Uma abordagem híbrida para a identificação e modelagem de componentes estocásticos e determinísticos presentes em séries temporais

Ricardo Araújo Rios

Uma abordagem híbrida para a identificação e modelagem de componentes estocásticos e determinísticos presentes em séries temporais

Ricardo Araújo Rios

Orientador: Prof. Dr. Rodrigo Fernandes de Mello

Monografia apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, para o Exame de Qualificação, como parte dos requisitos para obtenção do título de Doutor em Ciências - Ciências de Computação e Matemática Computacional.

USP – São Carlos
Setembro de 2010

Resumo

O estudo do comportamento de sistemas tem sido realizado por diversos domínios da ciência, os quais visam conhecer suas características intrínsecas a partir do monitoramento de informações. Essas abordagens de monitoramento, em geral, adicionam interferências no sistema em estudo. Uma maneira de evitar essas interferências se dá por intermédio da organização e estudo de saídas, produzidas por sistemas, utilizando séries temporais. Há abordagens de modelagem de séries temporais que permitem compreender o comportamento de sistemas. Dentre essas abordagens estão as estatísticas, que representam componentes estocásticos, e aquelas baseadas em sistemas dinâmicos e teoria do caos, as quais caracterizam componentes determinísticos. No entanto, há situações em que séries temporais apresentam uma mistura de ambos componentes. Nessa situação, o descarte de qualquer um deles pode depreciar a acurácia do modelo obtido. Diante disso, este projeto visa identificar segmentos determinísticos e estocásticos de séries temporais, a fim de segmentá-los e representá-los por meio de um modelo híbrido que ofereça maior acurácia. Estudos iniciais foram conduzidos a fim de comprovar os benefícios de se modelar, individualmente, componentes estocásticos e determinísticos de séries temporais. Em seguida, outros experimentos foram realizados com o objetivo de estudar a segmentação de séries segundo seus componentes. Resultados fornecem indícios de que componentes estocásticos e determinísticos podem ser analisados de maneira separada e, em seguida, modelados, a fim de descrever, com maior acurácia, o comportamento de séries.

Este texto foi redigido segundo o novo acordo ortográfico da língua Portuguesa.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ACF	*Autocorrelation Function*
ACVF	*Autocovariance Function*
AM	Aprendizado de Máquina
AMI	*Auto Mutual Information*
ANN	*Artificial Neural Networks*
CWT	*Complex Wavelet Transform*
DWT	*Discret Wavelet Transform*
FNN	*False Nearest Neighbors*
IBL	*Instance-Based Learning*
LOI	*Line of Identity*
MSE	*Mean Squared Error*
RBF	*Radial Basis Function*
RP	*Recurrence Plot*
RQA	*Recurrence Quantification Analysis*

CAPÍTULO

1

Introdução

Um sistema é definido por um conjunto de entidades, chamadas de componentes ou subsistemas, mutuamente interconectadas (Wangler e Backlund, 2005; Zampa e Arnost, 2002). Essas interconexões definem os diferentes graus de relações e dependências entre os componentes que formam um determinado sistema. O termo sistema é utilizado em diversos domínios tais como Economia, Computação, Astronomia, Mecânica e Biologia. Em geral, essas áreas buscam estudar sistemas a fim de compreender suas operações, o comportamento de seus subsistemas, suas relações com outros sistemas e seus resultados produzidos (Wangler e Backlund, 2005; Zampa e Arnost, 2002).

Segundo Wangler e Backlund (2005), o estudo de sistemas pode ser realizado de acordo com três etapas bem definidas. Na primeira etapa, o sistema é segmentado em diferentes partes ou componentes. Na segunda, modelos são usados para explicar, separadamente, comportamentos e propriedades de cada componente. Por fim, na última etapa, é realizada uma agregação de todos os modelos com o objetivo de descrever o comportamento global do sistema.

No entanto, de acordo com Ackoff (1981), essa abordagem nem sempre fornece bons resultados, pois os modelos agregados na última etapa, quando ajustados com alta precisão

aos componentes, podem não descrever, de forma adequada, o comportamento do sistema analisado. Exemplos de dificuldades na compreensão desses comportamentos podem ser encontrados na área de Sistemas Emergentes (Holland, 2000). Segundo essa área, elementos que apresentam interações simples podem, em conjunto, constituir um comportamento além da soma de suas partes (Holland, 2000).

Além da complexidade em modelar o comportamento completo de um sistema por meio de seus componentes individuais, há ainda a dificuldade em monitorá-los sem causar interferência na sua execução (Morrison, 1990). Diante desses problemas encontrados, pesquisadores tipicamente realizam modelagens do comportamento de sistemas analisando seus resultados produzidos sem quaisquer interferências de monitoramento.

A análise dos resultados produzidos por um sistema, comumente denominados dados experimentais, tem atraído a atenção de diversos pesquisadores, os quais buscam desenvolver métodos e técnicas que permitam modelar e compreender o comportamento de sistemas. Conhecendo-se esse comportamento, espera-se tomar decisões mais precisas (ex: decisões gerenciais em empresas), simular situações futuras (ex: drogas interagindo em um organismo vivo), predizer operações (ex: mercado de ações), estimar estados de um sistema (ex: estudos de clima) e suas interações com outros sistemas (ex: efeitos climáticos na agricultura e na produção), detectar a ocorrência de falhas (ex: problemas em linha de produção mecânica) e demais eventos que venham a influenciar ou alterar seu funcionamento.

Um outro fator importante no estudo de sistemas reais é que os dados analisados tendem a apresentar dependência temporal, ou seja, o comportamento de um sistema em dado instante está relacionado com seus resultados em instantes anteriores. Essa relação de dependência é observada, por exemplo, nas temperaturas de determinada região, as quais, apesar de variarem, apresentam relações com períodos anteriores.

A presença de dependência temporal motivou pesquisadores a organizarem os resultados produzidos por sistemas reais na forma de sequências ordenadas de observações, também denominadas séries temporais. Dessa maneira, abordagens são aplicadas a fim de modelar e, consequentemente, estudar séries temporais, permitindo compreender o comportamento de sistemas reais. Diversas áreas da ciência têm buscado alternativas para estudar séries, essas frentes deram origem à área de Análise de Séries Temporais

(Box *et al.*, 1994; Shumway e Stoffer, 2006).

Essa área surgiu da necessidade de se analisar as dependências existentes entre observações de séries temporais, as quais não são consideradas por demais áreas tais como a Estatística Descritiva e a Inferência Estatística (Shumway e Stoffer, 2006). De maneira geral, além da dependência temporal entre observações, o estudo de séries temporais permite, ainda, analisar comportamentos implícitos nos dados.

As contribuições oriundas da área de análise de séries temporais, bem como estudos prévios sobre a dinâmica comportamental de sistemas realizados pelo grupo de pesquisa no qual este trabalho está inserido, motivaram a elaboração deste plano de pesquisa. A seguir, esse contexto motivacional é apresentado.

1.1 Contextualização e motivação

O grupo de pesquisa Biocom (Computação Bioinspirada) do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo tem uma frente de trabalho voltada para a área de Computação Autônoma (Parashar e Hariri, 2006). Essa área visa estudar e projetar sistemas capazes de se autogerenciar. O principal componente de um sistema autônomo é denominado gerente autônomo, o qual é responsável por extrair dados do componente monitorado (software ou hardware), analisá-los, planejar e executar decisões a fim de aumentar o desempenho final desse sistema, tolerar falhas, detectar intrusões, etc.

Nessa linha de pesquisa, os primeiros trabalhos desenvolvidos consideraram conhecimentos sobre o comportamento de aplicações a fim de melhorar a tomada de decisões em ambientes distribuídos (Senger *et al.*, 2007). Esses trabalhos permitiram concluir que, ao modelar o comportamento de uma aplicação, pode-se aumentar a acurácia de decisões envolvidas em atividades tais como o escalonamento de tarefas distribuídas (Mello *et al.*, 2007b, 2009; Mello, 2009; Dodonov e Mello, 2007; Ishii *et al.*, 2007), otimizar modelos de comunicação entre tarefas (Dodonov *et al.*, 2005, 2006) e acessos a dados distribuídos (Mello *et al.*, 2007a; Ishii e Mello, 2009).

Durante o desenvolvimento dessas pesquisas, o grupo buscou aprofundar-se na análise da dependência temporal existente no comportamento de aplicações (Mello, 2009; Ishii, 2010; Dodonov, 2009; Albertini e Mello, 2007). Nesses estudos, o grupo empregou e ava-

liou ferramentas de Teoria do Caos sobre dados experimentais a fim de analisar e modelar a dependência temporal entre observações. No entanto, concluiu-se que apesar do emprego de Teoria do Caos gerar bons modelos para estudar o comportamento de aplicações, observou-se, em dados experimentais, grande influência de componentes estocásticos. Esses componentes dificultavam e, até mesmo, impossibilitavam a modelagem e a compreensão do comportamento de aplicações (Mello *et al.*, 2008). Isso se deve ao fato de que técnicas de modelagem de séries temporais foram desenvolvidas para tratar aspectos específicos presentes em séries. Assim, uma abordagem baseada em Teoria do Caos tende a considerar apenas componentes determinísticos, enquanto demais abordagens, tais como as estatísticas, apresentam maior foco na estocasticidade das observações (Alligood *et al.*, 1997; Box *et al.*, 1994).

Essa limitação presente em técnicas atuais motivou este plano de pesquisa de doutorado, o qual visa projetar e avaliar uma abordagem para modelar, de forma híbrida, o comportamento de séries temporais, considerando tanto seus componentes estocásticos quanto determinísticos.

1.2 Hipótese e Objetivo

Séries temporais tendem a apresentar componentes determinísticos e estocásticos, conforme pode ser observado nos experimentos realizados por (Marwan *et al.*, 2007). Quando uma série é totalmente determinística, indica-se modelos baseados em Sistemas Dinâmicos, mais precisamente aqueles provenientes da área de Teoria do Caos (Mello, 2009). Em contrapartida, quando uma série apresenta comportamento estocástico, modelos estatísticos demonstram-se mais adequados (Box *et al.*, 1994). No entanto, séries temporais geradas a partir de sistemas reais tendem a apresentar uma mistura de ambos componentes. Nessa situação, o descarte de qualquer um deles pode depreciar a qualidade do modelo obtido.

Esse problema foi apresentado em (Mello, 2009), no qual o comportamento reconstruído de séries temporais, mesmo após o emprego de técnicas baseadas em Sistemas Dinâmicos e Teoria do Caos, apresentou componentes estocásticos, que prejudicaram a modelagem proveniente da abordagem determinística proposta. Nesse caso, a estocasticidade presente

nessas séries prejudicou tanto a modelagem quanto estimações e predições decorrentes, devido a perturbações adicionadas ao modelo.

Essa conclusão motivou este trabalho a investigar mecanismos que permitam considerar tanto componentes determinísticos quanto estocásticos e propôr uma abordagem híbrida para modelar séries temporais. Pretende-se, com essa abordagem, obter maior acurácia no modelo gerado, visto que nenhuma observação da série é desconsiderada. Dessa maneira, a hipótese que rege este trabalho é:

> *Séries temporais podem ser segmentadas, segundo seus componentes estocásticos e determinísticos, a fim de que todas suas observações sejam modeladas por meio de uma abordagem híbrida que ofereça maior acurácia.*

Para tanto, faz-se necessário identificar segmentos determinísticos e estocásticos de dada série temporal, a fim de separá-los com o objetivo de obter modelos híbridos. Existem diversas técnicas bem conhecidas na literatura que visam separar e quantificar cada um desses segmentos. Dentre elas, pode-se citar as transformadas de Fourier e *Wavelet*, bem como abordagens baseadas em Sistemas Dinâmicos, tais como a *Recurrence Plot* (Apêndice A).

A transformada de Fourier produz uma função com informações espectrais sobre o comportamento de um sistema na forma de ondas. No contexto deste plano, pode-se empregá-la com o objetivo de distinguir os comportamentos determinísticos e estocásticos, por meio do estudo de amplitudes e fases das ondas resultantes.

Além da transformada de Fourier, transformadas *Wavelet*[1] também podem ser utilizadas para descrever séries temporais (Nason, 2008; Tang, 2009). Essas transformadas, em geral, resultam em duas sequências de diferentes escalas de onda, a primeira denominada aproximação e a segunda, detalhe. A sequência de detalhe representa o grau de diferença ou variação da série em estudo, a qual auxilia na descrição de componentes estocásticos.

Por fim, com relação à quantificação de segmentos estocásticos e determinísticos, pode-se empregar abordagens para a análise de recorrência em séries temporais. A área de Sistemas Dinâmicos conta com uma ferramenta capaz de prover tal análise, denominada *Recurrence Plot* (RP) (Marwan *et al.*, 2007). RP agrupa observações de uma série e, com base na frequência de pontos recorrentes representados em uma matriz bidimensional,

[1] O termo *Wavelet* tem origem na literatura de geofísica e significa "onda pequena" (Nason, 2008).

quantifica níveis de determinismo.

Essa etapa de segmentação impacta, fortemente, na modelagem de cada componente individual. Ao representar, separadamente, as influências de segmentos determinísticos e estocásticos de séries temporais, pretende-se evitar desvios no modelo obtido, usualmente observados sob o emprego de ferramentas específicas para cada componente. Esse fato motiva, no contexto deste plano, estudos e avaliações das abordagens de segmentação de componentes de séries temporais apresentadas acima, bem como de outras presentes na literatura. Após esse estudo, deve-se projetar uma abordagem híbrida de modelagem, a qual deve gerar um modelo de maior acurácia.

Este trabalho de doutorado está organizado da seguinte forma: No Capítulo 2 são apresentados conceitos sobre séries temporais e suas principais ferramentas de análise e modelagem. No Capítulo 3, são apresentados detalhes sobre o plano de pesquisa de doutorado e o cronograma proposto. O Capítulo 4 apresenta estudos iniciais realizados, com o objetivo de confirmar a necessidade da abordagem proposta e entender um dos mecanismos de segmentação de séries apresentados. Os apêndices A e B apresentam, respectivamente, conceitos sobre ferramentas de análise de recorrência e de processamento de sinais. Por fim, são apresentadas as referências utilizadas no desenvolvimento deste trabalho.

CAPÍTULO

2

Análise de Séries Temporais

2.1 Considerações iniciais

Séries ãçaõ temporais organizam uma sequência de observações ocorridas no tempo, permitindo modelar e analisar o comportamento de sistemas (Box *et al.*, 1994). A importância do emprego de séries temporais na análise de comportamentos é observada em diversas áreas de estudo como, por exemplo, Economia, Ciência da Computação, Biologia, Medicina e Clima (Morettin e Toloi, 2004; Shumway e Stoffer, 2006; Chatfield, 2004).

No caso da área econômica, pode-se citar o emprego de séries temporais na análise do comportamento do mercado de ações, o qual busca modelar flutuações a fim de predizer preços de produtos e valores de ações com o intuito de evitar, por exemplo, quebras na bolsa de valores[1] (Guhathakurta *et al.*, 2010; LeBaron *et al.*, 1999; Choi *et al.*, 1999).

Além do uso de séries temporais na modelagem de sistemas econômicos, pode-se citar, também, seu uso na otimização de processos computacionais. Mello (2009) apresentou uma abordagem orientada a sistemas dinâmicos para modelar, por meio de séries temporais, o comportamento de aplicações em execução. Ao conhecer o comportamento de

[1] Do termo original: Stock Market Crashes

tarefas de uma aplicação, pode-se obter seu modelo, ou regra geradora, a qual suporta, com boa acurácia, a predição de eventos. Essa predição é utilizada para melhorar resultados de otimização de escalonamento de tarefas em ambientes distribuídos.

Um outro exemplo do emprego de séries temporais sobre sistemas computacionais é encontrado no trabalho desenvolvido por Wu *et al.* (2010). Nesse trabalho, os autores apresentam uma técnica de separação de sinais de áudio, representadas como séries temporais, a fim de destacar um determinado sinal a partir de uma combinação de múltiplos sinais.

Em áreas Biológicas e Médicas, séries temporais são amplamente utilizadas para analisar, por exemplo, sintomas de doenças e a evolução de determinados tratamentos (Tschacher e Kupper, 2002; Hosokawa *et al.*, 2003), tendências de surgimento de câncer em indivíduos (Summa *et al.*, 2007) e variações de batimentos cardíacos (Zhuang *et al.*, 2008; Ponomarenko *et al.*, 2005). A fim de melhor caracterizar essas áreas, considere o trabalho desenvolvido por Raiesdana *et al.* (2009), no qual séries temporais são utilizadas para determinar o comportamento do cérebro humano na transição do estado normal para o epiléptico. Conforme resultados obtidos, comprovou-se que o cérebro no estado normal tem comportamento caótico enquanto que no estado epiléptico, seu comportamento é conservativo.

Por fim, o uso de séries temporais tem atraído a atenção de diversos pesquisadores na tentativa de modelar o comportamento de mudanças climáticas do planeta (Koçak *et al.*, 2004; Yu *et al.*, 2008; Kärner, 2009). Recentemente, essas mudanças têm sido amplamente discutidas em diversas áreas de pesquisa, principalmente devido à preocupação mundial com o aquecimento global. Nesse sentido, pesquisadores têm utilizado séries temporais com o objetivo de predizer as causas e o impacto desse aquecimento no clima mundial (Ko *et al.*, 1993) e regional, como no caso da pesquisa desenvolvida no Brasil por Fearnside (1999).

Além dessas áreas, há diversas outras que empregam análise de séries temporais como forma de estudar, modelar, estimar e predizer o comportamento de sistemas reais.

A seção a seguir apresenta uma definição formal de séries temporais e, ainda, uma discussão sobre seus componentes. O estudo desses componentes é importante para o desenvolvimento deste trabalho, cujo principal objetivo é fornecer um modelo híbrido capaz

de descrever, com alta acurácia, o comportamento presente em séries temporais.

2.2 Séries temporais

Uma série temporal é definida por um conjunto de observações ordenadas de acordo com seus instantes de coleta (Morettin e Toloi, 2004) na forma $X_t = \{x_0, x_1, x_2, \ldots, x_t\}$. Segundo Box *et al.* (1994), o uso de séries temporais pode ser classificado, de acordo com sua aplicação, em quatro principais áreas: predição, estimação de funções de transferência, análise de efeitos na intervenção de eventos e sistemas de controle.

O uso de séries para predição visa analisar n observações passadas ($\{x_{t-n-1}, \ldots, x_{t-2}, x_{t-1}, x_t\}$) para estimar l observações futuras no tempo ($\{x_{t+1}, x_{t+2}, \ldots, x_{t+l}\}$). Essas observações futuras são conhecidas como "*lead time*" e são estimadas por uma função de predição. O principal objetivo dessa função é reduzir, ao máximo, a diferença entre os valores observados $\{x_{t+1}, x_{t+2}, \ldots, x_{t+l}\}$ e os preditos $\{\hat{x}_{t+1}, \hat{x}_{t+1}, \ldots, \hat{x}_{t+l}\}$. Para encontrar funções com maior precisão, é necessário definir riscos associados à cada decisão de predição, os quais são expressos por limites, inferior e superior, de probabilidade (Figura 2.1). Quanto maior o número de observações preditas, maior a distância entre os limites inferior e superior, reduzindo, assim, a acurácia da função de predição. Em geral, a escolha do limite e do total de observações a serem preditas, l, depende do sistema em estudo.

Por outro lado, o uso de séries para estimar funções de transferência visa compreender o comportamento dinâmico de um sistema, cujos valores de entrada são representados por meio de uma série temporal. Ao contrário da abordagem preditiva, o estudo e a estimação de funções de transferência visa, de maneira geral, compreender o comportamento atual de uma observação x_t com base em um conjunto de observações passadas $\{x_{t-k}, \ldots, x_{t-2}, x_{t-1}\}$.

No caso da análise de efeitos na intervenção de eventos, séries temporais podem ser utilizadas para entender o impacto que um determinado sistema exerce sobre outro como, por exemplo, a tentativa de compreender o efeito da poluição sobre o aumento da temperatura no planeta.

Por fim, o uso de séries temporais em sistemas de controle visa monitorar um determi-

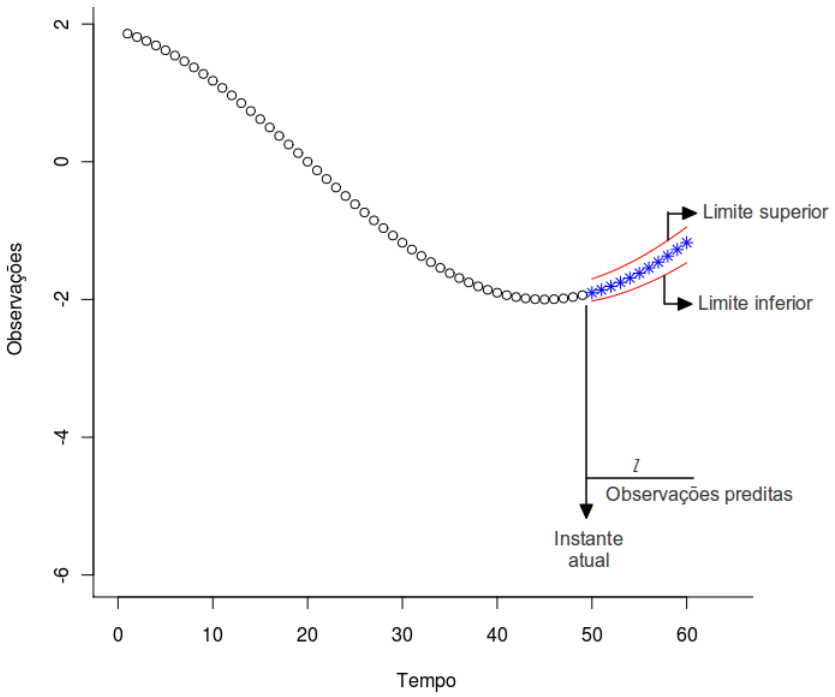


Figura 2.1: Exemplo de uma função de predição, cuja acurácia é delimitada pelas probabilidades inferior e superior (Adaptada de (Box *et al.*, 1994)).

nado processo de interesse a fim de detectar possíveis desvios do comportamento normal. À medida que esses desvios são detectados, ajustes são realizados para compensar a saída do sistema de forma que se aproxime do comportamento esperado.

Essas aplicações de séries temporais necessitam de modelos ajustados às características de suas observações. Dentre essas características, pode-se citar o número de variáveis utilizadas para compor uma observação e o intervalo de tempo entre coletas consecutivas de observações.

Com relação ao número de variáveis, séries temporais podem ser classificadas em univariadas ou multivariadas. Séries univariadas são compostas por valores escalares, sequencialmente coletados, na forma $\{x_1, x_2, \ldots, x_t\}$. Entretanto, quando k variáveis são observadas a cada instante de tempo t, diz-se que a série temporal é multivariada, denotada por $\{x_{1t}, x_{2t}, \ldots, x_{kt}\}$ (Hamilton, 1994).

Em relação ao intervalo entre coletas, subdivide-se séries temporais em duas classes (Morettin e Toloi, 2004): I) Discretas: cuja análise é feita no domínio temporal, segundo intervalos de tempo Δt, periódicos e fixos, em $\mathbb{N}$; II) Contínuas: cuja análise é realizada no

domínio de frequências (tempo em $\mathbb{R}+$).

Além de conhecer a quantidade de variáveis que compõem cada observação e o intervalo de tempo entre suas coletas, o processo de modelagem envolve, ainda, o conhecimento dos componentes responsáveis por definir o comportamento de séries temporais. Nesse sentido, pode-se denotar uma determinada série temporal X_t pela soma de três componentes não-observáveis $X_t = T_t + S_t + \varepsilon_t$, no qual T_t representa a tendência, S_t a sazonalidade e ε_t um componente aleatório (Morettin e Toloi, 2004). Cabe ressaltar que o componente T_t, ou tendência, representa variações no comportamento da série. O componente S_t, ou sazonalidade, indica se um determinado comportamento presente na série temporal tende a se repetir em intervalos de tempo Δt. Esses componentes, $\{T_t, S_t, \varepsilon_t\}$, são chamados de não-observáveis, pois não são coletados diretamente de um sistema, mas sim inferidos por meio de relações temporais entre observações.

Ao compreender esses componentes, pode-se avaliar aspectos globais de séries temporais como, por exemplo, a estocasticidade, a estacionariedade e a linearidade. A compreensão correta desses aspectos é importante para determinar modelos mais precisos para séries temporais.

Séries estocásticas são constituídas por observações e relações aleatórias que seguem funções de densidade de probabilidade e podem se modificar ao longo do tempo, dificultando a modelagem de seus eventos. Em contrapartida, séries determinísticas, predominantemente, apresentam observações com estrita dependência em relação aos valores passados.

Séries temporais estocásticas podem, ainda, ser subdivididas em estacionárias e não-estacionárias. Séries estacionárias encontram-se em um estado particular de equilíbrio estatístico (Box *et al.*, 1994), ou seja, elas se desenvolvem, aleatoriamente no tempo, em torno de uma média constante (Morettin e Toloi, 2004). Séries estritamente estacionárias são aquelas cujas propriedades não são afetadas de acordo com mudanças na origem do tempo, ou seja, a função de distribuição de probabilidade conjunta, associada com t observações ($\{x_1, x_2, \ldots, x_t\}$), não se altera sob o deslocamento d de suas observações no tempo ($\{x_{1+d}, x_{2+d}, \ldots, x_{t+d}\}$) (Box *et al.*, 1994). Por outro lado, uma série estocástica X_t é denominada fracamente estacionária, ou estacionária de segunda ordem, quando sua média e variância são constantes, $E(X_t) = \mu$ e $var(X_t) = \sigma^2$, e a sua covariância é definida

por $cov(X_t, X_{t+\tau}) = \gamma(\tau)$ (Box *et al.*, 1994). Geralmente, na análise e modelagem de séries temporais, tanto séries estritamente estacionárias quanto fracamente estacionárias são tratadas como estacionárias.

Séries temporais cujas observações são modeladas por processos estocásticos, que não satisfazem condições de estacionariedade, são denominadas estocásticas não-estacionárias. Geralmente, tais séries apresentam comportamento evolucionário, devido à presença de tendência, e são comumente modeladas por meio de processos explosivos (Morettin e Toloi, 2004).

Séries temporais podem, ainda, ser classificadas, segundo a regra que define suas observações, em lineares e não-lineares. Séries temporais lineares são aquelas cujas observações são compostas por uma combinação linear de ocorrências e ruídos passados. Portanto, a linearidade de uma série está presente no modelo, mapa, ou processo que a originou. Por sua vez, séries não-lineares são formadas por processos de combinação não-linear de observações e ruídos passados.

Após compreender esses aspectos essenciais de séries temporais, pode-se selecionar um subconjunto de técnicas mais adequadas para modelar e compreender seus comportamentos mais representativos. Nesse sentido, a próxima seção apresenta os principais modelos, definidos pela Estatística e pela área de Sistemas Dinâmicos, utilizados para compreender e analisar o comportamento de séries temporais.

2.3 Análise de Séries Temporais

Diversos sistemas produzem saídas ou resultados ao longo do tempo. Ao analisar essas saídas ou resultados, pode-se compreender o sistema em estudo, seu comportamento, suas transições, suas relações com outros sistemas, bem como estimar e predizer observações futuras. A área responsável por tal estudo é denominada análise de séries temporais, a qual busca, com base no conhecimento de uma regra (ou função geradora), determinar o comportamento das observações que compõem a série estudada. No entanto, para se obter essa regra, deve-se estudar características da série e buscar por seus componentes implícitos. Essa etapa é complexa e, em geral, depende da opinião subjetiva de um especialista (Shumway e Stoffer, 2006).

Além da opinião de um especialista, esse processo de análise de séries temporais pode ser conduzido com o auxílio de técnicas que realizam testes sobre séries e forneçam indicativos sobre o nível de ajuste de diferentes modelos. Nesse sentido, pode-se citar o trabalho desenvolvido por Ishii (2010), o qual apresenta uma abordagem para analisar, de maneira automática, séries temporais e sugerir modelos que atendam a acurácia desejada.

Os modelos utilizados para análise e compreensão do comportamento de séries temporais, sejam eles selecionados de forma automática ou com ajuda de especialistas, necessitam de uma certa quantidade de parâmetros que devem ser estimados a partir das observações disponíveis (Box *et al.*, 1994). Todavia, na prática, a quantidade muito elevada de parâmetros torna a modelagem complexa e até mesmo inadequada em determinadas circunstâncias. Isso ocorre devido à variação desses parâmetros (graus de liberdade do modelo) quando utilizados para representar séries temporais.

Diante dessa variação de parâmetros, Box *et al.* (1994) discutem o uso do princípio da parcimônia apresentado por Tukey (1961), o qual define que modelos adequados devem ser obtidos utilizando a menor quantidade possível de parâmetros. Dessa forma, modelos para séries temporais podem ser classificados em paramétricos, quando apresentam um número finito de parâmetros e a análise é feita no domínio temporal, e não-paramétricos, os quais não apresentam limites para o número de parâmetros e cuja análise é feita no domínio das frequências (Morettin e Toloi, 2004).

Uma outra tarefa realizada no processo de análise de séries, independentemente da quantidade de parâmetros utilizados, é a classificação segundo a aleatoriedade do processo observado. Essa classificação permite, por exemplo, diferenciar séries determinísticas e estocásticas. Séries determinísticas apresentam observações com estrita dependência em relação a observações passadas. Algumas áreas conduzem pesquisas nessa linha, dentre elas a Matemática e a Física, as quais empregam conceitos de Sistemas Dinâmicos e Teoria do Caos (Alligood *et al.*, 1997).

Em contrapartida, a área Estatística assume que séries temporais apresentam componentes estocásticos em sua formação. Esses componentes seguem funções de densidade de probabilidade e podem sofrer interferências temporais, o que dificulta estimação e predição de eventos futuros. Percebe-se, portanto, que há duas linhas principais de caracterização e estudo de séries temporais, uma voltada para seus aspectos determinísticos e outra para

os estocásticos.

De maneira geral, essas linhas permitem obter modelos para séries temporais, os quais, associados a demais estudos conduzidos pelo grupo de pesquisa, motivaram a proposta deste plano de pesquisa que visa integrar ambas abordagens com o intuito de aumentar a acurácia na modelagem de séries temporais. Nesse sentido, as próximas seções apresentam as principais técnicas que permitem modelar séries estocásticas e determinísticas, utilizando abordagens disponibilizadas pela Estatística e pela área de Sistemas Dinâmicos e Teoria do Caos.

2.3.1 Abordagem baseada em Estatística

As abordagens estatísticas de modelagem são, em sua maioria, voltadas para análise de séries temporais lineares estacionárias e não-estacionárias. Nesse contexto, define-se, a seguir, alguns dos principais processos utilizados para modelar séries.

Uma série temporal linear X_t composta apenas por uma sequência de ruídos ($x_t = \varepsilon_t$) é dita ser puramente aleatória (*Purely Random Process*) ou ruído branco. Tal série pode ser modelada como um processo estacionário, onde seus ruídos são mutuamente independentes e identicamente distribuídos, cuja média é zero, $E(X_t) = 0$, e variância é constante, $var(X_t) = \sigma^2$ (Box *et al.*, 1994).

Por outro lado, se as observações de uma série temporal X_t são definidas pela soma de um valor passado e um ruído branco, $x_t = x_{t-1} + \varepsilon_t$, então a série é dita ser um passeio aleatório (*Random Walk*). Nessa situação, um valor futuro depende tanto de uma observação passada quanto de um valor aleatório. Dessa forma, essa série não é considerada um processo estacionário, devido à mudança da média $E(X_t) = t \cdot \mu$ e da variância $var(X_t) = t \cdot \sigma^2$ em função do tempo t (Box *et al.*, 1994).

Considere ainda a série temporal X_t, tal que suas observações são definidas com base em q valores aleatórios (ruídos brancos) passados. Essa série pode ser modelada por um processo de média móvel (*Moving Average Processes*) de ordem q, MA(q), o qual é definido por $x_t = x_{t-1} + \theta_1 \cdot \varepsilon_{t-1} + \theta_2 \cdot \varepsilon_{t-2} + \cdots + \theta_q \cdot \varepsilon_{t-q}$, tal que $\{\theta_q\}$ são constantes e $\{\varepsilon_t\}$ são elementos formados por um processo puramente aleatório com média $E(X_t) = 0$ e variância $var(X_t) = \sigma^2$. Processos modelados por MA(q) são estacionários, pois possuem médias constantes e a variância não se altera em função do tempo (Box *et al.*, 1994). O termo

média móvel origina-se do fato de que a observação x_t é calculada em função da média e da soma ponderada de ruídos passados (Hamilton, 1994).

Modelos de média móvel são geralmente utilizados nas áreas econômica e financeira, nas quais seus indicadores são diretamente afetados por variáveis aleatórias como, por exemplo, decisões governamentais e falta de matérias-primas. Essas variáveis nem sempre afetam imediatamente o valor de um produto, mas funcionam como funções de transferência, gerando variações em observações subsequentes (Hamilton, 1994).

Por outro lado, se as ocorrências de uma série temporal X_t são definidas com base em p valores de observações anteriores, essa série pode ser modelada por um processo autoregressivo (*Autoregressive Processes*) de ordem p, AR(p). Tal modelo é definido por $x_t = \phi_1 \cdot x_{t-1} + \phi_2 \cdot x_{t-2} + \cdots + \phi_p \cdot x_{t-p} + \varepsilon_t$, tal que $\{\phi_p\}$ são as constantes e ε_t é um processo puramente aleatório com média $E(X_t) = 0$ e variância $var(X_t) = \sigma^2$ (Box *et al.*, 1994). Um processo autoregressivo de primeira ordem, AR(1), também chamado de processo de Markov, é denotado por $x_t = c + \phi_1 \cdot x_{t-1} + \varepsilon_t$.

Ao contrário do modelo de média móvel apresentado anteriormente, cuja modelagem de séries temporais é feita em função de ruídos, o modelo autoregressivo busca modelar uma série e predizer realizações futuras em função de suas observações passadas. Esse último modelo considera que o valor atual de uma observação x_t está relacionado às observações passadas $\{x_{t-1}, x_{t-2}, \cdots, x_{t-p}\}$, onde p determina o *lag* (ou atraso), o qual é utilizado para representar relações entre observações (Shumway e Stoffer, 2006). Esse modelo é comumente adotado para representar fenômenos naturais, devido à relação de recorrência e influência entre observações.

Uma série temporal X_t pode, ainda, ser modelada segundo p observações e q ruídos passados. Nesse caso, a área Estatística sugere utilizar o modelo ARMA (*Autoregressive and Moving Average* – Autoregressivo e de Médias Móveis), o qual combina modelos autoregressivos e de média móvel, conforme demonstrado na Equação 2.1 (Box *et al.*, 1994).

$$x_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \tag{2.1}$$

Esse modelo é formalmente definido por ARMA(p,q), onde p representa a ordem da parte autoregressiva, AR(p), e q refere-se à ordem da média móvel, MA(q). De maneira geral, o modelo autoregressivo permite relacionar valores futuros com informações históricas, en-

quanto que a média móvel mensura erros na análise de valores históricos. A relevância do modelo ARMA consiste em sua habilidade de modelar uma série temporal estacionária com menos parâmetros do que a aplicação, isolada, de um modelo puramente autoregressivo e de média móvel (Chatfield, 2004).

ARMA, conforme apresentado anteriormente, modela séries temporais estacionárias. Na prática, contudo, grande parte das séries temporais reais não são estacionárias. Nesse sentido, uma generalização do modelo ARMA, a fim de modelar séries temporais não-estacionárias, é definida pelo modelo ARIMA (modelo ARMA Integrado).

No modelo ARIMA, uma série temporal não-estacionária é inicialmente integrada, a fim de que seja removida a fonte não-estacionária de variação, para torná-la uma série temporal estacionária. Formalmente, o componente x_t do modelo ARMA, Equação 2.1, é substituído por w_t, tal que $w_t = \nabla^d x_t$, ou seja, w_t é uma diferença de x_t (então, x_t é uma integral de w_t). Nesse sentido, o modelo ARIMA é dito ser integrado devido à ∇^d, a qual permite ajuste no modelo estacionário para que observações distintas sejam integradas, visando fornecer um modelo para séries originalmente não-estacionárias (Equação 2.2) (Morettin e Toloi, 2004).

$$x_t = c + \phi_1 w_{t-1} + \phi_2 w_{t-2} + \ldots + \phi_{p+d} w_{t-p-d} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \tag{2.2}$$

Como pode ser observado, diante de tantas características e comportamentos, a modelagem de séries temporais não é uma tarefa simples de ser realizada. Na prática, antes da aplicação de um modelo, geralmente, é realizada uma observação visual dos dados. Uma ferramenta muito importante para auxiliar nessa observação é a função de autocorrelação (*Autocorrelation Function* – ACF), a qual permite visualizar o grau de dependência e características implícitas nos dados (Chatfield, 2004).

Para definir formalmente a ACF, considere a série $X_t = \{x_1, x_2, \ldots, x_n\}$, sendo $\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$ o valor médio dessa amostra. A autocorrelação, $\hat{\rho}(h)$, de X_t é dada pela razão entre a autocovariância de X_{t+h}, considerando um deslocamento h, e a autocovariância de

X_t, ou seja,

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \tag{2.3}$$

sendo que a função de autocovariância (*Autocovariance Function* – ACVF) é calculada por meio da seguinte equação:

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x}) \tag{2.4}$$

A fim de exemplificar o uso da ACF, considere três casos apresentados nas Figuras 2.2, 2.3 e 2.4. Para cada caso apresentado, há dois gráficos: o primeiro (acima) é a série temporal conforme os dados foram coletados e o segundo (abaixo) é a ACF dessa série.

A série temporal da Figura 2.2 é composta por valores independente e identicamente distribuídos (iid) gerados a partir de uma distribuição normal com média zero e variância igual a 1. O gráfico da função de autocorrelação foi obtido considerando o deslocamento $h = 1, 2, \ldots, 40$. Como pode ser visto no gráfico da ACF, em séries compostas apenas por ruídos (valores aleatórios), os valores de autocorrelação tendem a ser próximos de zero, $\hat{\rho}(h) \simeq 0$. Nesse sentido, espera-se que aproximadamente 95% dos valores obtidos estejam dentro do intervalo $\pm 1.96/\sqrt{n}$ [2]. Portanto, cerca de $40 \times (0.05) = 2$ valores devem ficar fora desse intervalo (Brockwell e Davis, 2002).

Uma outra característica que é facilmente identificável com o uso da ACF é a tendência. Séries que possuem tendência são consideradas não-estacionárias devido à dependência temporal, conforme descrito anteriormente. A função de autocorrelação, nesse caso, apresenta valores de $\hat{\rho}(h)$ reduzindo suavemente até zero à medida que o valor de h aumenta (Chatfield, 2004), como pode ser visto na Figura 2.3. A série temporal apresentada nessa figura é composta por dados sobre o aquecimento global e encontra-se disponível em (Shumway e Stoffer, 2006).

Por fim, a Figura 2.4 apresenta um outro exemplo de identificação de características presentes em séries por meio da ACF. Nesse exemplo, foi utilizada uma série temporal composta por dados sobre a relação entre mortes e a variação de temperatura diária na região de Los Angeles (Shumway e Stoffer, 2006). Como pode ser visto, há um comporta-

[2] O valor 1.96 refere-se ao quantil da distribuição normal padrão $z_{0.975}$

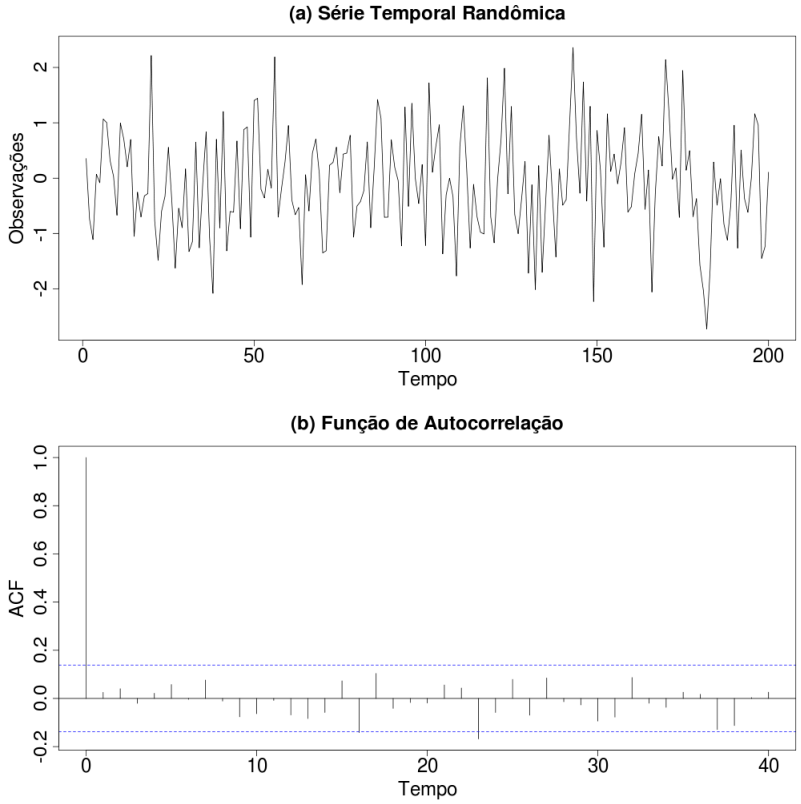


Figura 2.2: (a) Série temporal aleatória, $X_t \sim IID(0, 1)$. (b) Função de autocorrelação (ACF) da série.

mento sazonal nessa série, o qual se reflete no gráfico da ACF, fazendo-o oscilar na mesma frequência.

Conforme discutido anteriormente, os modelos de análise de séries temporais apresentados são indicados para situações nas quais os dados analisados são provenientes de sistemas estocásticos. A fim de complementar essas abordagens, a próxima seção apresenta modelos cujo objetivo é compreender a dinâmica comportamental de séries determinísticas.

2.3.2 Abordagem baseada em Sistemas Dinâmicos e Teoria do Caos

As ferramentas estatísticas apresentadas neste capítulo podem ser utilizadas para modelar e predizer observações futuras de séries temporais. Essas ferramentas são muito úteis e computacionalmente simples de serem implementadas. Afora isso, esses modelos apresentam baixos erros médios de predição para séries lineares (Box *et al.*, 1994), contudo geram resultados insatisfatórios para séries mais complexas, as quais motivaram aproxi-

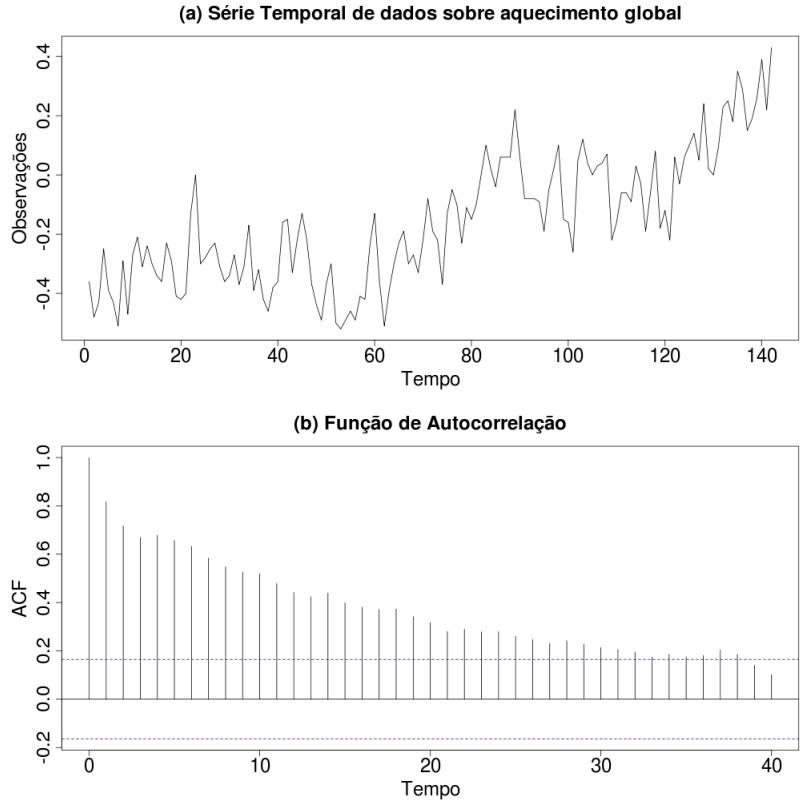


Figura 2.3: (a) Série temporal de dados sobre o aquecimento global. (b) Função de autocorrelação (ACF) da série.

mações locais (Casdagli, 1989) e o estudo de regularidades internas por meio de sistemas dinâmicos (Alligood *et al.*, 1997).

Nesse sentido, visando maior acurácia na modelagem e predição em séries temporais mais complexas, uma nova frente de pesquisa tem sido explorada, a qual emprega conceitos de Sistemas Dinâmicos e Teoria do Caos (Mello, 2009). Essa frente de pesquisa motiva a adoção desses conceitos, os quais utilizam ferramentas estatísticas e matemáticas, a fim de modelar comportamentos de séries temporais com maior precisão, visto que tais ferramentas são menos sensíveis a ruídos e *outliers* (Alligood *et al.*, 1997).

Para analisar e melhor compreender o comportamento de séries temporais, pesquisadores da área de Sistemas Dinâmicos concentraram esforços na busca de métodos para avaliar a estabilidade e correlação de suas observações. Dentre esses métodos, destacam-se os expoentes de Lyapunov e Hurst. O primeiro é utilizado para avaliar a estabilidade assintótica de séries, o qual é calculado por meio da taxa de variação de trajetórias vizi-

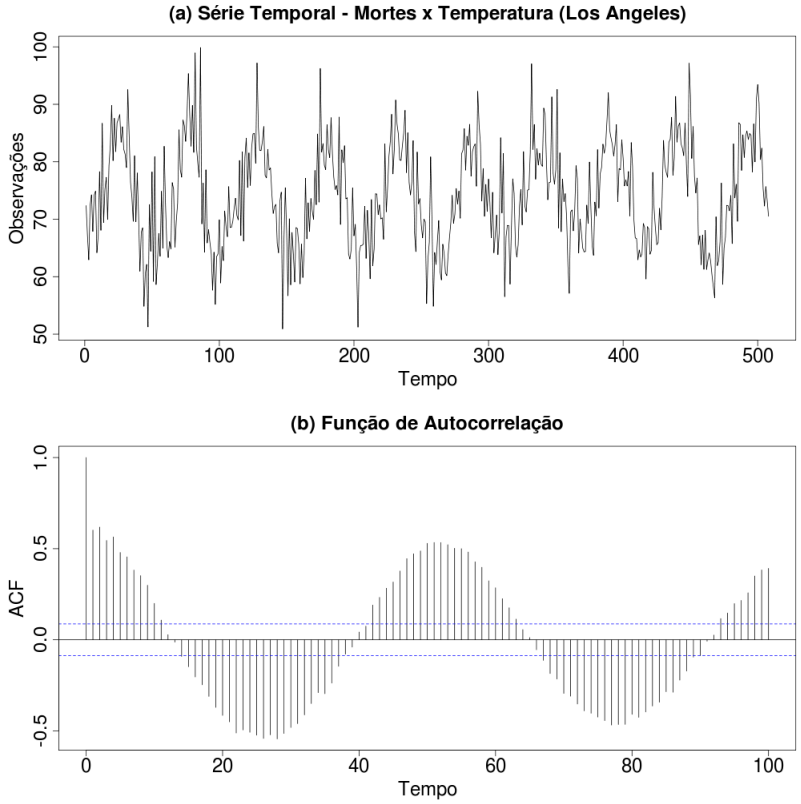


Figura 2.4: (a) Série temporal de dados sobre a relação entre mortes e temperatura diária na região de Los Angeles. (b) Função de autocorrelação (ACF) da série.

nhas, considerando uma separação inicial $\delta \mathbf{Z}_0$. Essa divergência dá-se por $|\delta \mathbf{Z}(t)| \approx e^{\lambda t} |\delta \mathbf{Z}_0|$, onde t é o instante de tempo e λ é o expoente de Lyapunov.

$$\lambda = \lim_{t \to \infty} \frac{1}{t} \ln \frac{|\delta \mathbf{Z}(t)|}{|\delta \mathbf{Z}_0|} \tag{2.5}$$

O resultado desse expoente (λ) permite concluir sobre a estabilidade de séries: 1) quando $\lambda < 0$, a série é atraída para um ponto fixo estável (onde pontos da órbita se repetem no tempo na mesma ou em diferentes escalas); 2) para $\lambda = 0$, a série comporta-se como um sistema conservativo; e, 3) quando $\lambda > 0$, a série tem características caóticas e instáveis, isso significa que a distância entre pontos da trajetória irá sempre divergir, em média, à uma taxa exponencial definida pelo expoente de Lyapunov (Rosenstein *et al.*, 1993; Eckmann e Ruelle, 1985).

O segundo expoente, denominado expoente de Hurst, mede a aleatoriedade de um con-

junto de dados e é limitado ao intervalo $[0, 1]$. Quando próximo de 1, indica comportamento persistente, ou seja, há correlação entre um evento e a ocorrência de outro no futuro. Caso o valor do expoente seja próximo de zero, a série temporal apresenta comportamento anti-persistente, ou seja, há uma autocorrelação negativa entre pares de eventos. Nesse caso, um acréscimo em um valor passado da série gera decréscimo em outra observação futura e vice-versa. Valores próximos de $0,5$ indicam que a série temporal comporta-se como uma caminhada aleatória (*Random Walk*, Seção 2.3.1), dessa forma, valores futuros não dependem do histórico. Observa-se, portanto, que o expoente de Hurst é uma ferramenta importante para avaliar a relevância de dados históricos.

Além da interpretação oferecida pelos expoentes de Lyapunov e Hurst, pesquisadores desenvolveram técnicas para estimar regras, ou funções, geradoras de sistemas dinâmicos, as quais permitem compreender relações entre estados. Essas relações são mais simples de serem modeladas para sistemas dinâmicos determinísticos, uma vez que suas regras são mais facilmente obtidas.

Regras geradoras podem ser obtidas por meio da modelagem do conhecimento embutido em sistemas dinâmicos, as quais permitem, por exemplo, conhecer tendências, realizar predições e classificar operações (Alligood *et al.*, 1997). Nesse sentido, Whitney (1936) aplicou variedades diferenciáveis como forma de reconstruir funções utilizando transformações para espaços multidimensionais.

Whitney (1936) observou que esse desdobramento da série em um número maior de dimensões permitiria a compreensão de comportamentos não observáveis ou pouco representativos quando descritos sob número reduzido de dimensões. A partir disso, ele propôs seu teorema de imersão, segundo o qual qualquer variedade em n dimensões pode ser mapeada em um espaço de $2n + 1$ dimensões.

Baseado nos estudos de Whitney (1936), Takens (1980) prova que, ao invés de mapear os estados de um sistema dinâmico em espaço de $2n + 1$ dimensões, pode-se reconstruí-lo considerando deslocamentos no tempo. Segundo o teorema de imersão de Takens (1980), uma série temporal $\{x_0, x_1, \ldots, x_{n-1}\}$ pode ser reconstruída em espaço multidimensional $x_n(m, \tau) = \{x_n, x_{n+\tau}, \ldots, x_{n+(m-1)\tau}\}$, ou de coordenadas de atraso, onde m é a dimensão embutida e τ representa um *time delay* ou dimensão de separação. Essa técnica de mapeamento ou reconstrução permite transformar as saídas produzidas por sistemas dinâmi-

cos, geralmente representadas por séries temporais unidimensionais, em um conjunto de pontos em espaço (em geral, Euclidiano) de m dimensões.

Essa reconstrução auxilia na obtenção de regras geradoras de sistemas dinâmicos, simplificando o estudo de comportamentos e, consequentemente, seu emprego na classificação e predição de operações (Alligood *et al.*, 1997).

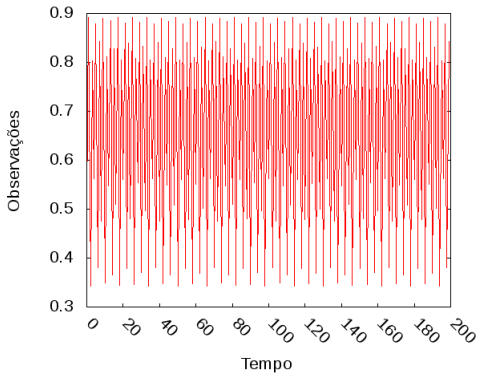
Dimensão embutida e de separação

Para exemplificar a obtenção de funções geradoras, que definem os comportamentos de sistemas dinâmicos, considere, inicialmente, a equação Logística $x_{t+1} = b \cdot x_t \cdot (1,0 - x_t)$, com condições iniciais $t \in [0; 4000]$, $b = 3,8$ e $x_0 = 0,5$. A Figura 2.5(a) apresenta as saídas, ou órbita percorrida no tempo, para essa regra, a qual tem expoente de Hurst $H = 0,40535$ e Lyapunov $\lambda = 0,447039$ (Alligood *et al.*, 1997). O expoente de Hurst evidencia pequeno grau de anticorrelação, contudo a série apresenta, em geral, comportamento aleatório. O expoente de Lyapunov aponta seu comportamento caótico e instável (a distância entre pontos da trajetória tende sempre a divergir, o que dificulta a modelagem). Esses expoentes permitem concluir que ao aplicar, diretamente, uma técnica de predição sobre tal série, tende-se a obter resultados ruins.

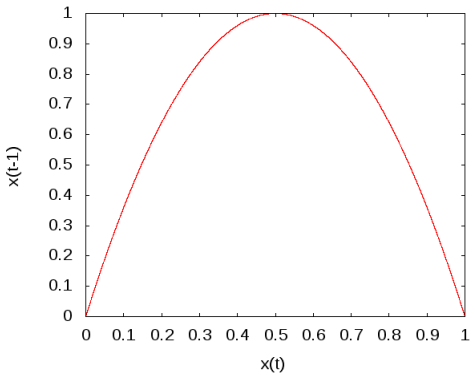
Contudo, pode-se aplicar uma forma alternativa de reconstrução dessa órbita, utilizando a dimensão de separação e embutida, a qual permite observar regularidades internas e simplificar a compreensão do sistema em estudo. Para melhor compreender as dimensões embutida e de separação, considere as saídas da função logística reconstruídas em um espaço multidimensional onde $m = 2$ e $\tau = 1$, a qual resulta em pares de pontos (x_t, x_{t+1}) (Figura 2.5(b)). Após a reconstrução, o comportamento da função logística, que era aparentemente uma caminhada aleatória (Figura 2.5(a)), pode ser estudado, compreendido e modelado de forma mais simples. Ao realizar uma regressão dos pontos resultantes, pode-se obter a regra do sistema dinâmico e determinar seus comportamentos futuros. Tendo essa regra e o valor de uma observação x_t, pode-se, por exemplo, definir o próximo valor da série, x_{t+1}, o qual pode ser retroalimentado para gerar x_{t+2}, e assim sucessivamente.

A dimensão embutida define, basicamente, o número de eixos, do espaço de coordenadas de atraso, necessários para desdobrar o comportamento reconstruído da série. Nesse

Figura 2.5: Função logística reconstruída em dimensão embutida 2 e de separação 1



(a) Saídas da função logística



(b) Saídas da função logística desdobrada

caso a série necessitou de duas dimensões, outras podem requerer espaços com mais eixos. A dimensão de separação, por outro lado, auxilia na extração de comportamentos periódicos de séries, ou seja, informa o deslocamento de valores históricos que devem ser avaliados a fim de predizer comportamento futuro, ou seja, essa dimensão permite destacar sazonalidades presentes na série estudada.

Observa-se que as dimensões embutida e de separação auxiliam no estudo de séries determinísticas. Contudo, é necessário estimar essas dimensões para quaisquer séries oriundas de dados experimentais. Nesse sentido, Abarbanel *et al.* (1993) afirma que a utilização da função de autocorrelação auxilia na determinação da dimensão de separação de séries. Entretanto, essa técnica apresenta bons resultados apenas para séries lineares e com comportamento estacionário. A presença, por exemplo, de tendência nas observações que compõem uma série temporal, tornando-a não-estacionária, faz com que valores de autocorrelação reduzam suavemente até próximo de zero (Figura 2.3), ou seja, quando há tendência numa série, a função de autocorrelação não consegue destacar semelhanças entre as observações, impossibilitando a determinação da dimensão de separação.

A fim de superar essa limitação, Fraser e Swinney (1986) estudaram e confirmaram que a técnica de autoinformação mútua (*Auto Mutual Information* – AMI) apresenta melhores resultados na estimativa de dimensões de separação. Para tanto, aplica-se essa técnica considerando diferentes deslocamentos no tempo. Em seguida, traça-se uma curva em função dos deslocamentos (iniciando em 1 e incrementando) e adota-se seu primeiro mí-

nimo como dimensão de separação.

A informação mútua média é definida pela Equação 2.6, onde X e Y seguem, respectivamente, as funções de distribuição de probabilidades P_X e P_Y, e X e Y ocorrem em pares com distribuição conjunta P_{XY} (Kennel *et al.*, 1992).

$$I(X;Y) = \int P_{XY}(x,y) \log_2 \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} dxdy \tag{2.6}$$

A estimação da dimensão embutida foi estudada por Takens (1980) e Mañé (1980), os quais confirmaram que o limite superior da dimensão embutida $D_e \in \mathbb{N}$ pode ser estimado utilizando a dimensão fractal D_f, de acordo com a equação $D_e > 2,0 \cdot D_f$. Contudo, os estudos realizados comprovam que a dimensão resultante dessa equação é, em geral, maior que o necessário e trabalhando-se, desnecessariamente, com mais dimensões, adiciona-se complexidade e tempo de processamento à modelagem e à análise de resultados (Kennel *et al.*, 1992).

Para solucionar esse problema e encontrar a dimensão embutida mínima, pode-se utilizar o cálculo de invariantes do sistema (tais como o expoente de Lyapunov (Alligood *et al.*, 1997)) para diferentes valores de dimensão, observando a saturação dos resultados. A complexidade dessa abordagem motivou Kennel *et al.* (1992) a propor o método de falsos vizinhos (*False Nearest Neighbors* – FNN), que calcula os vizinhos mais próximos de cada ponto, no espaço de coordenadas de atraso (iniciando com dimensão embutida igual a 1). Em seguida, uma nova dimensão é adicionada e a distância entre vizinhos mais próximos é novamente calculada. Caso haja acréscimo nessa distância, os pontos são considerados falsos vizinhos, o que evidencia a necessidade de mais dimensões para reconstruir o comportamento da série.

Kennel *et al.* (1992) consideram uma dimensão embutida d onde o r-ésimo vizinho mais próximo de $y(n)$ é dado por $y^{(r)}(n)$. A distância Euclidiana (geralmente adotada) entre o ponto $y(n)$ e seu r-ésimo vizinho mais próximo é dada pela Equação 2.7. Ao adicionar uma nova dimensão, reconstrói-se a série em $d+1$ e adiciona-se a coordenada $(d+1)$ em cada vetor $y(n)$, a qual é incluída na equação de distância Euclidiana (termo $x(n+dT)$ da Equação 2.8). Dessa forma, o critério mede a variação de distância ao adicionar uma nova

dimensão, conforme descrito pela Equação 2.9.

$$R_d^2(n,r) = \sum_{k=0}^{d-1} (x(n+kT) - x^{(r)}(n+kT))^2 \tag{2.7}$$

$$R_{d+1}^2(n,r) = R_d^2(n,r) + (x(n+dT) - x^{(r)}(n+dT))^2 \tag{2.8}$$

$$V_{n,r} = \sqrt{\frac{R_{d+1}^2(n,r) - R_d^2(n,r)}{R_d^2(n,r)}} = \frac{|x(n+dT) - x^{(n)}(n+dT)|}{R_d^2(n,r)} \tag{2.9}$$

Segundo os autores, se $V_{n,r} > R_{tol}$ então os pontos são considerados falsos vizinhos, onde R_{tol} é um limiar. Eles ainda concluem, empiricamente, que $R_{tol} \geq 10,0$ é um bom limite para a geração de resultados.

Após definir as duas dimensões, aplica-se a teoria de imersão de Takens (1980) e reconstrói-se a série, desdobrando-se, completamente, o comportamento da regra desse sistema dinâmico. Para exemplificar esse desdobramento, considere o conjunto de dados de Lorenz (Boker, 2010) apresentado na Figura 2.6. O atrator de Lorenz, utilizado nesse exemplo é uma estrutura fractal caótica que corresponde a um comportamento coletado a longo prazo de um sistema de três dimensões (Alligood *et al.*, 1997).

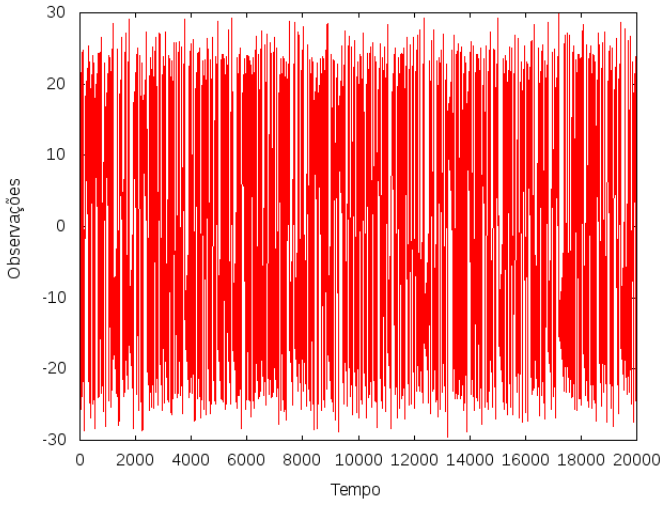


Figura 2.6: Atrator de Lorenz.

O cálculo da dimensão de separação foi realizada utilizando a técnica AMI, resultando na Figura 2.7. Como pode ser visto na figura, o valor estimado da dimensão de separação

é $\tau = 5$, o qual representa o primeiro mínimo obtido.

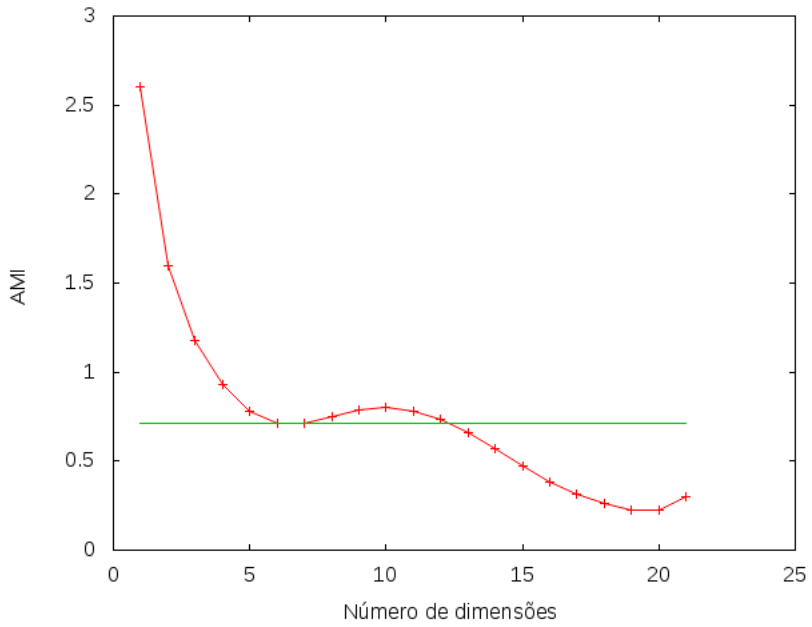


Figura 2.7: Atrator de Lorenz: Cálculo da dimensão de separação utilizando a técnica AMI. A linha tracejada representa o primeiro mínimo obtido.

A dimensão embutida, por sua vez, foi calculada usando a técnica FNN. Estudos desenvolvidos por Liebert *et al.* (1991) propõem que a escolha do número de dimensões embutidas pode ser feita quando encontrado o primeiro valor da fração de falsos vizinhos abaixo de 30%. Nesse sentido, considerando a Figura 2.8, pode-se afirmar que o número de dimensões embutidas para a série de Lorenz é $m = 3$.

Tomando como base os valores das dimensões obtidas, a série temporal original pode ser reconstruída, desdobrando o seu comportamento e obtendo sua regra, ou seja, a função que define sua órbita no tempo. Obtendo tal função, pode-se estudar seus pontos fixos estáveis e instáveis, e compreender suas tendências. Como exemplo, considere a série de Lorenz analisada, cujas as observações devem correlacionadas segundo um espaço de coordenadas de atraso com $m = 3$ e $\tau = 5$, ou seja, para estudar ou predizer cada observação x_t deve-se considerar as observações (x_{t-5}, x_{t-10}). Esse desdobramento é apresentado na Tabela 2.1 e pode ser visualizado na Figura 2.9.

Na seção a seguir são apresentados trabalhos relacionados ao problema de modelagem de séries temporais. Dentre esses estudos, são apresentadas abordagens que visam a

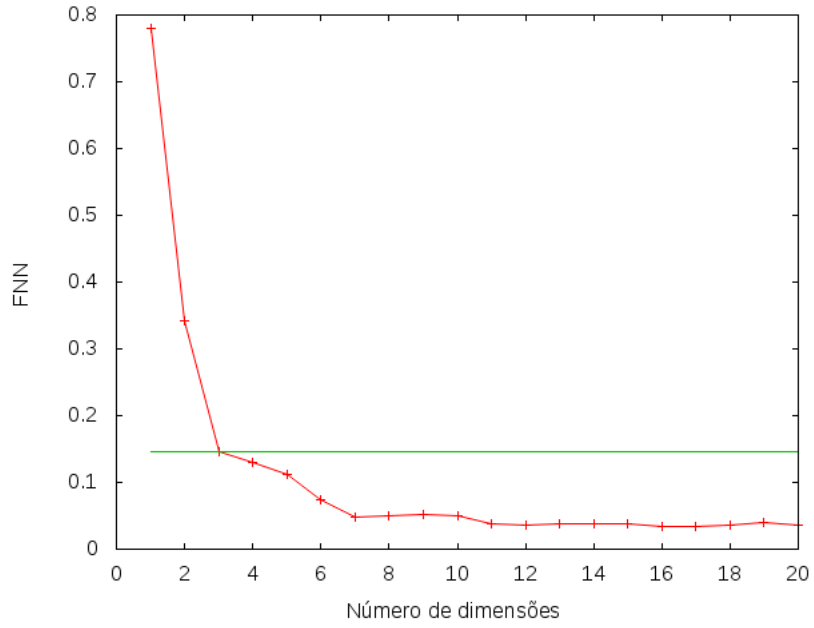


Figura 2.8: Atrator de Lorenz: Cálculo da dimensão embutida usando a técnica FNN.

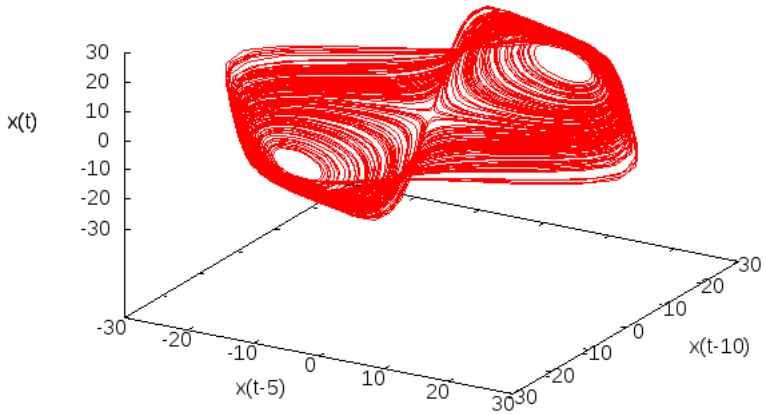


Figura 2.9: Série de Lorenz desdobrada em 3 dimensões.

modelagem direta de séries estudadas e abordagens que, assim como neste trabalho, visam segmentar componentes, segundo a estocasticidade e o determinismo presente em séries temporais.

Tabela 2.1: (a) Conjunto de dados original do atrator de Lorenz; (b) Conjunto de dados do atrator de Lorenz reconstruído segundo as dimensões embutida e de separação encontradas ($m = 3$ e $\tau = 5$)

(a)

Dimensão 1
-9.6559617
-6.9902085
-4.9834927
-3.5773619
-2.6589215
-2.1120568
-1.8411753
-1.7784935
-1.8834828
-2.1397586
-2.5521791
-3.1453527
-3.9638112
-5.0733551
-6.5619076
-8.5356685
-11.100864
-14.311700
-18.056232
-21.873802
-24.819411

(b)

Dimensão 1	Dimensão 2	Dimensão 3
-9.6559617	-2.1120568	-2.5521791
-6.9902085	-1.8411753	-3.1453527
-4.9834927	-1.7784935	-3.9638112
-3.5773619	-1.8834828	-5.0733551
-2.6589215	-2.1397586	-6.5619076
-2.1120568	-2.5521791	-8.5356685
-1.8411753	-3.1453527	-11.100864
-1.7784935	-3.9638112	-14.3117
-1.8834828	-5.0733551	-18.056232
-2.1397586	-6.5619076	-21.873802
-2.5521791	-8.5356685	-24.819411

2.4 Trabalhos Relacionados

A regularidade presente em séries temporais possibilitou a criação de diversas abordagens capazes de modelar e compreender sistemas que deram origem às suas observações. Além de abordagens que realizam essa tarefa no domínio temporal, conforme apresentado anteriormente, existe uma outra linha de pesquisa que estuda o comportamento de séries no domínio das frequências. Nesta seção são discutidas algumas das abordagens desenvolvidas para esse propósito e, ainda, trabalhos relacionados que utilizam conceitos de Aprendizado de Máquina para, também, modelar séries temporais e extrair informações sobre seu comportamento.

Com relação às pesquisas desenvolvidas no domínio das frequências, pode-se destacar a transformada de Fourier. Essa transformada é realizada com base na variação de amplitudes e frequências obtidas pela soma de funções seno e cosseno (Graps, 1995). A

representação da série em diferentes frequências permite destacar componentes, os quais podem ser isolados por meio de técnicas de filtragem.

Além da aplicação de transformada de Fourier diretamente na série temporal, existe ainda a possibilidade de aplicá-la nas estruturas geradas pelas matrizes da ferramenta RP (apresentada na Seção 1.2). A aplicação da transformada 2D de Fourier possibilita utilizar outros filtros sobre a série no domínio das frequências, a fim de remover pontos isolados nas matrizes de recorrência. Essa aplicação permite segmentar séries temporais de acordo com seus componentes estocásticos e determinísticos.

Uma desvantagem da transformada de Fourier é a utilização de um único conjunto de funções que se baseia apenas em senos e cossenos (Tang, 2009). Essa limitação impede o tratamento de séries temporais quando há presença de comportamentos irregulares ou grandes variações temporais (Minerva, 2010). Por essa razão, a utilização dessa transformada, apesar de possibilitar a realização de regressões na série e a remoção de ruídos, dificulta a criação de modelos que evidenciem a dependência temporal entre observações.

Uma forma alternativa, e mais eficiente, à transformada de Fourier é dada pela utilização de transformadas *Wavelet*. *Wavelets* são funções matemáticas que separam dados em diferentes escalas e resoluções (Graps, 1995). Essa separação possibilita que os dados sejam analisados de maneira temporal ou, tal como na transformada de Fourier, de acordo com frequências. Além disso, ao contrário da transformada de Fourier que utiliza senos e cossenos como funções base, a transformada *Wavelet* utiliza funções de ondas e coeficientes. Os coeficientes são ordenados utilizando dois padrões dominantes: coeficientes que funcionam como filtro de suavização dos dados e coeficientes que fornecem detalhes sobre os dados (Graps, 1995). Esses detalhes representam variações no vetor de informações, as quais podem ser utilizadas para, por exemplo, detectar a presença de ruídos nos dados (cheng Huang e Cressie, 1998).

Um outro exemplo do uso de *Wavelets* na análise de séries temporais é apresentado por Minerva (2010). Nesse trabalho, o autor utiliza uma transformada *Wavelet* Discreta (*Discret Wavelet Transform* – DWT) para segmentar componentes estocásticos e determinísticos presentes em séries temporais. De maneira geral, a aplicação da DWT sobre uma determinada série temporal gera um sinal resultante composto de ondas de alta escala e baixa frequência, que representam as informações determinísticas, e ondas de baixa es-

cala e alta frequência, que representam comportamentos estocásticos presentes na série. O autor desse trabalho discute, ainda, as vantagens de compreender o comportamento separado de cada sistema, o que confirma a hipótese desta pesquisa de doutorado.

Atualmente, existem diversas abordagens que utilizam *Wavelets* para a remoção de ruídos em sinais, i.e., *Wavelets* de Haar e *Complex Wavelet Transform* (CWT) (Nason, 2008), e em imagens bidimensionais (Gonzalez e Woods, 2006), podendo ser aplicadas, assim como a transformada de Fourier, para remover ruídos em estruturas geradas pela ferramenta RP. Essas variações nas abordagens produzem diferentes informações sobre os componentes presentes nas séries, o que dificulta a escolha de qual tipo de transformada *Wavelet* usar na geração do modelo híbrido proposto neste plano de trabalho. Apesar de existirem indicativos favoráveis à utilização dessas transformadas, é necessária uma investigação detalhada a fim de determinar como utilizá-las na etapa de segmentação de séries temporais.

Além dessas transformações no domínio temporal e de frequência, a área de Aprendizado de Máquina tem proposto técnicas e ferramentas para análise de séries temporais (Wang *et al.*, 2009). Segundo Mitchell (1997), Aprendizado de Máquina (AM) busca auxiliar na modelagem de conhecimento de um sistema de acordo com suas experiências prévias. Essa área emprega técnicas projetadas por diversas áreas de conhecimento como, por exemplo, Inteligência Artificial, Estatística, Teoria da Informação e Teoria de Controle.

Conceitos, abordagens e ferramentas utilizadas pela área de Aprendizado de Máquina têm-se provado úteis para a análise de séries temporais (Wang *et al.*, 2009; Liao, 2005; Jain e Kumar, 2007; Mello, 2010; Albertini e Mello, 2007). No contexto de análise *online* de observações, alguns trabalhos têm utilizado técnicas de AM com treinamento não-supervisionado para modelar *Data Streams* (Marsland *et al.*, 2002; Albertini e Mello, 2007). Essas técnicas têm sido aplicadas devido à capacidade de aprendizado *online*, que tende a ser mais rápido para atender aplicações com restrições temporais. Além disso, essas técnicas criam estruturas para representar o comportamento de séries, eliminando, em alguns casos, a necessidade de armazenamento de todas as observações históricas de um sistema em estudo.

Sistemas de aprendizado de máquina podem ser classificados de acordo com os seguintes paradigmas (Monard e Baranauskas, 2003): simbólico, estatístico, baseado em

exemplos, evolutivo e conexionista.

No aprendizado de máquina simbólico são utilizadas técnicas que exploram estruturas gráficas ou lógicas, que representam conhecimento de alto nível, o qual é facilmente compreensível por seres humanos. Dentre as diversas técnicas disponibilizadas por esse paradigma, destacam-se as árvores de decisão. No entanto, no contexto de séries temporais, esse paradigma, geralmente, não é adotado devido à dificuldade em modelar uma série temporal segundo uma estrutura hierárquica de suas observações. De maneira geral, essa organização hierárquica das observações apresenta bom desempenho quando os dados analisados são gerados a partir de processos independentes e identicamente distribuídos. Por essa razão, a principal limitação desse paradigma na análise de séries temporais é a dificuldade em gerar modelos hierárquicos que mantenham evidentes as dependências temporais entre observações.

No paradigma estatístico, técnicas tomam decisões e aprendem com base em inferências e probabilidades estatísticas obtidas a partir dos dados analisados. Técnicas baseadas em estatísticas têm sido amplamente utilizadas na modelagem de séries temporais, conforme citado na Seção 2.3.1.

Por outro lado, o aprendizado de máquina baseado em exemplos é caracterizado pela presença de uma base de instâncias com exemplos anteriormente produzidos pelo sistema em estudo. Novas instâncias são comparadas aos casos similares presentes nessa base. Um exemplo de técnica baseada nesse paradigma é o aprendizado baseado em instâncias (*Instance-Based Learning* – IBL).

Além desses paradigmas, existem ainda técnicas baseadas no paradigma evolutivo, as quais possuem uma analogia direta à teoria evolucionista. Nesse caso, o aprendizado é baseado na seleção de indivíduos mais aptos (os quais podem representar observações e suas relações). Como exemplo desse paradigma, pode-se citar os algoritmos genéticos. No contexto de séries temporais, diversas abordagens utilizam essa técnica (Baragona e Vitrano, 2006) como, por exemplo: I) abordagens que definem cada série temporal como sendo um indivíduo, a fim de se obter relações entre diferentes indivíduos (ou séries); II) abordagens em que cada observação de uma série temporal é modelada como sendo um indivíduo da população, cujo objetivo é obter regras geradoras a partir de indivíduos mais aptos.

Considerando esses conceitos de identificação de comportamento de séries utilizando os paradigmas evolutivo e baseado em instâncias, Mello *et al.* (2007b) desenvolveram a política chamada RouteGA (*Route with Genetic Algorithm Support*). Essa política emprega um algoritmo genético para realizar a primeira etapa do escalonamento de aplicações sobre ambientes distribuídos. As informações sobre o comportamento das aplicações executadas e dos recursos utilizados são organizadas como séries temporais e armazenadas em uma base de experiências, a qual é analisada por um algoritmo IBL a fim de estimar o comportamento esperado de novas aplicações. No entanto, o maior problema na utilização do IBL no contexto deste trabalho é que e, de fato, essa abordagem não modela uma série temporal, apenas fornece estimativas, a curto prazo, sobre observações futuras.

Por fim, existe ainda o paradigma conexionista, o qual é inspirado nas redes neurais que constituem o cérebro de seres vivos e no aprendizado biológico. Esse paradigma deu origem às redes neurais artificiais (*Artificial Neural Networks* – ANN), as quais têm sido amplamente utilizadas para modelar séries temporais (Wang *et al.*, 2009; Liao, 2005; Jain e Kumar, 2007; Mello, 2010; Albertini e Mello, 2007; Dodonov e Mello, 2007). Grande parte dessas técnicas utilizam observações como entrada da rede e, baseando-se no aprendizado obtido, estimam funções capazes de modelar a ocorrência de observações de séries.

Um exemplo de uso da abordagem conexionista na modelagem de séries temporais é encontrado no trabalho desenvolvido por (Lee *et al.*, 1993). Nesse trabalho foi proposto um teste estatístico, baseado em (White, 1990), chamado WNN (*White Neural Network*). O principal objetivo desse teste é verificar se uma determinada série temporal analisada foi gerada a partir de um sistema linear.

De maneira geral, a utilização exclusiva de ANN, no contexto deste trabalho, possui algumas dificuldades como, por exemplo, determinar a quantidade de neurônios que devem ser utilizados para modelar uma série e o treinamento. A escolha de um número muito grande de neurônios pode levar o modelo a *overfitting* e, por outro lado, uma quantidade insuficiente pode fazer com que o modelo generalize de maneira insatisfatória, fazendo com que, em ambos os casos, aumente o erro entre os valores esperados e preditos. Para resolver esse problema, existem diversas pesquisas que tentam estimar essa quantidade. Por outro lado, surgem novos questionamentos sobre como a variação no número de neurônios pode afetar o modelo resultante (Mello, 2010). Quanto ao treinamento, a maioria das

abordagens baseadas em ANN requerem treinamento para atualizar seu conhecimento, ou seja, há a necessidade periódica de ajuste completo do modelo. Além disso, com relação à segmentação de componentes, que é uma importante etapa no desenvolvimento deste trabalho, faz-se necessário determinar, ainda, quais funções de ativação poderiam ser utilizadas (em neurônios) para representar o comportamento estocástico e o determinístico.

2.5 Considerações finais

Neste capítulo foram apresentados os principais conceitos de séries temporais e componentes que descrevem o comportamento de suas observações. Com base nesses comportamentos, foram discutidos modelos disponibilizados pela área Estatística e de Sistemas dinâmicos, cujo principal objetivo é compreender o comportamento de séries temporais.

Além disso, foram discutidos trabalhos que modelam séries temporais no domínio de frequências e técnicas que utilizam ferramentas disponibilizadas pela área de Aprendizado de Máquina. Como pôde ser observado, esses trabalhos e técnicas possuem limitações, principalmente quando o objetivo é realizar a modelagem segundo os componentes estocásticos e determinísticos. Essas limitações motivaram o desenvolvimento deste trabalho, que é detalhado no próximo capítulo.

CAPÍTULO

3

Plano de Pesquisa

3.1 Considerações iniciais

A importância do uso de séries temporais na análise do comportamento de sistemas é observada em trabalhos realizados nas mais diversas áreas da ciência (Capítulo 2). Conforme apresentado nos capítulos anteriores, pesquisadores propuseram técnicas e ferramentas que permitem analisar séries e encontrar modelos que descrevem seu comportamento (Box *et al.*, 1994; Alligood *et al.*, 1997). Uma vez encontrado o modelo adequado, pode-se determinar a regra (ou função) geradora da série, a qual é utilizada para compreender o comportamento das observações geradas pelo sistema em estudo.

Em geral, ferramentas de modelagem de séries temporais, quando aplicadas sobre dados experimentais sintéticos, tendem a produzir bons resultados visto que o mecanismo de geração das observações possui comportamento bem caracterizado. No entanto, quando aplicados sobre dados coletados a partir de sistemas reais, há uma grande divergência em relação aos resultados gerados pelo modelo utilizado. Essa divergência, geralmente, acontece devido à presença de ruídos na série, a qual possui parte do comportamento determinístico e parte estocástico.

Atualmente, existem diversas técnicas que permitem modelar o comportamento determinístico ou estocástico de séries temporais. No entanto, essas técnicas são aplicadas de maneira independente. Logo, em séries com presença de ambos comportamentos, a aplicação dessas técnicas não provê bons resultados. Nesse sentido, durante os estudos realizados e com base nos trabalhos apresentados ao longo deste plano de pesquisa, verificou-se que a utilização de uma abordagem híbrida, a qual modela tanto o comportamento determinístico quanto o estocástico, pode auxiliar na análise eficiente de séries temporais que apresentam ambos comportamentos.

3.2 Objetivo

O objetivo deste plano de pesquisa é comprovar a hipótese apresentada na Seção 1.2, a qual afirma que séries temporais podem ser segmentadas, segundo seus componentes estocásticos e determinísticos, a fim de que todas suas observações sejam compreendidas por meio de uma abordagem de modelagem híbrida. Espera-se que, com a aplicação dessa abordagem, modelos com maior acurácia sejam obtidos na análise de séries temporais.

A etapa de segmentação (Figura 3.1) deve analisar uma série temporal X_t por intermédio de uma função $F(.)$ de filtragem, a qual fornece como resultado dois conjuntos $O_d = F(X_t, d)$ e $O_e = F(X_t, e)$ de observações com comportamento determinístico e estocástico, respectivamente. Após segmentar a série, as observações que compõem os componentes estocásticos e determinísticos devem ser analisadas a fim de compreender seus comportamentos individuais, que combinados devem compor um modelo híbrido, possibilitando, assim, obter maior acurácia na modelagem de séries temporais.

A seção a seguir apresenta uma descrição formal do problema que motivou esse trabalho de doutorado e como resolvê-lo usando a abordagem proposta.

3.3 Descrição do problema

Considere um sistema S composto por um conjunto de componentes C, tal que $S = \{c_1, c_2, \ldots, c_i\}$, onde $i \in \mathbb{N}$. Considere, também, um conjunto imagem $M = \{m_1, m_2, \ldots, m_j\}$, onde $j \in \mathbb{N}$, que representa todos os possíveis modelos usados para descrever o sistema S. Os elementos dos conjuntos S e M são relacionados por meio de uma função $f(.)$ de

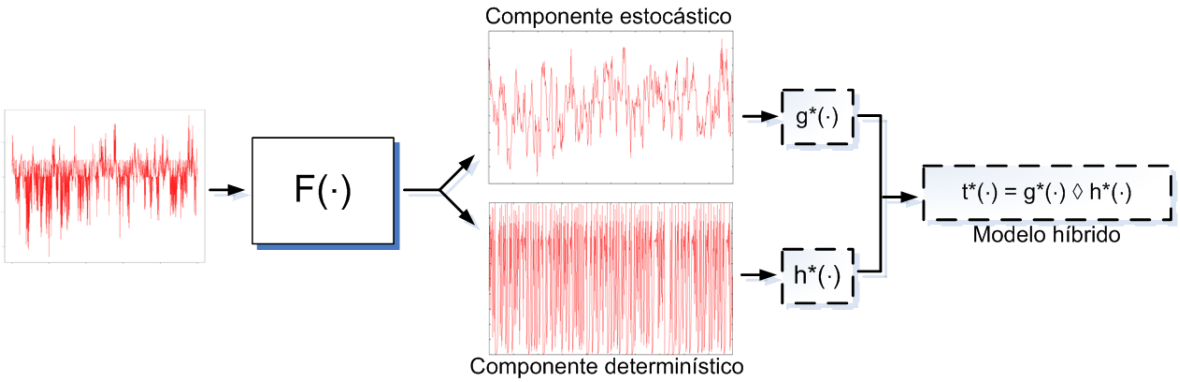


Figura 3.1: Execução do projeto.

modelagem, onde $f : S \rightarrow M$. É importante destacar, ainda, que para todo $m \in M$, existe ao menos um grupo de componentes $C' = \{c_p, c_{p+1}, \ldots, c_n\} \in S$ tal que $f(c_p, c_{p+1}, \ldots, c_n) = m_k$, onde $p, n \in [1, i]$ e $k \in [1, j]$.

Assume-se que com a aplicação da função $f(.)$ de modelagem sobre diferentes grupos de componentes C', obtém-se diferentes modelos m, que descrevem um sistema S com diferentes graus de acurácia. Esses graus de acurácia são medidos por uma função $\Lambda(.)$. Há, portanto, modelos $m \in M$ com baixa, média e alta acurácia. A partir disso, pode-se definir um modelo mínimo capaz de representar S com acurácia aceitável. Dessa maneira, determinados modelos, de baixa acurácia ou mesmo sem qualquer representatividade, podem, eventualmente, ser descartados. Os demais modelos representam diferentes níveis de precisão e podem ser aplicados em diferentes contextos para representar, estimar, predizer ou estudar comportamentos decorrentes de S.

Além da classificação de modelos com base na acurácia resultante, cada modelo $m \in M$ é construído com base em um conjunto de componentes C', o qual está contido no conjunto original de componentes C em S, ou seja, $C' \subseteq C$.

A estrutura de cada modelo m, $\forall\ m \in M$, pode ser definida por uma função $\Omega(.)$, a qual calcula a complexidade, em termos de espaço, necessária para representar S. Dessa maneira, seria ideal para qualquer sistema S encontrar um modelo m com a menor complexidade de espaço, $min(\Omega(m))$, $\forall\ m \in M$, e com maior acurácia resultante, $max(\Lambda(m))$, $\forall\ m \in M$. Em outras palavras, dado um conjunto de componentes $C = \{c_1, c_2, \ldots, c_i\}$ e $C' = \{c_p, c_{p+1}, \ldots, c_n\}$, para $p, n \in [1, i]$, o ideal é que $f(C') \simeq f(C)$. Esse modelo satisfaria o princípio da navalha de Occam (Thornburn, 1918), o qual também rege o princípio da

parcimônia comumente considerado pela área de Análise de Séries Temporais (Box *et al.*, 1994; Ariew, 1976).

A partir dessas definições apresentadas, considere um sistema S composto por componentes estocásticos e determinísticos em diferentes proporções. Dependendo dessas proporções, pode-se obter um modelo $m \in M$ que descarte um desses componentes e, mesmo assim, seja capaz de modelar o sistema S com acurácia mínima e com baixa complexidade de espaço, portanto, $f(C') \simeq f(C)$. Por outro lado, assumindo uma mistura desses componentes na mesma proporção, o descarte de um deles pode prejudicar a acurácia de m a ponto de não produzir resultados satisfatórios, mesmo aumentando a complexidade de seu espaço, ou seja, buscando adicionar mais componentes em C'.

Para solucionar esse problema, este trabalho de doutorado visa desenvolver uma função $f'(.)$ de modelagem que garanta a utilização dos componentes estocásticos e determinísticos, visando assegurar o princípio da parcimônia e fornecer um modelo m' com alta acurácia. De maneira geral, espera-se obter um modelo definido por $m' = f'(c_e, c_d)$, tal que c_e e c_d representam os componentes estocásticos e determinísticos, respectivamente, onde a acurácia do modelo híbrido seja maior e a complexidade de espaço seja semelhante aos modelos que consideram, isoladamente, ambos componentes.

Para exemplificar, considere um sistema que tenha produzido saídas segundo dois processos diferentes: o primeiro gera n observações segundo um processo estocástico $X'_t = \text{AR}(1)$, e o segundo gera, também, n observações, mas segundo um processo determinístico $X''_t = (b - x_{t-1}) \cdot x_{t-1}$, tal que b é uma constante. Considere, ainda, que o mecanismo de monitoramento desses processos agrupa, em uma única série temporal, as n observações de cada um desses processos. Em outras palavras, pode-se afirmar que a série temporal resultante X_t é uma mistura de observações de X'_t e X''_t, ou seja, $X_t = \Xi(A_{2n}^{X'_t \cup X''_t})$, sendo que A é um arranjo simples entre os diferentes processos X'_t e X''_t, e $\Xi(\cdot)$ é uma função que retorna uma única combinação das observações resultante desse arranjo.

Conforme descrito anteriormente, técnicas convencionais, geralmente, descartam um dos processos geradores da série X_t, o que pode resultar em modelos com baixa acurácia. No entanto, este trabalho de doutorado visa compreender cada processo de maneira individual e, então, gerar um modelo m' que considere o conjunto de componentes $C' = \{c_e, c_d\}$, tal que $c_e = X'_t$ e $c_d = X''_t$, ou seja, $m' = f'(X'_t, X''_t)$.

A seguir são apresentados os materiais e métodos, que serão utilizados no desenvolvimento deste projeto, os mecanismos de análise dos resultados obtidos e de validação da abordagem proposta.

3.4 Materiais, métodos e análise de resultados

A primeira etapa do desenvolvimento do projeto está relacionada à composição de uma base formada por diversas séries temporais sintéticas, cujos conceitos foram discutidos no Capítulo 2. Essa base de séries será utilizada para validar a abordagem proposta, a qual visa segmentar os componentes estocásticos e determinísticos.

Diversas técnicas serão estudadas e empregadas sobre a base de séries sintéticas, a fim de avaliar a segmentação resultantes dos componentes estocásticos e determinísticos. Em estudos iniciais empregando Cadeias de Markov, ACF, abordagens estatísticas, de sistemas dinâmicos e redes neurais artificiais, observou-se que a ferramenta *Recurrence Plot* apresentou resultados satisfatórios. Isso se deve à sua capacidade de gerar estruturas, cujos formatos fornecem informações sobre a taxa de estocasticidade, com a presença de pontos isolados, e determinismo, com a presença de linhas diagonais. Com a geração de estruturas pela ferramenta RP, foram realizados estudos na área de Filtragem de Imagens, os quais buscaram identificar ferramentas para remover pontos isolados dessas estruturas, aumentando a taxa de determinismo. Com a aplicação dessas ferramentas, pode-se segmentar séries temporais, destacando seus componentes estocásticos e determinísticos.

Após essa segmentação, serão avaliadas técnicas para modelar componentes estocásticos e determinísticos resultantes da etapa anterior. A partir disso, serão criados modelos híbridos para analisar as séries em estudo. Finalmente, pretende-se avaliar as influências que componentes estocásticos e determinísticos causam na acurácia dos modelos híbridos obtidos.

Essa avaliação será realizada utilizando uma função $\Lambda(.)$, que forneça informações sobre a acurácia, e uma função $\Omega(.)$, que quantifique a complexidade de espaço. De maneira geral, há abordagens que atuam sobre o componente estocástico (c_e), as quais permitem obter um modelo denotado por $m_e = f(c_e)$, tal que $m_e \in M$ (em que M é o conjunto de todos os possíveis modelos). Da mesma forma, há outras abordagens voltadas para modelar o

componente determinístico (c_d), as quais geram um modelo definido por $m_d = f(c_d)$, tal que $m_d \in M$. Por sua vez, a modelagem híbrida visa considerar ambos componentes para gerar um outro modelo denotado por $m_k = f(c_e, c_d)$, tal que $m_k \in M$. Com a utilização das funções $\Lambda(.)$ e $\Omega(.)$ de avaliação, espera-se encontrar modelos híbridos com maior acurácia do que modelos de componentes individuais ($\Lambda(m_k) > \Lambda(m_z)$, onde $z = \{e, d\}$), cuja complexidade de espaço não seja muito maior, ou seja, $\Omega(m_k) \simeq \Omega(m_z)$, onde $z = \{e, d\}$.

Espera-se validar o modelo obtido com a base de séries temporais sintéticas e, posteriormente, empregá-los sobre séries reais tais como climáticas, epidemiológicas, etc. Os resultados obtidos serão comparados às demais abordagens propostas na literatura.

3.5 Cronograma

A execução deste plano de pesquisa está dividida nas seguintes etapas:

1. Séries temporais: embora essa etapa já tenha sido iniciada, é importante realizar um estudo contínuo sobre fundamentos de séries temporais, bem como os principais modelos para descrevê-las e analisá-las;

2. Sistemas Dinâmicos: deve-se continuar os estudos realizados sobre as principais ferramentas da área de Sistemas Dinâmicos e Teoria do Caos, que permitem descrever e modelar séries temporais;

3. Segmentação: nessa etapa, pretende-se realizar um estudo detalhado sobre abordagens de segmentação de séries temporais;

4. Modelagem Híbrida: nessa etapa, pretende-se estudar técnicas utilizadas para modelar o comportamento de séries temporais. O principal objetivo é propor uma nova abordagem que permita caracterizar os componentes estocásticos e determinísticos de séries temporais;

5. Avaliação do Modelo Híbrido: o modelo híbrido será avaliado utilizando séries sintéticas e, posteriormente, empregado sobre séries reais. Nessa etapa, pretende-se elaborar um estudo em parceria com outros institutos a fim de obter séries temporais reais, cuja modelagem e compreensão forneçam contribuição relevante para diferentes áreas de pesquisa.

Na Tabela 3.1 é apresentado o cronograma deste plano, sendo que cada símbolo representa um bimestre e a coluna dos anos é dividida em semestres. O símbolo • significa que a tarefa foi iniciada, enquanto que o símbolo ◦ representa tarefas por fazer. Cabe ainda ressaltar que o desenvolvimento deste trabalho foi feito com base em um cronograma de 3 anos, no entanto as tarefas realizadas no ano de 2009 são, também, apresentadas, a fim de prover uma visão geral sobre o andamento do projeto.

Tabela 3.1: Cronograma de atividades

Atividades	2009		2010		2011		2012		2013
1)Cumprimento de créditos de disciplinas	●●	●●	●●						
2)Estudo sobre séries temporais	●●	●●	●●●						
3)Estudo sobre Sistemas Dinâmicos	●	●●●	●●						
4)Segmentação de componentes de séries temporais			●●	○○○	○○○	○○○	○○○		
5)Preparação da Qualificação			●●●						
6)Obtenção e análise de séries temporais			●●●	○○○	○○○				
7)Desenvolvimento de uma abordagem de modelagem híbrida					○○○	○○○	○○○	○○	
8)Experimentos e validação da abordagem						○○○	○○○	○○○	
9)Estudo de caso com séries temporais reais						○○	○○○	○○○	
10)Escrita de artigos com os resultados obtidos					○	○	○○	○○○	○
11)Escrita da tese							○	○○○	○
12)Defesa da tese									○

CAPÍTULO

4

Atividades Desenvolvidas

4.1 Considerações iniciais

O estudo do comportamento de sistemas ao longo do tempo possibilitou o desenvolvimento de uma nova área chamada Análise de Séries Temporais. Essa área distingue-se da estatística clássica, pois, assume relações temporais entre observações geradas por sistemas. Diversas pesquisas foram desenvolvidas buscando modelar tais relações, a fim de descrever como observações se comportam e como predizê-las (Shumway e Stoffer, 2006).

Nesse sentido, o estudo apresentado no Capítulo 2 forneceu uma visão geral sobre as principais áreas que fornecem técnicas de modelagem e compreensão do comportamento de sistemas, cujas observações são organizadas por meio de séries temporais. A aplicação de técnicas estatísticas tem se mostrado eficiente na modelagem de séries com comportamento estocástico. Por outro lado, séries temporais com comportamento determinístico têm sido modeladas eficientemente com o uso de técnicas de sistemas dinâmicos e teoria do caos. No entanto, entende-se que modelos mais precisos poderiam ser obtidos se os componentes estocásticos e determinísticos fossem ponderados de acordo com suas

influências sobre o comportamento global do sistema analisado. Para tanto, faz-se necessário segmentar séries temporais, visando conhecer, separadamente, as observações que influenciam na formação desses componentes. Após essa segmentação, cada componente pode ser modelado sem que suas observações sofram interferência do comportamento do outro componente.

Visando alcançar esse objetivo, foram realizados dois estudos experimentais. No primeiro, abordou-se um caso ideal (*toy problem*), no qual uma série temporal sintética foi construída considerando dois processos geradores distintos. Esses processos não interagem na criação das observações, resultando, assim, em uma série com dois comportamentos diferentes. Entende-se que essa série não é, normalmente, encontrada em situações reais, mas a sua utilização foi importante na realização de um dos primeiros experimentos, a fim de comprovar a ideia de que a modelagem individual de componentes produz bons resultados.

Por outro lado, o segundo estudo experimental foi realizado com o objetivo de comprovar a possibilidade de segmentação de séries segundo seus componentes estocásticos e determinísticos. Nesse estudo, diversas séries temporais foram segmentadas usando ferramentas de filtragem em conjunto com técnicas de sistemas dinâmicos. As seções seguintes apresentam em detalhes esses estudos experimentais.

4.2 Estudo de caso

Nesta seção é apresentado um estudo de caso ideal, cujo principal objetivo foi comprovar se a modelagem individual dos componentes estocásticos e determinísticos produzia bons resultados. Os experimentos executados nesse estudo são divididos em duas fases. Na primeira, é realizada uma análise das observações, buscando obter um modelo que descreva seu comportamento. Esse modelo é utilizado na fase seguinte, a qual visa predizer observações futuras.

Para execução desse estudo, uma série temporal foi construída utilizando dois processos diferentes. O primeiro processo P_e, que possui comportamento estocástico, é um modelo autoregressivo de ordem $p = 1$ e o segundo processo P_d, que possui comportamento determinístico, é um mapa logístico. Foram geradas 500 observações de cada um desses

processos, as quais foram agrupadas, de maneira alternada, resultando em uma única série temporal X_t (Figura 4.1).

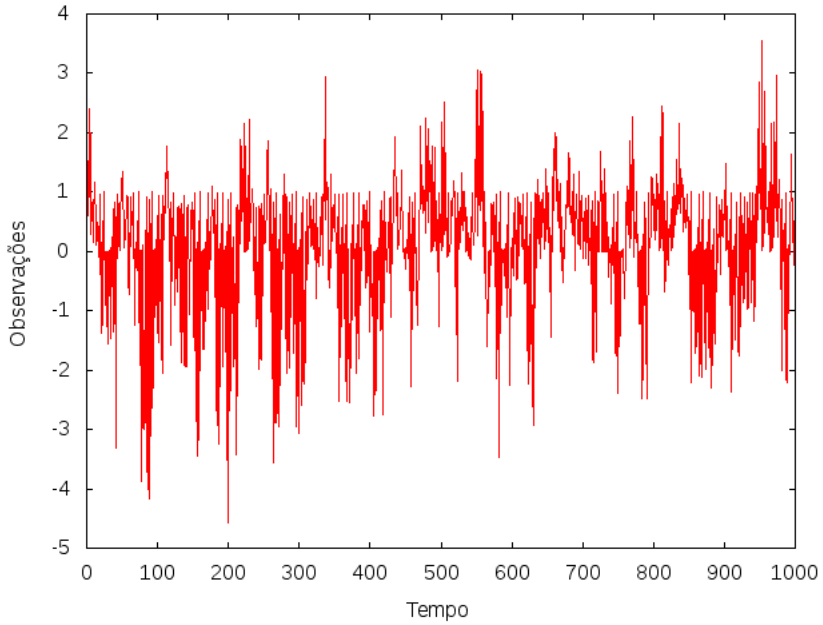


Figura 4.1: Série temporal gerada a partir de dois processos: estocástico e determinístico.

O primeiro passo executado para entender o comportamento da série X_t foi desdobrá-la no espaço de coordenadas de atraso (seguindo a abordagem típica de Sistemas Dinâmicos e Teoria do Caos – Capítulo 2). Para tanto, a série foi analisada por intermédio das ferramentas *Auto Mutual Information* (AMI) (Kennel *et al.*, 1992) e *False Nearest Neighbors* (FNN) (Kennel *et al.*, 1992) (maiores detalhes sobre essas abordagens na Seção 2.3.2), as quais forneceram, respectivamente, os seguintes valores para dimensões de separação e embutida: $\tau = 1$ e $m = 3$. Como pode ser visto na Figura 4.2, mesmo com o desdobramento da série no espaço de coordenadas de atraso, não é possível identificar um comportamento regular (tal como observado nos exemplos apresentados na Seção 2.3.2).

Diante da ausência de regularidade na série, a predição de observações futuras, utilizando técnicas de modelagem tradicionais, tendem a fornecer resultados insatisfatórios. Para comprovar essa afirmação, a série X_t foi analisada por 3 diferentes técnicas de modelagem. A primeira análise foi realizada pelo mesmo modelo utilizado na criação das observações originadas do processo P_e, i.e., o modelo AR(p)[1], com $p = 1$. A segunda e ter-

[1]As etapas de ajuste do modelo e predição de pontos futuros foram desenvolvidas na ferramenta R, baseando-se

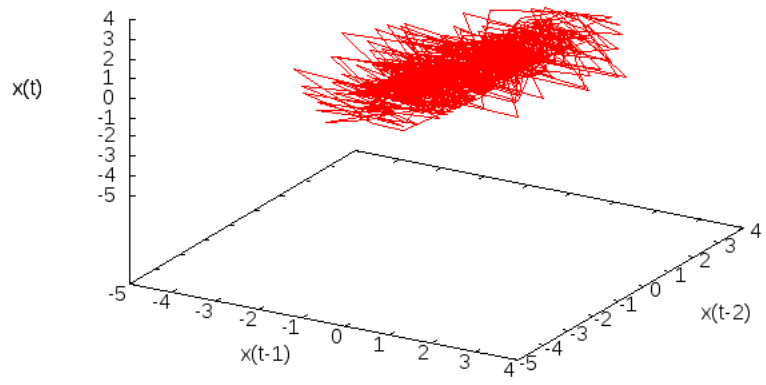


Figura 4.2: Desdobramento da série no espaço de coordenadas de atraso com $m = 3$ e $\tau = 1$.

ceira análises foram realizadas por modelos utilizados na modelagem de comportamento determinístico, tal como o gerado pelo processo P_d. Para tanto, foram adotados os modelos *Radial Basis Function* (RBF) e *Polynomial*[2], os quais são explicados a seguir. Para as três análises, as 900 observações iniciais da série foram utilizadas no ajuste do modelo e as 100 observações finais foram utilizadas para validação da predição.

Sobre a análise com o modelo estatístico AR(1), os ajustes sobre as 900 observações iniciais foram realizados pelo método Yule-Walker (Box *et al.*, 1994). Os resultados das observações preditas a partir dessa análise possuem baixa acurácia, como visto na Figura 4.3. Nessa figura, a linha tracejada vertical indica o instante em que as observações começaram a ser preditas e a linha em azul representa os valores das observações preditas. Apenas com uma análise visual, pode-se identificar que os valores preditos seguiram o valor médio das observações da série.

A segunda análise foi realizada utilizando uma Função de Base Radial (*Radial Basis Function* – RBF) (Mitchell, 1997). RBF é uma função de aproximação que realiza uma regressão de acordo com uma distância ponderada das observações analisadas. De maneira

no processo de predição apresentado por Shumway e Stoffer (2006)

[2]Para execução desses experimentos, foram utilizadas as implementações disponíveis na ferramenta TISEAN (Hegger *et al.*, 1999)

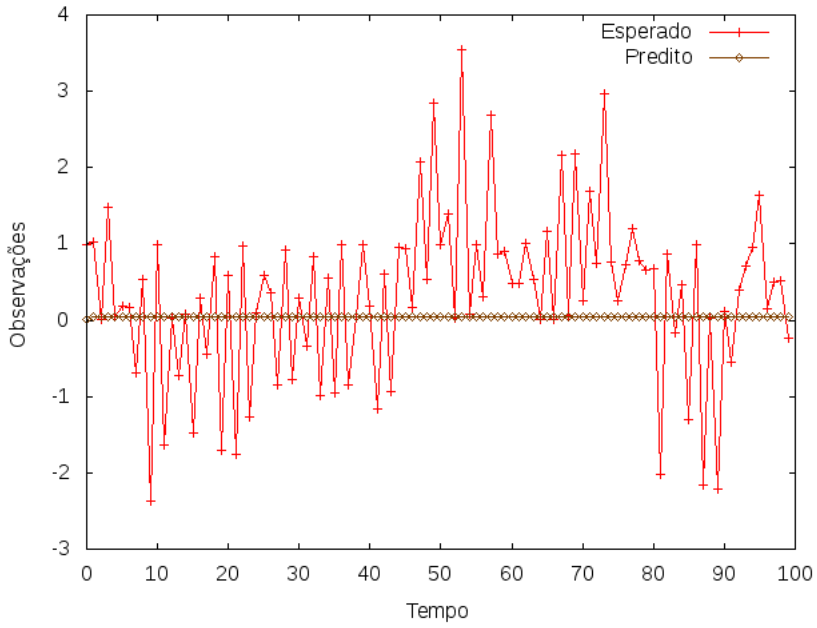


Figura 4.3: Predição de pontos futuros utilizando o modelo AR(1) sem segmentação. A linha tracejada vertical representa o instante em que as observações começaram a ser preditas. A linha em azul representa as observações preditas e as linhas em verde e vermelho são os erros padrão de predição.

geral, a RBF visa estimar uma função $\hat{f}(.)$ que produza saídas semelhantes à função $f(.)$, a qual originou a série analisada. Essa função estimada é definida pela Equação 4.1.

$$\hat{f}(x_n) = w_0 + \sum_{i=1}^{k} w_i K(d(x_i, x_n)) \tag{4.1}$$

Nessa equação, k representa a quantidade de centros, w_i é um conjunto de pesos associados a cada centro e $K(.)$ representa uma função[3], cujo valor é reduzido à medida que a distância $d(.)$ entre as observações x_i e x_n aumenta. De maneira geral, o objetivo da RBF é determinar quais grupos de observações estão associados a cada centro, sendo que uma observação x_n é associada àquele centro x_i mais próximo.

No entanto, o número de centros utilizados pela função RBF não é simples de ser estimado. Para esse estudo de caso, esse número foi escolhido com base em uma análise empírica. Para tanto, as fases de modelagem e predição da série X_t foram executadas diversas vezes com a RBF, variando o número de centros ($k \in [2, 100]$). Em seguida, para cada predição realizada, foi calculado o Erro Quadrático Médio (*Mean Squared Error* – MSE)

[3]Emprega-se, geralmente, uma função Gaussiana.

(Mood *et al.*, 1974). Contudo, para evitar que o resultado dessa medida de erro fosse influenciado pela diferença de amplitude das observações, foi realizada uma normalização conforme apresentado na Equação 4.2, sendo $\{x_i\} \in X_t$ as observações da série modelada, $\{\hat{x}_i\} \in \hat{X}_t$ as observações preditas, e $min(X_t)$ e $max(X_t)$ o maior e menor valor, respectivamente, das observações de X_t.

$$\text{NMSE}(\hat{X}_t, X_t) = \frac{\sum_{i=1}^{N}(\hat{x}_i - x_i)^2}{n \times (max(X_t) - min(X_t))} \tag{4.2}$$

Os resultados obtidos do cálculo do NMSE estão representados, graficamente, na Figura 4.4. Para a série X_t analisada, a melhor estimativa para o número de centros foi 31.

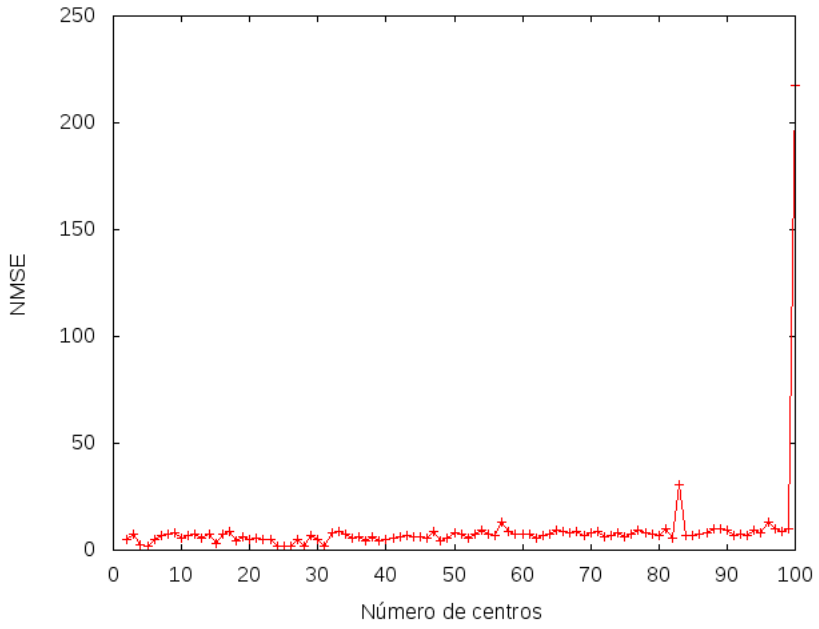


Figura 4.4: Cálculo do NMSE utilizado na escolha do melhor número de centros da RBF.

Baseando-se na melhor estimativa de centros, a fase de predição com a RBF foi realizada, visando prever 100 observações futuras da série X_t (Figura 4.5). A fim de facilitar a visualização, a Figura 4.5 apresenta apenas as observações preditas e as esperadas, omitindo as observações utilizadas no processo de treinamento. Como pode ser visto, a acurácia deste modelo, também, é baixa, resultando em grande divergência entre os pontos preditos e os observados.

A terceira análise foi realizada por intermédio da ferramenta *Polynomial* (Casdagli,

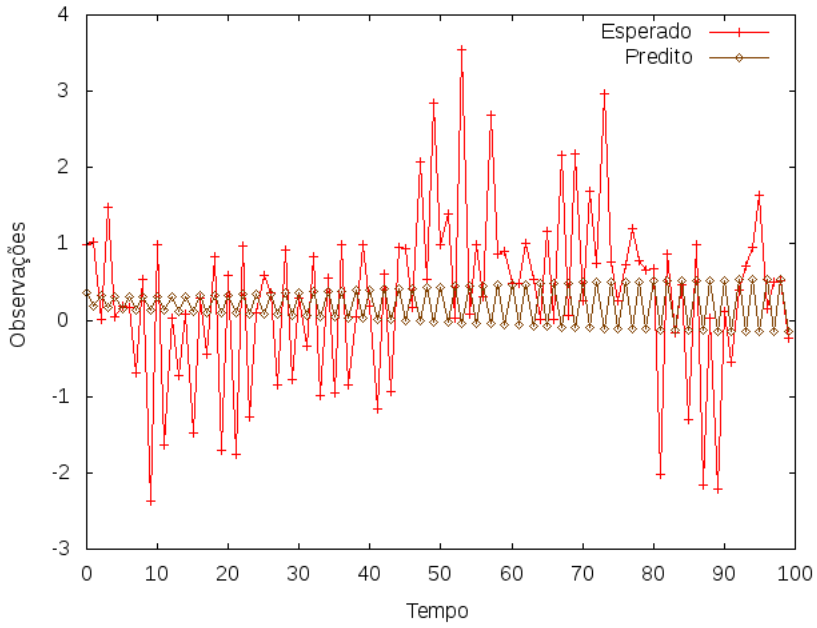


Figura 4.5: Predição de pontos futuros utilizando RBF sem segmentação.

1989). Essa ferramenta é um preditor polinomial, cuja função base é definida por $\pi_i f_N \colon \mathbb{R}^m \to \mathbb{R},\ i = 1, \ldots, m$, onde m representa as variáveis de grau (no máximo d). Os parâmetros livres do polinômio $\binom{m+d}{m} \equiv \frac{(m+d)!}{m!d!}$ são escolhidos a fim de reduzir a equação $\sum_{n=1}^{N-1}(\pi_i x_{n+1} - \pi f_N(x_n))^2$ (Casdagli, 1989). De maneira geral, o principal objetivo da ferramenta *Polynomial* é fornecer um preditor f_N, cujos valores de saída sejam próximos da função que deu origem às observações da série (Casdagli, 1989).

Assim como nas demais análises com as ferramentas AR(1) e RBF, os resultados obtidos com o preditor polinomial, Figura 4.6, apresentaram grande divergência face aos valores esperados, tendendo, também, ao valor médio das observações.

Dados os resultados com baixa acurácia obtidos com as modelagens e predições anteriores, novas análises foram conduzidas considerando a segmentação de séries, tal como discutido nos capítulos anteriores. Com essa segmentação, as observações que foram originadas do processo P_e foram separadas das demais observações geradas a partir do processo P_d. As observações foram separadas manualmente, dado que se conhecia o processo gerador de cada observação da série. É importante ressaltar que, nesse momento, o objetivo dos experimentos é realizar uma análise da modelagem individual dos componentes, ou seja, dos processos P_e e P_d, e não avaliar a segmentação de tais componentes.

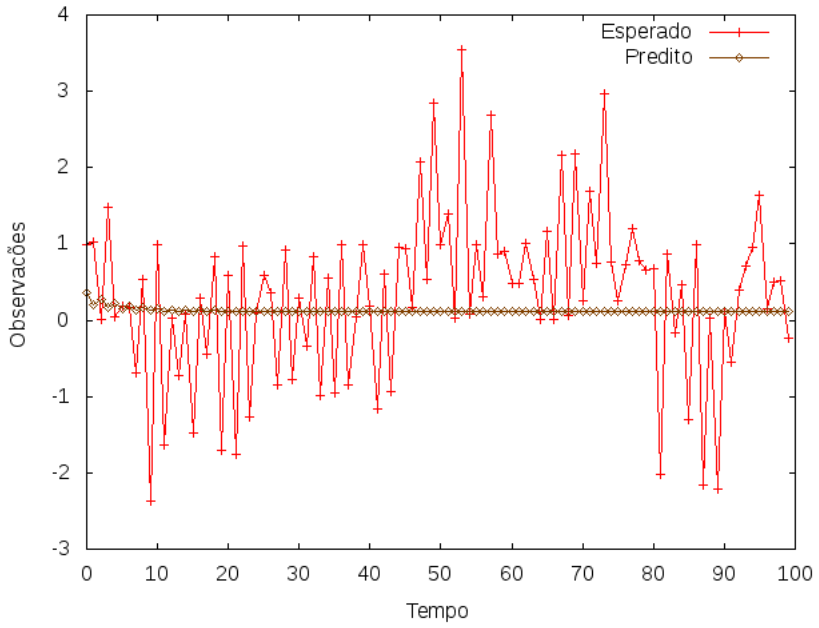


Figura 4.6: Predição de pontos futuros utilizando *Polynomial* sem segmentação.

Esses experimentos, assim como os anteriores, são divididos em duas fases. Na primeira, as 450 observações iniciais de cada componente foram utilizadas no ajuste do modelo e as 50 observações finais foram utilizadas na validação.

Na análise do processo P_e, cujo comportamento é caracterizado pela presença de estocasticidade, utilizou-se o modelo AR(1), o qual foi ajustado por meio do método Yule-Walker Box *et al.* (1994). Para esse experimento, contudo, mesmo utilizando o modelo e a ordem adotados na criação das observações, os resultados preditos não foram satisfatórios (Figura 4.7). Isso acontece devido à natureza do modelo AR(p), necessitando, assim, de uma busca mais detalhada por um modelo que represente as observações estocásticas de forma mais apropriada.

Por outro lado, na análise do processo P_d, cujo comportamento de suas observações possuem características determinísticas, os resultados dos experimentos realizados, tanto com a RBF quanto com a ferramenta *Polynomial*, apresentaram maior acurácia na fase de predição.

No caso da RBF, assim como discutido anteriormente, deve-se, inicialmente, escolher a quantidade adequada de centros. Para a modelagem do processo P_d, a quantidade que forneceu maior acurácia na predição foi 41. Esse valor foi estimado com base no cálculo do

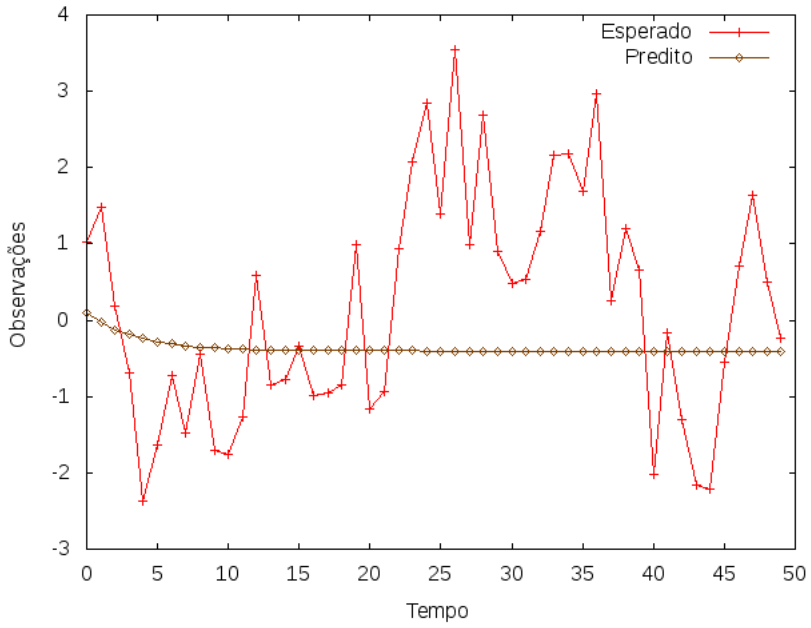


Figura 4.7: Predição de pontos estocásticos utilizando o modelo AR(1).

NMSE (Figura 4.8).

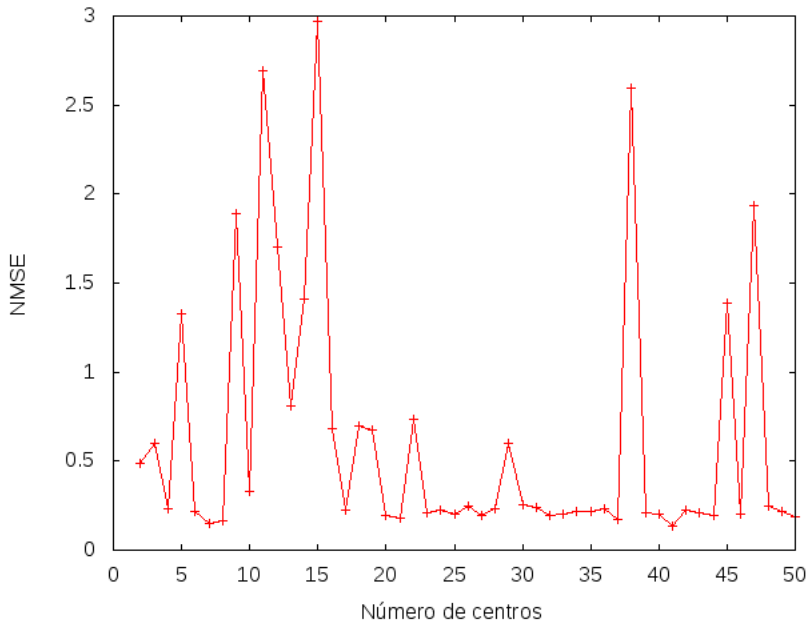


Figura 4.8: Cálculo do NMSE utilizado na estimação do número de centros da técnica RBF.

A Figura 4.9 apresenta os valores esperados e preditos por meio da abordagem RBF.

Conforme visto, o modelo obtido a partir dessa função forneceu resultados satisfatórios, ou seja, apresentou melhores resultados com a série segmentada do que com a série original.

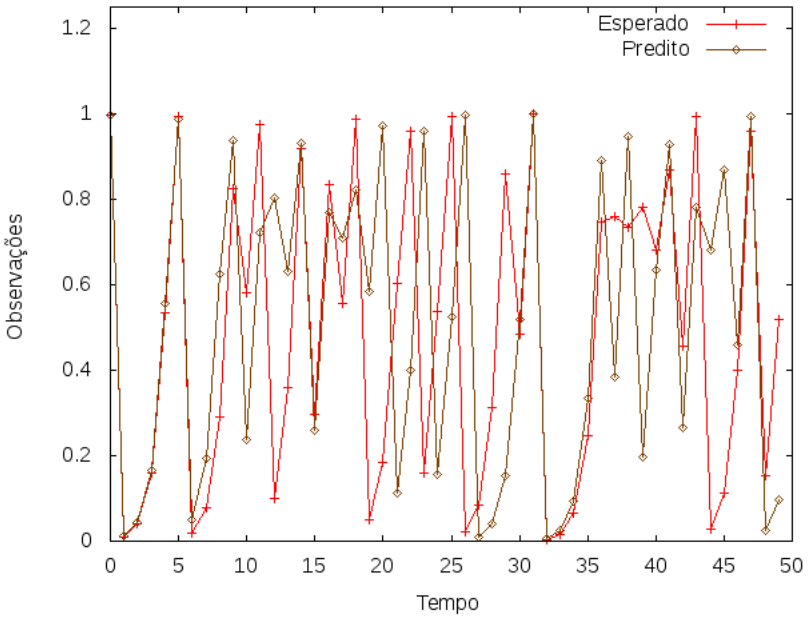


Figura 4.9: Predição de pontos determinísticos utilizando RBF.

Por fim, o componente definido pelo processo P_d foi analisado, ainda, com a ferramenta *Polynomial* (Figura 4.10). Como pode ser observado, os resultados preditos são muito precisos em relação ao conjunto de observações utilizadas na validação. Isso demonstra que o modelo utilizado foi capaz de representar o componente determinístico com alta precisão. Uma outra consideração acerca desse experimento é o aumento do erro de predição à medida que se aumenta o número de observações preditas. Esse aumento deve-se aos limites inferior e superior de predição, que foram discutidos na Seção 2.2.

Com base nos experimentos iniciais apresentados nesta seção, pode-se concluir que, apesar do estudo de um caso com dois comportamentos bem distintos, ao se segmentar de maneira adequada uma série e analisar, individualmente, seus componentes, pode-se obter modelos de maior acurácia. Além disso, é importante ressaltar que esse experimento exemplificou a execução da abordagem proposta em uma situação onde as observações da série pertencem somente a um grupo, ou seja, cada observação possui apenas comportamento estocástico ou determinístico. No entanto, entende-se que em situações reais, uma mesma observação, poderia ser influenciada, simultaneamente, pelos dois compo-

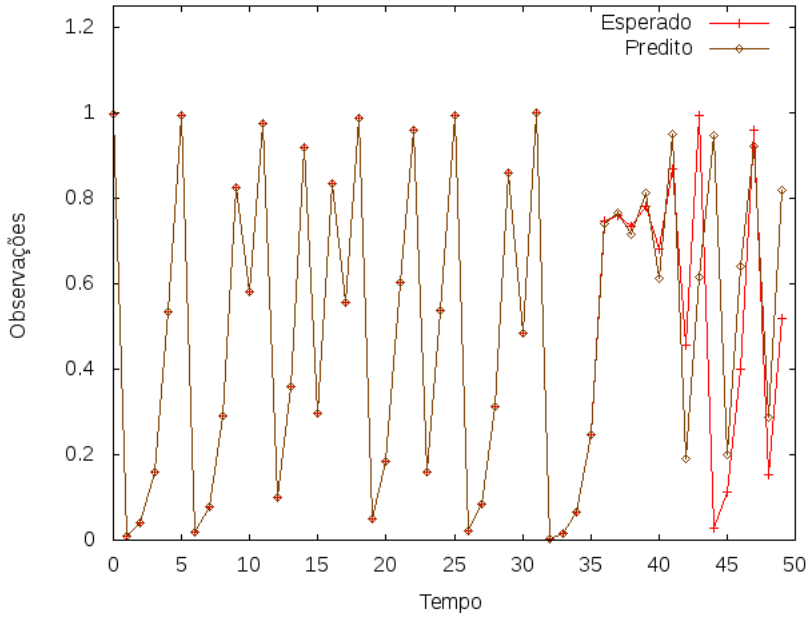


Figura 4.10: Predição de pontos determinísticos utilizando *Polynomial*.

nentes. Nessa situação, deveria-se segmentar cada observação individual segundo esses componentes.

4.3 Experimentos Preliminares

Um segundo estudo experimental foi realizado com o objetivo de avaliar a viabilidade da segmentação de componentes estocásticos e determinísticos presentes em séries temporais. Nesse sentido, estudos indicam que a ferramenta *Recurrence Plot* (RP) (maiores detalhes no Apêndice A) permite compreender esses componentes de maneira isolada. Essa ferramenta analisa uma série temporal desdobrada em seu espaço de coordenadas de atraso, formando uma matriz de recorrência. Com base nas estruturas formadas por essa matriz, RP estima o nível de estocasticidade e determinismo presente na série. De maneira geral, essas estruturas podem ser interpretadas da seguinte forma: pontos isolados representam ruídos, enquanto que linhas diagonais representam comportamento determinístico.

Nesse sentido, partindo do pressuposto que ao se remover pontos isolados da matriz de recorrência aumenta-se o nível de determinismo da série, foi desenvolvido um estudo inicial utilizando ferramentas de processamento de imagens (Apêndice B). Esse estudo for-

neceu indícios que a utilização dessas ferramentas em conjunto com técnicas de sistemas dinâmicos pode auxiliar na segmentação de componentes presentes em séries analisadas.

Sendo assim, esta seção apresenta experimentos preliminares, os quais foram desenvolvidos visando avaliar a acurácia de técnicas de processamento de imagens na filtragem de pontos isolados presentes nas estruturas geradas pela ferramenta RP. A metodologia adotada nesses experimentos consistiu da execução da ferramenta RP sobre uma série, a qual gerou uma matriz M_{RR} de recorrência e forneceu valores sobre a taxa de determinismo. Em seguida, a matriz M_{RR} foi analisada por meio do emprego de filtros individuais de processamento de imagens ($F(\cdot)$), em que cada filtro gera uma nova matriz ($M^D_{RR} = F(M_{RR})$). Essa matriz resultante é composta apenas de trajetórias recorrentes que representam o comportamento determinístico da série. Por outro lado, a matriz composta apenas pelas trajetórias que representam o comportamento estocástico (M^E_{RR}) pode ser obtida pela diferença entre a matriz de recorrência original e a matriz determinística obtida pelo processo de filtragem, ou seja, $M^E_{RR} = M_{RR} - M^D_{RR}$.

Atualmente existem diversas implementações da ferramenta RP disponíveis na literatura como, por exemplo, (Marwan, 2010, 2009; Narzo, 2010; Webber, 2010). Entretanto, para os testes apresentados neste capítulo, essas ferramentas não se mostraram adequadas, pois foram desenvolvidas com o objetivo de executar diretamente sobre séries temporais. Devido ao processo de filtragem realizado nos experimentos apresentados nesta seção, foi necessário desenvolver uma nova implementação da ferramenta RP para processar, também, matrizes de recorrência, além do processamento aplicado sobre séries temporais. Essa implementação foi validada com base nos resultados obtidos do processamento de diversas séries temporais com as ferramentas de análise de recorrência disponibilizadas em (Marwan, 2009, 2010), as quais foram desenvolvidas com base nos estudos apresentados em (Marwan *et al.*, 2007).

As séries utilizadas nesses experimentos e na validação da implementação da ferramenta RP são sintéticas, cujo objetivo principal é observar a presença de estocasticidade e determinismo, e visualizar o processo de identificação de cada um desses componentes. As seções a seguir apresentam resultados obtidos e considerações acerca do uso das técnicas de filtragem utilizadas.

4.3.1 Experimentos com filtro de mediana

Um primeiro conjunto de experimentos foi desenvolvido com base no estudo realizado com o filtro de mediana. Conforme discutido no Apêndice B, a execução desse filtro consiste na substituição de *pixels* pelo valor da mediana de sua vizinhança. No entanto, além de verificar a eficácia da aplicação desse filtro na remoção de ruídos de séries, os experimentos realizados mostraram-se importantes para identificar o tamanho da vizinhança utilizada. Essa vizinhança, no caso do filtro de mediana, é composta de valores ímpares e, nos experimentos, seu tamanho variou de acordo com $n = (3 \times 3, 5 \times 5, 7 \times 7, \ldots, 23 \times 23)$, cujo valor máximo utilizado foi obtido de forma empírica.

Os experimentos apresentados na Tabela 4.1 foram obtidos com a análise de uma série temporal, composta por 1000 observações, gerada a partir de um processo autoregressivo de ordem $p = 0,7$, discutido na Seção 2.3.1. Com base nos resultados das ferramentas FNN e AMI, apresentadas na Seção 2.3.2, estimou-se que essa série possui dimensão de separação igual a $\tau = 8$ e embutida igual a $m = 4$.

A taxa de determinismo obtida pela aplicação direta da ferramenta RP sobre essa série desdobrada foi TxDET $= 0,384375$. Na Tabela 4.1, n representa o tamanho da vizinhança e TxEST é o valor da da taxa de estocasticidade, ou seja, TxEST $= (1 -$ TxDET$)$. Esse valor apresentado em TxEST é resultado do processo de filtragem da matriz de recorrência da série e posterior aplicação da ferramenta RP sobre a matriz filtrada. É importante ressaltar que após a aplicação do filtro, espera-se obter uma matriz com comportamento estocástico e outra com comportamento determinístico. A aplicação da RP para obtenção de TxEST é realizada na matriz estocástica.

Como pode ser observado na Tabela 4.1, a aplicação do filtro possibilitou separar, de maneira mais adequada, o componente determinístico. Além disso, esse experimento mostra que para essa série temporal a vizinhança de tamanho 3×3 gera melhores resultados.

Um segundo experimento foi realizado com uma série temporal composta por 1000 observações geradas a partir de um processo autoregressivo e de média móvel com ordens iguais a $p = 0,7$ e $q = 0,2$. A aplicação das ferramentas AMI e FNN para estimação das dimensões de separação e embutida forneceram valores iguais a $\tau = 4$ e $m = 6$, respectivamente. A taxa de determinismo obtida pela aplicação direta da ferramenta RP sobre

Tabela 4.1: Filtragem da Série Temporal $AR(p = 0,7)$ utilizando filtro de mediana

n	TxEST	TxDET
3×3	0,874225	0,125775
5×5	0,660318	0,339682
7×7	0,622525	0,377475
9×9	0,618894	0,381106
11×11	0,618352	0,381648
13×13	0,618594	0,381406
15×15	0,619006	0,380994
17×17	0,619133	0,380867
19×19	0,619136	0,380864
21×21	0,619280	0,380720
23×23	0,619549	0,380451

essa série desdobrada foi TxDET $= 0,078347$ e os resultados de filtragem dessa série são apresentados na Tabela 4.2.

Tabela 4.2: Filtragem da Série Temporal $ARMA(p = 0,7; q = 0,2)$ utilizando filtro de mediana

n	TxEST	TxDET
3×3	0,939611	0,060389
5×5	0,922074	0,077926
7×7	0,922066	0,077934
9×9	0,922120	0,077880
11×11	0,922899	0,077101
13×13	0,923358	0,076642
15×15	0,923411	0,076589
17×17	0,924085	0,075915
19×19	0,924138	0,075862
21×21	0,924399	0,075601
23×23	0,924659	0,075341

Nesse caso, como visto na tabela, a aplicação do filtro de mediana não forneceu melhores resultados na separação dos componentes em relação ao que foi obtido com a aplicação direta da ferramenta RP na série. Além disso, a variação no tamanho da vizinhança não forneceu resultados significativamente diferentes.

Os experimentos apresentados foram realizados utilizando séries temporais com comportamento predominantemente estocástico. O próximo experimento (Tabela 4.3), no entanto, apresenta resultados da aplicação do filtro sobre uma série temporal determinística. Essa série é composta por um conjunto de dados gerados pelo atrator de Lorenz (Boker, 2010), a qual foi apresentada na Figura 2.6 (Seção 2.3.2). Essa série é composta de 1000 observações e os valores das dimensões de separação e embutida são $\tau = 5$ e $m = 3$, respec-

tivamente. Nesse caso, a taxa de determinismo obtida pela aplicação direta da ferramenta RP sobre essa série desdobrada foi TxDET $= 0,926170$.

Tabela 4.3: Filtragem da Série Temporal Lorenz (Boker, 2010) utilizando filtro de mediana

n	TxEST	TxDET
3×3	0,181373	0,818627
5×5	0,079763	0,920237
7×7	0,074774	0,925226
9×9	0,0750000	0,925000
11×11	0,075200	0,924800
13×13	0,075400	0,924600
15×15	0,075974	0,924026
17×17	0,078049	0,921951
19×19	0,078764	0,921236
21×21	0,079104	0,920896
23×23	0,079797	0,920203

Os resultados permitem concluir que a filtragem de séries com grande parte do comportamento determinístico utilizando filtro de mediana não provê bons resultados. Isso indica que a matriz de recorrência gerada pela ferramenta RP, aplicada diretamente sobre a série, tende a gerar poucos pontos isolados e o processo de remoção desses pontos por intermédio do filtro de mediana tende a gerar mais ruídos do que, de fato, removê-los. Com relação ao tamanho da vizinhança, para essa série analisada, valores acima de 7×7 não apresentam melhoria na separação dos componentes estocásticos e determinísticos.

4.3.2 Experimentos com Detecção de linhas

Um segundo conjunto de experimentos foi desenvolvido utilizando um filtro de detecção de linhas diagonais. O principal objetivo desse filtro é destacar comportamentos determinísticos presentes em séries, enfatizando as linhas diagonais nas estruturas geradas pela ferramenta RP. Conforme discutido no Apêndice B, a máscara utilizada para alcançar esse objetivo foi a máscara diagonal, apresentada na Figura B.6, cujo ângulo é de -45^o. Além disso, antes da aplicação do filtro, as últimas linhas e colunas da matriz foram duplicadas e os valores das novas linhas e colunas foram substituídos por zero, a fim de permitir que os valores das bordas fossem calculados. Além disso, assim como a necessidade de definição do tamanho da vizinhança para o filtro de mediana, o filtro de detecção de linhas necessita de um valor de *threshold*. Esse *threshold* permite transformar os valores da ma-

triz filtrada em binário, tal como a saída da ferramenta RP. A escolha inadequada desse *threshold* pode adicionar ruídos à série, comprometendo a utilização dessa ferramenta. Para os experimentos executados nesta seção, foi definido, de forma empírica, um valor de *threshold* igual a 3.

No primeiro experimento realizado com esse filtro, foi analisada a série temporal de Lorenz. Essa série, assim como nos experimentos com filtro de mediana, foi desdobrada de acordo com as dimensões de separação ($\tau = 5$) e embutida ($m = 3$). Para essa série, os valores das taxas de recorrência e de determinismo encontrados pela ferramenta RP foram, respectivamente, TxRR $= 0,005911$ e TxDET $= 0,926170$. Após a aplicação do filtro de detecção de linhas o valor da taxa de recorrência diminuiu para TxRR $= 0,004936$, enquanto que o valor da taxa de determinismo aumentou para TxDET $= 0,974056$. No primeiro caso, a taxa de recorrência diminuiu porque houve uma remoção dos pontos isolados, consequentemente, reduzindo a taxa de trajetórias recorrentes. A redução nessa taxa influenciou diretamente na taxa de determinismo, que aumentou devido à remoção dos pontos considerados ruidosos.

Um segundo experimento preliminar com o filtro de detecção de linhas foi executado sobre a série temporal $AR(p = 0,7)$. Nessa série, as dimensões de separação e embutidas estimadas foram $\tau = 8$ e $m = 4$, respectivamente. Os valores das taxas de recorrência e de determinismo encontrados por meio da aplicação da ferramenta RP sobre essa série desdobrada foram, respectivamente, TxRR $= 0,092744$ e TxDET $= 0,384375$. Após a aplicação do filtro de detecção de linhas, de maneira similar à análise da série de Lorenz, o valor da taxa de recorrência diminuiu para TxRR $= 0,012564$ e o da taxa de determinismo aumentou para TxDET $= 0,488748$.

4.4 Considerações finais

Neste capítulo, foram desenvolvidos dois estudos experimentais, visando comprovar os benefícios e a possibilidade de se modelar, individualmente, componentes estocásticos e determinísticos de séries temporais. No primeiro estudo, foi gerada uma série sintética, com dois comportamentos distintos e facilmente separáveis, os quais representam a presença de estocasticidade e determinismo. Os experimentos executados nesse estudo comprova-

ram que ao se realizar, de forma adequada, a identificação desses componentes, pode-se obter modelos de maior acurácia frente à utilização de um único modelo para descrever todas as observações dessa série.

Por outro lado, no segundo estudo, os experimentos comprovaram que é possível identificar os componentes estocásticos e determinísticos que compõem séries temporais. Essa identificação fornece indícios de que o desenvolvimento de um mecanismo híbrido de modelagem, o qual, tendo conhecimento específico sobre os componentes que descrevem a série, deve obter a regra geradora com maior acurácia. Além disso, experimentos demonstraram que a técnica de detecção de linhas apresenta resultados melhores e mais estáveis do que utilizando o filtro de mediana para separar comportamentos estocásticos e determinísticos. Todavia, percebe-se ainda que esses resultados são iniciais. Espera-se avaliar demais abordagens tanto baseadas em filtragem quanto oriundas de outras áreas, tais como, *Wavelets*. Ao se determinar uma abordagem adequada de segmentação de séries temporais, novos experimentos, utilizando séries mais complexa do que a utilizada no primeiro estudo, serão executados a fim de comprovar a hipótese apresentada nesse trabalho.

APÊNDICE

$\mathcal{A}$

Recurrence Plot

A análise de recorrência de eventos de um dado sistema não é exclusiva de áreas da ciência, mas é algo natural do ser humano. As pessoas fazem, diariamente, análises de recorrência, tomando como base conhecimentos passados e as frequências com que eles se repetem para, por exemplo, determinar os riscos que envolvem uma atividade ou, simplesmente, para afirmar sobre fenômenos da natureza. Baseado nesse conceito intuitivo, a análise de recorrência começou a ser formulada matematicamente por Henri Poincaré em 1890, conforme discutido em (Marwan *et al.*, 2007).

Essa análise matemática de recorrência aplicada em séries temporais é uma ferramenta disponibilizada pela área de Sistemas Dinâmicos, a qual visa caracterizar o comportamento de sistemas de acordo com sua reconstrução no espaço de coordenadas de atraso. Esse espaço, também denominado espaço fase, representa todos os possíveis estados do sistema, ou seja, considerando que os possíveis estados de um dado sistema no instante t podem ser especificados por d componentes, então, a coordenada de atraso $d-$dimensional do sistema é definida por $\vec{x}(t) = (x_1(t), x_2(t), \ldots, x_d(t))^T$ (Marwan *et al.*, 2007).

Nesse sentido, define-se *Recurrence Plot* (RP) como sendo uma ferramenta que determina recorrências de trajetórias ($\vec{x}_t \in \mathbb{N}^d$) no espaço fase (Eckmann e Ruelle, 1985; Marwan

et al., 2007). As trajetórias recorrentes, comumente chamadas de pontos recorrentes, são definidas por $\vec{x}_t = \{x_t, x_{t+\tau}, x_{t+2\tau}, \ldots, x_{t+((m-1)\times\tau)}\}$, tal que m é a dimensão embutida e τ é a dimensão de separação. Quando a coordenada de atraso de uma série possui dimensão embutida $m = 1$, a trajetória $\vec{x_t}$ representa a própria observação x_t.

De maneira geral, a ferramenta RP visa analisar uma série temporal de entrada e retornar como saída uma matriz, cujos valores determinam se suas observações são recorrentes. Essa análise de recorrência é realizada através da aplicação da Equação A.1.

$$R_{i,j} = \Theta(\varepsilon - ||\vec{x_i} - \vec{x_j}||) \tag{A.1}$$

Nessa equação, ε é um $threshold$ de distância, $||\cdot||$ representa a norma utilizada para calcular a distância entre as observações i e j, e $\Theta(\cdot)$ é uma função degrau (*Heaviside Function*), a qual é definida por:

$$\Theta(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \tag{A.2}$$

De maneira geral, a primeira etapa da execução da ferramenta RP está relacionada à reconstrução da série em seu espaço de coordenadas de atraso, com dimensão de separação τ e dimensão embutida m. Em seguida, cada ponto é analisado em função dos demais pontos da série, para que aqueles correlacionados sejam considerados recorrentes. A Figura A.1 apresenta um exemplo de dois pontos separados por uma distância d. Esses pontos são considerados recorrentes, pois a distância entre eles é menor que o raio de vizinhança definido pelo *threshold* de distância ε.

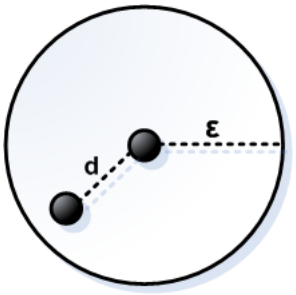


Figura A.1: Distância entre dois pontos considerados recorrentes pela ferramenta RP.

Um exemplo dessas etapas iniciais da execução da ferramenta RP pode ser vista na

Figura A.2, a qual apresenta uma série temporal, reconstruída em seu espaço de coordenadas de atraso, obtida a partir do sistema de Rössler (Rössler, 1976). Pode-se notar que os pontos • em destaque representam duas trajetórias consideradas recorrentes, pois estão dentro do raio de vizinhança, enquanto que uma terceira trajetória ◦ não é considerada um ponto recorrente, pois se encontra fora desse raio.

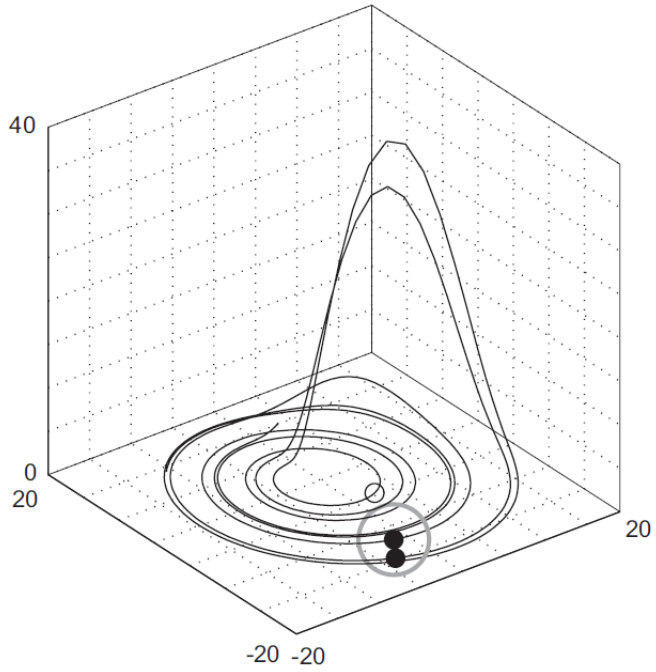


Figura A.2: Análise de recorrência da série gerada a partir do atrator de Rössler (Marwan *et al.*, 2007).

Por fim, depois que todas as observações são analisadas, as trajetórias são organizadas em uma matriz de recorrência bidimensional, cujos valores são definidos pela Equação A.1. Essa matriz pode ser visualizada através de estruturas de recorrência, as quais são obtidas por meio de gráficos que utilizam cores diferentes para os valores binários da matriz. Por exemplo, um ponto preto é utilizado para a coordenada (i, j), caso $R_{i,j} \equiv 1$, ou branco, caso $R_{i,j} \equiv 0$. A Figura A.3 (Marwan *et al.*, 2007) ilustra representações gráficas de matrizes de recorrência de séries simuladas. Ambos os eixos x e y de cada gráfico representam o tempo. Como pode ser visto, em todos os gráficos há uma linha diagonal, chamada linha de identidade (*Line of Identity* – LOI), a qual possui, por definição, valores de recorrência iguais a 1, pois $R_{i,j} \equiv 1, \forall i = j$.

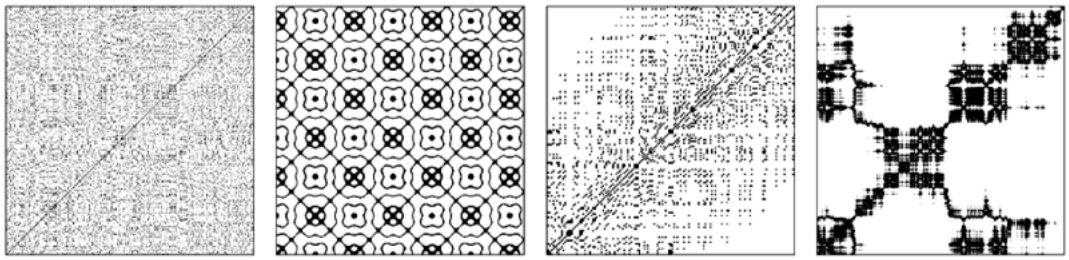


Figura A.3: *Recurrence Plot* das séries: (a) Ruído branco, (b) Oscilador harmônico, (c) Mapa logístico e (d) Movimento Browniano (Marwan *et al.*, 2007).

As estruturas geradas por esses gráficos podem fornecer diversas informações a respeito da série analisada como, por exemplo:

- Estruturas homogêneas: a presença de pontos homogêneos na estrutura de recorrência indica que a série analisada foi gerada a partir de um sistema estacionário como, por exemplo, Figura A.3(a);
- Pontos acumulados: tais estruturas são formadas por séries cujos valores das observações variam suavemente no tempo, o que representa o comportamento típico de sistemas não-estacionários. Um exemplo desse comportamento pode ser visto na Figura A.3(c), cujos pontos se acumulam próximo da LOI;
- Mudanças bruscas: estruturas com comportamento que se altera sensivelmente no tempo apresentam grandes áreas sem pontos recorrentes, como pode ser visto na Figura A.3(d). Nesses caso, o uso da ferramenta RP permite encontrar anomalias (Albertini e Mello, 2007) no comportamento de séries temporais;
- Pontos isolados: a presença de muitos pontos isolados no gráfico de recorrência indica que estados do sistema se repetem raramente. Em outras palavras, quanto maior o número de pontos isolados, maior a estocasticidade da série, como pode ser notado na Figura A.3(a). A série que originou essa figura é composta por observações geradas apenas a partir de uma sequência de ruídos;
- Linhas diagonais: essas linhas são definidas por $R_{i+k,j+k} \equiv 1|_{k=1}^{l}$, sendo que l representa o seu comprimento, e ocorrem sempre que há comportamento persistente na

série, ou seja, quanto maior o número de linhas diagonais, mais determinística é a série como, por exemplo, no caso apresentado na Figura A.3(b);

- Linhas verticais: a presença dessas linhas indicam intervalos de tempo no qual um estado não muda ou varia lentamente;

- Estruturas periódicas ou quasi-periódicas: são estruturas formadas pela presença de linhas diagonais e estruturas bem definidas. Essas estruturas indicam que a série estudada foi gerada a partir de sistemas periódicos ou quasi-periódicos como, por exemplo, observações geradas por um oscilador harmônico (tal como apresentado na Figura A.3(b)).

A análise dessas estruturas, no entanto, nem sempre são simples de serem realizadas somente de maneira visual. A fim de tornar essa análise mais simples e de forma automática, foram desenvolvidas diversas métricas chamadas de análise de quantificação de recorrência (*Recurrence Quantification Analysis* – RQA) (Marwan *et al.*, 2007).

Dentre essas métricas, destaca-se a taxa de recorrência (RR), a qual é considerada a medida mais simples de análise. Essa taxa de recorrência determina a densidade dos pontos recorrentes em uma matriz fornecida pela ferramenta RP. Essa medida é definida pela Equação A.3. É importante ressaltar que o intuito dessa métrica é avaliar a quantidade de observações distintas que são consideradas recorrentes, portanto, a linha de identidade não deve ser contabilizada nesse cálculo.

$$RR(\varepsilon) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} R_{i,j}(\varepsilon) \tag{A.3}$$

Além dessa métrica, RQA fornece outras medidas que são baseadas nas estruturas que formam as linhas diagonais. Em geral, essas métricas utilizam um histograma, o qual contabiliza a frequência de cada linha diagonal de comprimento l em uma determinada estrutura gerada pela ferramenta RP. Por exemplo, considere a Figura A.4, a qual apresenta um histograma da série de Lorenz que foi desdobrada considerando a dimensão embutida $m = 3$ e a de separação $\tau = 5$, conforme apresentado na Figura 2.9. Como pode ser visto nesse histograma, a linha diagonal com comprimento $l = 3$ aparece mais frequentemente na RP.

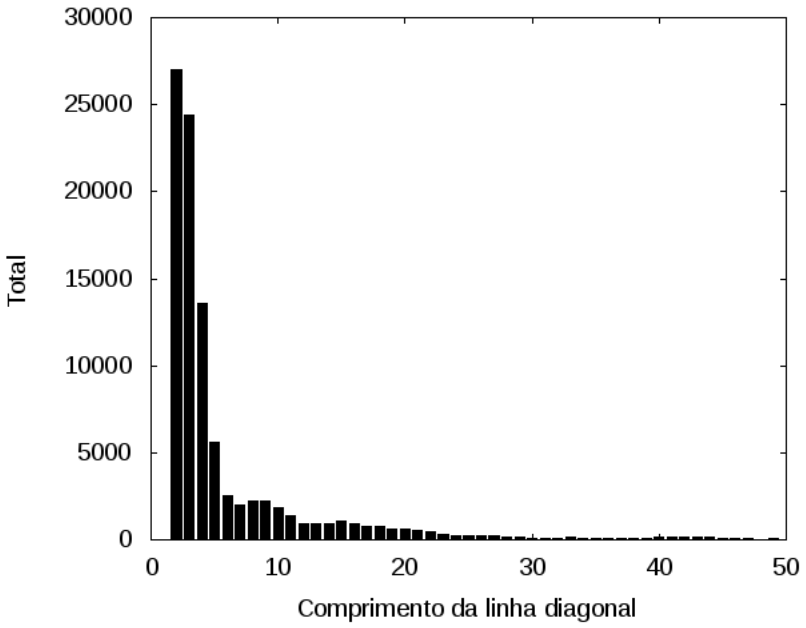


Figura A.4: Histograma da ferramenta RP gerado a partir da análise da série de Lorenz.

A frequência das linhas diagonais utilizadas pelas métricas de RQA é calculada pela Equação A.4:

$$P(\varepsilon, l) = \sum_{i,j=1}^{N} (1 - R_{i-1,j-1(\varepsilon)})(1 - R_{i+l,j+l(\varepsilon)}) \prod_{k=0}^{l-1} R_{i+k,j+k(\varepsilon)} \tag{A.4}$$

Histogramas que apresentam poucas ou nenhuma linha diagonal, ou ainda uma alta frequência apenas de linhas diagonais muito curtas, indicam que foram gerados a partir de processos fracamente correlacionados, com comportamento estocástico ou caótico. Por outro lado, a alta frequência de linhas diagonais longas e a presença de um número reduzido de pontos isolados indicam que a série analisada tem um comportamento determinístico (Marwan *et al.*, 2007). Nesse sentido, a taxa de determinismo de uma série temporal pode ser quantificada por meio da medida DET, a qual é calculada pela Equação A.5 e refere-se à taxa de pontos recorrentes que formam estruturas diagonais de tamanho $l \geq l_{min}$.

$$DET = \frac{\sum_{l=l_{min}}^{N} lP(\varepsilon, l)}{\sum_{i=1}^{N} \sum_{j=1}^{N} R_{i,j}, \forall i \neq j} \tag{A.5}$$

Uma outra medida muito importante, também definida a partir do cálculo das frequências das linhas diagonais, é o cálculo da entropia. Essa medida (Equação A.6) determina

o nível de complexidade das linhas diagonais de uma determinada série analisada, a qual se refere à probabilidade de encontrar linhas diagonais de tamanho exatamente igual a l. Séries compostas apenas por ruídos não-correlacionados, por exemplo, em casos em que não há presença de muitas linhas diagonais longas, possuem baixo valor de entropia, indicando baixa complexidade (Marwan *et al.*, 2007).

$$ENTR = -\sum_{l=l_{min}}^{N} P(\varepsilon, l) log_2 P(\varepsilon, l) \tag{A.6}$$

Segundo Marwan *et al.* (2007), existem diversas medidas de quantificação que podem ser aplicadas às matrizes de recorrência. No entanto, como o objetivo principal deste trabalho é realizar uma segmentação de séries temporais de acordo com seus componentes estocásticos e determinísticos, serão adotadas apenas as medidas anteriormente citadas: RR, DET e $ENTR$.

Uma consideração importante a respeito da ferramenta RP está relacionada à escolha do *threshold* de distância ε. A escolha do valor desse *threshold* é uma etapa que exige atenção especial, pois se escolhido um valor alto demais, pontos que não possuem nenhuma correlação podem ser considerados semelhantes. Por outro lado, a escolha de um valor muito baixo impede que pontos com comportamento semelhante sejam considerados correlacionados.

Segundo Marwan *et al.* (2007), existem diversas abordagens que auxiliam na escolha do valor mais apropriado para o *threshold* de distância ε. Uma dessas abordagens defende que o valor de ε não pode ultrapassar 10% do valor médio da série. Uma outra abordagem defende que para dados provenientes de séries não-estacionárias, o valor de ε deve ser aproximadamente 1% da densidade de pontos recorrentes. Em grande parte dos estudos realizados com RP, o valor de ε é definido de maneira empírica, ou seja, é realizada uma intercalação entre diferentes valores de ε e a visualização de cada estrutura resultante, sendo escolhido o valor de ε para a estrutura que melhor se ajusta, visualmente, aos dados analisados.

APÊNDICE

B

Filtros de processamento de imagem

A aplicação de técnicas de processamento de imagens visa, de maneira geral, utilizar recursos computacionais a fim de melhorar a percepção visual humana de imagens, destacando aspectos relevantes, ou, ainda, processar imagens para armazenamento, transmissão e representação para máquinas autônomas (Gonzalez e Woods, 2006).

Segundo Gonzalez e Woods (2006), a área de Processamento Digital de Imagens pode ser dividida em 3 classes de acordo com os níveis de processos computacionais utilizados. Processos de baixo nível envolvem operações primitivas como, por exemplo, preprocessamento de imagens para remoção de ruídos ou alteração de contraste e nitidez. Por outro lado, processos de nível médio visam segmentar imagens em regiões ou objetos para serem utilizados em reconhecimento e classificação de padrões. Ao contrário dos processos de baixo nível que resultam em imagens transformadas, processos de nível médio geram atributos extraídos de imagens. Por fim, existem ainda os processos de alto nível, os quais visam gerar conhecimento a partir de conjuntos de objetos e padrões extraídos de imagens.

Considerando as definições de classes apresentadas acima, esta seção apresenta um estudo realizado com algumas das principais técnicas disponibilizadas pelos processos de baixo e médio nível. Além disso, experimentos apresentados na Seção 4.3 foram conduzidos com o objetivo de validar a utilização dessas técnicas na filtragem de ruídos presentes em estruturas geradas pela ferramenta RP.

B.1 Filtro de Mediana

O filtro de mediana, ferramenta disponibilizada pela classe de processos de baixo nível, é um dos mais conhecidos filtros de ordem estatística e, geralmente, é utilizado para remoção de ruídos. Filtros dessa ordem são chamados de filtros não-lineares porque a imagem resultante do processamento é gerada por meio de substituições de *pixels* por novos valores calculados a partir de uma vizinhança de *pixels*. No caso do filtro de mediana, o valor de um *pixel* é substituído pela mediana, $\xi(\cdot)$, dos valores de sua vizinhança (Gonzalez e Woods, 2006).

A primeira etapa na aplicação de um filtro de mediana sobre uma imagem é a escolha dos limites de vizinhança. Considerando a representação da imagem analisada por meio de uma matriz, pode-se afirmar que uma vizinhança n de um *pixel* p é uma submatriz $n \times n$, na qual p encontra-se no centro da matriz, dado que n é um número ímpar. Para exemplificar, considere a Figura B.1, a qual apresenta uma matriz contendo os *pixels* de uma imagem. Nesse caso, o filtro da mediana utiliza uma vizinhança 3×3 e a substituição do *pixel* em destaque, $p(2,3) = 15$, é realizada pela mediana da sua vizinhança, ou seja, $p(2,3) = \xi(10, 20, 20, 20, 15, 20, 20, 25, 100) = 20$.

	1	2	3	4	5
1	30	10	20	20	10
2	40	20	15	20	30
3	10	20	25	100	80
4	20	60	10	30	70
5	10	60	10	10	20

Figura B.1: Exemplo de aplicação do filtro de mediana.

Esse procedimento, que foi realizado no *pixel* $p(2,3)$, é repetido para todos os demais

pixels da imagem. Como pode ser visto, *pixels* da borda da matriz[1], que não possuem vizinhos suficientes para aplicação do filtro podem ser desconsiderados, ou seja, novos valores não são calculados para substituir os *pixels* dessa região. Contudo, quando utiliza-se uma vizinhança muito grande, a quantidade de *pixels* desconsiderados é alta. Para solucionar esse problema, as linhas e colunas afetadas podem ser repetidas ou novas linhas e colunas com zeros podem ser adicionadas à matriz (Gonzalez e Woods, 2006). Por exemplo, uma nova linha poderia ser adicionada no final da matriz apresentada na Figura B.1 com os mesmos valores da linha 5 para que essa linha pudesse ser analisada.

Uma aplicação prática do uso do filtro de mediana é apresentada na Figura B.2. A Figura B.2(a) apresenta uma imagem, na qual foram adicionados ruídos aleatórios. Em seguida, aplicou-se o filtro de mediana, resultando em uma nova imagem, Figura B.2(b), livre de ruídos. Conforme pode ser visto nessa figura, há indícios de que a aplicação do filtro de mediana sobre as estruturas geradas pela ferramenta RP pode produzir resultados satisfatórios no processo de remoção dos pontos isolados.

Figura B.2: (a) Imagem com ruído. (b) Remoção de ruídos com o filtro de mediana.

B.2 Detecção de Bordas

Os modelos de detecção de bordas são disponibilizados pela classe de processos de nível médio, cujo objetivo é segmentar imagens com base em mudanças de intensidades locais. Em outras palavras, ao contrário do filtro de mediana, que visa identificar e remover ruídos, o objetivo principal desses modelos é extrair de uma imagem todos os *pixels* que, potencialmente, podem compor uma borda (Gonzalez e Woods, 2006). No contexto de

[1] *Pixels* que se encontram na primeira e na última linha ou na primeira e na última coluna

filtragem de estruturas geradas pela ferramenta RP, a principal motivação para aplicação desse modelo é a possibilidade de destacar as linhas diagonais que representam a taxa de determinismo da série analisada.

Modelos de detecção de bordas utilizam operadores de gradiente para identificar direção e intensidade de bordas em imagens. Um dos mais conhecidos operadores de gradiente é *Robert-Cross*, o qual visa identificar bordas diagonais em imagens. De maneira geral, um operador de gradiente é uma máscara bidimensional que é aplicada sobre a imagem analisada. O operador *Robert-Cross* utiliza duas máscaras, as quais são apresentadas na Figura B.3 (Gonzalez e Woods, 2006).

-1	0
0	1

0	-1
1	0

Figura B.3: Operador *Robert-Cross* utilizado por modelos de detecção de bordas.

A fim de exemplificar o uso desse operador, considere a Figura B.4 como sendo uma região 3×3 de uma imagem. A transformação do *pixel* P_5 é obtida por meio da aplicação do operador *Robert-Cross*, resultando em uma nova coordenada (g_x, g_y), tal que $g_x = (P_9 - P_5)$ e $g_y = (P_8 - P_6)$ (Gonzalez e Woods, 2006).

P_1	P_2	P_3
P_4	P_5	P_6
P_7	P_8	P_9

Figura B.4: Região 3×3 de uma imagem.

A aplicação de modelos de detecção de bordas, no entanto, necessita de uma etapa de pré-processamento da imagem analisada para remover ruídos (Gonzalez e Woods, 2006). Para ilustrar os efeitos da aplicação desses modelos em imagens com ruídos, considere a Figura B.5. A imagem apresentada na Figura B.5(a) é composta por uma linha diagonal e dois *pixels* adicionados aleatoriamente para representar ruídos na imagem. Enquanto que a imagem apresentada na Figura B.5(b) representa a saída da execução do operador de gradiente *Robert-Cross*. Como pode ser visto, a linha diagonal foi detectada, mas o ruído próximo à linha foi anexado ao objeto segmentado.

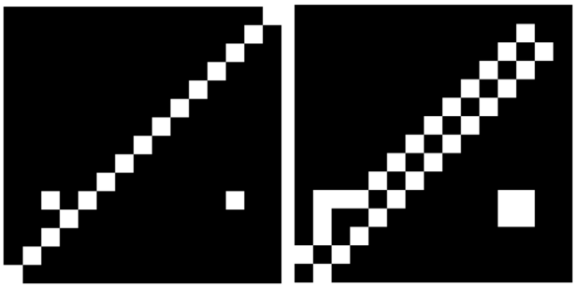


Figura B.5: Aplicação do operador *Robert-Cross* sem a etapa de remoção de ruídos.

No contexto de filtragem de estruturas geradas a partir da ferramenta RP, a necessidade de uma etapa de pré-processamento para remoção de ruídos inviabiliza a aplicação desse modelo de detecção de bordas. Isso se deve ao fato de que o principal objetivo na aplicação de técnicas de filtragem é justamente destacar as linhas diagonais descartando os pontos isolados. Se o ruído é removido anteriormente, torna-se desnecessária a utilização dessa técnica.

B.3 Detecção de Linhas

O problema com a aplicação de modelos de detecção de bordas, apresentado na seção anterior, motivou a busca de novas técnicas disponibilizadas por processos de nível médio, que permitissem a detecção de segmentos presentes em estruturas geradas a partir da ferramenta RP, sem que fossem influenciadas pela presença de ruídos. Nesse sentido, um dos principais modelos utilizados para esse propósito são os modelos de detecção de linhas.

A aplicação desse modelo é realizada por meio de convoluções entre a imagem original e máscaras, as quais são apresentadas na Figura B.6. A aplicação de cada uma dessas máscaras destaca a direção das linhas presentes nas imagens analisadas. Como pode ser observado nessa figura, a primeira máscara apresentada detecta linhas horizontais (Figura B.6(a)), enquanto que a segunda (Figura B.6(b)) e a terceira (Figura B.6(c)) máscaras detectam linhas diagonais e, por fim, a última máscara (Figura B.6(d)) é utilizada para detectar linhas verticais.

A convolução de uma imagem $f(x, y)$ com um máscara $w(x, y)$ é denotada pela Equação

-1	-1	-1
2	2	2
-1	-1	-1

(a) Horizontal

2	-1	-1
-1	2	-1
-1	-1	2

(b) $+45^o$

-1	-1	2
-1	2	-1
2	-1	-1

(c) -45^o

-1	2	-1
-1	2	-1
-1	2	-1

(d) Vertical

Figura B.6: Máscaras utilizadas pelos modelos de detecção de linhas.

B.1, sendo que $a = (\frac{m-1}{2})$ e $b = (\frac{n-1}{2})$.

$$w(x,y)\bigstar f(x,y) = \sum_{s=-a}^{a}\sum_{t=-b}^{b} w(x,y)f(x-s,y-t) \tag{B.1}$$

Para as máscaras apresentadas na Figura B.6, tem-se que $m = 3$ e $n = 3$. A fim de facilitar o entendimento da equação de convolução, suponha que se deseja saber qual o novo valor do *pixel* P_5 da matriz M apresentada na Figura B.4. Então, o valor de P_5 na matriz resultante M' é o valor máximo da aplicação das convoluções de cada máscara. O valor da convolução da aplicação da máscara horizontal é obtido pela resolução da equação: $M'(P_5) = M(P_1) \times (-1) + M(P_2) \times (-1) + M(P_3) \times (-1) + M(P_4) \times (2) + M(P_5) \times (2) + M(P_6) \times (2) + M(P_7) \times (-1) + M(P_8) \times (-1) + M(P_9) \times (-1)$

A utilização de todas essas quatro máscaras não é necessária no caso de detecção de linhas em estruturas geradas pela ferramenta RP, pois o principal objetivo é a detecção apenas de linhas diagonais. Nesse sentido faz-se necessário realizar apenas a convolução que utiliza a máscara diagonal, cujo ângulo é de -45^o.

A aplicação desse modelo é muito favorável para filtrar as estruturas geradas pela ferramenta RP devido, principalmente, à capacidade de detectar linhas diagonais com alta precisão e por ser tolerante a ruídos. Com a aplicação desse modelo, espera-se destacar comportamentos determinísticos presentes em séries temporais, separando-os de componentes estocásticos.

Referências Bibliográficas

Abarbanel, H. D. I.; Brown, R.; Sidorowich, J. J.; Tsimring, L. S. (1993). The analysis of observed chaotic data in physical systems. *Reviews of Modern Physics*, v.65, p.1331–1392.

Ackoff, R. (1981). *Creating the corporate future*. John Wiley and Sons Inc.

Albertini, M. K.; Mello, R. F. (2007). A self-organizing neural network for detecting novelties. *SAC '07: Proceedings of the 2007 ACM symposium on Applied computing*, p. 462–466, New York, EUA. ACM.

Alligood, K. T.; Sauer, T. D.; Yorke, J. A. (1997). *Chaos: An Introduction to Dynamical Systems*. Springer.

Ariew, R. (1976). *Ockham's Razor: A Historical and Philosophical Analysis of Ockham's Principleof Parsimony*. Tese (Doutorado), University of Illinois at Urbana-Champaign.

Baragona, R.; Vitrano, S. (2006). Genetic algorithms-based approaches for clustering time series. Zani, S.; Cerioli, A.; Riani, M.; Vichi, M., editores, *Data Analysis, Classification and the Forward Search*, Studies in Classification, Data Analysis, and Knowledge Organization, p. 3–10. Springer Berlin Heidelberg.

Boker, S. M. (2010). Lorenz - Série temporal dinâmica não-linear. Disponível em:

http://people.virginia.edu/ smb3u/psych611/lor63.dat. Acessado em: 17 de julho de 2010.

Box, G.; Jenkins, G. M.; Reinsel, G. (1994). *Time Series Analysis: Forecasting & Control*. Prentice Hall, 3^a edição.

Brockwell, P. J.; Davis, R. A. (2002). *Introduction to Time Series and Forecasting*. Springer.

Casdagli, M. (1989). Nonlinear prediction of chaotic time series. *Physica D: Nonlinear Phenomena*, v.35, n.3, p.335 – 356.

Chatfield, C. (2004). *The Analysis of Time Series: An Introduction*. CRC Press LLC.

cheng Huang, H.; Cressie, N. (1998). Deterministic/stochastic wavelet decomposition for recovery of signal from noisy data. *Technometrics*, v.42, p.262–276.

Choi, J. J.; Hauser, S.; Kopecky, K. J. (1999). Does the stock market predict real activity? time series evidence from the g-7 countries. *Journal of Banking and Finance*, v.23, n.12, p.1771 – 1792.

Dodonov, E. (2009). *Uma abordagem para prover autonomia a ambientes distribuídos heterogêneos por meio do estudo e predição da dinâmica de padrões comportamentais de processos*. Tese (Doutorado), Universidade de São Paulo.

Dodonov, E.; Mello, R. F. (2007). A model for automatic on-line process behavior extraction, classification and prediction in heterogeneous distributed systems. *Seventh IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2007)*, p. 899–904. IEEE Computer Society.

Dodonov, E.; Mello, R. F.; Yang, L. T. (2005). A network evaluation for LAN, MAN and WAN grid environments. *EUC*, v. 3824 de *Lecture Notes in Computer Science*, p. 1133–1146. Springer.

Dodonov, E.; Mello, R. F.; Yang, L. T. (2006). Adaptive technique for automatic communication access pattern discovery applied to data prefetching in distributed applications using neural networks and stochastic models. *The International Symposium on Parallel and Distributed Processing and Application – ISPA*, p. 292–303.

Eckmann, J.-P.; Ruelle, D. (1985). Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics*, v.57, p.617–656.

Fearnside, P. M. (1999). Forests and global warmingnext term mitigation in Brazil: opportunities in the brazilian forest sector for responses to previous termglobal warmingnext term under the clean development mechanism. *Biomass and Bioenergy*, v.16, n.3, p.171 – 189.

Fraser, A. M.; Swinney, H. L. (1986). Independent coordinates for strange attractors from mutual information. *Physical Review A*, v.33, n.2, p.1134–1140.

Gonzalez, R. C.; Woods, R. E. (2006). *Digital Image Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, EUA, 3^a edição.

Graps, A. (1995). An introduction to wavelets. *IEEE Computational Science and Engineering*, v.2, p.50–61.

Guhathakurta, K.; Bhattacharya, B.; Chowdhury, A. R. (2010). Using recurrence plot analysis to distinguish between endogenous and exogenous stock market crashes. *Physica A: Statistical Mechanics and its Applications*, v.389, n.9, p.1874–1882.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hegger, R.; Kantz, H.; Schreiber, T. (1999). Practical implementation of nonlinear time series methods: The TISEAN package. *Chaos*, v.9, n.2, p.413–435.

Holland, J. H. (2000). *Emergence: From Chaos to Order*. Oxford University Press.

Hosokawa, Y.; Shirato, H.; Nishioka, T.; Tsuchiya, K.; Chang, T.-C.; Kagei, K.; Ohomori, K.; ichi Obinata, K.; Kaneko, M.; Miyasaka, K.; Nakamura, M. (2003). Effect of treatment time on outcome of radiotherapy for oral tongue carcinoma. *International Journal of Radiation Oncology, Biology, and Physics*, v.57, n.1, p.71–78.

Ishii, R. P. (2010). *Antecipação de operações de entrada e saída visando otimizar o tempo de resposta de aplicações distribuídas que manipulam grandes volumes de dados*. Tese (Doutorado), Universidade de São Paulo.

Ishii, R. P.; Mello, R. F. (2009). A history-based heuristic to optimize data access in distributed environment. *21st IASTED International Conference Parallell and Distributed Computing and Systems (PDCS2009)*, Cambridge, MA, EUA.

Ishii, R. P.; Mello, R. F.; Yang, L. T. (2007). A complex network-based approach for job scheduling in grid environments. *High Performance Computing and Communications, Third International Conference, HPCC 2007, Houston, EUA, setembro 26-28, 2007, Proceedings*, v. 4782 de *Lecture Notes in Computer Science*, p. 204–215. Springer.

Jain, A.; Kumar, A. M. (2007). Hybrid neural network models for hydrologic time series forecasting. *Applied Soft Computing*, v.7, n.2, p.585–592.

Kennel, M. B.; Brown, R.; Abarbanel, H. D. I. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A*, v.45, n.6, p.3403–3411.

Ko, M. K.; Sze, N. D.; Molnar, G.; Prather, M. J. (1993). Global warming from chlorofluorocarbons and their alternatives: Time scales of chemistry and climate. *Atmospheric Environment. Part A. General Topics*, v.27, n.4, p.581–587.

Koçak, K.; Saylan, L.; Eitzinger, J. (2004). Nonlinear prediction of near-surface temperature via univariate and multivariate time series embedding. *Ecological Modelling*, v.173, n.1, p.1 – 7.

Kärner, O. (2009). Arima representation for daily solar irradiance and surface air temperature time series. *Journal of Atmospheric and Solar-Terrestrial Physics*, v.71, n.8-9, p.841 – 847.

LeBaron, B.; Arthur, W. B.; Palmer, R. (1999). Time series properties of an artificial stock market. *Journal of Economic Dynamics and Control*, v.23, n.9-10, p.1487 – 1516.

Lee, T.-H.; White, H.; Granger, C. W. J. (1993). Testing for neglected nonlinearity in time series models: a comparison of neural network methods and alternative tests. *Journal of Econometrics*, v.56, p.269–290.

Liao, T. W. (2005). Clustering of time series data – a survey. *Pattern Recognition*, v.38, n.11, p.1857–1874.

Liebert, W.; Pawelzik, K.; Schuster, H. G. (1991). Optimal embeddings of chaotic attractors from topological considerations. *Europhysics Letters*, v.14, n.6, p.521–526.

Marsland, S.; Shapiro, J.; Nehmzow, U. (2002). A self-organising network that grows when required. *Neural Networks*, v.15, p.1041 – 1058.

Marwan, N. (2009). Commandline recurrence plots. Disponível em: http://www.agnld.uni-potsdam.de/~marwan/6.download/rp.php. Acessado em: 27 de julho de 2010.

Marwan, N. (2010). The toolbox contains MATLAB routines for computing recurrence plots and related problems. Disponível em: http://www.agnld.uni-potsdam.de/~marwan/toolbox/index.html. Acessado em: 27 de julho de 2010.

Marwan, N.; Romano, M.; Thiel, M.; Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, v.438, n.5-6, p.237–329.

Mañé, R. (1980). *On the dimension of the compact invariant sets of certain nonlinear maps*. Springer.

Mello, R. F. (2009). *Sistemas Dinâmicos e Técnicas Inteligentes para a Predição de Comportamento de Processos: Uma Abordagem para Otimização de Escalonamento em Grades Computacionais*. Tese (Doutorado), Universidade de São Paulo. Disponível em: http://www.icmc.usp.br/ mello/teseld.pdf. Acessado em: 16 de dezembro de 2009.

Mello, R. F. (2010). Improving the performance and accuracy of time series modeling based on autonomic computing systems. *Journal of Ambient Intelligence and Humanized Computing*, p. 1–23.

Mello, R. F.; Andrade Filho, J. A.; Dodonov, E.; Ishii, R. P.; Yang, L. T. (2007a). Optimizing distributed data access in grid environments by using artificial intelligence techniques. *The International Symposium on Parallel and Distributed Processing and Application – ISPA*, p. 1–12.

Mello, R. F.; Andrade Filho, J. A.; Senger, L. J.; Yang, L. T. (2007b). RouteGA: A grid load balancing algorithm with genetic support. *The IEEE 21th International Conference on Advanced Information Networking and Applications (AINA 2007)*, p. 885–892. IEEE Computer Society.

Mello, R. F.; Andrade Filho, J. A.; Senger, L. J.; Yang, L. T. (2008). Grid job scheduling using route with genetic algorithm support. *Telecommunication Systems*, v.38, n.3-4, p.147–160.

Mello, R. F.; Dodonov, E.; Bertagna, R.; Senge, L. J. (2009). Extracting and predicting the communication behaviour of parallel applications. *International Journal of Parallel, Emergent and Distributed Systems*, v.24, n.3, p.225–242.

Minerva, T. (2010). Wavelet filtering for prediction in time series analysis. *Non-Linear Systems & Wavelet Analysis*, p. 89–94.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York, EUA.

Monard, M. C.; Baranauskas, J. A. (2003). Conceitos sobre aprendizado de máquina. Rezende, S. O., editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, chapter 4, p. 89–114. Manole, 1^a edição.

Mood, A. M.; Graybill, F. A.; Boes, D. C. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill Companies, 3^a edição.

Morettin, P. A.; Toloi, C. M. C. (2004). *Análise de Séries Temporais*. Editora Edgard Blücher Ltda., São Paulo.

Morrison, M. A. (1990). *Understanding Quantum Physics: A User's Manual*. Prentice-Hall, Inc.

Narzo, A. F. D. (2010). tserieschaos: Analysis of nonlinear time series. Disponível em: http://cran.r-project.org/web/packages/tseriesChaos/index.html. Acessado em: 27 de julho de 2010.

Nason, G. (2008). *Wavelet Methods in Statistics with R*. Springer.

Parashar, M.; Hariri, S. (2006). *Autonomic Computing: Concepts, Infrastructure, and Applications*. CRC Press.

Ponomarenko, V.; Prokhorov, M.; Bespyatov, A.; Bodrov, M.; Gridnev, V. (2005). Deriving main rhythms of the human cardiovascular system from the heartbeat time series and detecting their synchronization. *Chaos, Solitons & Fractals*, v.23, n.4, p.1429–1438.

Raiesdana, S.; Golpayegani, S. M. R. H.; Firoozabadi, S. M. P.; Mehvari Habibabadi, J. (2009). On the discrimination of patho-physiological states in epilepsy by means of dynamical measures. *Computers in Biology and Medicine*, v.39, n.12, p.1073–1082.

Rosenstein, M. T.; Collins, J. J.; Luca, C. J. D. (1993). A practical method for calculating largest lyapunov exponents from small data sets. *Physica D*, v.65, p.117–134.

Rössler, O. E. (1976). An equation for continuous chaos. *Physics Letters A*, v.57, n.5, p.397 – 398.

Senger, L. J.; Mello, R. F.; Santana, M. J.; Santana, R. H. C. (2007). Aprendizado baseado em instâncias aplicado à predição de características de execução de aplicações paralelas. *Revista de Informática Teórica e Aplicada*, v.14, p.44–68.

Shumway, R. H.; Stoffer, D. S. (2006). *Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics)*. Springer, 2^a edição.

Summa, M. G.; Steyaert, J.; Vautrain, F.; Weitkunat, R. (2007). A new clustering method for time series to discover geographical cancer trends from 1960 to 2000. *Annals of Epidemiology*, v.17, n.9, p.744–744.

Takens, F. (1980). Detecting strange attractors in turbulence. *Dynamical Systems and Turbulence*, p. 366–381. Springer.

Tang, Y. Y. (2009). *Wavelet theory approach to pattern recognition*, v. 74. World Scientific Publishing Co. Pte. Ltd., 2^a edição.

Thornburn, W. M. (1918). The myth of occam's razor. *Mind*, v.27, p.345–353.

Tschacher, W.; Kupper, Z. (2002). Time series models of symptoms in schizophrenia. *Psychiatry Research*, v.113, n.1-2, p.127–137.

Tukey, J. (1961). *Discussion, emphasizing the connection between analysis of variance and spectrum analysis*, v. 3. Technometrics.

Wang, W.-C.; Chau, K.-W.; Cheng, C.-T.; Qiu, L. (2009). A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *Journal of Hydrology*, v.374, n.3-4, p.294 – 306.

Wangler, B.; Backlund, A. (2005). Information systems engineering: What is it? *CAiSE Workshops (2)*, p. 427–437.

Webber, C. L. (2010). Recurrence Quantification Analysis (rqa) tool. Disponível em: http://homepages.luc.edu/˜cwebber/. Acessado em: 27 de julho de 2010.

White, H. (1990). Connectionist nonparametric regression: multilayer feedforward networks can learn arbitrary mappings. *Neural Network*, v.3, n.5, p.535–549.

Whitney, H. (1936). Differentiable manifolds. *The Annals of Mathematics*, v.37, n.3, p.645–680.

Wu, B.-H.; Too, G.-P.; Lee, S. (2010). Audio signal separation via a combination procedure of time-reversal and deconvolution process. *Mechanical Systems and Signal Processing*, v.24, n.5, p.1431 – 1443. Special Issue: Operational Modal Analysis.

Yu, S.; Clark, O. G.; Leonard, J. J. (2008). A statistical method for the analysis of nonlinear temperature time series from compost. *Bioresource Technology*, v.99, n.6, p.1886–1895.

Zampa, P.; Arnost, R. (2002). A new approach to system structure definition. *Systems, Man and Cybernetics, 2002 IEEE International Conference on*, v. 7, p. 5 pp. vol.7.

Zhuang, J. J.; Ning, X. B.; He, A. J.; Zou, M.; Sun, B.; Wu, X. H. (2008). Alteration in scaling behavior of short-term heartbeat time series for professional shooting athletes from rest to exercise. *Physica A: Statistical Mechanics and its Applications*, v.387, n.26, p.6553–6557.