglomeração e o expoente característico.

3.1.2 ANÁLISE EM LARGA ESCALA DAS VAZÕES DE RIOS NO BRASIL À PARTIR DO USO DAS TÉCNICAS DE DFA(DETRENDED FLUCTUATION ANALYSIS), ENTROPIA E COMPLEXIDADE DE PERMUTAÇÃO

Para a análise apresentada nesta seção, será empregado os mesmos dados usados na subseção anterior. Por questão de conveniência, nesta seção, a vazão será definida por uma função $x(t)$ (em m^3/s) e $t = 1, 2, \ldots$ sendo os dias do ano. Uma vez mais, removeu-se o último ponto dos dados de todas as séries temporais de anos bissextos, garantindo que todas as séries temporais tenham o mesmo comprimento.

A figura 16 mostra exemplos da evolução temporal de três estações diferentes ao longo de quatro anos. Cada gráfico mostra a vazão natural $x(t)$ em unidades de m^3/s ao longo do ano (indicado nos gráficos) com taxa de amostragem de um dia. O painel à esquerda mostra medida no rio Paraná nas proximidades da barragem de Itaipu; o painel da direita mostra os dados do rio Tocantins obtidos nas proximidades da barragem de Tucuruí; finalmente, o painel da direita mostra dados do rio Paraná medido nas proximidades da barragem de Ilha Solteira. Uma das característica marcantes apresentadas

nessas figuras, é a sazonalidade natural da evolução de $x(t)$, o que reflete fortemente a associação das descargas fluviais com o sistema climático.

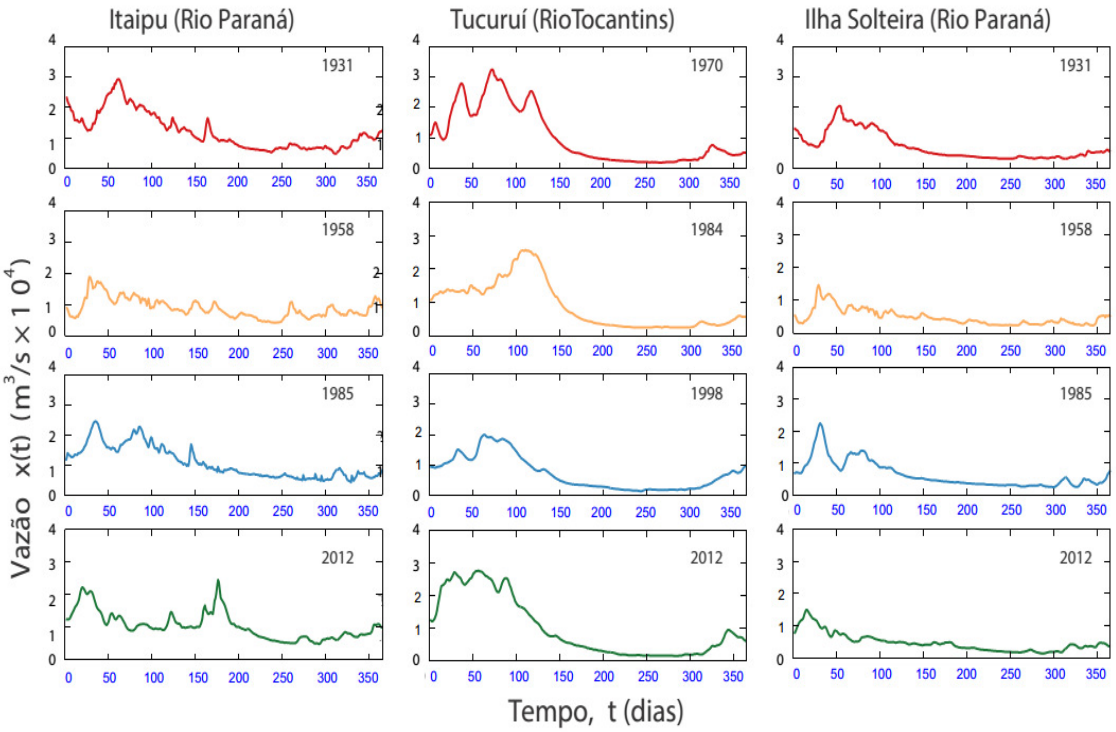

Figura 16 – Exemplo de séries temporais das vazões naturais em rios brasileiros.

A partir da análises dessas figuras pode-se inferir que grandes descargas pluviométricas ocorrem durante os primeiros e os últimos meses do ano (verão) e, por outro lado, verifica-se que as baixas descargas aparecem durante o inverno (em torno do meio do ano). Nesse contexto, nasce um questionamento natural acerca de como se pode definir o período T associado à essas séries temporais, bem como, se este período T tem algum aspecto evolutivo ao longo dos anos.

Afim de se obter uma resposta a essa pergunta, foi aplicado o chamado espectro de permutação, que consiste em avaliar a entropia H e complexidade C como função do *embedding delay* τ, em um intervalo particular. Semelhantemente ao que ocorre com a *embedding dimension* d (veja a seção 2.3 na página 41), grandes valores de τ exigem longas séries temporais. Assim para ter boas estatísticas, escolheu-se o valor máximo de τ em torno de 20% do comprimento N da série temporal.

Esse espectro de permutação provou ser bem sucedido na identificação de fenômenos de *delay* e mecanismos de *"feedback"* em séries temporais (ZUNINO et al., 2010; KULP; ZUNINO, 2014). Em geral, as relações H em função de τ e C em função de τ exibem picos ou vales que correspondem a harmônicos e sub-harmônicos associados à série temporal.

A fim de identificar os períodos T associados às descargas fluviais, agrupou-se a série temporal $x(t)$ em intervalos de tempo de 20 anos e, para cada conjunto, calculou-se o espectro de permutação para H e C. A figura 17(A) mostra um exemplo do espectro de permutação das descargas do rio Paraná na estação de Itaipu, tanto para H (painel superior) e C (painel central). Observou-se que a entropia H apresenta picos, enquanto a complexidade C tem vales, espaçados em cerca de 365 dias.

Para estimar o período T associado à série temporal, analisou-se a diferença entre H e C, como ilustrado no painel inferior da figura 17(A). Isso garante o uso de mais informações sobre a série temporal e pode ajudar a aumentar a precisão no valor estimado de T. Nessa relação entre $(H - C)$ e τ, identificou-se numericamente a localização dos picos, isto é, os valores de $\tau = \tau_i^*$ em que os picos ocorrem, mediante a imposição de que um pico deve ser o maior valor em torno de 180 valores menores à esquerda e à direita.

A performance deste procedimento simples foi verificada manualmente, sendo que não foram encontrados erros de identificação. O período T associado a um intervalo de tempo particular de uma série temporal é, assim, definido como o valor médio da diferença entre dois picos consecutivos τ_i^*. O painel inferior da figura 17(A) mostra os picos identificados na relação de $H - C$ em função de τ e também ilustra a definição de T.

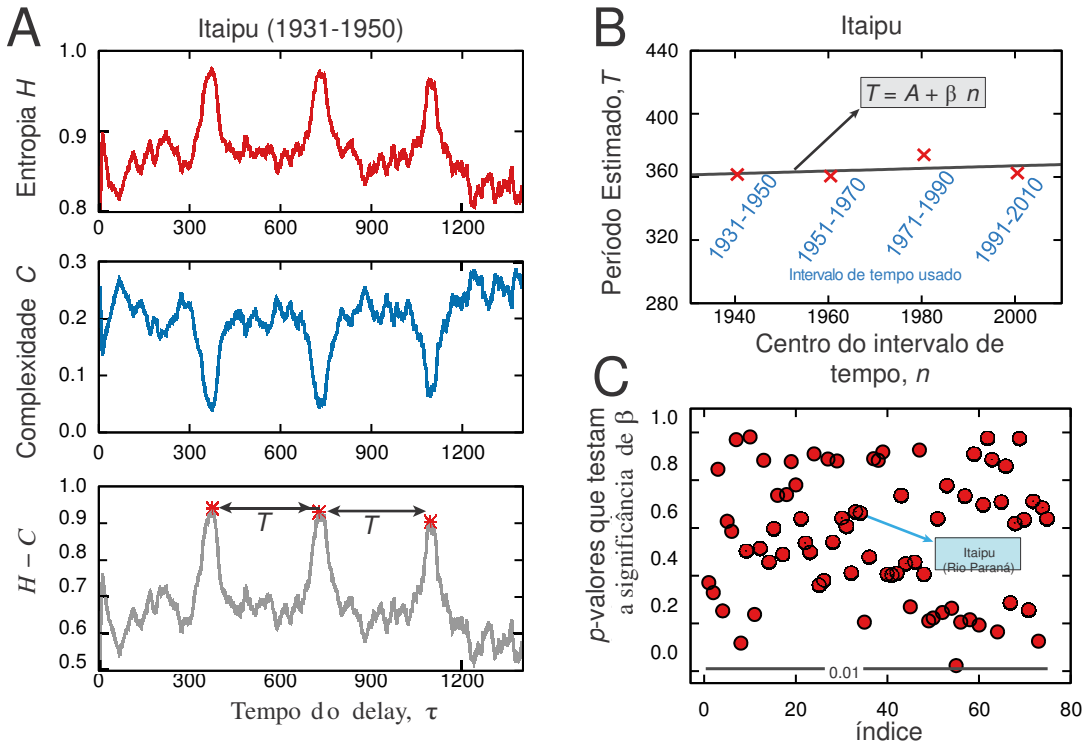

Figura 17 – Obtendo o período das séries temporais através da entropia de permutação e complexidade estatística.

Estimou-se o valor de T associado a cada intervalo de tempo de 20 anos para todas as séries temporais, como pode ser verificado na figura 17(B) na qual, são mostrados, os valores estimados de T de quatro intervalos de tempo para as descargas do rio Paraná na

estação de Itaipu. Note que o T estimado é de cerca de 365 dias, no entanto, esse valor também exibe pequenas flutuações. Em particular, a partir do exemplo da figura 17(B), pode-se supor que T tem tido incrementos com o decorrer do tempo. Para ter acesso, de forma sistemática, se T apresenta algum aspecto evolutivo, pode-se ajustar um modelo linear a relação entre T e o centro do intervalo temporal usado para obter T para todas as séries temporais com comprimento mínimo de 80 anos, tal que

$$T = A + \beta\, n \tag{3.7}$$

em que A e β são constantes, e n representa o centro do intervalo de tempo usado para estimar T. Na figura 17(B), a linha contínua mostra o modelo ajustado, em que o coeficiente linear β foi considerado não-significativo ($\beta = 0.08 \pm 0.16$; p-valor $= 0.62$), sugerindo que uma função de valor constante é uma descrição melhor para estes dados.

Foi verificada a significância estatística de β para todas as estações através dos seus p-valores, conforme mostrado na figura 17(C). O p-valor indica a probabilidade do parâmetro β ser diferente de zero, de forma que p-valores menores do que 0.01 (linha horizontal no gráfico) significa que rejeita-se a hipótese nula de que os valores de β sejam nulos a um nível de confiança de 99%, isto é, que existe apenas 1% de chance de β não ser diferente de zero. A figura 17(C) mostra todos os p-valores e deixa claro que β não tem significado estatístico e, portanto, T não apresenta uma tendência linear crescente ou decrescente.

Naturalmente, esse resultado não exclui outras possibilidades de evolução para T, como um comportamento oscilatório ou dependências temporais mais complicadas. Ainda assim, quatro valores para T não são suficientes para responder essas perguntas de forma objetiva. Além disso, outra questão importante (provavelmente uma das mais importantes) está relacionada às flutuações de T, por exemplo, essas flutuações são crescentes? Essa hipótese é muito mais difícil de testar, pois as flutuações de τ_i não tem qualquer razão para ser diretamente associadas às flutuações em T.

Agora, a análise será na versão normalizada $z(t)$ da série temporal $x(t)$. A figura 18 ilustra essa definição e mostra que esse procedimento remove a principal sazonalidade presente em $x(t)$, do mesmo modo como realizado na subseção anterior. Essa mostra o valor médio da vazão, $\mu(t)$, em função do dia t no rio Paraná próximo à Usina de Itaipu. Note que $\mu(t)$ deixa claro a sazonalidade natural das vazões. Já na figura 18(B) tem-se o desvio-padrão da vazão, $\sigma(t)$, em função do dia do ano t no mesmo local. Observa-se, ainda, que grandes flutuações são observadas quando a média $\mu(t)$ também é grande. Na figura 18(C), mostra-se exemplos de séries temporais normalizadas, $z(t) = [x(t) - \mu(t)]/\sigma(t)$, para as vazões no rio Paraná próximo à Usina de Itaipu em quatro anos (como indicado nos gráficos). Novamente, nota-se que a normalização praticamente elimina as tendências associadas à sazonalidade natural.

Uma questão importante a respeito de $z(t)$ está relacionada à sua distribuição

de probabilidade. Diversos trabalhos tem apontado que uma distribuição universal pode descrever a distribuição empírica de $z(t)$, independentemente das particularidades do rio (BRAMWELL; HOLDSWORTH; PORTELLI, 2002; DAHLSTEDT; JENSEN, 2005; DOMENICO; LATORA, 2011). O grande número de séries temporais no conjunto de dados pode fornecer uma resposta definitiva para esta questão. Para isso, avaliou-se a distribuição agregada de $z(t)$ para todas as séries temporais.

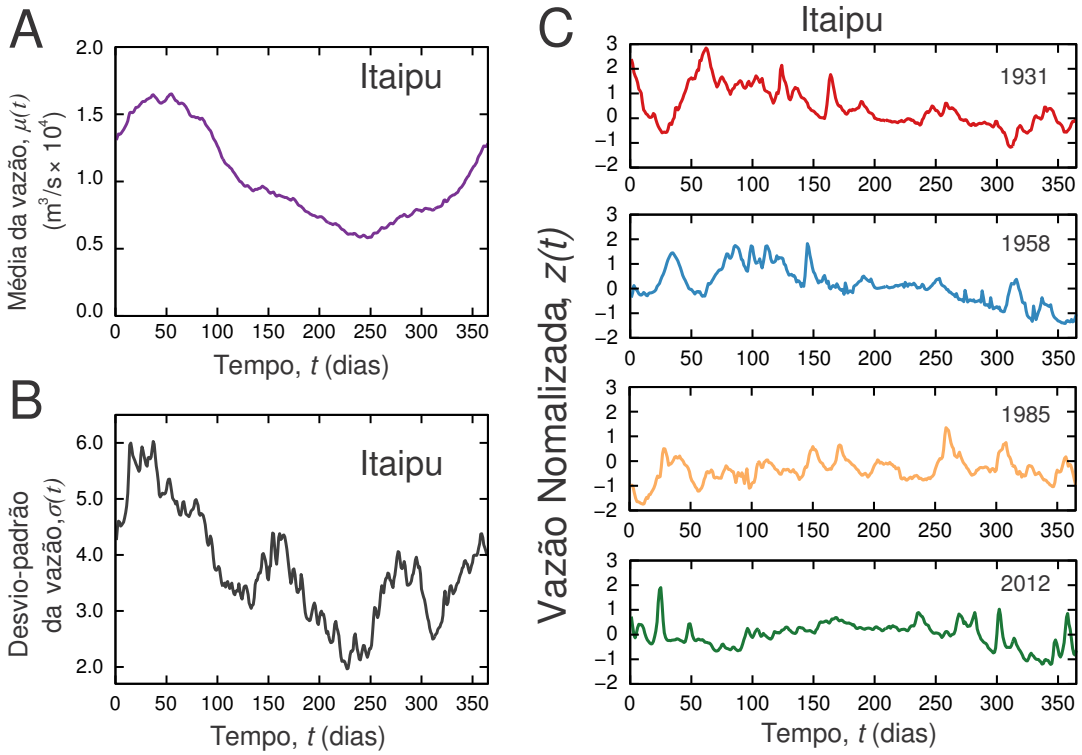

Figura 18 – Construção da série temporal normalizada.

A Figura 19 mostra essas distribuições, e um bom colapso é observado, corroborando ainda mais para a hipótese de que a distribuição empírica de $z(t)$ é universal entre diferentes rios.

Outra pergunta é se essa distribuição empírica universal pode ser descrita por alguma forma funcional. Nesse contexto, o modelo Bramwell-Holdsworth-Pinton (BHP) (BRAMWELL; HOLDSWORTH; PINTON, 1998), que descreve as flutuações magnéticas no modelo-XY clássico nas proximidades da criticalidade, bem como, a distribuição de valor extremo de primeira ordem (ou a distribuição de Gumbel (GUMBEL, 1958)),

$$f(z) = \frac{1}{\theta} e^{\frac{\lambda - z}{\theta} - e^{\frac{\lambda - z}{\theta}}}, \qquad z \in \mathbb{R} \tag{3.8}$$

em que $\lambda \in \mathbb{R}$ e $\theta > 0$ são parâmetros de ajuste, foram utilizados para ajustar estas distribuições empíricas.

Na figura 19, compara-se as distribuições empíricas com a forma de Gumbel e uma boa concordância é, de fato, observada (concordância semelhante é obtida para o

modelo BHP). A distribuição de Gumbel está relacionada ao máximo do conjunto de $n \to \infty$ números aleatórios extraídos de uma distribuição que assintoticamente decai mais rapidamente do que qualquer lei de potência. A concordância com essa distribuição sugere, assim, que as flutuações normalizadas $z(t)$ podem ser também modeladas como um processo de valor extremo.

Os pontos cinzentos na figura 19 mostram as distribuições dos valores agregados das flutuações normalizadas para cada série temporal do conjunto de dados. Nota-se um bom colapso dessas distribuições, sugerindo que as distribuições podem ter uma forma universal. Os círculos vermelhos são os valores médios ao longo de todas as distribuições e a linha tracejada é a distribuição de Gumbell (Equação (3.8)) ajustados para os valores médios através do método dos mínimos quadrados (os melhores parâmetros de ajuste são mostrados no gráfico).

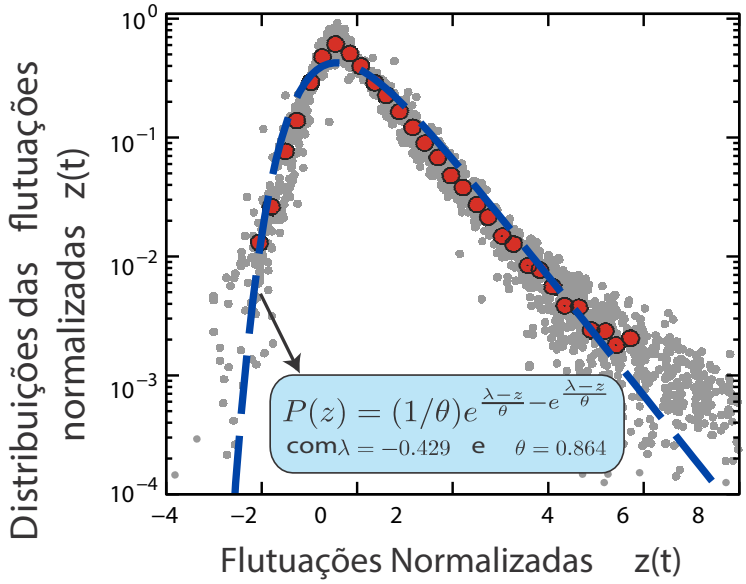

Figura 19 – Comportamento universal das distribuições de probabilidades das vazões normalizadas.

Para finalizar esse trabalho, uma última análise foi feita. Estudou-se se há ou não memória de longo alcance na série temporal das descargas dos rios. Para investigar essa hipótese, considerou-se as séries temporais normalizadas $z(t)$ agrupadas por ano e empregou-se o detrended fluctuation analysis (DFA) (PENG et al., 1994; KANTELHARDT et al., 2001). Pode-se resumir esse procedimento em quatro etapas (como apresentado anteriormente): *i*) Definimos o perfil integrado

$$Z(t) = \sum_{t'=1}^{t} z(t') \,; \tag{3.9}$$

ii) Divide-se $z(t)$ em $N_m = N/m$ partições não-superpostas de tamanho m, $Z_{\nu,m}(t)$, onde N é o comprimento da série e ν um índice para as partições; *iii*) Para cada partição, uma tendência polinomial local (aqui usa-se um polinômio de primeira ordem, mas ordens

superiores não alteram os resultados) é calculada e subtraída de $Z_{\nu,m}(t)$, definindo

$$(\Delta Z_{\nu,m})^2 = \frac{1}{m-1}\sum_{t=1}^{m}[Z_{\nu,m}(t) - p_\nu(t)]^2 \tag{3.10}$$

onde $p_\nu(t)$ representa a tendência local na ν-ésima partição; *iv*) Finalmente, a função de flutuação

$$F(m) = \left[\frac{1}{N_m}\sum_{\nu=1}^{N_m}(\Delta Z_{\nu,m})^2\right]^{1/2} \tag{3.11}$$

é calculada.

Conforme já visto, se $z(t)$ é auto-similar, a função de flutuação $F(m)$ apresenta uma dependência tipo lei de potência sobre o tempo de escala m, ou seja, $F(m) \sim m^h$, onde h é o expoente de Hurst. Para $h > 0.5$ ou $h < 0.5$ a série é correlacionada de longo alcance, enquanto que para $h = 0.5$ é do tipo não correlacionada ou apresenta correlação de curto alcance (tipo exponencial). O valor de h também deve ser de cerca de 0.5 para versões embaralhadas de $z(t)$, caso contrário, o DFA pode levar a falsas correlações associadas com uma possível natureza livre de escada da distribuição de $z(t)$ (o que não acontece no caso estudado).

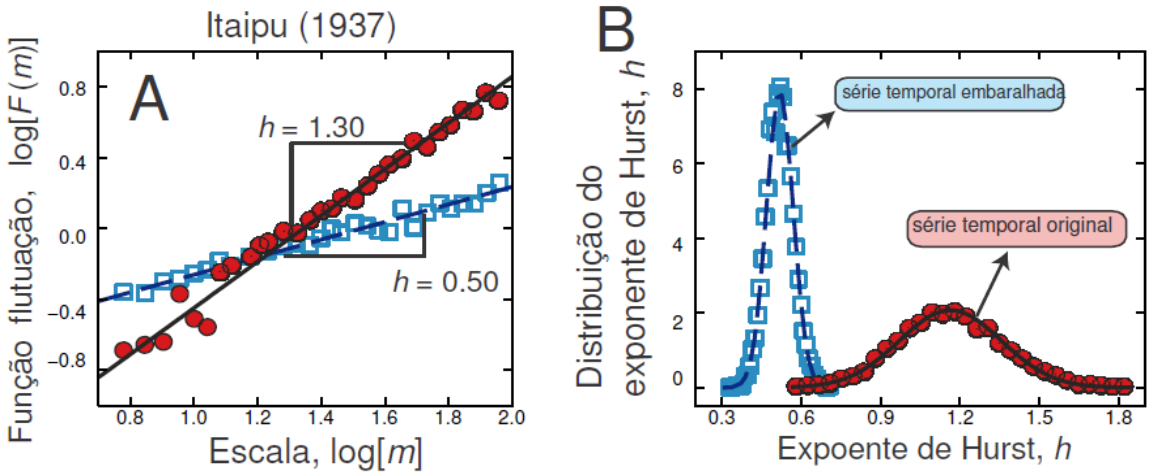

Figura 20 – Correlações de longo alcance nas vazões normalizadas.

A figura 20(A) mostra um exemplo prático de DFA para as flutuações normalizadas $z(t)$ do rio Paraná, na estação de Itaipu e no ano de 1937. Esse gráfico mostra o logaritmo de flutuação $F(m)$ em função do logaritmo de escala temporal m para a série original (círculos vermelhos) e para uma versão embaralhada aleatória (quadrados azuis). Nota-se que o gráfico log-log lineariza o relacionamento entre $F(m)$ e m, onde h é o expoente de Hurst, de modo que a inclinação dessa curva é o valor de h. Assim, estimou-se o valor de h por meio de regressão linear sobre a relação log-log e, para esse caso, $h = 1.3$ para a série original e $h = 0.5$ para versão embaralhada (como seria de esperar). Esse resultado indica a presença de correlações de longo alcance na evolução temporal de $z(t)$.

Empregando o mesmo procedimento, estimou-se h para todas as séries agrupadas por ano. Na figura 20(B), mostra-se a distribuição de probabilidades de h para a série temporal original (círculo vermelho) em comparação com uma distribuição gaussiana de média 1.16 e desvio-padrão 0.19 (linha contínua). Mostra-se ainda, a distribuição h para a versão embaralhada da série (quadrados azuis) em comparação com a distribuição Gaussiana de média 0.50 e desvio-padrão 0.05 (linha pontilhada). Confirmando que as correlações de longo alcance estão presentes em todas as séries temporais no conjunto de dados.

Assim, os resultados obtidos aqui estão em acordo com trabalhos anteriores (DOLGONOSOV; KORCHAGIN; KIRPICHNIKOVA, 2008; TESSIER et al., 1996; KANTELHARDT et al., 2001; MOVAHED; HERMANIS, 2008; ZHANG; WANG; W., 2015; HAJIAN; MOVAHED, 2010; YU et al., 2014; ZHANG; XU; YANG, 2009) sobre a existência de correlações de longo alcance na série temporal normalizada relacionadas às descargas fluviais. No entanto, os valores de h mostram flutuações importantes, que podem ser úteis para uma classificação de rios e também podem levar a outras investigações relacionando h com particularidades do rio e/ou do ano em análise.

3.2 SOBRE A DETECÇÃO DE CORRELAÇÕES EM SÉRIES TEMPORAIS: UMA COMPARAÇÃO OBJETIVA ENTRE DFA, TRANSFORMAÇÕES WAVELET E DEA.

Nos Fundamentos Teóricos apresentou-se alguns aspectos relacionados à função de correlação com ênfase em séries temporais. Discutiu-se a relação entre a existência de invariância de escala e correlações de longo alcance. Em particular, verificou-se como essa relação surge no contexto do movimento Browniano fracionário. Usando o movimento Browniano fracionário, foi-se capaz de construir séries temporais com propriedades fractais. Com essas séries, constatou-se que o cálculo da função de correlação usando diretamente a definição requer séries muito grandes. Na tentativa de contornar esse problema, revisitou-se três dos principais métodos existentes na literatura que conseguem identificar correlações de longo alcance. Nominalmente: o método DFA, as transformações Wavelet e o método DEA.

Nessa seção, apresenta-se uma comparação objetiva entre os três métodos apresentados anteriormente. Mais especificamente, investiga-se como é a convergência[1] desses métodos com relação ao tamanho (número de termos) das séries.

Para isso, gerou-se séries fractais usando o movimento Browniano fracionário com diferentes valores de h e diferentes tamanhos ($n \sim 10$ até $n \sim 10^5$). Aplicou-se os três métodos em cada uma das séries, obtendo os valores do expoente h para cada método. Uma média sobre 2000 realizações desse procedimento foi levada em conta, visando obter o

[1] No sentido de proximidade com o valor esperado.

valor médio de h fornecido por cada método, para cada tamanho de série. Esses resultados são mostrados na figura 21.

Nessa figura, observou-se que para valores pequenos de n, todos os métodos conduzem a valores médios distantes dos verdadeiros valores de h. Naturalmente, ao aumentar o tamanho das séries, todos os métodos convergem para o valor teórico de h. Contudo, pode-se ver que o DFA converge por valores superiores ao verdadeiro valor de h, enquanto as transformações Wavelet e o DEA, o fazem por valores inferiores. No que diz respeito à especificidade de cada método notou-se que:

- Para valores de $h < 0.3$ o DEA apresenta uma convergência superior aos demais. Porém, para valores de $h > 0.5$ sua convergência é consideravelmente mais lenta;
- Comparado com as transformações Wavelet, o método DFA apresenta uma convergência ligeiramente mais rápida para séries persistentes ($h > 0.5$) e ligeiramente mais lenta para séries anti-persistentes ($h < 0.5$);
- Para séries com mais de 10000 termos os três métodos produzem valores de h que praticamente coincidem com os verdadeiros valores.

A simetria (aproximada) existente entre as transformações Wavelet e DFA, em relação ao valor de h da série, sugere que os métodos podem conduzir a melhores resultados se forem aplicados simultaneamente.

Para testar essa ideia, aplicou-se DFA e Wavelet em uma mesma série e calculou-se o valor médio do expoente h obtido para cada método. Na Figura 21, mostrou-se também como se dá a convergência para o verdadeiro valor de h, utilizando essa combinação de métodos. A linha tracejada indica o valor verdadeiro de h, isto é, o valor de h usado para gerar as trajetórias do movimento Browniano fracionário. Veja que há uma considerável melhora na convergência, sendo que mesmo para séries pequenas ($n \sim 100$) o valor obtido por essa combinação é muito próximo ao valor verdadeiro de h. Embora seja um resultado preliminar, essa combinação de métodos pode ajudar a obter valores mais confiáveis do expoente de Hurst para séries temporais com poucos termos.

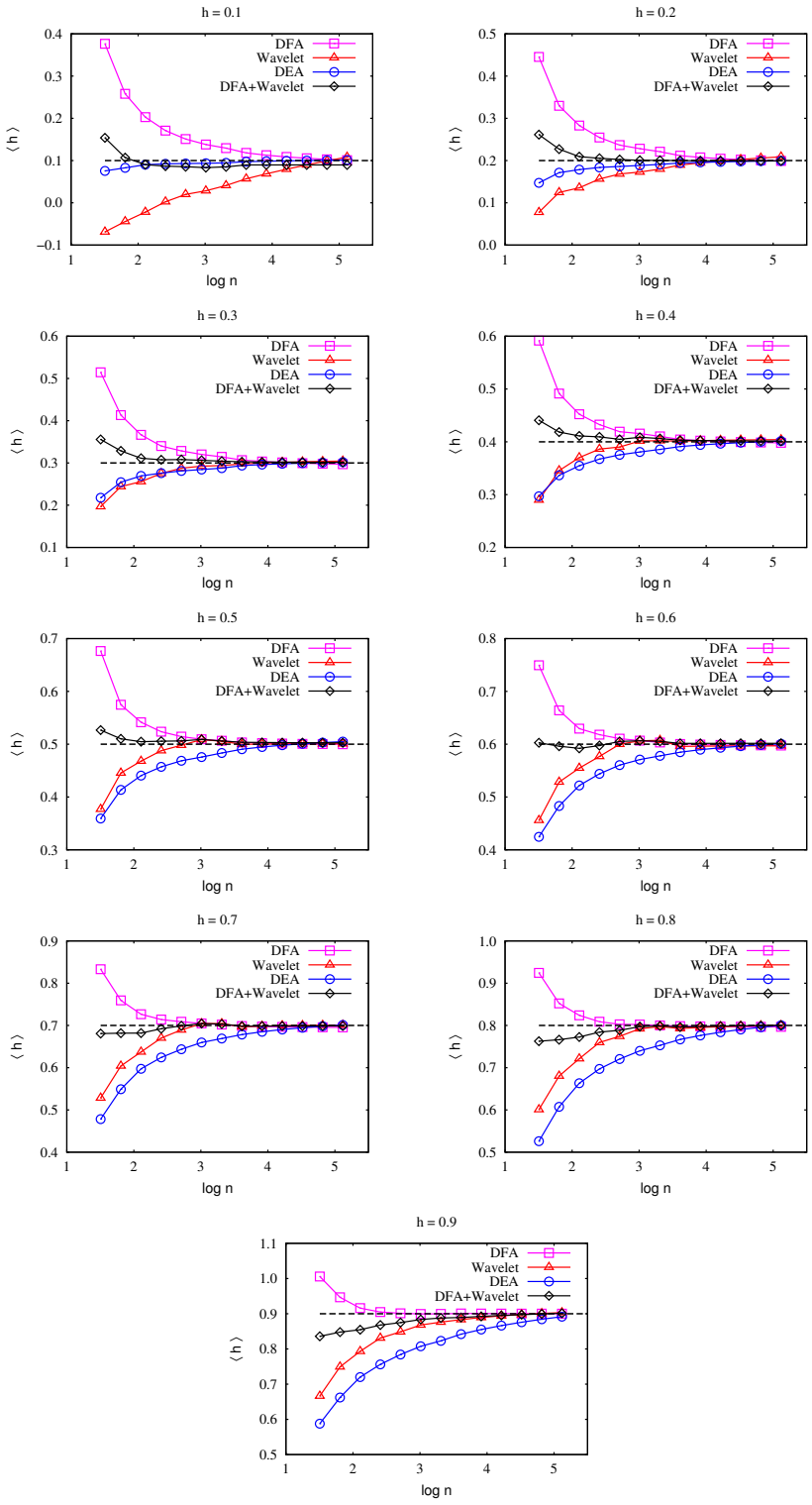

Figura 21 – Valor médio do expoente h calculado para diferentes valores de n, usando DFA, Wavelet, DEA e combinação média entre DFA e Wavelet.

4 CONCLUSÃO

Foram estudadas as descargas fluviais a partir de um conjunto de dados relativamente grande, composto de quase cento e cinquenta estações que cobrem mais de cinquenta rios brasileiros em períodos de tempo de mais de 80 anos. Esse trabalho propõe, empregar a estrutura de rede do método de visibilidade horizontal para caracterização das flutuações das vazões dos rios.

A abordagem adotada pelo método de visibilidade horizontal apontou algumas características evolutivas intrigantes das descargas fluviais e mostrou que as descargas fluviais em certas estações de medição estão se tornando mais ou menos correlacionadas, bem como exibindo estruturas de redes internas mais ou menos complexas.

Apesar de ser difícil testar esta hipótese a partir dos dados disponíveis, os resultados obtidos levam a crer que essas características evolutivas podem estar relacionadas as mudanças no sistema climático, particularmente em relação ao regime de chuvas. Outros fenômenos provocados pelo homem, como o uso em grande escala de água em atividades agrícolas podem também ter contribuído para estas características evolutivas. Talvez, outras investigações em nível mais local poderiam ajudar a elucidar os mecanismos subjacentes aos aspectos evolutivos das descargas fluviais aqui apresentados. Dessa forma, a caracterização em larga escala das flutuações das vazões de rios através de grafos de visibilidade horizontal leva a crer que esse trabalho lança uma nova luz sobre a dinâmica das vazões dos rios e abre a possibilidade para outras investigações diretas com base no método de visibilidade horizontal.

Em outro contexto, a utilização da entropia e complexidade de permutação, mostrou que é possível associar um período T à sazonalidade natural das séries temporais estudadas, por meio da qual, investigou-se a possibilidade de T apresentar tendências evolutivas, estimando seu valor dentro de diferentes intervalos de tempo da série temporal. Uma regressão linear mostrou que T não apresenta uma tendência linear de aumento ou diminuição para todas as séries temporais. O estudo das versões normalizadas da série temporal resultou na confirmação de uma distribuição universal, apesar de toda complexidade e diferenças dos rios em análise. Foi visto que a distribuição de Gumbel pode ser ajustada aos dados empíricos, o que, de alguma forma, pode conectar as flutuações normalizadas com processos de valor extremo.

Ao fim desta etapa, analisou-se a memória de longo alcance presente nessas séries temporais via DFA, por meio da qual, mostrou-se que o expoente de Hurst realmente confirma a existência de correlações de longo alcance, no entanto, seu valor exibe uma faixa de variação, dependendo do rio e/ou do ano em análise. Esse achado pode ser, eventualmente, usado para promover uma classificação dos rios e outras questões sobre possíveis relações entre h e particularidades do rio e/ou do ano em análise. As descobertas

relatadas nessa análise contribuem para os resultados citados na literatura associadas às distribuições das flutuações normalizadas (BRAMWELL; HOLDSWORTH; PORTELLI, 2002; DAHLSTEDT; JENSEN, 2005; DOMENICO; LATORA, 2011) e às correlações de longo alcance (DOLGONOSOV; KORCHAGIN; KIRPICHNIKOVA, 2008; TESSIER et al., 1996; KANTELHARDT et al., 2001; MOVAHED; HERMANIS, 2008; ZHANG; WANG; W., 2015; HAJIAN; MOVAHED, 2010; YU et al., 2014; ZHANG; XU; YANG, 2009) e também lançam novas possibilidades para investigar essas séries que podem encontrar implicações para modelagem e previsão da vazão de rios.

Assim conclui-se que o uso de elementos teóricos abordados pela Física de Sistemas Complexos, constitui uma ferramenta poderosa na caracterização de um grande número de rios contidos em diferentes bacias hidrográficas do Brasil, podendo ser extrapolada para análises em toda a gama de rios que contém a mesma característica dos sistemas analisados.

Por fim, fez-se uma comparação entre os três métodos mais comuns na investigação de correlação de longo alcance quanto a sua convergência para o verdadeiro valor do expoente de Hurst, em função do tamanho das séries geradas. Nessa comparação, observou-se algumas peculiaridades de cada método. Por exemplo, que o DFA converge por valores superiores de h, enquanto Wavelet e DEA o fazem por valores inferiores. Com base nesse achado empírico, propôs-se aplicar simultaneamente DFA e Wavelet. Isso fez com que a convergência para o valor verdadeiro de h fosse alcançada para séries razoavelmente pequenas.

4.1 SUGESTÕES PARA TRABALHOS FUTUROS

Os resultados encontrados neste trabalho foram comparados com os resultados obtidos de séries temporais das vazões naturais dos rios existentes na literatura, impulsionando novas possibilidades para investigar essas séries que podem encontrar implicações para modelagem e previsão da vazão de rios.

De uma maneira mais geral, acredita-se que esses estudos iniciais forneceram várias técnicas e métodos para análise de séries temporais. Essas ferramentas possibilitam a investigação de muitos outros sistemas e dados.

Por fim, como trabalhos futuros, pretende-se analisar dados relacionados à Usina Hidrelétrica de Itaipu, tais como dados dos sismógrafos e acelerômetros usados no monitoramento de usinas, bem como, das vazões de vertedouros.

REFERÊNCIAS

AGUIAR, S. G. *Contribuição ao Estudo de Redes Complexas: Modelo de Afinidade com Métrica*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Norte, 2012.

AHMADLOU, M.; ADELI, H.; ADELI, A. New diagnostic EEG markers of the Alzheimer's disease using visibility graph. *J. Neural. Transm.*, v. 117, p. 1099–1109, 2010.

ALBERT, R.; BARABÁSI, A.-L. Statistical mechaniscs of complex networks. *Rev. Mod. Phys.*, v. 74, p. 47, 2002.

ARNEODO, A. et al. Charactering long-range correlations in DNA sequences from wavelet analysis. *Phys. Rev. Lett.*, v. 74, p. 3293–3296, 1995.

AUYANG, S. Y. *Foundations of complex-systems*. Cambridge: Cambridge University Press, 1998.

BANDT, C.; POMPE, B. Permutation entropy: a natural complexity measure for time series. *Phys. Rev. Lett.*, v. 88, p. 174102, 2002.

BIGACHEV, M. I.; BUND, E. A. Universality in the precipitation and river runoff. *EPL*, v. 97, p. 48011, 2012.

BOCCARA, N. *Modeling complex systems*. New York: Springer-Verlag, 2004.

BOETTLE, M.; RYBSKI, D.; P., K. J. How changing sea level extremes and protection measures alter coastal flood damages. *Water Resources Research*, v. 49, p. 1199–1210, 2013.

BORDIGNON, S.; LISI, F. Nonlinear analysis and prediction of river flow time series. *Environmetrics*, v. 11, p. 463–477, 2000.

BRAGA, A. C. et al. Characterization of river flow fluctuations via horizontal visibility graphs. *Physica A*, v. 444, p. 1003–1011, 2016.

BRAMWELL, S.; HOLDSWORTH, P.; PINTON, J. F. Universality of rare fluctuations in turbulence and critical phenomena. *Nature*, v. 396, p. 552–554, 1998.

BRAMWELL, S. T.; HOLDSWORTH, P. C. W.; PORTELLI, B. Universal fluctuations of the danube water level: A link with turbulence, criticality and company growth. *EPL*, v. 57, p. 310–314, 2002.

CHEN, Z. et al. Effect of nonstationarities on detrended fluctuation analysis. *Phys. Rev. E*, v. 65, p. 041107, 2002.

DAHLSTEDT, K.; JENSEN, H. J. Fluctuation spectrum and size scaling of river flow and level. *Physica A*, v. 348, p. 596–610, 2005.

DIEKER, A. B.; MANDJES, M. On spectral simulation of fractional Brownian motion. *Probability in the Engineering and Informational Sciences*, v. 17, p. 417–434, 2003.

DOLGONOSOV, B. M.; KORCHAGIN, K. A.; KIRPICHNIKOVA, N. V. Modeling of annual oscillations and 1/f - noise of daily river discharges. *Journal of Hydrology*, v. 357, p. 174–187, 2008.

DOMENICO, M. D.; LATORA, V. Scaling and universality in river flow dynamics. *EPL*, v. 94, p. 5802, 2011.

DOVE, M. R.; KAMMEN, D. M. *Science, society and the environment: applying anthropology and physics to sustainability*. New York: Routledge, 2015.

EFRON, B.; TIBSHIRANI, R. *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.

ELSNER, J. B.; JAGGER, T. H.; FOGARTY, E. A. Visibility network of United States hurricanes. *Geophys. Res. Lett.*, v. 36, p. L16702, 2009.

FEDERS, J. *Fractals*. New York: Plenum Publishers, 1988.

GABARDO, A. C. *Análises de REDES SOCIAIS: Uma Visão Computacional*. São Paulo: Novatec, 2015.

GAO, Z. K.; JIN, N. D. Characterization of chaotic dynamic behavior in the gas-liquid slug flow using directed weighted complex network analysis. *Physica A*, v. 391, p. 3005–3016, 2012.

GRIGOLINI, P.; PALATELLA, L.; RAFFAELLI, G. Asymmetric anomalous diffusion: An efficient way to detect memory in time series. *Fractals*, v. 9, p. 439–449, 2001.

GROSSE, I. et al. Analysis of symbolic sequences using the Jensen-Shannon divergence. *Phys. Rev. E*, v. 65, p. 041905, 2002.

GUMBEL, E. J. *Statisticw of Extremes*. New York: Columbia University Press, 1958.

HAJIAN, S.; MOVAHED, M. S. Multifractal detrended cross-correlation analysis of sunspot numbers and river flow fluctuations. *Physica A*, v. 389, p. 4942–4957, 2010.

HAKEN, H. *Information and self-organization*. Berlin: Springer, 2006.

HOGG, R. V.; CRAIG, A. *Introduction to Mathematical Statistics*. New York: Prentice Hall, 1995.

HU, K. et al. Effect of trends on detrended fluctuation analysis. *Phys. Rev. E*, v. 64, p. 011114, 2001.

HURST, H. E. Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, v. 116, p. 770–799, 1951.

JÁNOSI, I. M.; GALLAS, J. A. C. Growth of companies and water-level fluctuations of the river danube. *Physica A*, v. 271, p. 448–457, 1999.

JENSEN, H. J. *Self-organized criticality*. Cambridge: Cambridge University Press, 1998.

JIANG, S. et al. Visibility graph analysis on heartbeat dynamics of meditation training. *Appl. Phys. Lett.*, v. 102, p. 253702, 2013.

KANTELHARDT, J. W. et al. Detecting long-range correlations with detrended fluctuation analysis. *Physica A*, v. 295, p. 441, 2001.

KANTELHARDT, J. W. et al. Multifractality of river runo and precipitation: comparison of fluctuation analysis and wavelet methods. *Physica A*, v. 330, p. 240–245, 2003.

KOUTSOYIANNIS, D. The Hurst phenomenon and fractional Gaussian noise made easy. *Hydrological Sciences Journal*, v. 47, p. 573–595, 2002.

KULP, C. W.; ZUNINO, L. Discriminating chaotic and stochastic dynamics through the permutation spectrum test. *Chaos*, v. 24, p. 033116, 2014.

LACASA, L. et al. From time series to complex networks: The visibility graph. *Proc. Natl. Acad. Sci. USA*, v. 105, p. 4972–4975, 2008.

LACASA, L. et al. The visibility graph: A new method for estimating the hurst exponent of fractional Brownian motion. *Europhys. Lett. EPL*, v. 86, p. 30001, 2009.

LACASA, L.; TORAL, R. Description of stochastic and chaotic series using visibility graphs. *Phys. Rev. E*, v. 82, p. 046103, 2010.

LIN, J. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theory*, v. 37, p. 145–151, 1991.

LIVINA, V. et al. A stochastic model of river discharge fluctuations. *Physica A*, v. 330, p. 283–290, 2003.

LÓPEZ-RUIZ, R. et al. A statistical measure of complexity. *Phys. Lett. A*, v. 209, p. 321, 1995.

LUQUE, B. et al. Horizontal visibility graphs: Exact results for random time series. *Phys. Rev. E*, v. 80, p. 046103, 2009.

MACHIWAL, D.; JHA, M. K. *Hydrologic time series analysis: theory and practice*. New Delhi: Springer, 2012.

MANDELBROT, B. B.; NESS, J. W. V. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, v. 10, p. 422–437, 1968.

MANIMARAN, P.; PANIGRAHI, P. K.; PARIKH, J. C. Wavelet analysis and scaling properties of time series. *Phys. Rev. E*, v. 72, p. 046120, 2005.

MANTEGNA, R. N.; STANLEY, H. E. *An introduction to ecnophysics*. Cambridge: Cambridge University Press, 1999.

MARTIN, M. T.; PLASTINO, A.; ROSSO, O. A. Generalized statistical complexity measures: Geometrical and analytical properties. *Physica A*, v. 369, p. 439, 2006.

MENDES, R. S. et al. Earthquake-like patterns of acoustic emission in crumpled plastic sheets. *EPL*, v. 92, p. 29001, 2010.

MENDES, R. S. et al. Universal patterns in sound amplitudes of songs and music genres. *Phys. Rev. E*, v. 83, p. 017101, 2011.

METZ, J. et al. *Redes Complexas: conceitos e aplicações*. São Carlos: Relatórios Técnicos do ICMC, 2007.

MIHAILOVIC, D. T. et al. Complexity analysis of the turbulent environmental fluid flow time series. *Physica A*, v. 395, p. 96–104, 2014.

MOVAHED, M. S.; HERMANIS, E. Fractal analysis of river flow fluctuations. *Physica A*, v. 387, p. 915–932, 2008.

MURKS, A.; PERC, M. Evolutionary games on visibility graphs. *Adv. Complex Syst.*, v. 14, p. 307–315, 2011.

MUZY, J. F.; BACRY, E.; ARNEODO, A. Wavelets and multifractal formalism for singular signals: application to turbulence data. *Phys. Rev. Lett.*, v. 67, p. 3515–3518, 1991.

NEWMAN, M. E. J. The structure and function of complex networks. *SIAM Rev.*, v. 45, p. 167–256, 2003.

NEWMAN, M. E. J.; STROGATZ, S. H.; WATTS, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, v. 64, p. 026118, 2011.

OLIVER, H. R.; OLIVER, S. A. *The role of water and the hydrological cycle in global change*. [S.l.]: Springer, 1995.

PENG, C. K. et al. Mosaic organization of DNA nucleotides. *Phys. Rev. E*, v. 49, p. 1685, 1994.

PICOLI JR., S. et al. Spreading patterns of the influenza a (H1N1) pandemic. *Plos One*, v. 6, p. e17823, 2011.

PORPORATO, A.; RIDOLFI, L. Nonlinear analysis of river flow time sequences. *Water Resources Research*, v. 33, p. 1353–1367, 1997.

RABASSA, P.; BECK, C. Superstatistical analysis of sea-level fluctuations. *Physica A*, v. 417, p. 18–28, 2015.

RIBEIRO, H. V. *Identificação e Modelagem de Padrões em Sistemas Complexos*. Tese (Doutorado) — Universidade Estadual de Maringá, 2012.

RIBEIRO, H. V. et al. Long-range spatial correlations and fluctuation statistics of lightning activity rates in brazil. *EPL*, v. 104, p. 69001, 2013.

RIBEIRO, H. V. et al. Analogies between the cracking noise of ethanol-dampened charcoal and earthquakes. *Phys. Rev. Lett.*, v. 115, p. 025503, 2015.

RIBEIRO, H. V. et al. Symbolic sequences and Tsallis Entropy. *Braz. J. Phys*, v. 39, p. 444, 2009.

RIBEIRO, H. V. et al. Anomalous diffusion in a symbolic model. *Physica Scripta*, v. 83, p. 045007, 2011.

RIBEIRO, H. V. et al. On the dynamics of bubbles in boiling water. *Chaos Solitons and Fractals*, v. 44, p. 178, 2011.

RIBEIRO, H. V. et al. Dynamics of tournaments: the soccer case. *Eur. Phys. J. B.*, v. 75, p. 327, 2010.

RIBEIRO, H. V. et al. The soundscape dynamics of human agglomeration. *New J. Phys.*, v. 13, p. 023028, 2011.

ROSSO, O. A. et al. Distinguishing noise from chaos. *Phys. Rev. Lett.*, v. 99, p. 154102, 2007.

RYBSKI, D.; HOLSTEN, A.; KROPP, J. P. . towards a unified characterization of phenological phases: fluctuations an correlations with temperature. *Physica A*, v. 390, p. 680–688, 2011.

SCAFETTA, N.; GRIGOLINI, P. Scaling detection in time series: Diffusion entropy analysis. *Phys. Rev. E*, v. 66, p. 036130, 2002.

SCAFETTA, N.; HAMILTON, P.; GRIGOLINI, P. The thermodynamics of social processes: The teen birth phenomenon. *Fractals*, v. 9, p. 193–208, 2001.

SHANNON, C. E. A mathematical theory of communication. *Bell. Syst. Tech. J.*, v. 27, p. 623–656, 1948.

SILVA, A. I. *Atividade Psicomotora, Epidemias e Lideranças como Sistemas Complexos*. Tese (Doutorado) — Universidade Estadual de Maringá, 2015.

SIMONSEN, I.; HANSEN, A. Determination of the Hurst exponent by use of wavelet transforms. *Phys. Rev. E*, v. 58, p. 2779, 1998.

SORNETTE, D. *Critical Phenomena in Natura Scienses*. Heidelberg: Springer, 2006.

TAQQU, M. S.; TEVEROVSKY, V.; WILLINGER, W. Estimators for long-range dependence: an empirical study. *Fractals*, v. 3, p. 785–798, 1995.

TELESCA, L.; LOVALLO; TOTH M., L. Visibility graph analysis of 2002-2011 Pannonian seismicity. *Physica A*, v. 416, p. 219–224, 2014.

TELESCA, L.; LOVALLO, M. Analysis of seismic sequences by using the method of visibility graph. *Europhys. Lett. EPL*, v. 97, p. 50002, 2012.

TESSIER, Y. et al. Multifractal analysis and modeling of rainfall and river flows and scaling, causal transfer functions. *Journal of Geophysical Research*, v. 101, p. 26,427–26,440, 1996.

TORRENCE, C.; COMPO, G. P. A practical guide to wavelet analysis. *Bulletin of American Meteorological Society*, v. 79, p. 61–78, 1997.

TURNER, D. L. et al. Explaining sudden losses of outer radiation belt electrons during geomagnetic storms. *Nature Physics*, v. 8, p. 208–212, 2012.

VJUSHIN, D. et al. Scaling analysis of trends using DFA. *Physica A*, v. 302, p. 234, 2001.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of small-word networks. *Nature*, v. 393, p. 440–442, 1998.

WMO. *World Meteorological Organization*. 2015. <http://www.wmo.int/pages)themes/climate/understanding_climate.php>. Acessado em Julho de 2015.

YANG, Y. et al. Visibility graph approach to exchange rate series. *Physica A*, v. 388, p. 4431–4437, 2009.

YU, Z. G. et al. Multifractal analyses of daily rainfall time series in pearl river basin of china. *Physica A*, v. 405, p. 193–202, 2014.

ZHANG, B.; WANG, J.; W., F. Volatility behavior of visibility graph EMD financial time series from ising interacting system. *Physica A*, v. 432, p. 301–314, 2015.

ZHANG, Q.; XU, C. Y.; YANG, T. Scaling properties of the runoff variations in the arid and semi-arid regions of china: a case study of the yellow river basin. *Stoch Environ Res Risk Assess*, v. 23, p. 1103–1111, 2009.

ZHUANG, E.; SMALL, M.; FENG, G. Time series analysis of the developed financial markets' integration using visibility graphs. *2014*, v. 410, p. 483–495, 2014.

ZOU Y., D. R. V. M. N. S. M.; KURTHS, J. Long-term changes in the north-south asymmetry of solar activity: a nonlinear dynamics characterization using visibility graphs. *Nonlinear Processes Geophys*, v. 21, p. 1113–1126, 2014.

ZUNINO, L. et al. Permutaion - information - theory approach to unveil delay dynamics from time series analysis. *Phys. Rev. E*, v. 82, p. 046212, 2010.

APÊNDICES

APÊNDICE A – BOOTSTRAP

O *bootstrap* é uma abordagem alternativa para encontrar intervalos de confiança e quantidades similares diretamente dos dados. Em vez de assumir informações sobre a distribuição de valores e, então, empregar argumentos teóricos, o *bootstrap*[1] se volta à ideia original: e se pudesse ter amostras adicionais da população estudada? Esse método provém uma simples forma para obter o intervalo de confiança mesmo em situações em que resultados teóricos não são disponíveis.

Podemos criar amostras adicionais por amostragem em substituição à série original. Para cada uma dessas amostras "sintéticas", podemos calcular a média, além de outras quantidades, e usar esse conjunto de valores para determinar uma medida do espalhamento dessa distribuição por meio de qualquer método padrão. O *bootstrap* não é um método para obter a melhor estimativa da própria quantidade original, para isso, é necessário obter uma grande amostra fazendo amostras adicionais da população original.

Para que o *bootstrap* funcione em um determinado conjunto de dados, é necessário que duas condições sejam satisfeitas:

1. A amostra original precisa prover uma boa representação de toda população;
2. A quantidade estimada precisa depender "suavemente" dos dados.

A primeira condição requer que a amostra original seja suficientemente grande e relativamente limpa, ou seja, não apresentar grandes oscilações em seus valores ou outras anomalias. Se a amostra for muito pequena, então, a série original estimada pela quantidade em questão (a média, por exemplo), não será muito boa. O *bootstrap*, de certa forma, piora esse problema, porque os dados têm grandes chances de serem usados repetidamente nas amostras "sintéticas". A amostra precisa ser relativamente limpa: resultados imensamente discrepantes, por exemplo, podem ser um problema. A menos que o tamanho da amostra seja muito grande, os resultados tem uma chance significativa de serem reutilizados na amostra do *bootstrap*, distorcendo os resultados.

A segunda condição sugere que o método não funciona muito bem para quantidades que dependem criticamente de apenas uns poucos pontos dos dados. Por exemplo, podemos desejar estimar o valor máximo de alguma distribuição, mas tal estimativa depende criticamente do maior valor observado, que é um único ponto. Para esse tipo de aplicação, o *bootstrap* não é apropriado.

O número de amostras a serem consideradas vai depender do tamanho da amostra original. Se os pontos na amostra original são poucos, então, criar muitas amostras do

[1] A palavra *bootstrap*, que em uma tradução livre significa "alça de botina", vem da frase "*To lift himself up by his bootstraps*" (erguer a si mesmo pelas alças da botina). Isso se refere a algo que é absurdo e impossível. Por mais que tente, uma pessoa não pode se erguer no ar puxando partes de sua botina. Uma referência ao método de "criar" novas amostras adicionais partindo de uma série original.

bootstrap irá gerar a mesma amostra "sintética" várias vezes. Entretanto, se a amostra original for razoavelmente grande, isso não será um problema, uma vez que o número de amostras do *bootstrap* cresce muito rápido com o número de dados da amostra original, de modo que será altamente improvável que a mesma amostra seja gerada mais que uma vez, mesmo ao gerarmos milhares de amostras.