

Universidade Federal de Minas Gerais
Programa de Pós-Graduação em Engenharia Elétrica
Centro de Pesquisa e Desenvolvimento em
Engenharia Elétrica

Análise de Determinismo em Séries Temporais
Quase-Periódicas

Mara Cristina da Silveira Coelho

Tese submetida à banca examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais, como parte dos requisitos necessários à obtenção do grau de Doutor em Engenharia Elétrica.

Orientadores: Luis Antonio Aguirre
Eduardo Mazoni Andrade Marçal Mendes

“Nobody said it was easy ...
No one ever said it would be this hard ...”

Coldplay, The scientist

AGRADECIMENTOS

Gostaria de agradecer a Deus, por vosso amor incondicional e por cada um de nós.

Ao prof. Luis Antonio Aguirre pela orientação, colaboração técnica, incentivo e “puxões de orelha”. Ao nos receber em sua sala, ele verdadeiramente nos acolhe com serenidade, carinho e palavras de incentivo que soam como bênçãos.

Ao prof. Eduardo Mazoni A. M. Mendes pela orientação de todos os meus trabalhos científicos, desde a graduação. Mazoni, é mais do que um orientador, apresentou-me o fantástico mundo de pesquisas em identificação e análise de sistemas. Sempre coerente e com senso de humor.

À Banca examinadora, Luiz A. Baccalá, Reinaldo Roberto Rosa, Carlos Júlio T. Criollo, Hani Camille Yehia, Murilo E. D. Gomes, pelas contribuições técnicas para a elaboração da tese. Em especial ao prof. Hani pelos conselhos como professor, coordenador do PPGEE. À Anete e Arlete, pela atenção e carinho com que recebem aos alunos.

Ao meu amor Júlio que me apoiou e me apoia em todos os momentos de minha vida com dedicação, amor, companheirismo. Sua presença no dia-a-dia, finais de semana e madrugadas foi de vital importância.

À minha família, em especial as minhas irmãs à Flávia e Mayara, por todo amor e pelas palavras de carinho.

Aos colegas e funcionários do CPDEE, CPH e do MACSIN, em especial Ana Cláudia, Clodualdo, Davidson L. Firmo, Dorianara Erivelton G. Nepomuceno, Gleison F. V. Amaral, Seu Hirton, Kétia S. Moreira, Marcelle, Marcelo Costa, Márcio F. S.

Barroso.

Aos amigos queridos: Bruno Otávio S. Teixeira (meu amigo de sala que me abandonou e que me mata de rir); Dair J. Oliveira (super-ultra-mega paciente) e Marcelo M. Duarte, pela convivência harmoniosa que tivemos e pelas valiosas discussões técnicas e “um tanto filosóficas”.

A todos que, direta ou indiretamente, contribuíram para a realização de mais uma etapa em minha vida.

Ao apoio financeiro do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), o qual foi de vital importância para a realização deste trabalho.

RESUMO

Séries temporais com comportamento quase-periódico são bastante comuns nas ciências naturais, exatas, engenharias, áreas social e médica. As técnicas disponíveis na literatura para séries temporais, nem sempre são adequadas para analisar e modelar séries quase-periódicas, uma vez que o comportamento quase-periódico tende a ocultar qualquer outro comportamento determinístico, além do padrão quase-periódico. Todavia, pouca atenção tem sido dada ao desenvolvimento de ferramentas específicas para lidar com esse tipo de séries.

Na análise, modelagem e, possivelmente, previsão de séries temporais quase-periódicas, um dos primeiros passos de grande interesse é verificar se há algum padrão determinístico na série temporal além do comportamento quase-periódico. Neste contexto, o teste de hipóteses baseado em séries sub-rogadas vem auxiliar na detecção de determinismo subjacente ao comportamento quase-periódico.

Este trabalho propõe um teste de hipóteses baseado em dados sub-rogados para verificar a presença de vínculo determinístico intra-ciclo em séries temporais quase-periódicas.

Neste sentido, procurou-se avaliar as técnicas disponíveis na literatura para a geração das séries sub-rogadas. Os resultados obtidos mostram que, na presença de comportamento quase-periódico, tais técnicas nem sempre são adequadas. São discutidas algumas das dificuldades encontradas e sugeridos procedimentos alternativos, tal como a redução dos ciclos e da tendência, para a execução do teste de determinismo usando dados sub-rogados. Uma vez reduzido o ciclo da série quase-periódica, obtém-se uma série residual correspondente.

O teste de hipótese baseado em séries sub-rogadas, proposto neste trabalho, consiste em gerar um conjunto de séries sub-rogadas que preservam algumas características estatísticas da série residual analisada, e destroem outras. Em seguida, ajusta-se um modelo às séries, residual e sub-rogadas, e realiza-se a previsão de curto-prazo. Foi proposto como fator estatístico discriminante, o índice mRSE (*mean Root Square Error*), o qual avalia a capacidade de previsão da série residual e das séries artificiais. Por meio do mRSE testa-se a hipótese nula. Se o valor do fator discriminante da série residual for significativamente diferente do obtido para o conjunto de séries sub-rogadas, então pode-se rejeitar a hipótese nula como sendo uma provável origem da série temporal, ou seja, de que a série quase-periódica foi gerado por um processo gerador de séries aleatórias. Porém, se o fator discriminante da série temporal não for significativamente diferente do obtido para o conjunto de dados sub-rogados, não se pode rejeitar a hipótese nula.

O teste de determinismo proposto é, então, aplicado em séries quase-periódicas reais de diferentes áreas, tais como: a) série de consumo de energia elétrica; b) série de produção mensal de leite; c) série de números de manchas solares (*sunspots*) e; d) série de casos mensais de caxumba.

ABSTRACT

Quasiperiodic time series are sufficiently common in natural and Physical Sciences, Engineering, Medicine and Social Sciences. The techniques available in the literature for time series are not always suitable for analyzing and modelling quasiperiodic time series, because this behavior tends to mask any other deterministic behavior, but the quasiperiodic pattern that possibly exists. However, little attention has been given to the development of specific tools to deal with this kind of time series.

In the analysis, modelling and possibly forecasting of quasiperiodic time series, one of the first steps, which is of great interest, is to verify if there is some deterministic pattern in the time series apart from the quasiperiodic one. In this context, the hypothesis testing based on surrogate data is useful to shed some light in the detection of determinism (apart from the quasiperiodic behavior).

The proposed of this work is to verify the presence of intra-cycle determinism in quasiperiodic time series, by means of surrogate data testing.

So, first, the available techniques in literature for the generation of surrogate data were studied. The obtained results show that, in the presence of quasiperiodic pattern, such techniques are not always proper.

Some of the difficulties found are discussed and suggested alternative procedures, just as the reduction of the cycles and of the trend, for the testing of determinism using surrogates data. Once reduced the cycle of the time series, is obtained a corresponding residual time series.

The hypothesis test based on surrogates, proposed in this thesis, it consists

of generating a surrogates data that preserve some characteristics statistics of the residual series and destroy others.

In order to evaluate the deterministic feature of the residual time series a discriminating statistic must be chosen. Since the objective of this work is to verify the existence of intra-cycle determinism in quasiperiodic time series based on the existence of some predictability, the residual time series and the surrogate data set are submitted to the NARMA (Nonlinear Autoregressive and Moving Average model) polynomial modeling and subsequent prediction. After building the models from both the original time series and surrogate data, 1-to- h step-ahead predictions are determined.

It was proposed as discriminating statistic, the index mRSE, which evaluates the capacity of forecast of the residual series and of the surrogates series, the null hypothesis is tested. If the mRSE of the residual time series is significantly different from the obtained for the group of surrogates, then the null hypothesis can be rejected as a probable origin of the time series, in other words, that the quasiperiodic series was generated by a generating process of aleatory time series. However, if the mRSE of the residual time series is not significantly different from the obtained for the group of surrogates data, it cannot reject the null hypothesis.

The test of proposed determinism is applied in real quasiperiodic time series of different areas, such as: a) electric load time series; b) monthly milk production; c) sunspots time series ; d) monthly reported number of cases of mumps.

CONTEÚDO

Agradecimentos	v
Resumo	vii
Abstract	ix
Conteúdo	xi
Lista de Tabelas	xv
Lista de Figuras	xvii
Lista de Símbolos	xxi
Nomenclatura	xxiii
1 Introdução	1
1.1 Apresentação do problema	2
1.1.1 Técnicas para detectar determinismo	3
1.2 Teste de determinismo usando dados sub-rogados	7
1.3 Redução da componente quase-periódica	9
1.3.1 Objetivos	10
1.4 Organização do texto	11
2 Modelagem para previsão de séries temporais	13
2.1 Modelagem matemática	15

2.1.1	Modelos lineares auto-regressivos	17
2.1.2	Modelos NARMAX polinomiais	23
2.2	Detecção de Estrutura	26
2.2.1	Estimação dos Parâmetros	27
2.3	Comentários finais	31
3	Técnicas para análise de séries temporais	33
3.1	Características das séries temporais	34
3.1.1	Determinismo versus previsibilidade	35
3.2	Análise linear de séries temporais	38
3.3	Estimação e eliminação da tendência e dos ciclos	39
3.3.1	Diferenciação	40
3.3.2	Decomposição Clássica	41
3.3.3	Métodos para estimação e extração de tendência	41
3.3.4	Estimação e eliminação da componente de tendência e cíclica	47
3.3.5	Aplicação dos métodos para redução dos ciclos	50
3.4	Análise não-linear	54
3.4.1	Imersão de séries temporais	54
3.4.2	Imersão de séries usando coordenadas de atraso	55
3.4.3	Parâmetros de imersão	56
3.4.4	Séries quase-periódicas	58
3.5	Análise de séries usando teste de hipóteses	59
3.6	Teste de hipóteses baseado em dados sub-rogados	63
3.6.1	Algoritmos para séries temporais sem ciclos	64
3.6.2	Algoritmos para séries periódicas	67
3.7	Comentários finais	70
4	Análise de séries baseada em inferência estatística	71
4.1	Estabelecimento do fator estatístico discriminante	72
4.2	Definição das hipóteses alternativas	76

4.3	Intervalos de confiança	83
4.4	Comentários Finais	91
5	Análise de determinismo em séries quase-periódicas	93
5.1	Tratamento dos dados	93
5.1.1	Teste de estacionariedade	96
5.2	Teste de determinismo proposto	98
5.2.1	Algoritmo proposto para a detecção de determinismo	99
5.3	Algoritmos para gerar dados sub-rogados	102
5.3.1	FMMC para redução da componente quase-periódica	107
5.4	Comentários	116
6	Determinismo intra-ciclo em séries reais	119
6.1	Série de carga	119
6.1.1	Teste de estacionariedade	121
6.1.2	Teste de determinismo na série quase-periódica	127
6.1.3	Redução da quase-periodicidade da série $y_Q(k)$	127
6.2	Índices mensais de caxumba	136
6.3	Produção mensal de leite	140
6.4	Série de números anuais de manchas solares	143
6.5	Série de números diários de manchas solares	148
6.6	Comentários Finais	152
7	Discussão e sugestões para a continuidade do trabalho	157
7.1	Discussões e contribuições	158
7.2	Sugestões de Trabalhos Futuros	161
	Bibliografia	176

LISTA DE TABELAS

3.1	Exemplo de uma série sub-rogada $s(k)$ gerada pelo algoritmo SSS. . .	70
4.1	Índices de erro de previsão: $\text{mRSE}(i)$, $U(i)$ e $\text{mRSE}_N(i)$, obtidos para a série simulada com assinatura determinística e os erros médios para o conjunto de 1.000 séries sub-rogadas.	74
6.1	Teste de determinismo usando o método SSS (Seção 3.6.2) nas janelas de dados da série de carga $y_Q(k)$	127
6.2	Modelos NARMA polinomiais usados na previsão e os resultados do teste de determinismo para cada uma das quatro janelas de dados da série residual $\hat{x}_{s_D}(k)$ testadas.	152

LISTA DE FIGURAS

1.1	Séries quase-periódicas simuladas: $y_1(k)$ com assinatura determinística subjacente à quase-periodicidade e, $y_2(k)$ com comportamento aleatório subjacente à quase-periodicidade.	4
1.2	Série temporal periódica. Em (a) $f(\cdot)$ relaciona as observações entre os ciclos. Em (b) $g(\cdot)$ relaciona as observações intra-ciclo.	8
2.1	Série de produção mensal de eletricidade.	20
3.1	Série temporal periódica. Em (a) a função $f(\cdot)$ relaciona as observações entre os ciclos (determinismo entre-ciclos). Em (b) a função $f(\cdot)$ relaciona as observações subjacentes ao ciclo (determinismo intra-ciclo).	37
3.2	Série de produção mensal de eletricidade na Austrália de janeiro de 1956 à agosto de 1995 ELEC.DAT.	45
3.3	Série de produção mensal de eletricidade na Austrália de janeiro de 1956 a agosto de 1995 ELEC.DAT.	46
3.4	Série temporal do mapa de Hénon dado por (3.24).	51
3.5	Componentes de tendência e cíclica para compor a série temporal $y(k)$ (3.27).	52
3.6	Mapa de Hénon estimado pela série residual.	53
3.7	Exemplos de órbitas de séries periódicas com $\omega_1/\omega_2 = 0,5$ e quase-periódicas com $\omega_1/\omega_2 = 1,314159265358\dots$	59
3.8	Exemplo da tomada de decisão de um teste de hipóteses.	61

4.1	Resultado do teste de determinismo para séries cíclicas simuladas: $s_A(k), s_B(k), s_C(k), s_D(k), s_E(k)$	79
4.2	Ilustração do resultado do teste de hipóteses para a aceitação e rejeição de H_0, H_1 e H_2	82
4.3	Histograma dos índices $\text{mRSE}(i)$ obtidos para a previsão de $k = 1 \cdots 6$ passos à frente para as séries sub-rogadas geradas a partir da série com assinatura determinística.	86
4.4	Histograma dos índices $\text{mRSE}(i)$ obtidos para a previsão de $i = 1 \cdots 6$ passos à frente para as séries sub-rogadas geradas a partir de uma série aleatória.	87
4.5	Histograma dos índices $\text{mRSE}(1)$ da previsão de 1 passo à frente das séries sub-rogadas, considerando a distribuição dos índices $\text{mRSE}(i)$ como uma Normal e como uma distribuição qualquer, ambas com 68,27% de confiança.	90
5.1	Teste de determinismo na série quase-periódica com vínculo deter- minístico de curto-prazo, $y_1(k)$, usando o método SSS (Nakamura & Small, 2005).	104
5.2	Teste de determinismo na série série quase-periódica com comporta- mento aleatório subjacente, $y_2(k)$ usando o método SSS (Nakamura & Small, 2005).	105
5.3	Série residual com assinatura determinística $\hat{x}_1(k)$	109
5.4	Série residual aleatória $\hat{x}_2(k)$	109
5.5	Resultados do teste de determinismo usando dados sub-rogados para a série residual com assinatura determinística $\hat{x}_1(k)$, usando os algo- ritmos AAFT, AAWFT, BOOT e CAAFT.	112
5.6	Resultados do teste de determinismo baseado em dados sub-rogados para a série residual com assinatura determinística $\hat{x}_1(k)$, usando os algoritmos FT, GAUSS, IID e WFT.	113

5.7	Resultados do teste de determinismo para uma série residual aleatória $\hat{x}_2(k)$ baseado nos dados sub-rogados usando os algoritmos AAFT, AAWFT, BOOT e CAAFT.	114
5.8	Resultados do teste de determinismo para uma série residual aleatória $\hat{x}_2(k)$ baseado nos dados sub-rogados usando os algoritmos FT, GAUSS, IID e WFT.	115
5.9	Cálculo do índice mRSE.	117
6.1	Série de demanda de carga da região de Queensland, Austrália.	121
6.2	Função de autocorrelação e periodograma da série de demanda de carga da região de Queensland, Austrália $y_Q(k)$	122
6.3	Testes de estacionariedade para a série de carga da região de Queensland, Austrália.	125
6.4	Teste de estacionariedade para as <i>janelas 1, 2 e 3</i>	126
6.5	Teste de determinismo usando o método SSS (Nakamura & Small, 2005) aplicado à série $y_Q(k)$	128
6.6	Série residual $\hat{x}_{Q_{24}}(k)$	131
6.7	Série residual $\hat{x}_{Q_{168}}(k)$	132
6.8	Teste de determinismo intra-ciclo da série de carga $y_{Q_{168}}(k)$, realizado a partir de sua respectiva série residual $\hat{x}_{Q_{168}}(k)$	135
6.9	Série temporal quase-periódica de casos mensais de caxumba $y_c(k)$	138
6.10	Teste de determinismo para a série de índices de caxumba $y_c(k)$, a partir de sua respectivamente série de resíduos $\hat{x}_c(k)$ dada em (6.10).	140
6.11	Série temporal periódica de produção mensal de leite $y_L(k)$ (<i>milk.dat</i>) e, sua correspondente série residual $\hat{x}_L(k)$	141
6.12	Teste de determinismo para a série de produção mensal de leite, $y_L(k)$, realizado a partir de sua correspondente série de resíduos $\hat{x}_L(k)$	143
6.13	Série quase-periódica, $y_{s_a}(k)$, de números anuais de manchas solares (<i>sunspots</i>) e sua respectiva série residual $\hat{x}_{2_{s_a}}(k)$	147

6.14	Teste de determinismo para a série quase-periódica, $y_{s_a}(k)$, de números anuais de manchas solares (<i>sunspots</i>) realizado a partir de sua respectiva série residual $\hat{x}_{2s_a}(k)$	148
6.15	Série de números diários de manchas solares (<i>sunspots</i>), $y_{s_D}(k)$, e sua respectiva série residual $\hat{x}_{s_D}(k)$	150
6.16	Teste de determinismo para a série de resíduos $\hat{x}_{s_D}(k)$	151

LISTA DE SÍMBOLOS

k	Tempo discreto
t	Tempo contínuo
$\text{var}(\cdot)$	Variância
$y(k - j)$	Regressor de Saída
$e(k)$	Ruído presente no sistema
n_y	Máximo atraso do sinal de saída
n_e	Máximo atraso do sinal de ruído
$f(\cdot)$	Função linear ou não-linear
ℓ	Grau de não-linearidade
n_θ	número de termos candidatos a um modelo NARX polinomial
N	Número de dados
N_i	Número de dados usados na identificação de um modelo
N_v	Número de dados usados na validação de um modelo
ξ	Resíduo
Θ	Vetor de parâmetros
$\phi_{y'y'}$	Função de autocorrelação linear do sinal de saída
$\phi_{y'^2y'^2}$	Função de autocorrelação não-linear do sinal de saída
$\Omega_{y^p u^{m-p}}$	Agrupamento de termos
$\Sigma_{y^p u^{m-p}}$	Coefficiente do agrupamento de termos $\Omega_{y^p u^{m-p}}$
\mathbf{Y}	Vetor Contendo os Regressores da Saída

$J(\cdot)$	Função Custo
r	coeficiente de correlação linear
d	período sazonal
$s(k)$	componente de sazonalidade
$m(k)$	componente de tendência
τ	atraso de tempo
d_i	dimensão de imersão

NOMENCLATURA

ARX	Auto-Regressivo com variáveis eXógenas
ARIMA	Auto-Regressivo Integrado e de Média Móvel
NARMAX	Não-linear Auto-Regressivo de Média Móvel com variáveis eXógenas
NARMAX	Não-linear Auto-Regressivo de Média Móvel com variáveis eXógenas
NARX	Não-linear Auto-Regressivo com Entrada eXógena
RNA	Rede Neural Artificial
ERR	taxa de redução de erro (<i>Error Reduction Ratio</i>)
mRSE	média da raiz do erro quadrático (<i>mean Root Square Error</i>)
RMSE	raiz do erro médio quadrático (<i>Root Mean Square Error</i>)
MAPE	erro percentual absoluto médio (<i>Mean Absolute Percentage Error</i>)
PPS	dados sub-rogados pseudo-periódicos (<i>PseudoPeriodic Surrogates</i>)
SSS	dados sub-rogados gerados pelo embaralhamento temporal de curto-prazo (<i>Small-Shuffle Surrogates</i>)
FAC	função de autocorrelação
<i>sunspots</i>	manchas solares
AAFT	<i>Amplitude adjusted Fourier transform algorithm</i>
IID	<i>independent and identically distributed algorithm</i>
BOOT	<i>Bootstrap algorithm</i>

FT	<i>Fourier Transform algorithm</i>
WFT	<i>Windowed Fourier transform algorithm</i>
GAUSS	<i>Gaussian algorithm</i>
FMMC	Filtro de Média Móvel para componentes Cíclicas
H_0	hipótese nula
H_1 e H_2	hipóteses alternativas

INTRODUÇÃO

Um dos principais interesses em análise de séries temporais é a previsão. Em muitas áreas, as tomadas de decisões são baseadas em vários aspectos, dentre eles, nos comportamentos futuros dos sistemas geradores das séries temporais. Para prever esses comportamentos futuros é necessário encontrar um modelo adequado (Makridakis, Andersen, Carbone, Fildes, Hibon, Lewandowski, Newton, Parzen & Winkler, 1982). Muitos trabalhos utilizam de técnicas mais sofisticadas (não-lineares) para modelar e prever séries temporais. No entanto, nem sempre é necessário o uso de tais técnicas (Darbellay & Slama, 2000).

Há algumas técnicas disponíveis na literatura para decidir se a série temporal é ou não modelável e, assim, apontar qual a técnica, linear ou não-linear, mais adequada para modelá-la. Dentre elas, pode-se citar as técnicas de detecção de determinismo e não-linearidade (Sugihara & May, 1990; Kaplan & Glass, 1992; Theiler, Eubank, Longtin & Farmer, 1992; Theiler, 1995; Small, Yu & Harrison, 2001; Nakamura & Small, 2005).

Ao detectar que a série temporal tem vínculo determinístico, tem-se uma indicação de que se pode procurar por um modelo adequado para tal série. Caso contrário, são dispensáveis quaisquer esforços para encontrar um modelo para a série, conseqüentemente, não é possível obter previsões para essa série temporal.

No princípio da década de oitenta, utilizava-se de forma indiscriminada, as invariantes dinâmicas para detectar determinismo e não-linearidade em séries tem-

porais. O principal interesse dos pesquisadores na época era verificar se a série tinha comportamento determinístico caótico (Nicolis & Nicolis, 1984; Grassberger, 1986; Nicolis & Nicolis, 1987; Kaplan & Cohen, 1990). Devido ao uso indiscriminado desses fatores, muitos resultados espúrios foram encontrados (Kanters, Holsteinrathlou & Agner, 1994; Theiler, 1995).

Muitas séries reais apresentam um comportamento quase-periódico,¹ que no jargão de séries temporais é denominado comportamento sazonal. A diferença essencial entre comportamento sazonal e cíclico é que a primeira possui padrões de oscilação que ocorrem em períodos anuais ou submúltiplos ao ano. No entanto, o comportamento cíclico apresenta oscilações com períodos não necessariamente anuais.

Prever os valores futuros e entender o mecanismo do sistema gerador de séries temporais com padrão de comportamento quase-periódico é de grande interesse em diversas áreas do conhecimento, como engenharia, economia e medicina (Lima, Rodrigues & Aguirre, 2002; Coelho, Silva, Fonseca & Aguirre, 2003; Coelho, Silva, Martinez & Aguirre, 2003; Small & Tse, 2003; Coelho, Oliveira & Mendes, 2004; Zhang & Qi, 2005; Small, 2005). Além disso, é desejável verificar se há alguma dinâmica determinística além do comportamento quase-periódico da série temporal. Há algumas técnicas disponíveis na literatura que podem ser usadas para inferir se uma série temporal tem ou não vínculo determinístico (Gomes, 2001; Small et al., 2001; Nakamura & Small, 2005).

1.1 Apresentação do problema

O comportamento quase-periódico de uma série temporal tende a mascarar qualquer outro determinismo além do próprio comportamento quase-periódico. Em princípio, nem sempre é possível detectar se há dinâmica determinística em uma

¹Informalmente, pode-se dizer que o comportamento quase-periódico da série temporal está relacionado a padrões que, apesar de parecidos, não se reproduzem exatamente em intervalos constantes de tempo. O comportamento quase-periódico de séries temporais será melhor definido na seção 3.4.4.

série temporal, além do comportamento quase-periódico, somente por meio de uma inspeção visual da série ou usando técnicas convencionais como a função de autocorrelação (FAC) e o espectro de potência das séries.

Sejam duas séries quase-periódicas por construção, $y_1(k)$ e $y_2(k)$. Por meio de uma inspeção visual das séries na Figura 1.1(a) e (b) é impossível distinguir se há comportamento dinâmico determinístico superposto à oscilação quase-periódica em ambas as séries. Como pode ser observado nas Figuras 1.1(c), (d), (e) e (f), as séries, $y_1(k)$ e $y_2(k)$, apresentam funções de autocorrelação e periodogramas muito parecidos, o que impossibilita a distinção entre uma série e a outra, utilizando-se somente essas ferramentas convencionais.

Contudo, sabe-se *a priori* que a série $y_1(k)$ apresenta vínculo determinístico linear (de curto-prazo) além da quase-periodicidade e, a série $y_2(k)$ tem comportamento aleatório subjacente ao padrão quase-periódico (ver equações (5.5) e (5.6) na página 103).

Este trabalho propõe um procedimento baseado em teste de hipóteses para decidir sobre a presença desse vínculo determinístico de curto-prazo em séries temporais quase-periódicas.

1.1.1 Técnicas para detectar determinismo

Inferir sobre o comportamento dinâmico de um sistema por meio de uma série temporal medida desse sistema é um problema de interesse em diversas ciências. O problema de se distinguir uma série temporal determinística de uma série aleatória começou a ser melhor estudado no início da década de oitenta. Uma vez que diversas séries reais, possivelmente caóticas, apresentavam um comportamento dinâmico que era confundido com o comportamento aleatório.

Essa distinção entre determinismo e aleatoriedade era feita até então por meio dos cálculos, via séries temporais, da dimensão de correlação, dos expoentes de Lyapunov, da entropia de Kolmogorov (Grassberger & Procaccia, 1983b,a; Wolf, Swinney & Vastano, 1985; Eckmann & Ruelle, 1985; Eckmann, Kamphorst, Ruelle

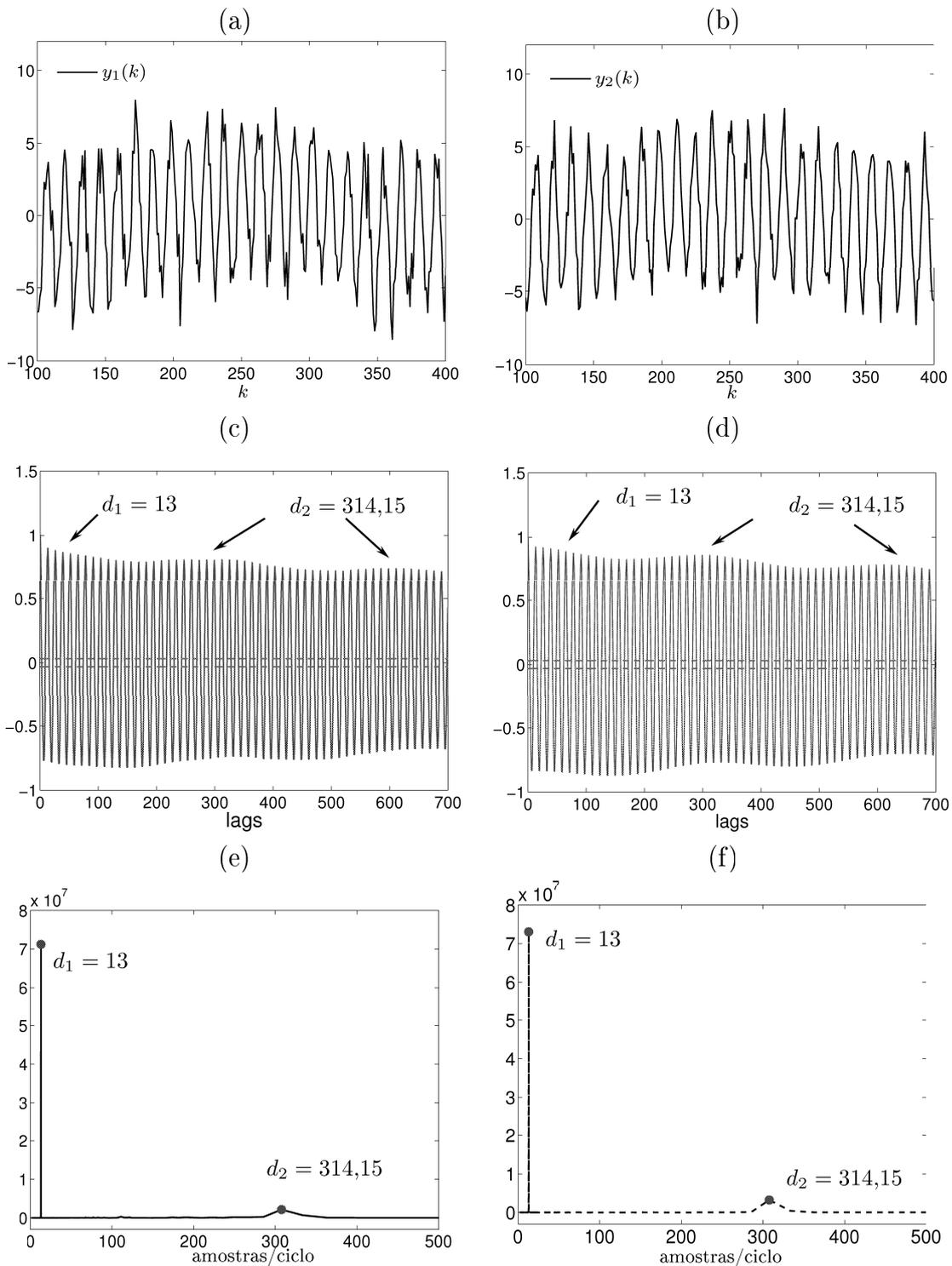


Figura 1.1: Séries quase-periódicas simuladas. Em (a) série quase-periódica com vínculo determinístico subjacente, $y_1(k)$. Em (b) série quase-periódica com comportamento aleatório subjacente, $y_2(k)$. Em (c) FAC de $y_1(k)$. Em (d) FAC de $y_2(k)$. Em (e) periodograma de $y_1(k)$. Em (f) periodograma de $y_2(k)$.

& Ciliberto, 1986).²

Baseado na teoria de que séries determinísticas têm dimensão finita enquanto séries aleatórias têm dimensão infinita, para uma quantidade de dados infinita, o método para o cálculo da dimensão de correlação de séries temporais, proposto por Grassberger e Procaccia (1983b), tornou-se o mais popular dentre todos. Na prática, verifica-se a necessidade de uma grande quantidade de dados para que o método numérico de cálculo da dimensão de correlação consiga distinguir se a série temporal é advinda de um processo aleatório ou de um processo determinístico (Osborne & Provenzale, 1989; Theiler, 1988, 1991). Porém, mesmo usando uma grande quantidade de dados, alguns algoritmos para o cálculo da dimensão de correlação indicam uma dimensão finita, mesmo quando os dados são aleatórios (Osborne & Provenzale, 1989; Kennel & Isabelle, 1992).

O fato da presença de ruído e de não estacionariedade nos dados poder imitar ou mascarar determinismo nos índices clássicos para caracterizar caos (expoente de Lyapunov, entropia de Kolmogorov e dimensão de correlação) fez com que a necessidade de novos métodos para distinguir determinismo de aleatoriedade, diretamente da série temporal, fossem especificados. Somente no início da década de noventa outros métodos foram propostos para a análise de determinismo (Sugihara & May, 1990; Kaplan & Glass, 1992; Wayland, Bromley, Pickett & Passamante, 1993; Salvino & Cawley, 1994; Salvino, Cawley, Grebogi & Yorke, 1995).

Em 1992 Theiler e colegas propuseram um método menos pretensioso do que o de estimar a dimensão de correlação da série temporal em questão e, assim, caracterizá-la como determinística ou aleatória. Para isso, Theiler e colegas (1992) sugeriram que a dimensão de correlação estimada para a série temporal medida fosse comparada com as dimensões de correlação estimadas para um conjunto de séries

²Alguns desses índices, tal como a dimensão de correlação que, em princípio, é invariante à imersão. Sabe-se que a dinâmica caótica é determinística. Porém, como às vezes o comportamento caótico é muito parecido com o comportamento aleatório a ponto de confundí-los, o uso desses índices auxilia na distinção de uma dinâmica determinística de um comportamento aleatório. Além disso, o avanço dos computadores no início da década de oitenta proporcionava uma maior facilitação no cálculo desses índices.

artificiais.³ Assim, com base nessa comparação, pode-se inferir sobre a presença de determinismo ou aleatoriedade na série medida.

Sugihara e May (1990) usaram a capacidade de previsão da série temporal como meio para distinguir determinismo de aleatoriedade. Os autores compararam o resultado da previsão da série medida aumentando o horizonte de previsão. Quando a série é determinística caótica, à medida que se aumenta o horizonte de previsão o erro de previsão aumenta, enquanto que para uma série aleatória o erro mantém-se praticamente constante (Sugihara & May, 1990). Vale destacar que, no trabalho de Sugihara e May, a maneira como se distingue uma série determinística de uma série aleatória é válida somente para séries caóticas que possuem essa característica de prever apenas poucos passos à frente, ou seja, o índice de erro calculado aumenta com o aumento do horizonte de previsão.

Dois anos mais tarde, Kaplan e Glass (1992; 1993) propuseram uma alternativa para a análise de determinismo em séries temporais. De forma sucinta, o método consiste em analisar as trajetórias (que são vetores) no espaço de fases da série temporal. Segundo os autores, quando o valor da dimensão e do atraso de imersão fossem corretos, se as trajetórias vizinhas possuírem direções iguais, ou similares, em pequenas regiões do espaço de fases, significaria que a dinâmica presente na série é determinística. Porém, quando o comportamento presente fosse aleatório, as trajetórias vizinhas no espaço possuiriam direções diferentes.⁴

Em suma, a técnica supracitada, usada para a detecção de determinismo por meio da análise de continuidade no espaço de fases não é aplicável diretamente na série temporal, é necessário antes, reconstruir o espaço de fases adequado para a série sob investigação. Dessa forma, a série temporal é imersa em um espaço de fases (Takens, 1980) e esse espaço é dividido em pequenas caixas, por meio das quais as direções das trajetórias vizinhas são analisadas. Deve-se atentar que a correta

³Theiler e colegas (1992) geraram um conjunto de 5 a 10 séries artificiais (*surrogates*) aleatórias por construção, as quais mantiveram algumas características estatísticas da série temporal medida.

⁴No espaço de fases de uma série determinística, o fato das trajetórias vizinhas terem a mesma direção assegura a singularidade de solução, que é uma das características de sistemas determinísticos (Kaplan & Cohen, 1990).

reconstrução do espaço de fases depende de uma escolha adequada da dimensão de imersão e do atraso de imersão (Fiedler-Ferrara & Prado, 1994; Monteiro, 2002).

Como é mostrado no artigo (Kaplan & Glass, 1992), o método apresenta algumas deficiências, como por exemplo, a necessidade de um grande número de observações. Conseqüentemente, o método necessita de um alto esforço computacional.

Ainda nessa linha de raciocínio de detecção de determinismo por meio da análise de continuidade no espaço de fases da série, destacam-se os trabalhos (Wayland et al., 1993; Salvino & Cawley, 1994; Ortega & Louis, 1998; Ortega, 2000), que são computacionalmente mais simples, podendo ser usados em uma pequena quantidade de dados.

Baseados na capacidade de previsão das séries para detectar determinismo destacam-se também os trabalhos (Kennel & Isabelle, 1992; Theiler et al., 1992; Gomes, 2001).

1.2 Teste de determinismo usando dados sub-rogados

Como mencionado na seção anterior, a análise de séries temporais usando dados sub-rogados foi introduzida, primeiramente, para assegurar a validade dos resultados das investigações de não-linearidade e caos determinístico (Theiler et al., 1992). A partir desse trabalho diversos pesquisadores desenvolveram métodos de análise de não-linearidade e determinismo em séries temporais baseados em dados sub-rogados (Schreiber & Schumitz, 2000; Gomes, 2001; Small, 2005).

No contexto deste trabalho, o interesse é detectar dinâmica determinística intra-ciclos por meio da previsibilidade da série temporal quase-periódica. As técnicas até então existentes eram para detectar determinismo em séries não periódicas e estacionárias (Sugihara & May, 1990; Theiler et al., 1992; Kaplan & Glass, 1992).

Por determinismo entre-ciclos entende-se que existe uma lei que relaciona a disposição dos ciclos sazonais e, por determinismo intra-ciclo entende-se que existe uma lei que governa a relação entre as observações subjacentes aos ciclos sazonais, ou seja, a relação de curto tempo entre as observações. Neste trabalho o interesse é

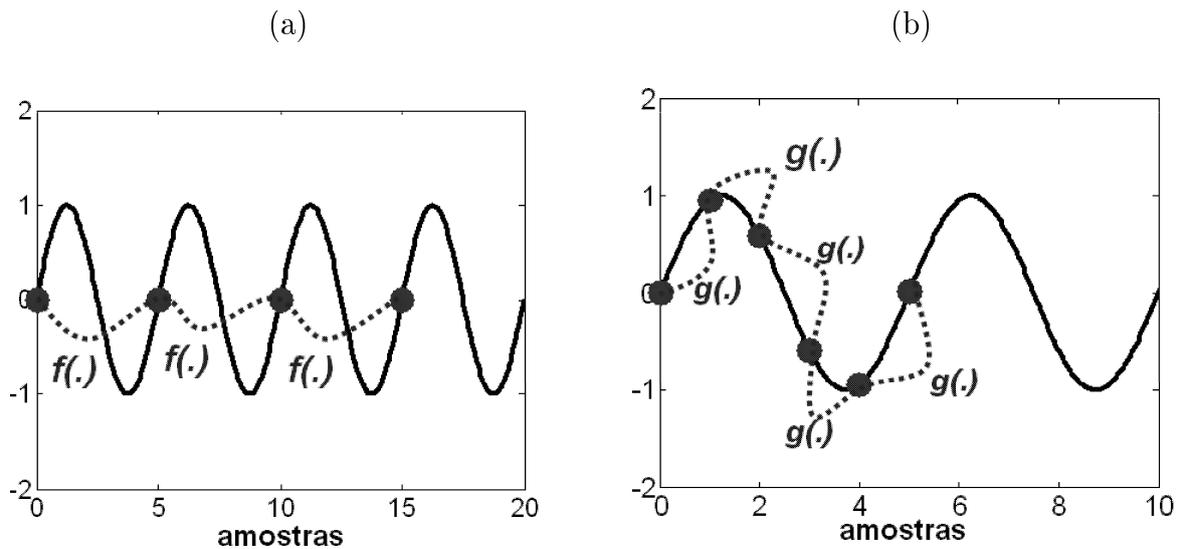


Figura 1.2: Série temporal periódica. Em (a) $f(\cdot)$ relaciona as observações entre os ciclos. Em (b) $g(\cdot)$ relaciona as observações intra-ciclo.

detectar se há determinismo intra-ciclos nas séries quase-periódicas. As Figuras 1.2 (a) e (b) ilustram o determinismo entre-ciclos e intra-ciclo, respectivamente.

Para séries periódicas apenas poucos métodos foram desenvolvidos para analisar determinismo, como o algoritmo CSS (*Cycle Shuffled Surrogate*) proposto por Theiler e colegas (1995), o qual procura por determinismo entre-ciclos por meio do embaralhamento dos ciclos. O algoritmo CSS divide a série pseudo-periódica em segmentos de mesmo tamanho. Cada segmento contém um número inteiro de ciclos. Por meio do embaralhamento aleatório desses segmentos obtém-se os dados sub-rogados. Esse algoritmo destrói a correlação entre-ciclos e mantém a correlação intra-ciclos.

Um problema prático em se implementar esse algoritmo é que é necessário conhecer anteriormente, de forma precisa, o período em que ocorrem os ciclos, que pode variar de um ciclo para outro. Conseqüentemente, gera-se uma descontinuidade entre os ciclos (Theiler & Rapp, 1996; Small, 2005).

A técnica PPS (*Pseudo-Periodic Surrogate*) proposta por Small e colegas (2001) também procura por determinismo entre-ciclos, usando para isso a imersão

da série temporal para gerar os dados sub-rogados. Esses métodos não são aplicáveis neste trabalho, pois procuram por determinismo entre-ciclos e não intra-ciclo.

A técnica SSS (*Small Shuffled Surrogate*) proposta por Nakamura e Small (2005) detecta determinismo intra-ciclo, em séries cíclicas, via embaralhamento temporal de curto-prazo da série temporal. Esse algoritmo é similar ao algoritmo *Shuffled* proposto em (Theiler et al., 1992), porém, faz um embaralhamento de curto-prazo mantendo os ciclos da série. Apesar dessa técnica ser focada para o mesmo objetivo deste trabalho, determinismo intra-ciclo, a mesma apresenta resultados espúrios quando submetido à séries semelhantes como mostrado na Seção 5.3.

Percebe-se que as três técnicas mencionadas - CSS, PPS e SSS - para testar determinismo em séries quase-periódicas consistem em uma nova maneira de gerar os dados sub-rogados. Destaca-se ainda que tais técnicas foram desenvolvidas para serem aplicadas diretamente à série quase-periódica, ou seja, não é necessário extrair a componente quase-periódica. O teste de determinismo proposto neste trabalho procura aliar o que já se tem disponível, e bem estabelecido, na literatura para gerar dados sub-rogados. Para isso, tem-se que adaptar as séries quase-periódicas para que se possa empregar tais algoritmos. Logo, é necessário extrair a componente quase-periódica e de tendência da série pseudo-periódica.

1.3 Redução da componente quase-periódica

Uma das técnicas mais usadas na área econômica para extrair as componentes sazonal e de tendência é conhecida como X-11-ARIMA (Shiskin, Young & Musgrave, 1967) e, sua versão adaptada o X-12-ARIMA (Findley, Monsell, Bell, Otto & Chen, 1998). Esse método foi desenvolvido tendo em vista o ajustamento sazonal em séries econômicas que, freqüentemente, apresentam sazonalidades mensais, trimestrais e semestrais, com respectivos períodos sazonais $d = 12, 4, 2$. No entanto, em aplicações empíricas constata-se que o X-11 não é aplicável a séries físicas, biológicas, as quais podem apresentar períodos cíclicos maiores do que 12.

Outro método bastante difundido para a redução da sazonalidade e ciclos é

o filtro de médias móveis cíclicas (FMMC), o qual é empregado neste trabalho para quaisquer períodos cíclicos e está descrito na Seção 3.3.4.

1.3.1 Objetivos

Inspirado nos trabalhos de (Theiler et al., 1992; Gomes, 2001), este trabalho tem como principal objetivo propor um procedimento para detectar determinismo intra-ciclo em séries quase-periódicas. Para isto, emprega-se o teste de hipóteses usando dados sub-rogados, baseado na idéia de que previsibilidade implica determinismo.

Inicialmente, é necessário estimar e reduzir as componentes cíclicas e de tendência da série quase-periódica, obtendo assim a série residual. A partir da série residual, gera-se um conjunto de séries artificiais, as quais mantêm algumas características estatísticas da série residual.

A fim de se avaliar o aspecto determinístico da série residual, deve-se utilizar um fator estatístico discriminante endereçado a esta característica. Tal índice discriminante é obtido após a submissão das séries residual e artificiais à modelagem não-linear e à predição de 1 até h passos à frente. Se o índice discriminante obtido para a série residual for qualitativamente diferente dos índices obtidos para os dados sub-rogados, pode-se rejeitar a hipótese nula (H_0) e aceitar uma hipótese alternativa H_1 . Caso o índice discriminante obtido para a série residual seja semelhante aos obtidos pelas séries artificiais, H_0 não poderá ser rejeitada. Nesse sentido este trabalho propõe-se a:

- Investigar o método adequado para a redução da quase-periodicidade nas séries temporais, de modo a obter a série residual.
- Definir qual algoritmo para gerar os dados sub-rogados é o mais adequado para se testar determinismo.
- Investigar quais índices de erros são adequados para avaliar a capacidade de

previsão das séries, residual e sub-rogadas, conseqüentemente, avaliar a assinatura determinística presente na série quase-periódica.

- Analisar a distribuição desses índices de erro e assim calcular o intervalo de confiança para tal índice de erro, conseqüentemente, atribuir um grau de confiança ao teste de determinismo intra-ciclo.
- Definir a hipótese nula e a(s) hipótese(s) alternativa(s) com respeito ao índice de erro empregado.

1.4 Organização do texto

Este trabalho é composto por sete capítulos. Neste capítulo, procurou-se localizar na literatura os métodos até então usados no teste de determinismo de séries temporais. No Capítulo 2 são apresentados as principais técnicas, lineares e não-lineares, de modelagem de séries temporais.

No Capítulo 3 são apresentadas algumas das técnicas para a análise de séries temporais. Define-se as principais características de uma série temporal, bem quando a mesma apresenta um comportamento quase-periódico. São mostrados, por meio de resultados simulados, os desempenhos de alguns dos métodos usados na literatura para a quantificação e redução das componentes cíclicas, sazonais e de tendência.

No capítulo 4, são apresentadas algumas das contribuições desta tese com respeito ao estudo e escolha do índice discriminante. No caso deste trabalho, o índice escolhido é o mRSE, em que foi feito um estudo da distribuição aproximada deste índice para o cálculo dos intervalos de confiança, e assim, atribuir um grau de confiança ao teste de determinismo. Com base no índice mRSE são definidas as hipóteses nula e alternativas.

No capítulo 5 é descrito o procedimento para testar determinismo intra-ciclo em séries quase-periódicas. Verifica-se também, por meio de séries quase-periódicas simuladas, qual o algoritmo adequado para gerar séries sub-rogadas para o teste de determinismo.

No capítulo 6 são apresentados alguns resultados da aplicação do teste de determinismo em séries quase-periódicas reais de: consumo diário de energia elétrica, casos mensais de caxumba, produção mensal de leite e números de manchas solares (*sunspots*).

No capítulo 7 são feitas algumas discussões e conclusões acerca do método proposto e de suas aplicações apontando algumas sugestões de trabalhos futuros.

MODELAGEM PARA PREVISÃO DE SÉRIES TEMPORAIS

Várias técnicas têm sido utilizadas na análise e previsão de séries temporais. A análise tradicional de séries baseia-se no espectro de potência e/ou na função de autocorrelação linear. Muito tem-se feito na tentativa de desenvolver técnicas mais robustas para a análise de séries reais, que em sua maioria, apresentam comportamento dinâmico não-linear (Weigend & Gershenfeld, 1994; Schreiber, 1999; Bagarinao, Pakdaman, Nomura & Sato, 1999; Soofi & Cao, 2002).

Por meio da modelagem de uma série temporal pode-se entender o mecanismo do sistema gerador da série e prever valores futuros, sendo que a análise da série, que visa sintetizar suas propriedades e caracterizar seu comportamento, pode identificar ou sugerir características para auxiliar na escolha do modelo adequado.

Um modelo procura descrever a relação entre as variáveis do sistema gerador da série temporal. O modelo é dito matemático quando for possível expressar tais relações em termos de equações matemáticas. Dentre as motivações para a construção de modelos pode-se citar que:

- algumas técnicas de controle avançado requerem um modelo do sistema;
- modelos podem ser usados para análise, simulação e previsão do comportamento do sistema.

A complexidade de um modelo matemático depende da aplicação desejada. O levantamento e a formulação matemática de todos os fenômenos que afetam o comportamento de um dado sistema é uma tarefa extremamente complexa. Com isso, é impossível o modelo reproduzir exatamente o comportamento do sistema original. Entretanto, o modelo deve ser capaz de reproduzir o comportamento original da melhor maneira possível pois, caso contrário, todos os esforços posteriores para a análise e previsão do sistema serão pouco eficientes. Dois grandes grupos de técnicas usadas para a obtenção de modelos matemáticos são: **a)** modelagem pela física do processo ou modelagem caixa-branca; **b)** modelagem a partir de dados ou identificação de sistemas que subdivide-se em identificação caixa-preta e caixa cinza;

A modelagem pela física do processo utiliza as leis físicas que descrevem os fenômenos envolvidos, além de conhecimento *a priori* sobre o sistema. Esta é feita somente quando se tem disponível uma visão global do comportamento do processo (Ogata, 1998; Garcia, 2006).

Quando o sistema é grande e complexo e há a necessidade de atuação de controle em tempo real, por exemplo, torna-se inviável a utilização de técnicas de modelagem caixa-branca. Uma alternativa é a utilização de técnicas de identificação de sistemas, por meio das quais o modelo é obtido a partir de dados de entrada e de saída e, perturbações (quando disponível), do sistema. Este tipo de identificação é denominado caixa-preta, uma vez que só utiliza as informações contidas nos dados. Por outro lado, a identificação caixa-cinza utiliza, além dos dados, alguma(s) informação(s) *a priori* para auxiliar na determinação da estrutura e/ou estimação dos parâmetros resultando em modelos mais concisos (Aguirre, 2004; Corrêa, 2001; Barroso, 2001, 2006).

O método mais proeminente, dentre as técnicas clássicas de modelagem em séries temporais, é a aproximação proposta por Box & Jenkins em 1976. Dentre as derivações lineares do modelo de Box & Jenkins vale destacar os modelos ARIMA

(*AutoRegressive Integrated Moving Average*), AR¹, ARMA², ARX³, ARMAX⁴. Ainda dentro dos modelos de regressão encontram-se os modelos NARMAX (*Nonlinear AutoRegressive Moving Average with eXogenous inputs*) polinomiais/rationais que utilizam técnicas de identificação não-lineares (Aguirre, 2004).

O teste de determinismo, aqui proposto, baseia-se na idéia de que previsibilidade implica determinismo. A partir da série temporal a ser analisada, gera-se um conjunto de séries artificiais e ajusta-se modelos às séries, original e sub-rogadas, e assim realiza-se as previsões de tais séries. Este capítulo apresenta a revisão de algumas das principais técnicas para a modelagem de séries temporais.

2.1 Modelagem matemática

De maneira geral, os modelos podem ser classificados como lineares (AR) e não-lineares (NARMAX, RNA⁵, RBF); locais (PL⁶) e globais (NARMAX, RNAs); paramétricos (NARMAX) e não-paramétricos (resposta ao impulso e resposta em frequência) e determinístico (AR, RNA) e estocástico (ARMA). Os tipos de modelos são inúmeros, sendo que maiores detalhes podem ser vistos em (Aguirre, 2004).

Os modelos globais oferecem uma alternativa à aproximação local. No entanto, é provável que um único modelo global não seja capaz de representar a dinâmica de um sistema. Conseqüentemente, a maioria dos estudos nesta área investigam modelos globais com comportamento localizado, tais como o PL, a rede *neuro-fuzzy* de Yamakawa (Caminhas & Gomide, 2000) e as redes RBFs (*Radial Basis Functions*) (Amaral, 2006).

A rede RBF é um exemplo de modelo global com aproximação local. O uso de redes RBFs em previsão de séries foi introduzido por Broomhead & Lowe (1988) e, no contexto de dinâmica não-linear, por Casdagli (1989). Na área de identificação

¹AR (*AutoRegressive*).

²ARMA (*AutoRegressive Moving Average*).

³ARX (*AutoRegressive with eXogenous variables*).

⁴ARMAX (*AutoRegressive Moving Average with eXogenous variables*).

⁵RNA (Rede Neural Artificial (Haykin, 2001)).

⁶PL (Preditor Local).

de sistemas o uso de RBFs foi popularizado a partir dos trabalhos de Chen e colegas (1990a; 1990b; 1991).

A rede RBF é uma classe especial de redes neurais *feedforward* com três camadas: a de entrada, que é completamente conectada à camada intermediária, sem pesos nas conexões. Os nodos da camada intermediária, os quais são compostos por RBFs de maneira análoga às funções de ativação em RNA, são então conectados à camada de saída com o uso de pesos nas conexões.

Nas últimas duas décadas, as redes RBFs têm sido aplicadas ao problema de previsão em diversas áreas das ciências biológicas (Naguib, Wayman, Bennett, Raimés & Griffin, 1998), econômicas (Guo & Luh, 2003) e das engenharias (Ranaweera, Hubele & Papalexopoulos, 1995; Mason & Kambhampati, 1996; Chang, Liang & Chen, 2001; Gontar, Sideratos & Hatziargyriou, 2004; Sinclair, Musavi & Qiao, 1995). Outras aplicações podem ser vistas em (Sze, 1995).

Os modelos fuzzy (nebulosos)(RNAs), redes neurais e os neuro-fuzzy são baseados em técnicas de inteligência artificial. Esta última tem sido muito utilizada devido a sua alta capacidade de adaptação. Porém, muitos trabalhos utilizam os modelos clássicos devido, principalmente, à sua fundamentação teórica (Hwarng & Ang, 2001).

Modelos nebulosos mapeiam um conjunto de entradas em uma saída, baseado em um sistema de inferência que utiliza uma base de raciocínio do tipo *if – then* e o valor predito pode ser obtido usando um método de desnebulização (ou *defuzzificação*) (Yang & Huang, 1998; Doveh, Feigin, Greig & Hyams, 1999).

Os modelos nebulosos realizam, por meio de regras nebulosas, um mapeamento de regiões nebulosas do espaço de entradas e saídas. As regiões do espaço de entradas e saídas são determinadas na etapa de identificação da estrutura destes modelos. Essa divisão do espaço de entradas e saídas é feita por meio do método de partição. A escolha do método de particionamento tem grande influência sobre o desempenho desse tipo de modelo. A rede proposta por Yamakawa é bastante atrativa para lidar com problemas de previsão *on-line* (Caminhas & Gomide, 2000), uma

vez que ela divide o espaço das entradas em regiões em que somente duas funções de pertinência complementares atuam simultaneamente. Isso reduz o seu esforço computacional.

Um método não-linear muito utilizado é o baseado em redes neurais artificiais (RNAs). As RNAs têm sido muito utilizadas devido a sua alta capacidade de adaptação e flexibilidade, sendo capazes de modelar uma vasta classe de funções. Porém, a sua flexibilidade tem um preço. Ainda não há nenhum método estabelecido para determinar a estrutura que melhor aproxima a função a ser mapeada por meio de entradas e saídas. Como resultado, são necessários experimentos exaustivos de tentativa e erro que consomem tempo. Esse esforço pode dificultar a aplicação da RNA, principalmente a do tipo MLP (*MultiLayer Perceptron*), em aplicações *on-line* (Oliveira, Machado & Wazlawick, 2000).

RNAs têm sido empregadas em diversos problemas de previsão, como por exemplo: em previsão de séries caóticas (Lapedes & Farber, 1987), de séries de manchas solares (Park, Murray & Chen, 1996), de séries financeiras (Makridakis & Hibon, 2000; Soofi & Cao, 2002), séries de carga (Reis, 1999; Silva, Rodrigues, RochaReis, Moulin & Neto, 2001; Silva & Moulin, 2000; Hippert, Pedreira & CastroSouza, 2001; Rodrigues, 2002), entre outras (Weigend & Gershenfeld, 1994; Zhang, Patuwo & Hu, 1998).

Apesar das RNAs apresentarem resultados bastante promissores, estes também incorporam incertezas e ainda não é evidente os efeitos da escolha de fatores chave (tais como o número de nodos na camada intermediária e as entradas) no desempenho da previsão usando RNAs (Zhang et al., 1998).

2.1.1 Modelos lineares auto-regressivos

Pode-se dizer que a modelagem de séries temporais teve início com o trabalho de Yule (1927), que utilizou modelos lineares auto-regressivos AR (*AutoRegressive*) no problema de previsão da série de números de manchas solares de Wolfer (*sunspots*), sendo esta uma série bastante conhecida devido a sua alta variabilidade

e dificuldade de predição (Mingoti & Neves, 1999). O modelo autônomo AR, o qual utiliza informações passadas da série temporal para estimar a saída atual, $y(k)$, é dado por:

$$y(k) = \sum_{j=1}^{n_y} a_j y(k-j) + e(k), \quad (2.1)$$

sendo $y(k)$ a série temporal, $a_0 = 1$ e a_j os parâmetros da parte auto-regressiva do modelo ($a_{n_y} \neq 0$), n_y é um valor inteiro que define o máximo atraso de y , conseqüentemente, a ordem dinâmica do modelo AR. $e(k)$ é o ruído.

Em Wold (1954) foi mostrado que qualquer série temporal discreta poderia ser representada por modelos AR e modelos de média móvel, MA (*Moving Average*). Porém, a implementação desses métodos só foi possível a partir da década de sessenta com o advento dos computadores (Morettin & Tolo, 2004).

Modelos ARMA

Uma classe mais geral de representações lineares autônomas, denominada modelos ARMA (*AutoRegressive Moving Average*), utiliza o resíduo no processo de estimação, sendo dado por:

$$y(k) = \sum_{j=1}^{n_y} a_j y(k-j) + \sum_{i=1}^{n_e} c_i e(k-i) + e(k). \quad (2.2)$$

$$A(q)y(k) = C(q)e(k). \quad (2.3)$$

Destaca-se que nos modelos ARMA as raízes dos polinômios:

$$\begin{aligned} A(q) &= 1 - \sum_{j=1}^{n_y} a_j q^{-j} \\ C(q) &= \sum_{j=1}^{n_e} c_j q^{-j}, \end{aligned} \quad (2.4)$$

devem ter módulo menor que a unidade (Tong, 1990). $A(q)$ e $C(q)$ são denominados funções geradoras auto-regressiva e de médias móveis, respectivamente, e q o operador de atraso de modo que $y(k)q^{-1} = y(k-1)$. Para assegurar estacionariedade $|A| < 1$ (Box & Jenkins (1976), pag. 87).

Algumas vantagens e limitações dos modelos ARMA podem ser encontradas em Tong (1990), tais como:

- matematicamente, os modelos lineares baseados em equações de diferenças lineares são considerados o tipo mais simples de representações discretas e possuem uma teoria bem fundamentada (Box & Jenkins, 1976);
- O tempo computacional requerido para obter um modelo parcimonioso ARMA é pequeno;
- salienta-se que, ao longo dos anos, esses modelos têm sido amplamente utilizados como ferramentas práticas em análise, previsão e controle de séries temporais e têm apresentado resultados considerados satisfatórios (Box & Jenkins, 1976). Conforme citado em (Tong, 1990): *“Eles não teriam sobrevivido setenta anos para nada!”*

Modelos ARIMA

Muitas séries temporais reais apresentam comportamento não-estacionário e em particular não variam em torno de um valor médio fixo. A Figura 2.1 ilustra o comportamento não estacionário de uma série de produção mensal de eletricidade na Austrália de janeiro de 1956 a agosto de 1995.⁷

Box & Jenkins então propuseram uma forma estendida do modelo ARMA, que lida com o comportamento não estacionário da série temporal, denominado modelo ARIMA (*AutoRegressive Integrated Moving Average*) de ordem (n_y, d, n_e) .

A série temporal não-estacionária, $y(k)$, é diferenciada de uma ordem d (normalmente, $d = 0, 1$ ou 2), tornando-se uma série estacionária, $z(k)$.⁸ Uma vez que a

⁷<http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/>

⁸Deve-se levar em consideração que a eliminação da não-estacionariedade de uma dada série

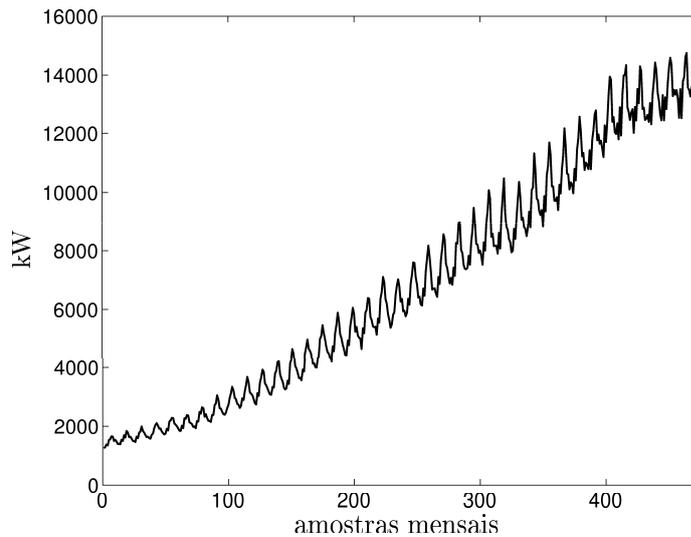


Figura 2.1: Série temporal de produção mensal de eletricidade na Austrália de janeiro de 1956 a agosto de 1995 (ELEC.DAT).

série foi transformada em estacionária, $z(k)$, esta é então modelada usando a representação ARMA (p e $q \neq 0$), ou AR ($q = 0$) ou MA ($p = 0$). De maneira geral, o modelo ARIMA, com $a_0 = 1$, é descrito por:

$$z(k) = \sum_{i=1}^{n_y} a_i z(k-i) + \sum_{j=1}^{n_e} c_j e(k-j) + e(k),$$

$$A(q)z(k) = C(q)e(k), \quad (2.5)$$

sendo que:

$$z(k) = y(k) - y(k-d), \quad (2.6)$$

ou,

$$z_k = \nabla^d y_k, \quad (2.7)$$

e d é a d -ésima diferença da série temporal de modo a reduzir ou eliminar a caracte-

temporal, só é possível se a mesma permitir essa transformação. Às vezes, a diferenciação pode somente reduzir a não-estacionariedade.

rística não estacionária da série. O significado do termo “integral”, na sigla ARIMA⁹, é devido à seguinte relação inversa à $z_k = \nabla^d y_k$, dada por:

$$\begin{aligned} y_k &= \nabla^{-d} z_k \\ y_k &= S^d z_k, \end{aligned} \tag{2.8}$$

sendo que,

$$\begin{aligned} y(k) &= Sz(k) + y(1) \\ &= \sum_{j=0}^{\infty} z(k-j) = z(k) + z(k-1) + z(k-2) + \dots + z(1) + y(1). \end{aligned}$$

Considere o seguinte exemplo:

$$\begin{aligned} z(1) = y(2) - y(1) &\implies y(2) = z(1) + y(1) \\ z(2) = y(3) - y(2) &\implies y(3) = z(2) + z(1) + y(1) \\ z(3) = y(4) - y(3) &\implies y(4) = z(3) + z(2) + z(1) + y(1) \\ &\quad \vdots \quad \quad \quad \vdots \\ z(k-1) = y(k) - y(k-1) &\implies y(k) = z(k-1) + z(k-2) + \dots + z(1) + y(1). \end{aligned}$$

A metodologia de Box & Jenkins (1976) consiste nas seguintes etapas :

- identificação (ou escolha) do modelo;
- estimação dos parâmetros do modelo;
- diagnóstico ou validação do modelo;
- previsão.

Na etapa de identificação procura-se determinar o valor do parâmetro d , ou seja, deve-se escolher quantas diferenciações devem ser realizadas na série para retirar

⁹ARIMA (*AutoRegressive Integred Moving Average*)

(reduzir) sua não-estacionariedade. Nessa etapa é feita, também, a escolha de qual tipo de representação será utilizada, visto que o modelo ARIMA engloba todas as outras representações lineares auto-regressivas.

A escolha da ordem do modelo da parte auto-regressiva (AR) é um dos aspectos mais difíceis na utilização da metodologia de Box & Jenkins. Algumas ferramentas podem auxiliar nessa escolha, tal como as funções de autocorrelação (Box & Jenkins, 1976), porém essa determinação da ordem do modelo ainda é uma questão em aberto.

O estimador de parâmetros usado por Box & Jenkins foi o de Máxima Verossimilhança. Porém, pode-se utilizar outras técnicas, como a dos mínimos quadrados e a Bayesiana. Aspectos gerais e a formulação matemática desses estimadores podem ser vistos em (Moon & Stirling, 2000).

A validação do modelo obtido, chamada de diagnóstico, analisa os resíduos de modelagem. Se o resíduo assemelha-se ao ruído branco o modelo estimado é considerado apropriado para a série temporal em questão. Caso contrário, o modelo deve ser reestimado (Brockwell & Davis, 1987). A última etapa é a da previsão, que consiste em gerar valores futuros da série através do modelo estimado.

De maneira geral, os modelos lineares discretos podem ser extraídos a partir do seguinte modelo geral (Aguirre, 2004):

$$A(q)y(k) = \frac{B(q)}{F(q)}u(k) + \frac{C(q)}{D(q)}\nu(k) \quad (2.9)$$

sendo q o operador de atraso, $\nu(k)$ o ruído branco e $A(q)$, $B(q)$, $C(q)$, $D(q)$ e $F(q)$ polinômios definidos como:

$$\begin{aligned} A(q) &= 1 + a_1q^{-1} + \dots + a_{n_y}q^{-n_y}; \\ B(q) &= b_1q^{-1} + \dots + b_{n_u}q^{-n_u}; \\ C(q) &= 1 + c_1q^{-1} + \dots + c_{n_\epsilon}q^{-n_\epsilon}; \\ D(q) &= 1 + d_1q^{-1} + \dots + d_{n_d}q^{-n_d}; \\ F(q) &= 1 + f_1q^{-1} + \dots + f_{n_f}q^{-n_f}. \end{aligned} \quad (2.10)$$

De acordo com a forma de cada um dos polinômios descritos em (2.10), pode-se obter modelos AR, ARMA, além dos sub-tipos não-autônomo, ARX, ARMAX (Aguirre, 2004).

2.1.2 Modelos NARMAX polinomiais

Uma ampla classe de sistemas não-lineares, discretos e invariantes no tempo, podem ser representados pelo modelo NARMAX (*Nonlinear AutoRegressive Moving Average with eXogenous variables*), que é uma derivação não-linear da representação linear ARMAX (AutoRegressive Moving Average with eXogenous variables). Os modelos NARMAX foram introduzidos em (Leontaritis & Billings, 1985a,b), e realizam o mapeamento não-linear das entradas e saídas passadas para a saída atual, $y(k)$, dada por:

$$y(k) = f[y(k-1), \dots, y(k-n_y), u(k-1), \dots, u(k-n_u+1), e(k-1), \dots, e(k-n_e)] + e(k), \quad (2.11)$$

sendo que os sinais de entrada, saída e o ruído aditivo são representados, respectivamente por $u(k)$, $y(k)$ e $e(k)$, com atrasos máximos de n_u , n_y e n_e . A forma funcional de $f(\cdot)$ em sistemas reais, geralmente, é desconhecida. A tarefa da identificação de sistemas consiste em aproximar $\hat{f}(\cdot)$ a partir de um modelo baseado em alguma classe de funções conhecida, tal como, a função racional, neural, polinomial. Essa última é tratada nesta seção. A justificativa teórica do uso de modelos NARMAX polinomiais pode ser encontrada em (Chen & Billings, 1989), onde é mostrado que existe uma função polinomial de grau ℓ que aproxima o sistema original com uma boa exatidão.

O modelo NARMAX polinomial é obtido expandindo a função $f(\cdot)$ como um

polinômio de grau ℓ , tal que:

$$y(k) = \theta_0 + \sum_{i_1=1}^n \theta_{i_1} x_{i_1}(k) + \sum_{i_1=1}^n \sum_{i_2=i_1}^n \theta_{i_1 i_2} x_{i_1}(k) x_{i_2}(k) + \dots + \sum_{i_1=1}^n \dots \sum_{i_\ell=i_{\ell-1}}^n \theta_{i_1 \dots i_\ell} x_{i_1}(k) \dots x_{i_\ell}(k) + e(k), \quad (2.12)$$

sendo que $n = n_y + n_u + n_e$, θ_i são os parâmetros a serem estimados e

$$x_1(k) = y(k-1), x_2(k) = y(k-2), \dots, x_{n_y+1}(k) = u(k-1), \dots, \\ x_{n_y+n_u+1}(k) = e(k-1), \dots, x_n(k) = e(k-n_e).$$

O número total de termos candidatos, n_θ , em um modelo NARMAX polinomial monovariável é dado pela relação (Korenberg, Billings, Liu & Mcilroy, 1988):

$$n_\theta = M + 1,$$

$$M = \sum_{i=1}^{\ell} n_i, \quad n_i = \frac{n_{i-1}(n_y + n_u + n_e + i - 1)}{i}, \quad n_0 = 1. \quad (2.13)$$

Verifica-se por (2.13) que o número de termos candidatos cresce rapidamente de acordo com o grau de não-linearidade ℓ e com os atrasos máximos da entrada, saída e do ruído, n_u , n_y e n_e , respectivamente.

Se todos os termos candidatos forem incluídos no modelo, como demonstra a equação (2.13), esse poderá conter muitos termos, proporcionando, na maioria das vezes, mau condicionamento da matriz de regressores. Uma maneira de contornar esse problema é determinar a estrutura que melhor se ajusta às informações disponíveis sobre o sistema em questão. Esse procedimento é chamado de detecção ou escolha da estrutura.

Reescrevendo o modelo (2.11) na forma de erro de predição, tem-se:

$$y(k) = \psi^T(k-1)\hat{\theta} + \xi(k). \quad (2.14)$$

em que $\xi(k) = y(k) - \hat{y}(k)$ é o resíduo de modelagem (ruído branco), e, a matriz de

regressores, os parâmetros estimados e a saída estimada são dados, respectivamente, por (2.15), (2.16) e (2.17):

$$\psi^T(k-1) = [\psi_{yu}^T(k-1) \quad \psi_{yu\xi}^T(k-1) \quad \psi_\xi^T(k-1)], \quad (2.15)$$

$$\hat{\theta} = [\hat{\theta}_{yu}^T \quad \hat{\theta}_{yu\xi}^T \quad \hat{\theta}_\xi^T], \quad (2.16)$$

$$\hat{y}(k) = \psi^T(k-1)\hat{\theta} = \Psi\hat{\theta}. \quad (2.17)$$

O argumento $(k-1)$ em (2.15) indica que o modelo utiliza a última informação da série temporal, assinalando a sua característica preditiva. A matriz de regressores Ψ é composta por termos de processo (entrada $(u(k))$ e saída $(y(k))$) e termos de resíduo $(\xi(k))$. Assim, $\psi_{yu}^T(k-1)$ indica todas as combinações lineares e não-lineares possíveis entre a saída e a entrada. Da mesma maneira, $\psi_{yu\xi}^T(k-1)$ indica todas as combinações lineares e não-lineares entre os termos de processo e os termos de resíduo. Por conseguinte, $\psi_\xi^T(k-1)$ indica todas as combinações lineares e não-lineares dos termos de resíduo de modelagem.

Neste trabalho, utiliza-se somente a série temporal $y(k)$, sem a componente de entrada. Logo, para se obter um modelo autônomo desconsidera-se a parte exógena do modelo (2.11), obtendo assim o modelo NARMA:

$$y(k) = f[(y(k-1), \dots, y(k-n_y), e(k-1), \dots, e(k-n_e))] + e(k). \quad (2.18)$$

Somente a parte determinística dos modelos (2.11) e (2.18), ou seja, a parte da matriz de regressores composta por termos de processo ($\psi_y^T(k-1)$), será usada na previsão, neste trabalho. A parte com média móvel, ou seja, termos compostos por combinações de termos de resíduo ($\psi_\xi^T(k-1)$), é usada na estimação de parâmetros objetivando reduzir a polarização destes provocada pelo ruído presente nos dados.

2.2 Detecção de Estrutura

A detecção da estrutura é uma etapa extremamente importante e difícil no processo de identificação de sistemas, visto que as características dinâmicas e estáticas que serão extraídas estão ligadas a estrutura do modelo estimado (Aguirre & Billings, 1995a).

A presença de termos desnecessários no modelo pode proporcionar instabilidade numérica e ocasionar comportamentos dinâmicos espúrios ao sistema. Além do número excessivo de termos, efeitos espúrios podem ocorrer devido a sobreparametrização do modelo em relação ao número de pontos fixos (Aguirre & Mendes, 1996; Mendes & Billings, 1998).

Existem técnicas bem estabelecidas para auxiliar na escolha da ordem dos modelos lineares (Box & Jenkins, 1976). No contexto de identificação caixa-preta, alguns métodos podem ser usados para auxiliar na determinação da estrutura, tais como: o critério de informação de Akaike, AIC (Akaike, 1974), a análise de agrupamento de termos (Aguirre, 1994; Aguirre & Billings, 1995b; Jácome, 1996; Cassini, 1999) e a taxa de redução de erro, ERR, (Billings, Chen & Korenberg, 1989; Chen, Billings & Luo, 1989). A composição desses três métodos aplicados na identificação de um sistema real pode ser visto em (Cassini, 1999).

O critério de informação de Akaike, AIC, é um critério estatístico, que verifica a redução na variância dos resíduos à medida que termos são acrescentados ao modelo, tal que:

$$AIC = N \ln(\text{var}(\xi(t))) + 2n_{\theta}. \quad (2.19)$$

sendo N o número de dados de treinamento, $\text{var}(\xi(t))$ a variância dos resíduos e n_{θ} o número de termos no modelo.

A taxa de redução de erro, ERR, atribui a cada termo candidato um valor correspondente à contribuição deste na explicação do valor quadrático médio dos dados de saída, ou seja, quantifica a redução no erro da saída do modelo devido à introdução de cada termo. Em identificação caixa-cinza, esses métodos podem

ser usados juntamente com alguma informação *a priori* sobre o sistema, tal como o número de pontos fixos ou o conhecimento da curva estática do sistema (Corrêa, 2001).

Definindo a variância do resíduo de modelagem, $\xi(k)$, como sendo:

$$\text{var}\{\xi(k)\} = \lim_{N \rightarrow \infty} \frac{1}{N} \left[\mathbf{y}^T \mathbf{y} - \sum_{i=1}^{n_\theta} g_i^2 w_i^T w_i \right], \quad (2.20)$$

sendo g_i os elementos do vetor de parâmetros, \mathbf{g} , e w_i indica os regressores ortogonais e \mathbf{y} é o vetor contendo os dados de saída¹⁰, no caso deste trabalho, a série temporal a ser modelada.

Supondo que nenhum termo fosse acrescentado ao modelo, a variância de $\xi(k)$ seria igual ao erro quadrático da saída $y(k)$. A cada termo acrescentado, a variância de $\xi(k)$ decresce de um fator igual a $\frac{1}{N}(g_i^2 w_i^T w_i)$, em que w_i indica o termo incluído e g_i o seu respectivo parâmetro. A redução no valor da variância pode ser normalizada com relação ao erro quadrático médio do sinal de saída. Assim, o ERR de cada termo é definido como sendo:

$$[ERR] = \frac{(g_i^2 w_i^T w_i)}{\mathbf{y}^T \mathbf{y}}. \quad (2.21)$$

O ERR pode ser utilizado na determinação da estrutura de modelos NARMA polinomiais. Escolhe-se o número de termos desejado por meio de uma ferramenta auxiliar, ou empiricamente, e consideram-se aqueles que possuem os maiores valores de ERR.

2.2.1 Estimação dos Parâmetros

Após determinar qual o tipo de representação e qual a sua estrutura, o próximo passo no processo de identificação de sistemas, consiste em estimar os parâmetros do modelo de modo a minimizar a diferença entre a predição estimada pelo

¹⁰Em Aguirre (2004), o autor descreve um algoritmo baseado na transformação de Householder para a ortogonalização dos regressores (Chen et al., 1989).

modelo e os dados medidos, ou seja, o procedimento para estimar os parâmetros do modelo NARX polinomial que minimiza a função custo do algoritmo dos mínimos quadrados.

Inicialmente, tem-se que o modelo dado pela equação (2.12), pode ser representado como:

$$y(k) = \sum_{i=1}^{n_\theta} \psi_i(k-1)\theta_i + e(k), \quad (2.22)$$

sendo $y(k)$ a saída atual do modelo, $\psi(k-1)$ são os termos que constituem a matriz de regressores Ψ tomadas até o instante $(k-1)$, θ os parâmetros a serem estimados e $e(k)$ o ruído. O número de termos no modelo é representado por n_θ .

Desprezando o termo de ruído, pode-se tomar (2.22), ao longo de uma janela de dados e, representar o conjunto resultante de equações (restrições) em forma matricial, tal que:

$$\mathbf{y} = \Psi\theta, \quad (2.23)$$

sendo que,

$$\mathbf{y} = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix}, \Psi = \begin{bmatrix} \psi^T(0) \\ \psi^T(1) \\ \vdots \\ \psi^T(N-1) \end{bmatrix}.$$

e θ o vetor de parâmetros. A saída do modelo, predita um passo à frente é dada por:¹¹

$$\begin{aligned} \hat{y}(k) &= \sum_{i=1}^{n_\theta} \psi_i(k-1)\hat{\theta}_i \\ &= \psi^T(k-1)\hat{\theta}. \end{aligned} \quad (2.24)$$

O resíduo, ou seja, a diferença entre o valor medido e o valor predito de um

¹¹O símbolo ($\hat{\quad}$) significa que são valores estimados.

passo à frente, é representado por:

$$\xi(k) = y(k) - \hat{y}(k) = y(k) - \psi^T(k-1)\hat{\theta}. \quad (2.25)$$

Utilizando-se um número maior de observações, N , do que o número de regressores n_θ , tem-se um sistema de equações sobre determinado, resultando em uma matriz Ψ não quadrada, o que impossibilita a estimação de θ a partir da inversão de Ψ em (2.23). Dessa forma, define-se um algoritmo para a estimação de $\hat{\theta}$ que minimize a soma dos quadrados do erro dado pela seguinte função custo:

$$J_{\text{MQ}} = \sum_{i=1}^N \xi(i)^2 = \xi^T \xi, \quad (2.26)$$

Substituindo (2.25) em (2.26), tem-se:

$$\begin{aligned} J_{\text{MQ}} &= (\mathbf{y} - \Psi\hat{\theta})^T (\mathbf{y} - \Psi\hat{\theta}) \\ &= \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \Psi \hat{\theta} - \hat{\theta}^T \Psi^T \mathbf{y} - \hat{\theta}^T \Psi^T \Psi \hat{\theta}. \end{aligned} \quad (2.27)$$

A minimização da função custo J_{MQ} em relação ao vetor de parâmetros $\hat{\theta}$ é feita resolvendo-se $\partial J_{\text{MQ}} / \partial \hat{\theta} = 0$, cuja solução é dada por:

$$\hat{\theta} = (\Psi^T \Psi)^{-1} \Psi^T \mathbf{y}. \quad (2.28)$$

A equação (2.28) é denominada estimador dos mínimos quadrados clássico.¹² Se a matriz de covariância dos regressores for singular ou mal condicionada, ela estará sujeita a problemas numéricos que podem afetar a estabilidade do algoritmo dos mínimos quadrados¹³. Para contornar este problema, pode-se utilizar métodos numéricos, os quais ortogonalizam a matriz Ψ , ou seja, os regressores não serão mais

¹²No presente trabalho a equação (2.28) não foi implementada, utilizou-se a função `pinv.m` do MATLAB.

¹³O mau condicionamento é devido à alta correlação entre os regressores que constituem as colunas da matriz Ψ .

correlacionados. Dentre esses, pode-se citar o método da transformação de Householder (Ljung, 1987).

Estimador de Mínimos quadrados estendido

A polarização no estimador MQ surge quando o erro que aparece na equação de regressão, $\mathbf{y} = \Psi\theta + \mathbf{e}$, for autocorrelacionado e ao mesmo tempo o modelo incluir regressores da saída. Se os resíduos de identificação forem modelados como um processo de média móvel, da seguinte maneira:

$$e(k) = c_i\nu(k - i) + \nu(k), \quad (2.29)$$

sendo $\nu(k)$ ruído branco, os termos $\nu(k - i)$ podem ser incorporados à matriz de regressores e os seus respectivos parâmetros ao vetor de parâmetros do modelo da seguinte forma:

$$\mathbf{y}^* = \Psi^*\theta^* + \mathbf{e}^*, \quad (2.30)$$

sendo que $\mathbf{y}^* = \mathbf{y}$, $\mathbf{e}^* = [\nu(k) \cdots \nu(k + N - 1)]^T$,

$$\Psi^* = \begin{bmatrix} \vdots & \nu(k - 1) \\ \vdots & \nu(k) \\ \Psi & \nu(k + 1) \\ \vdots & \vdots \\ \vdots & \nu(k + N - 2) \end{bmatrix} \quad (2.31)$$

e $\theta^* = [\theta \vdots c_i]^T$. A parte parametrizada por c_i da Equação (2.29) foi incorporada à matriz de regressores. Por este motivo, pode-se notar que \mathbf{e}^* é ruído branco, de forma que em (2.30) o termo \mathbf{e}^* é não-correlacionado com Ψ^* . Logo a estimativa por (MQ)

$$\theta^* = [\Psi^{*T}\Psi^*]^{-1}\Psi^{*T}\mathbf{y} \quad (2.32)$$

não apresentaria polarização, ou $E[\theta^*] = [\theta \quad \mathbf{c}]^T$.

Um cuidado deve ser tomado com os termos de ruído no modelo: de maneira prática, o ruído é modelado apenas com o intuito de se evitar a polarização do modelo NARX polinomial. O modelo final contém normalmente apenas termos de processo. Apenas em alguns casos, como em controle adaptativo, a parte estocástica também é usada.

Evidentemente, (2.32) não pode ser calculado como mostrado pois Ψ^* inclui regressores do tipo $\nu(k-i)$, que não são medidos. A fim de resolver este problema, tais regressores precisam ser estimados. Quando isso é feito, o resultado é um processo iterativo que é equivalente a resolver (2.32). Em (Aguirre, 2004) o autor descreve um algoritmo iterativo para a estimação de $\nu(k)$ da seguinte forma:

1. a partir da equação de regressão $y(k) = \psi^T(k-1)\theta + e(k)$ e dos dados disponíveis, monte a Equação matricial $\mathbf{y} = \Psi\theta + \mathbf{e}$, como no método de mínimos quadrados, e determine $\theta_{\text{MQ}} = [\Psi^T\Psi]^{-1}\Psi^T\mathbf{y}$;
2. calcule o vetor de resíduos $\xi_1 = \mathbf{y} - \Psi\theta_{\text{MQ}}$;
3. faça $i = 2$ (i indica o número da iteração);
4. com ξ_{i-1} , monte a matriz estendida de regressores, Ψ_i^* , e, em seguida, estime $\theta_{\text{MQE}}^* = [\Psi_i^{*\text{T}}\Psi_i^*]^{-1}\Psi_i^{*\text{T}}\mathbf{y}$;
5. determine o vetor de resíduos $\xi_i = \mathbf{y} - \Psi_i^*\theta_{\text{MQE}_i}^*$;
6. faça $i = i + 1$ e volte ao passo 4. Repita até convergir.

2.3 Comentários finais

A proposta do presente trabalho é estabelecer um procedimento para detectar vínculo determinístico em séries temporais quase-periódicas usando teste de hipóteses. Para inferir determinismo avalia-se a capacidade de previsão da série.

Se ela for previsível, conseqüentemente, pode-se conjecturar que a série tenha assinatura determinística. Se a série não for previsível, infere-se que esta não apresenta assinatura determinística. Para avaliar a capacidade de previsão da série temporal em questão deve-se ajustar um modelo para a série. Neste trabalho, emprega-se os modelos do tipo NARMA (2.18) polinomial, o qual procura ajustar uma função polinomial que seja capaz de representar a série temporal.

TÉCNICAS PARA ANÁLISE DE SÉRIES TEMPORAIS

Uma série temporal representa a evolução de uma variável ao longo do tempo (Morettin & Toloí, 2004). Uma grande quantidade de fenômenos observados na natureza, seja no campo da biologia, física, economia, engenharias, entre outros, são exemplos de observações ordenadas no tempo, as quais podem ser investigadas por meio de técnicas de análise e modelagem de séries temporais.

Os principais interesses em séries temporais estão relacionados à sua análise e previsão, ou seja, estando no instante (k) deseja-se prever o valor de uma dada variável no instante ($k+h$), sendo h o horizonte de previsão ($h > 0$). No entanto, para prever o comportamento futuro da série, antes é necessário analisá-la, sintetizando suas características e assinalando as mais importantes. Dentre os objetivos da análise de séries temporais destacam-se:

- descrever o comportamento dinâmico da série e, por meio de gráficos e/ou técnicas estatísticas, verificar a existência de tendências, ciclos e sazonalidades, ausência de medições (*missings*), *outliers* e mudanças abruptas;
- procurar periodicidades relevantes na série;
- investigar o mecanismo gerador da série;
- fazer previsões de valores futuros da série.

Uma primeira análise pode ser obtida visualmente por meio de uma representação gráfica da série temporal, a partir da qual é possível verificar algumas características tais como: a presença de componentes sazonais, tendência, ciclos e mudanças abruptas no comportamento. No entanto, a análise gráfica pode não ser suficiente para extrair as informações de uma dada série. Uma análise mais apurada pode ser feita por meio de técnicas estatísticas, sendo que estas podem necessitar de modelos.

3.1 Características das séries temporais

Algumas séries temporais podem ser representadas como a combinação de quatro componentes: de tendência, cíclica, de sazonalidade e residual (Brockwell & Davis, 1987, 2002; Makridakis, Wheelwright & Hyndman, 1998; Chatfield, 2003).

As componentes de *tendência* são, freqüentemente, mudanças graduais a longo prazo. São normalmente provocadas por fatores que produzem variações graduais e regulares ao longo tempo, mas não provocam mudanças repentinas.

As componentes *sazonais* são oscilações que ocorrem em períodos de 3, 4, 6 ou 12 meses de duração, ou seja, ocorrem em períodos relacionados às variações anuais. Essas flutuações são resultantes de eventos periódicos, ou quase-periódicos, que ocorrem anualmente. Por exemplo, a taxa de desemprego, em alguns países, é alta no inverno e baixa no verão.

As componentes *cíclicas* são todas as oscilações que não se encaixam nos padrões sazonais. Um exemplo é a variação diária da temperatura. A componente cíclica também pode ser periódica ou quase-periódica. A maioria das séries quase-periódicas usadas neste trabalho apresentam componentes cíclicas.

As componentes *residuais* podem ser interpretadas como o restante da série, ou seja, aquilo que não pode ser explicado pela tendência, pelo ciclo e, nem mesmo, pela sazonalidade. A série residual pode ser ou não aleatória.

Com a remoção das componentes de sazonalidade, cíclica e de tendência, a componente residual pode ser determinada. Alguns procedimentos para a de-

composição de séries temporais foram desenvolvidos, cada qual tentando isolar tais componentes da série o mais cuidadosamente possível (Shiskin et al., 1967; Findley et al., 1998; Brockwell & Davis, 1987, 2002). O objetivo desses procedimentos consiste em remover cada uma das componentes permitindo que o comportamento da série temporal seja melhor compreendido e, conseqüentemente, prever valores futuros mais apropriados.

3.1.1 Determinismo versus previsibilidade

As definições de determinismo e de previsibilidade segundo o dicionário Aurélio são as seguintes:

“Determinismo: Relação entre os fenômenos pela qual estes se acham ligados de modo tão rigoroso que, a um dado momento, todo fenômeno está completamente condicionado pelos que o precedem e acompanham, e condiciona com o mesmo rigor os que lhe sucedem. [Se relacionado a fenômenos naturais, o determinismo constitui o princípio da ciência experimental que fundamenta a possibilidade de busca de relações constantes entre os fenômenos; se se refere a ações humanas e a decisões da vontade, entra em conflito com a possibilidade da liberdade.]”

“Previsibilidade: Qualidade de previsível”, ou seja, cujo futuro pode ser previsto.

Na comunidade científica, inicialmente, o ato de prever o futuro estava intimamente ligado ao conceito de determinismo (Stewart, 1991). No entanto, definir determinismo é uma questão um tanto filosófica o que não será feito neste trabalho. Porém, é interessante expor alguns dos pensamentos a despeito desse assunto feito por alguns dos principais nomes da Ciência Mundial.

O físico francês Pierre Simon Laplace ficou conhecido, dentre outros trabalhos¹, pela sua teoria determinística segundo a qual Laplace afirmava, baseado em

¹Laplace também contribuiu na solução de equações diferenciais, com a Transformada de Laplace (Ogata, 1998).

experimentos mentais, que se for conhecida todas as equações que definem o estado do universo em um instante de tempo t , então, pode-se prever seu estado futuro em um instante de tempo futuro $t' > t$ (Laplace, 1812). A teoria determinística de Laplace ganhou mais força após apoiada às leis determinísticas de Newton. Em suma, para Laplace, determinismo implicava previsibilidade (Monteiro, 2002).

Alguns anos depois, Julis Henri Poincaré contestou a teoria de determinismo de Laplace. Poincaré contemplou que pequenas mudanças nas condições iniciais de alguns sistemas determinísticos acarretavam em diferentes soluções. Logo, Poincaré percebeu que nem sempre pode-se prever o comportamento futuro de sistemas determinísticos (Monteiro, 2002).

Em 1963, por meio de resultados empíricos com dados meteorológicos, Edward Lorenz (Lorenz, 1963) observou o que Poincaré já havia contemplado, dando origem ao fenômeno conhecido como efeito borboleta. Lorenz concluiu que, mesmo conhecendo as equações de um sistema, nem sempre é possível prever seu comportamento futuro. Este tipo de determinismo ficou conhecido como determinismo caótico. Lorenz foi o primeiro a mostrar que o fato do sistema ser determinístico não implica que ele seja previsível, a longo prazo.

Em séries temporais reais busca-se por vínculos determinísticos e funções que sejam capazes de modelar, ainda que parcialmente, os dados da série. Considera-se que uma série temporal $y(k)$ tem *vínculo determinístico* se existir uma função $g(\cdot)$ que explique a série no instante atual (k) baseada em informações de outras variáveis também no instante atual (k):

$$y(k) = g(x(k), z(k), \dots). \quad (3.1)$$

Verifica-se que a relação determinística entre as variáveis em (3.1) independem do instante de tempo, logo, denomina-se essa relação como determinismo estático.²

²Entende-se como comportamento estático quando não há uma relação temporal entre as variáveis do sistema.

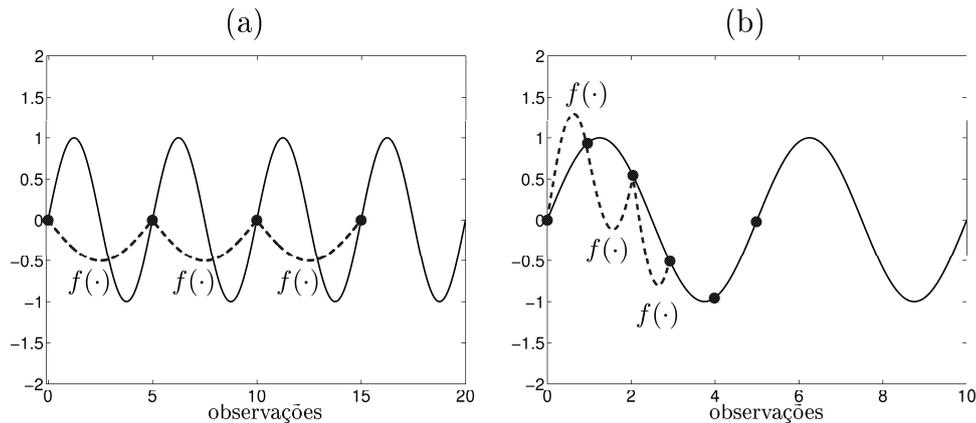


Figura 3.1: Série temporal periódica. Em (a) a função $f(\cdot)$ relaciona as observações entre os ciclos (determinismo entre-ciclos). Em (b) a função $f(\cdot)$ relaciona as observações subjacentes ao ciclo (determinismo intra-ciclo).

Pode-se considerar que uma série temporal, $y(k)$, possui vínculo determinístico dinâmico, quando existir uma relação $f(\cdot)$ que governe a evolução temporal da série $y(k)$, tal como:

$$y(k+1) = f(y(k), y(k-1), y(k-2), \dots), \quad (3.2)$$

Percebe-se que a função $f(\cdot)$ além de estabelecer uma relação dinâmica determinística, possibilita prever o comportamento futuro da série $y(k)$ por meio do seu comportamento passado.

Logo, pode-se dizer que a série temporal $y(k)$ é *previsível* se existir uma função $f(\cdot)$ que explique a série y no instante de tempo futuro $(k+1)$ a partir das observações passadas até o instante de tempo atual (k) , logo:

Por determinismo entre-ciclos entende-se que existe uma lei que relaciona a disposição dos ciclos e, por determinismo intra-ciclo entende-se que existe uma lei que governa a relação entre as observações subjacentes ao ciclo, ou seja, a relação de curto-prazo entre as observações. Neste trabalho, o interesse é detectar se há vínculo determinístico dinâmico intra-ciclo nas séries temporais quase-periódicas. As Figuras 3.1 (a) e (b) ilustram o determinismo entre-ciclos e intra-ciclo, respectivamente.

3.2 Análise linear de séries temporais

Na análise de séries temporais existem, basicamente, dois caminhos que o analista pode seguir: a análise no domínio do tempo e/ou a análise no domínio da frequência. Ambas as formas de análise são mais complementares do que concorrentes, visto que pode-se processar as mesmas informações de formas diferentes fornecendo diferentes visões da natureza da série temporal. Tais métodos têm como objetivo principal auxiliar na construção de modelos para as séries temporais com os propósitos determinados pelo analista. Em ambos os enfoques, tempo e frequência, existem várias técnicas lineares e não-lineares (Wavelets, espectrogramas, funções de coerência) que atendem às mais diversas aplicações no âmbito das ciências físicas, biológicas, das engenharias e da economia (Daubechies, 1990; Reis, 1999).

Uma outra característica importante que deve ser considerada na análise de uma série temporal é com relação a estacionariedade. Algumas formas de identificar a presença de não-estacionariedade,³ em uma série temporal pode requerer um conhecimento do processo envolvido (tais como estação do ano, valor de mercado), e/ou, uma análise estatística dos dados (função de auto correlação, teste de raízes unitárias).

Se for confirmada a presença de tais características, o previsor humano tem duas opções:

- usar um método que leve em consideração a não-estacionariedade, tal como o modelo ARIMA;
- extrair ou reduzir a não-estacionariedade da série, construir um previsor e então adicionar seus respectivos efeitos antes da previsão.

De forma simplificada, pode-se verificar se a série é não-estacionária na média ou na variância da seguinte maneira:⁴

³Uma definição formal de não-estacionariedade pode ser encontrada em Papoulis (1991).

⁴Para uma definição formal de estacionariedade ver (Box & Jenkins, 1976) pag. 26 e (Papoulis & Pillai, 2002).

- Se por meio do gráfico da série temporal for possível verificar que a média da série muda com o decorrer do tempo, pode-se inferir que tal série é não-estacionária na média;
- Se por meio do gráfico da série temporal for possível verificar que há uma mudança na variância da série com o decorrer do tempo, pode-se inferir que tal série é não-estacionária na variância.

Dessa forma, pode-se conjecturar que a série temporal que apresenta componentes de tendência pode ser considerada não-estacionária, uma vez que seu valor médio não se mantém constante ao longo do tempo. Vale ressaltar que, em alguns casos, somente uma inspeção visual do gráfico da série temporal não possibilita identificar a presença de tendência, sazonalidade, ciclos e não-estacionariedade em uma série. Outras ferramentas para detectar a presença de não-estacionariedade são a análise da função de autocorrelação e o teste de raízes unitárias (Dickey & Fuller, 1979).

3.3 Estimação e eliminação das componentes de tendência e cíclica

O objetivo desta seção é apresentar, de maneira sucinta, algumas técnicas para a estimação e extração das componentes determinísticas de: tendência ($m(k)$), cíclica ($c(k)$) e, de sazonalidade ($s(k)$), de modo a obter uma série residual estacionária ($x(k)$). Pode-se obter um modelo de séries estacionárias para $x(k)$, a partir do qual pode-se analisar as suas propriedades e usá-las em conjunção com $m(k)$, $c(k)$ e $s(k)$ para a previsão e simulação da série temporal quase-periódica ($y(k)$).

Quando os ciclos ocorrerem em períodos diferentes de 12, 6, 4, 3, 2 meses, denomina-se componente cíclica. Os métodos para a estimação e extração da componente sazonal, descritos na literatura, são mencionados ao longo deste trabalho em termos de componente cíclica, visto que, a componente sazonal apresenta ciclos an-

uais. Portanto, neste trabalho considera-se que a componente cíclica abrange todos os ciclos, inclusive os com período sazonal.

Usualmente, as componentes de tendência e de ciclo são detectadas por meio de inspeção visual do gráfico da série temporal. Entretanto, estas também podem ser detectadas por meio da função de autocorrelação da série temporal, a qual apresenta um decaimento lento (na presença de tendência) e próximo de periódico (na presença de um ciclo).

A eliminação da tendência e do ciclo quase-periódico pode ser feita por meio de dois métodos distintos:

- (i) pela diferenciação das observações, ou;
- (ii) pela decomposição clássica da série temporal ($y(k)$) em componentes de tendência ($m(k)$), componente cíclica ($c(k)$) e componente residual ($x(k)$).

3.3.1 Diferenciação

Uma das técnicas de pré-processamento de séries temporais, conceitualmente simples e útil, é conhecida por diferenciação de observações adjacentes ou simplesmente, diferenciação. O processo de diferenciar uma série temporal consiste em computar a diferença entre observações sucessivas. Assim, ao invés de trabalhar com as observações originais da série temporal, trabalha-se com suas variações.

Uma série temporal não-estacionária, $y(k)$, é diferenciada de uma ordem p (normalmente, $p = 1$ ou 2), tornando-se uma série estacionária $z(k)$.⁵ Na prática, as séries não-estacionárias podem se tornar estacionárias (ou quase estacionárias) por meio de uma única diferenciação. No entanto, algumas séries podem requerer dois ou mais estágios de diferenciação. Um outro tipo de diferenciação a ser considerado neste texto é a diferenciação de observações não adjacentes ou diferenciação cíclica. Neste tipo, uma nova série é obtida pela diferença entre uma observação no instante

⁵Deve-se levar em consideração que a eliminação da não-estacionariedade de uma série temporal, só é possível se a mesma permitir essa transformação. Às vezes, a diferenciação pode somente reduzir a não-estacionariedade e não eliminá-la.

(k) e outra observação no instante $(k + d)$, em que d é o período cíclico da série $y(k)$. Quando d for igual a 12, esse tipo de diferenciação pode ser referida como diferenciação sazonal. A série gerada $z(k)$ terá d observações a menos do que a série original $y(k)$.

3.3.2 Decomposição Clássica

A inspeção visual do gráfico da série temporal $y(k)$ pode inferir a presença de componentes de tendência e cíclicas, sugerindo a possibilidade de representar a série original, $y(k)$, como uma realização do modelo de decomposição clássico:⁶

$$y(k) = m(k) + c(k) + x(k), \quad (3.3)$$

para $k = 1, \dots, N$, em que N é o número de observações da série temporal $y(k)$. Sendo que, $m(k)$ é a componente de tendência; $c(k)$ é a componente cíclica com período cíclico d qualquer, ou seja, abrange também os ciclos inclusive os sazonais. $x(k)$ é a componente de resíduo.

3.3.3 Métodos para estimação e extração de tendência

Em várias séries temporais, pode-se verificar nitidamente uma tendência nos dados, como na série de produção mensal de eletricidade na Austrália, registradas de janeiro de 1956 a agosto de 1995, representada em tracejado na Figura 3.2 (a). Na ausência da componente cíclica, $c(k)$, o modelo (3.3) torna-se:

$$y(k) = m(k) + x(k), \quad (3.4)$$

para $k = 1, \dots, N$.

Para construir um modelo para esse tipo de série (com apenas tendência,

⁶Este trabalho limita-se a decompor a série como sendo originada por um modelo do tipo aditivo descrito na equação (3.3). Porém, existem outros modelos, tal como o modelo de decomposição clássico multiplicativo dado como: $y(k) = m(k) \times c(k) \times x(k)$.

sem ciclos) dois métodos são recomendados:

- a) o primeiro consiste em estimar a componente de tendência estimada, $\hat{m}(k)$, e então, subtraí-la da série original $y(k)$, e assim, obter a série residual por meio do modelo de decomposição clássico aditivo: $\hat{x}(k) = y(k) - \hat{m}(k)$;
- b) o segundo método consiste em eliminar a tendência da série $y(k)$ pela diferenciação dos dados.

Este último método tem a vantagem de requerer a estimação de poucos parâmetros e não se basear na suposição de que a tendência permaneça fixa ao longo das observações.

a) Estimação da tendência, $\hat{m}(k)$

Filtros especiais podem ser projetados para estimação e remoção das componentes de tendência e cíclica. A escolha deste filtro requer um certo julgamento subjetivo do analista.

a.1) Suavização com filtro de média móvel:

Considerando q um número inteiro não negativo e os dois lados do filtro de média móvel com sendo:

$$w(k) = \frac{1}{(2q+1)} \sum_{j=-q}^q y(k-j), \quad (3.5)$$

e, por meio de (3.4), para $q+1 \leq k \leq N-q$, tem-se:

$$w(k) = \frac{1}{(2q+1)} \sum_{j=-q}^q m(k-j) + \frac{1}{(2q+1)} \sum_{j=-q}^q x(k-j) \cong m(k). \quad (3.6)$$

Assumindo que $\hat{m}(k)$ é aproximadamente linear no intervalo $[k-q, k+q]$ e que o erro médio neste intervalo é próximo de zero, tem-se a componente de tendência

estimada:

$$\hat{m}(k) = \frac{1}{(2q+1)} \sum_{j=-q}^q y(k-j). \quad (3.7)$$

Assim, a série residual estimada, $\hat{x}(k)$, é obtida extraíndo a tendência estimada, $\hat{m}(x)$, da série original, $y(k)$:

$$\hat{x}(k) = y(k) - \hat{m}(k), \quad (3.8)$$

para $q+1 \leq k \leq N-q$.

A componente de tendência da série temporal de produção de eletricidade ilustrada na Figura 3.2 foi aproximada por um filtro de média móvel para $q = 36$, como pode ser visto nas Figuras 3.2 (a) e (b). Vale salientar que, a série residual $\hat{x}(k)$ possui $2q = 72$ observações a menos do que série original $y(k)$, como mostrado nos extremos da Figura 3.2 (a). Verifica-se também, pela comparação da FAC na Figura 3.2 (b) que a série residual $\hat{x}(k)$ não mais apresenta a componente cíclica da série $y(k)$.

Por meio de resultados de simulações verificou-se que a escolha de q interfere no desempenho deste método, o qual apresenta dificuldades em séries com componentes cíclicas. Uma vez que, as componentes cíclicas também são extraídas juntamente com a componente de tendência.

a.2) Ajuste polinomial:

Um método mais indicado na literatura para estimar a componente de tendência, $m(k)$, é o ajuste de uma função polinomial ao longo de toda a série temporal, ou seja, uma aproximação polinomial da tendência global da série (Brockwell & Davis, 2002):

$$m(k) = a_0 + a_1k + a_2k^2 + \dots + a_jk^j, \quad (3.9)$$

para os dados $[y(1), \dots, y(N)]$, sendo N o tamanho da série temporal, $y(k)$, e j o grau do polinômio. Deve-se estimar os parâmetros a_0, \dots, a_j do polinômio de modo

a minimizar a função custo $\sum_{k=1}^n (y(k) - \hat{m}(k))^2$, em que:

$$\hat{m}(k) = \hat{a}_0 + \hat{a}_1 k + \hat{a}_2 k^2 + \dots + \hat{a}_j k^j. \quad (3.10)$$

para $k = 1, \dots, N$.

Dessa forma, a função polinomial ajustada à série $y(k)$ é dada por:

$$\hat{m}(k) = -7,29 \times 10^{-5} k^3 + 0,08 k^2 + 6,11 k + 1.439,04, \quad (3.11)$$

cujos grau 3 do polinômio foi escolhido via inspeção visual da série temporal $y(k)$, mostrada na Figura 3.2(a).

Assim, a série residual estimada, que está ilustrada na Figura 3.2(c), foi obtida extraindo a componente de tendência estimada, $\hat{m}(k)$, da série $y(k)$, ou seja, $\hat{x}(k) = y(k) - \hat{m}(k)$. Uma comparação da FAC é mostrada na Figura 3.2(d), pela qual percebe-se que a série residual $\hat{x}(k)$ não apresenta mais a componente de tendência, porém, manteve a componente cíclica da série $y(k)$.

b) Eliminação de tendência por diferenciação

Diferentemente dos métodos apresentados na Seção 3.3.3a), em que a componente residual ($\hat{x}(k)$) é obtida estimando e extraindo a componente de tendência ($\hat{m}(k)$) da série original, $\hat{x}(k) = y(k) - \hat{m}(k)$, o método apresentado nesta seção tem como proposta a eliminação da tendência por meio da diferenciação da série $y(k)$. Considerando a diferença ($lag - 1$) como sendo o operador ∇ , tem-se:

$$\nabla^1 y(k) = y(k) - y(k-1) = (1 - B)y(k), \quad (3.12)$$

em que B é o operador de atrasos usado no método de diferenciação dos dados:

$$B^1(y(k)) = y(k-1).$$

A potência dos operadores ∇ e B é definida como: $B^j(y(k)) = y(k-j)$ e

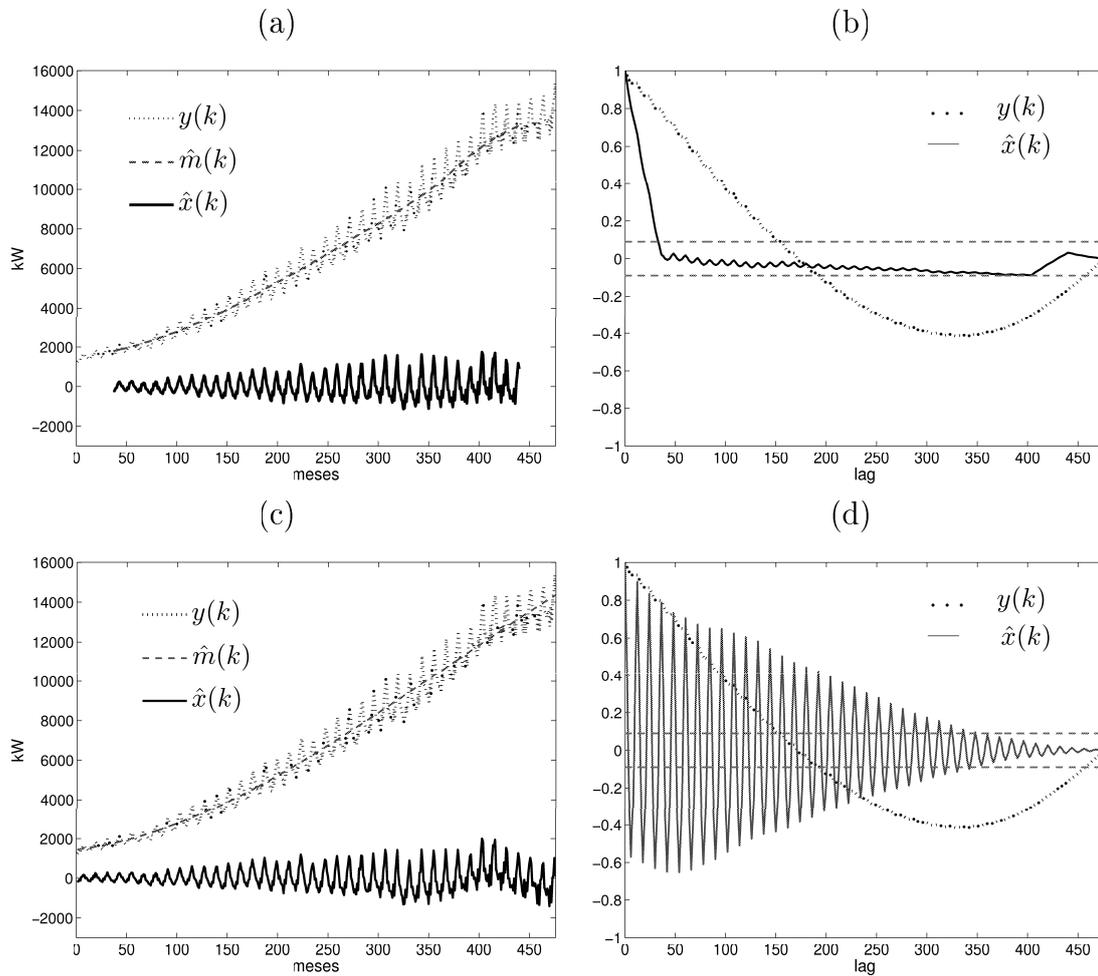


Figura 3.2: Série temporal de produção mensal de eletricidade na Austrália de janeiro de 1956 à agosto de 1995 ELEC.DAT. Em (a) $y(k)$ em (...) e componente de tendência ($\hat{m}(k)$) em (- -) ajustada por meio de filtro de média móvel finito (3.7) com $q = 36$ e, a componente residual ($\hat{x}(k)$) em (—). Em (b) função de autocorrelação: de $y(k)$ em (...) e de $\hat{x}(k)$ em (—). Em (c) a série $y(k)$ em (...), a componente de tendência $\hat{m}(k)$ em (- -) estimada via polinômio (3.11) e a série residual ($\hat{x}(k)$) em (—). Em (d) função de autocorrelação: de $y(k)$ em (...) e de $\hat{x}(k)$ em (—).

$\nabla^j(y(k)) = \nabla(\nabla^{j-1}(y(k)))$, $j \geq 1$ com $\nabla^0(y(k)) = y(k)$. Polinômios em B e em ∇ são manipulados da mesma maneira que funções polinomiais de variáveis reais. Como, por exemplo:

$$\begin{aligned}\nabla^2 y(k) &= \nabla(\nabla(y(k))) = (1 - B)(1 - B)y(k) = (1 - 2B + B^2)y(k), \\ &= y(k) - 2y(k-1) + y(k-2).\end{aligned}\quad (3.13)$$

Na prática, a ordem de diferenciação requerida é muito pequena, freqüentemente, ordem um ou dois, levando-se em consideração que muitas funções podem ser bem aproximadas, em um intervalo finito, por uma função polinomial de grau razoavelmente pequeno. Na Figura 3.3 pode-se verificar o resultado da estimação da série diferenciada por meio da diferenciação de ordem 2 dos dados originais, $y(k)$. Vale ressaltar que, apesar de eliminar a tendência, este método tem como principal desvantagem o fato da série diferenciada de ordem 2, por exemplo, conter duas observações a menos que a série original, $y(k)$.

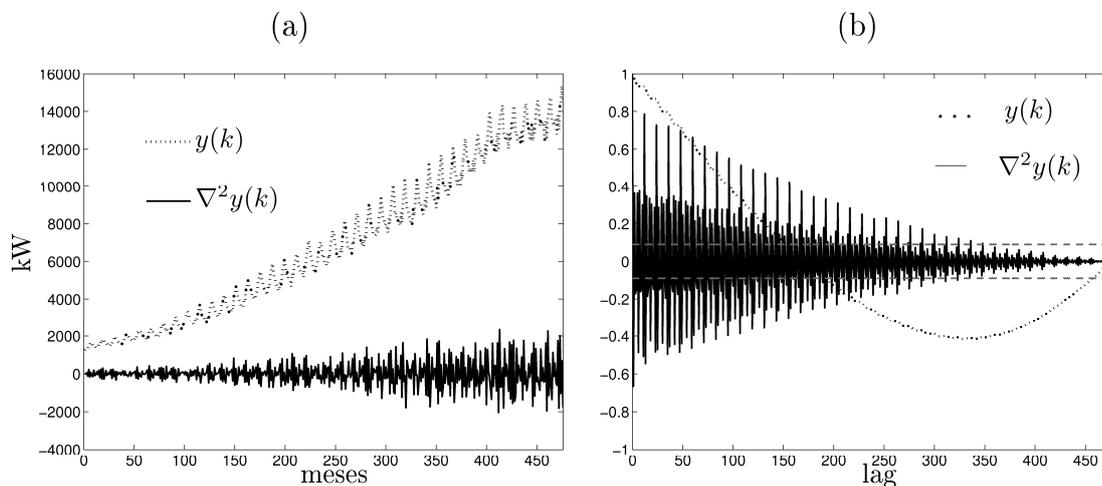


Figura 3.3: Série temporal de produção mensal de eletricidade na Austrália de janeiro de 1956 a agosto de 1995 ELEC.DAT. Em (a) a série $y(k)$ em (...) e a série residual $\nabla^2 y(k)$ em (—) obtida via diferenciação de ordem 2 da série $y(k)$. Em (b) a função de autocorrelação de $y(k)$ em (...) e da série residual $\nabla^2 y(k)$ (—).

3.3.4 Estimação e eliminação da componente de tendência e cíclica

O método de filtro de média móvel pode ser adaptado para eliminar a componente cíclica, de um modelo de decomposição clássico aditivo especificado como se segue:

$$y(k) = m(k) + c(k) + x(k), \quad k = 1, \dots, N, \quad (3.14)$$

em que $c(k)$ é a componente cíclica, sendo $c(k+d) \approx c(k), \forall d$. $m(k)$ é a componente de tendência, $x(k)$ a componente residual e N o número de observações da série.

Método I: Filtro de média móvel para componentes cíclicas (FMMC)

Considerando a série temporal, $y(k)$, com N observações, a componente de tendência $m_0(k)$ ⁷ pode ser estimada aplicando um filtro de média móvel escolhido desconsiderando as componentes $x(k)$ e $c(k)$. Se o período cíclico d for par, então, $d = 2q$ e:

$$\hat{m}_0(k) = \frac{[0,5y(k-q) + y(k-q+1) + \dots + y(k+q-1) + 0,5y(k+q)]}{d}, \quad (3.15)$$

sendo $q+1 \leq k \leq (N-q)$. Se o período for ímpar, então, $d = 2q+1$ e utiliza-se um filtro de média móvel simples (3.7).

O passo seguinte consiste em estimar a componente cíclica, $\hat{c}(k)$. Para cada $k_0 = 1, \dots, d$, calcula-se o valor médio da diferença entre a série $y(k)$ e $\hat{m}_0(k)$, da seguinte forma:

$$\hat{w}(k_0) = \left\{ \left(y(k_0 + jd) - \hat{m}_0(k_0 + jd) \right), \quad q+1 \leq k_0 + jd \leq (n-q) \right\}. \quad (3.16)$$

⁷Na verdade, a componente de tendência $\hat{m}_0(k)$ são os valores médios obtidos nas janelas móveis de dados. Essa “tendência” não incorpora a tendência global da série $y(k)$.

Uma vez que esses desvios médios, \hat{w} , não precisam, necessariamente, ter um somatório igual a zero, a componente cíclica, $c(k)$, pode ser estimada como:

$$\hat{c}(k_0) = \hat{w}(k_0) - \frac{1}{d} \sum_{i=1}^d \hat{w}(i), \quad (3.17)$$

$k_0 = 1, \dots, d$ e $\hat{c}(k_0) = \hat{c}(k_0 + d)$, $k_0 > d$. Logo, pode-se constituir a componente cíclica, $c(k)$, para $k = 1, \dots, N$

Os dados sem a componente cíclica são definidos extraindo a componente cíclica estimada, $\hat{c}(k)$, da série $y(k)$:

$$\hat{D}(k) = y(k) - \hat{c}(k), \quad k = 1, \dots, N. \quad (3.18)$$

Porém, como $\hat{m}_0(k)$ incorpora somente a tendência das janelas deslizantes dos dados com tamanho igual a d e não a tendência da série como um todo, $m(k)$, esta última deve ser estimada por meio da série sem a componente cíclica, $\hat{D}(k)$, usando um dos métodos descritos na seção 3.3.3. Em termos das componentes cíclica e de tendência estimadas, $\hat{c}(k)$ e $\hat{m}(k)$, respectivamente, a série residual $x(k)$ é obtida como:

$$\hat{x}(k) = y(k) - \hat{m}(k) - \hat{c}(k), \quad k = 1, \dots, N. \quad (3.19)$$

Método II: Diferenciação cíclica

A técnica apresentada na seção 3.3.3 para séries sem componente cíclica pode ser adaptada para lidar com séries cíclicas de período d , a partir da introdução da diferença ($lag - d$) no operador de diferenciação ∇_d :

$$\nabla_d y(k) = y(k) - y(k - d) = (1 - B^d)y(k), \quad (3.20)$$

ressaltando que o operador $\nabla_d = (1 - B^d)$ não pode ser confundido com o operador definido anteriormente na Seção 3.3.3.b), $\nabla^d = (1 - B)^d$.

Logo, a série sem ciclos obtida a partir da diferenciação cíclica de período d é

dada por:

$$z(k) = y(k) - y(k-d) = (1 - B^d)y(k), \quad (3.21)$$

ressaltando que, $z(k)$ terá d observações a menos que a série $y(k)$.

Método III: Média dos padrões cíclicos

Uma outra maneira de obter a série de resíduos, $x(k)$, é por meio do cálculo da média dos padrões referentes ao período cíclico d . A série residual pode ser obtida extraindo de $y(k)$ uma série de valores médios cíclicos, $\bar{y}_d(k)$, da seguinte maneira:

$$x(k) = y(k) - \bar{y}_d(k), \quad (3.22)$$

em que, para $k = 1 \dots d$:

$$\bar{y}_d(k) = \sum_{j=0}^{n_d-1} y(i + jd), \quad (3.23)$$

sendo d o período cíclico e n_d o número de períodos cíclicos que ocorrem ao longo da série $y(k)$, com N observações, dado por $n_d = N/d$. Por exemplo, para uma série temporal $y(k)$ com 1.500 observações e período cíclico $d = 10$, o número de períodos cíclicos que ocorrem ao longo da série é $n_d = 15$.⁸

Vale salientar, que apesar de retirar a média da série temporal, este método para reduzir a componente cíclica não incorpora a componente de tendência, $m(k)$. Logo, a tendência deve ser estimada usando um dos métodos descritos na seção 3.3.3.

Método IV: Agrupamento de padrões usando k -médias

Este método procura agrupar padrões similares de uma série temporal por meio da distância entre os padrões. Apesar de ser bastante simples de implementar

⁸Quando n_d não for um número inteiro este deve ser arredondado para o próximo valor inteiro, por exemplo, $n_d = 3,2 \approx 4$.

e de custo computacional reduzido, a escolha de alguns parâmetros que auxiliam na divisão desses padrões em agrupamentos é feita pelo analista de forma empírica.

3.3.5 Aplicação dos métodos para redução dos ciclos

Os métodos descritos na Seção 3.3.4, para estimação e redução da componente cíclica, são aplicados ao mapa estudado por Hénon (1976), dado pelas seguintes equações:

$$\begin{aligned}x_{\text{H}}(k+1) &= 1 - ax_{\text{H}}(k)^2 + y_{\text{H}}(k) \\y_{\text{H}}(k+1) &= bx_{\text{H}}(k),\end{aligned}\tag{3.24}$$

sendo a e b valores constantes. Para $a = 1,4$ e $b = 0,3$, foram gerados $N = 2.000$ observações, sendo que uma faixa desta série e o mapa de primeiro retorno correspondente à série estão ilustrados na Figura 3.4. Foi acrescentado à série temporal $x_{\text{H}}(k)$ em (3.24) uma componente de tendência ($m(k)$) e uma componente cíclica ($c(k)$), dadas respectivamente por:

$$m(k) = \frac{11}{2000}k + 1,\tag{3.25}$$

$$c(k) = 4\text{sen}(\omega k).\tag{3.26}$$

sendo $1 \leq k \leq 2.000$. Para um período $T = 48$, e frequência $\omega = 2\pi/T$.

Logo, a nova série $y(k)$ será composta pela adição de (3.24), (3.25) e (3.26), tal que:

$$y(k) = x_{\text{H}}(k) + m(k) + c(k).\tag{3.27}$$

As Figuras 3.5(b) e (c) mostram, respectivamente, a série temporal periódica, $y(k)$ dada por (3.27), e o mapa de primeiro retorno correspondente à $y(k)$.

Neste cenário, pretende-se que os métodos descritos na Seção 3.3.4 sejam capazes de estimar a série residual $\hat{x}_H(k) \approx x_H(k)$ em (3.24) de modo que se possa reconstruir adequadamente o mapa de Hénon original mostrado na Figura 3.4(b).

O mapa de primeiro retorno reconstruído utilizando-se a série residual estimada $\hat{x}_H(k)$ obtida por cada um dos métodos de redução de ciclos, apresentados na Seção 3.3.4, estão representados na Figura 3.6. Destaca-se que a componente de tendência de $y(k)$ foi extraída ajustando-se um polinômio de ordem 1 à série.

Utilizando somente a inspeção visual para avaliar os mapas reconstruídos na Figura 3.6, observou-se que os métodos I, III e IV apresentaram um desempenho qualitativo satisfatório. Porém, o desempenho do método II não foi adequado. Sendo, os índices RMSE calculados entre a série original ($x_H(k)$) e as séries residuais ($\hat{x}_H(k)$) estimadas pelos métodos: I (RMSE=0,40); II (RMSE=2,22); III (RMSE=0,39) e IV (RMSE = 0,40).⁹

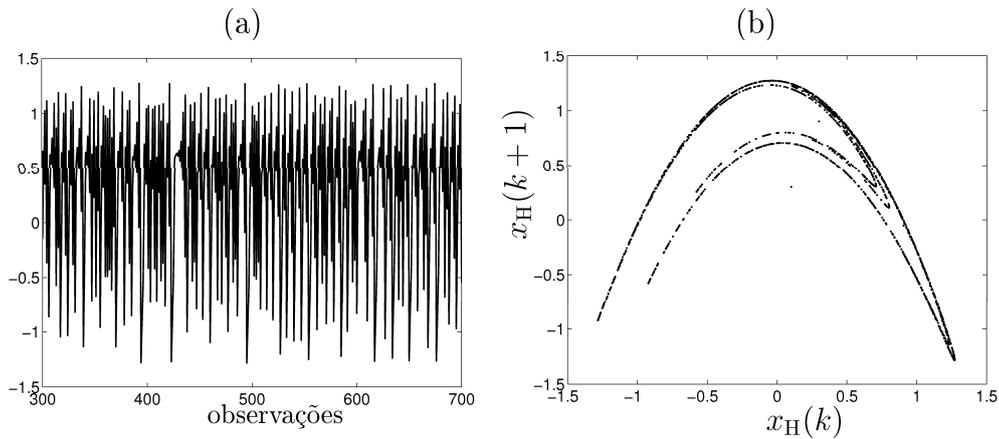


Figura 3.4: Em (a) faixa da série temporal do mapa de Hénon, $x_H(k)$, dado por (3.24). No eixo das abscissas o número de observações e no eixo das ordenadas a amplitude da série $x_H(k)$. Em (b) mapa de primeiro retorno da série $x_H(k)$.

⁹Índice de erro RMSE (raiz do erro médio quadrático):

$$\text{RMSE} = \frac{\sqrt{\sum_{k=1}^N (x_H(k) - \hat{x}_H(k))^2}}{\sqrt{\sum_{k=1}^N (x_H(k) - \bar{x}_H)^2}}$$

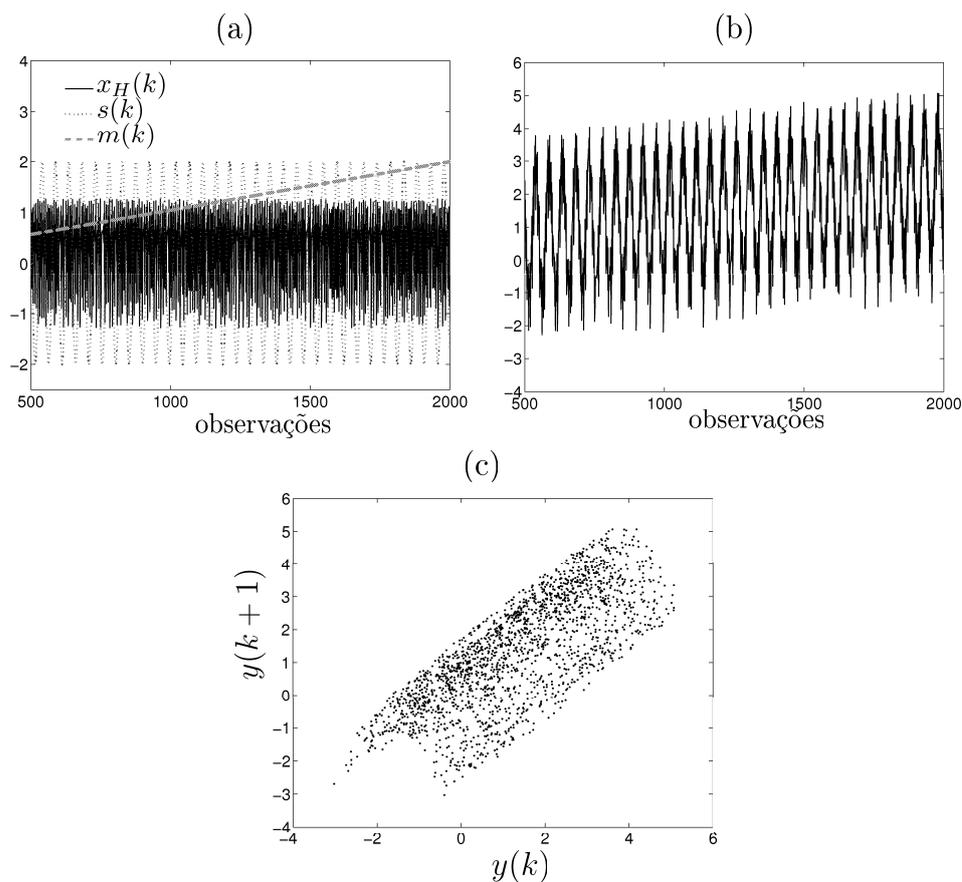


Figura 3.5: Em (a) série temporal do mapa de Hénon $x_H(k)$ (3.24) em (—), a componente de tendência $m(k)$ (3.25) em (---) e, a componente cíclica $c(k)$ (3.26) em (...). Em (b) a série $y(k)$ dada em (3.27). Em (c) mapa de primeiro retorno da série $y(k)$ dada em (3.27).

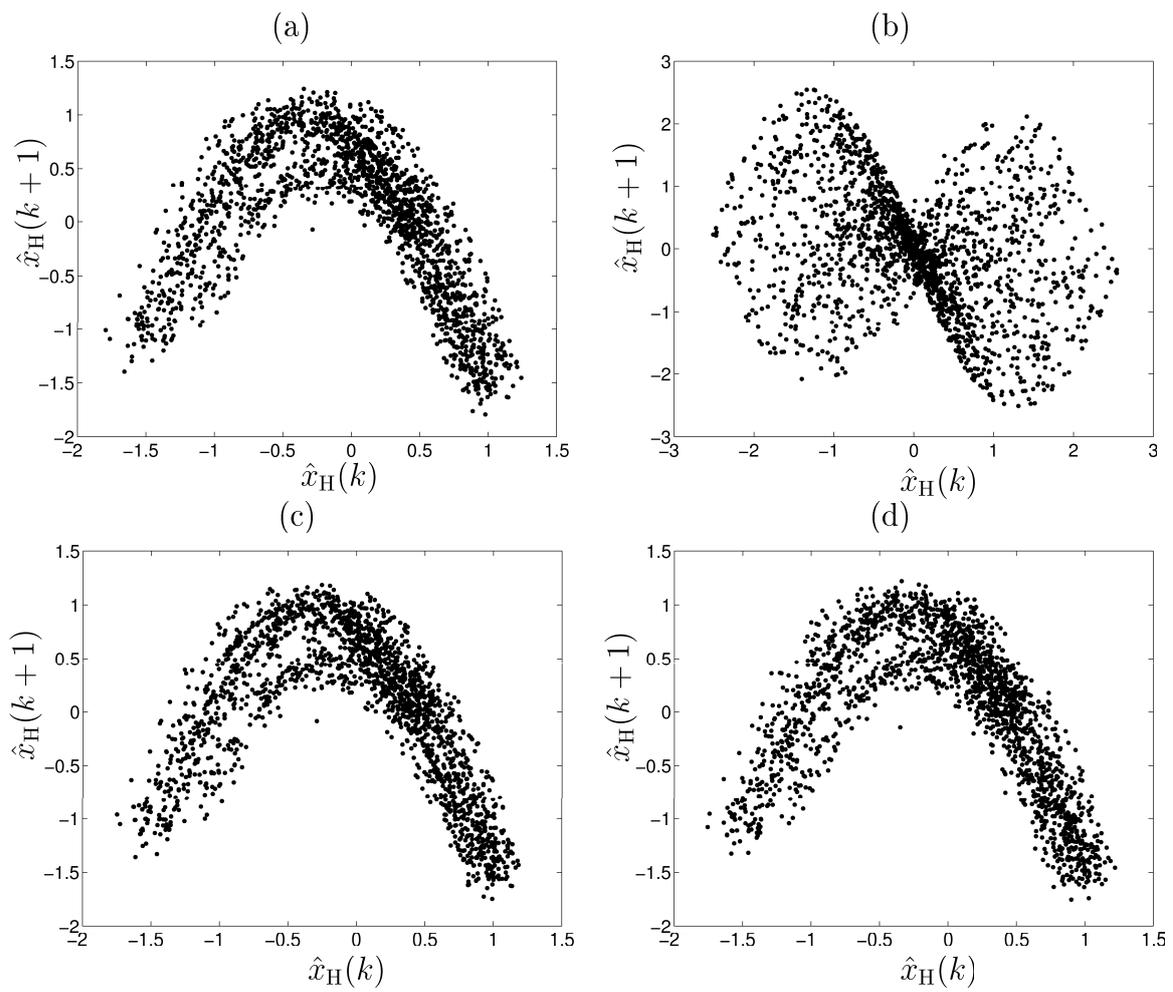


Figura 3.6: Mapa de primeiro retorno estimado pela série residual $\hat{x}_H(k)$ recuperada: em (a) pelo método I com $d = 48$, em (b) pelo método II, em (c) pelo método III e, em (d) pelo método IV.

3.4 Análise não-linear

A aplicação de métodos de análise não-linear em séries temporais, em alguns casos, tem sido direcionada para a detecção de caos determinístico em diversas áreas das ciências econômicas, médicas, entre outras (Hsieh, 1991; Goldberger, 1996; Cao & Soofi, 1999; Soofi & Cao, 2002).

Pode-se quantificar as propriedades de um dado sistema por meio do cálculo das suas invariantes (expoentes de Lyapunov, dimensão de correlação) que podem ser úteis na caracterização do seu atrator, o qual procura relacionar a evolução temporal do sistema dinâmico no espaço de estados.¹⁰

Para reconstruir o espaço de estados é necessário fazer a imersão da série temporal baseado no método das derivadas sucessivas ou no método das coordenadas de atraso (Packard, Crutchfield, Farmer & Shaw, 1980; Takens, 1980).

3.4.1 Imersão de séries temporais

O processamento e a análise de séries temporais caóticas não recebeu bastante atenção até o início da década de 80 devido, principalmente, a ausência de métodos matemáticos para caracterizar se o sistema sob investigação era ou não caótico, e, a deficiência computacional da época para a implementação das técnicas investigadas.

A partir da década de oitenta, com o advento dos computadores, vários pesquisadores começaram a utilizar técnicas e conceitos da teoria do caos, como a reconstrução do espaço de estados,¹¹ para analisar e compreender diversos outros fenômenos hidrodinâmicos, epidemiológicos, químicos e geofísicos, os quais apresentam comportamento dinâmico similar ao comportamento caótico.

O encorajamento dos resultados obtidos em tais estudos, relacionados à complexidade de diversos outros processos, serviu de estímulo para o desenvolvimento e a aplicação de novos procedimentos para a análise, controle e previsão de sistemas

¹⁰Para uma definição completa de atratores veja (Fiedler-Ferrara & Prado, 1994; Monteiro, 2002).

¹¹Na literatura, espaço de estados é mais comumente usado na área das engenharias, enquanto espaço de fases é mais comum na física (Abarbanel, 1993).

caóticos (Casdagli, 1989).

Vários trabalhos em dinâmica não-linear estão concentrados em aprender como classificar os sistemas não-lineares baseados na análise da saída (série temporal) do sistema. A análise de séries caóticas fundamenta-se no uso do teorema de imersão, o qual permite a reconstrução do espaço de estados por meio de uma das variáveis do sistema (Packard et al., 1980).¹²

A proposta inicial para a reconstrução do espaço de estados a partir de séries temporais foi apresentada por Packard e colegas em 1980. Naquele trabalho, foi sugerido o uso de derivadas sucessivas da série temporal de uma única variável do sistema, como sendo as coordenadas para a reconstrução do espaço de estados. Desta forma, para uma série temporal contínua $y(t)$ contendo N amostras, as coordenadas podem ser representadas em um espaço de imersão com dimensão d_i da seguinte maneira:

$$\mathbf{z}(t) = [y(t) \ \dot{y}(t) \ \dots \ d^{d_i}y/dt^{d_i}]. \quad (3.28)$$

Segundo os resultados apresentados no trabalho de Packard e colegas em 1980, as derivadas sucessivas de uma das variáveis do sistema contêm informações sobre a evolução temporal das demais variáveis do sistema. Porém, o cálculo de derivadas amplifica os erros presente nos dados e faz com que o algoritmo torne-se pouco prático, quando o número de variáveis for grande (Fiedler-Ferrara & Prado, 1994).

A dimensão do espaço de imersão d_i é o número mínimo de coordenadas necessárias para a localização de cada ponto do conjunto de dados da série temporal.

3.4.2 Imersão de séries usando coordenadas de atraso

No início da década de oitenta, um grupo da Universidade da Califórnia (Packard et al., 1980) e o pesquisador belgo David Ruelle, simultaneamente, porém,

¹²Na prática, é possível monitorar apenas uma variável do sistema, em função do tempo, no entanto, outras variáveis podem estar envolvidas.

individualmente, introduziram a idéia do uso de coordenadas de atraso para reconstruir os espaço de estados da série temporal (Abarbanel, 1993). Desta forma, o vetor de coordenadas de atraso, $\mathbf{z}(k)$, pode ser representado por:

$$\mathbf{z}(k) = [y(k) \quad y(k - \tau) \quad \dots \quad y(k - (d_i - 1)\tau)],$$

sendo d_i a dimensão de imersão e τ o atraso de tempo. Esse é o método mais usado em imersão de séries temporais e foi fundamentado matematicamente por Takens (1980) (e posteriormente estendido em (Sauer, Yorke & Casdagli, 1991)). Takens (1980) mostrou que é possível reconstruir atratores topologicamente equivalentes ao original usando apenas uma das variáveis coletadas do sistema, desde que a dimensão de imersão seja suficientemente grande. O método estabelecido por Takens é conhecido como método dos atrasos temporais e permite a reconstrução do atrator da série temporal $y(k)$ (Fiedler-Ferrara & Prado, 1994).

3.4.3 Parâmetros de imersão

O teorema de Takens é verdadeiro para casos em que o número de observações é infinito e na ausência de ruído. Takens (1980) não determina nenhuma restrição com relação a escolha do atraso τ e estabelece para a escolha da dimensão de imersão uma condição suficiente, mas não necessária, que $d_i \geq 2D_0 + 1$, sendo D_0 a dimensão fractal do atrator.

Todavia, em aplicações reais, as séries temporais são finitas e contaminadas com ruído. Portanto, a escolha adequada de τ e o cálculo da dimensão de imersão (d_i) é fundamental para inicializar a reconstrução do espaço de estados da série temporal. De fato, muitos dos pesquisadores em reconstrução do espaço de estados tem-se concentrado no problema da escolha do atraso de tempo e da dimensão de imersão, as quais podem ser chamadas de parâmetros de imersão para a coordenadas de atraso (Grassberger & Procaccia, 1983b).

Um dos critérios mais simples e usuais sugere que τ seja escolhido como sendo

a metade do primeiro zero da função de autocorrelação linear (FAC). Outro critério para a escolha de τ é o da informação mútua, baseado em teoria da informação, proposto por Fraser & Swinney (1986), que é mais sistemático e preciso para o cálculo de τ , podendo detectar dependências não-lineares entre as coordenadas de atraso. Este método sugere que ao escolher $\tau \approx \tau'$, pode-se garantir a reconstrução de vetores com o menor nível de informação redundante (vetores linearmente independentes), mas ainda correlacionados.¹³

A escolha da mínima dimensão de imersão pode ser feita:

- por inspeção visual.
- por meio do cálculo de alguma invariante no atrator reconstruído, como por exemplo a dimensão de correlação. Aumentando a dimensão de imersão para a reconstrução do atrator e calculando a invariante para cada valor de d_i , pode-se verificar quando o valor da invariante parar de mudar. Este método subjetivo, requer uma grande quantidade de dados e um alto esforço computacional.
- pelo método dos falsos vizinhos (Kennel & Abarbanel, 1992), o qual é bastante usado na escolha de d_i . Este método determina a dimensão de imersão mínima contando o número de falsos vizinhos. Quando o número de falsos vizinhos tende à zero, é possível identificar o valor mínimo de d_i . Tal critério fornece uma boa aproximação, mas é sensível à escolha subjetiva dos parâmetros limites.

O método dos falsos vizinhos modificado proposto por Cao (1997) tem sido muito usado, e apresenta algumas vantagens com relação ao anterior, dentre elas, pode-se destacar a não necessidade da escolha subjetiva de nenhum dos parâmetros (a não ser o atraso para as coordenadas τ).

¹³ τ' é o intervalo de tempo que minimiza a informação mútua contida em vetores vizinhos ao longo de sua evolução.

3.4.4 Séries quase-periódicas

Considere uma série temporal $y(k)$ composta por duas frequências ω_1 e ω_2 . Pode-se caracterizar uma série temporal como periódica quando a relação entre as frequências que a compõe (ω_1/ω_2) for um número racional. Quando a série é quase-periódica, essa relação entre as frequências é um número irracional. Além disso, o atrator de uma série periódica é diferente do atrator de estados de uma série quase-periódica.

Como mencionado, para reconstruir o espaço de estados da série $y(k)$, basta montar a matriz de coordenadas de atraso, $\mathbf{z}(k)$, da seguinte maneira:

$$\mathbf{z}(k) = [y(k) \quad y(k - \tau) \quad \dots \quad y(k - (d_i - 1)\tau)],$$

sendo d_i e τ a dimensão e o atraso de imersão, respectivamente.

Quando a série $y(k)$ é periódica, a evolução da série no espaço de estados retorna ao ponto inicial quando se atinge o período referente à frequência fundamental. Um exemplo de órbita periódica está mostrado na Figura 3.7(a). A relação entre as frequências $\omega_1/\omega_2 = 0,5$.

No entanto, se a série $y(k)$ for quase-periódica, a evolução da série nunca retornará ao mesmo ponto. Logo, os pontos referentes à frequência fundamental são ε -vizinhos, ou seja:

$$\|y(k) - y(k + T_f)\| \leq \varepsilon, \quad (3.29)$$

sendo T_f o período referente à frequência fundamental ω_f e, ε um valor muito menor do que o tamanho do atrator. Dessa forma, pode-se assumir que a órbita formada por $\mathbf{z}(k)$ é uma órbita ε -periódica da série quase-periódica, $y(k)$, com periodicidade d . A Figura 3.7(b) ilustra um exemplo de órbita ε -periódica, com a relação entre as frequências $\omega_1/\omega_2 = 1,314159265358\dots$

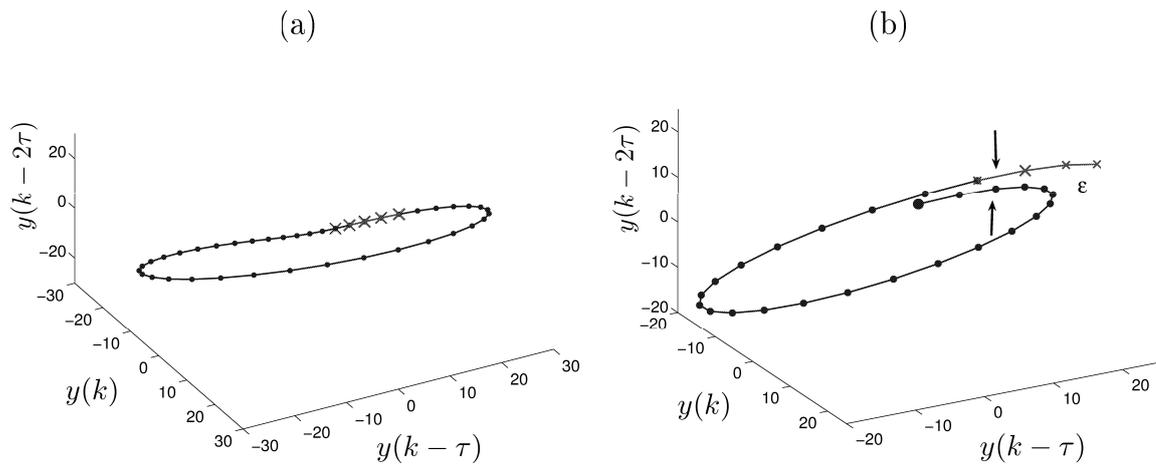


Figura 3.7: Exemplos de órbitas de séries periódicas e quase-periódicas. Em (a) órbita periódica com $\omega_1/\omega_2 = 0,5$. Em (b) órbita quase-periódica com $\omega_1/\omega_2 = 1,314159265358\dots$. Sendo que $(-\bullet-)$ representa a evolução da série até T_f e, em $(-x-)$ a evolução da série a partir de $(T_f + 1)$.

3.5 Análise de séries usando teste de hipóteses

Em muitos problemas reais é interessante inferir sobre um grande grupo de indivíduos, objetos ou parâmetros a partir de uma pequena amostra. Em vez de examinar o grupo inteiro, chamado de população, o que pode ser difícil ou até mesmo impossível de ser feito, pode-se examinar apenas uma pequena parte representativa dessa população, chamada de amostra, e dela pode-se inferir certas características sobre a população. Esse procedimento é conhecido como inferência estatística (Petruccelli, Nandram & Chen, 1999).

Segundo Casella & Berger (2001) e, Petruccelli e colegas 1999, pode-se definir:

Definição 3.5.1 (Inferência estatística) *Inferência estatística é o uso de um subconjunto de uma população, chamada amostra, para extrair conclusões sobre a população. São dois os tipos de inferência estatística: a estimação de parâmetros e o teste de hipóteses.*

Definição 3.5.2 (Hipótese estatística) *Hipótese estatística é qualquer informação sobre a população.*

Definição 3.5.3 (Hipóteses complementares) *As duas hipóteses complementares em um teste de hipóteses são chamadas de hipótese nula e hipótese alternativa, as quais são representadas por H_0 e H_1 , respectivamente.*

Definição 3.5.4 (Teste de hipóteses) *O procedimento para testar hipóteses ou teste de hipóteses é uma regra para decidir:*

- para quais valores do fator estatístico¹⁴ H_0 é aceita como verdadeira;
- para quais valores do fator estatístico H_1 é aceita como verdadeira.

Na análise de séries temporais usando inferência estatística deve-se aceitar ou rejeitar uma informação a respeito de algum parâmetro da série. Essa informação é a hipótese nula (H_0) enquanto que a decisão a respeito da hipótese é denominado teste de hipóteses.¹⁵ A rejeição da hipótese nula sempre leva a aceitação de uma hipótese alternativa (H_1).

Para testar a hipótese pode-se atribuir um parâmetro à H_0 . Esse parâmetro assume um valor ou um intervalo de valores chamado de região de aceitação e seus valores limites são denominados valores críticos. Dessa forma, se o parâmetro testado estiver dentro do intervalo de valores assumido por H_0 não se pode rejeitar H_0 . A faixa de valores referentes ao parâmetro atribuído à H_1 é denominada região crítica ou região de rejeição de H_0 .

A Figura 3.8 ilustra, de forma geral, a tomada de decisão em um teste de hipóteses. Em que se verifica que H_0 não é rejeitada quando, por exemplo, a média amostral \bar{x} , ou qualquer outra informação a respeito da série estiver dentro da região de aceitação e, rejeita-se H_0 em favor da hipótese alternativa H_1 , quando \bar{x} estiver dentro da região de rejeição (ou região crítica).

¹⁴Fator estatístico é um valor atribuído as hipóteses de modo a discriminá-las. Esse fator é obtido por meio das amostras. Um exemplo de fator estatístico discriminante é a média da amostra (\bar{x}).

¹⁵É importante lembrar que hipóteses são sempre informações sobre a população e não sobre a amostra. Neste contexto, as hipóteses testadas são informações sobre o processo gerador da série temporal quase-periódica.

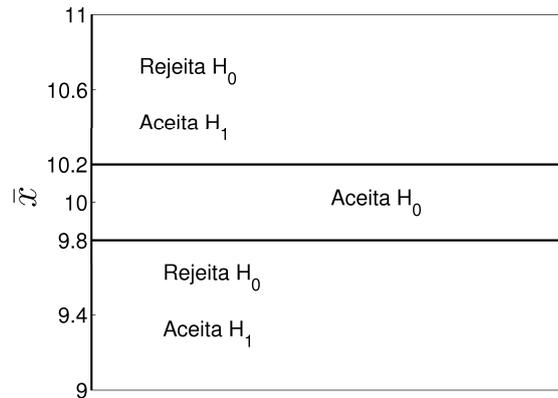


Figura 3.8: Decisão de um teste de hipóteses cujo parâmetro relacionado à H_0 é a média amostral \bar{x} de uma série temporal.

Em suma, se a informação for consistente com a hipótese nula, então não se rejeita H_0 . No entanto, se essa informação for inconsistente com H_0 , rejeita-se H_0 com um dado nível de confiança.

A veracidade ou falsidade de uma hipótese nula nunca será conhecida com certeza, a menos que toda a população seja examinada. Desse modo, um procedimento de teste de hipóteses deve ser desenvolvido tendo-se em mente a possibilidade de alcançar uma conclusão errada, que em termos estatísticos é definida por meio de dois tipos de erro:

- erro do tipo I: rejeita-se H_0 quando essa é verdadeira, ou,
- erro do tipo II: aceita-se H_0 quando essa é falsa.

A probabilidade de cometer o erro do tipo I é também denominada nível de significância do teste de hipóteses (α), o qual é a soma da área da distribuição de probabilidade da região de aceitação da(s) hipótese(s) alternativa(s). O nível de significância será usado para quantificar o grau de confiança do teste de determinismo ao longo deste trabalho.

Uma outra informação importante a ser definida para a execução de um teste de hipóteses é escolher se este será bilateral, quando a região de rejeição de H_0 divide-se em duas partes, ou unilateral, quando somente uma parte da distribuição estatística do teste compõe a região de rejeição de H_0 . Para o teste bilateral pode-se

ter até duas hipóteses alternativas H_1 e H_2 . Assim, usando como fator estatístico a média amostral \bar{x} , define-se:

$$\left. \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \\ \text{ou,} \\ H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{array} \right\} \Rightarrow \text{Teste de hipóteses unilateral}$$

$$\left. \begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \\ \text{ou,} \\ H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \\ H_2 : \mu > \mu_0 \end{array} \right\} \Rightarrow \text{Teste de hipóteses bilateral}$$

em que μ é a média da população.

Em alguns problemas reais, rejeitar H_0 é uma conclusão um tanto forte, uma vez que o pesquisador é quem escolhe o nível de significância do teste, α . Logo, ao estabelecer a(s) hipótese(s) alternativa(s) deve-se levar em consideração o ponto de vista e a experiência do pesquisador com relação à série temporal analisada.

Em suma, pode-se descrever os principais passos para a elaboração de um teste de hipóteses da seguinte maneira:

- 1^o - a partir do contexto do problema, identificar a informação de interesse a ser testada e assim estabelecer H_0 ;
- 2^o - escolher o nível de significância α do teste de hipóteses;
- 3^o - estabelecer o fator estatístico discriminante apropriado;
- 4^o - especificar a(s) hipótese(s) alternativa(s);

5^o - estabelecer a região de rejeição e aceitação de H_0 ao longo do fator estatístico;

6^o - testar a hipótese nula H_0 , ou seja, rejeitá-la ou aceitá-la.

No cenário deste trabalho, a informação a ser testada é a previsibilidade, conseqüentemente, pode-se inferir sobre a assinatura determinística das séries temporais quase-periódicas. A hipótese nula a ser testada é se a série temporal de interesse é aleatória, ou seja, se a série provém de um processo aleatório. A maneira como são especificados e escolhidos o fator estatístico, o nível de significância, as hipóteses alternativas e as regiões de aceitação e de rejeição de H_0 é apresentada na Seção 4.1.

3.6 Teste de hipóteses baseado em dados sub-rogados

O teste de hipótese baseado em dados sub-rogados foi inserido por Theiler e colaboradores em 1992 no campo da física e das engenharias. Tal teste é bastante utilizado para detecção de não-linearidade e determinismo em séries temporais (Theiler, 1995; Palus, 1996; Stam, Pijn & Pritchard, 1998; Schreiber, 2000; Dolan & Spano, 2001; Small & Tse, 2002; Rodrigues, 2002; Small & Tse, 2003).

Séries sub-rogadas (*surrogates*) formam um conjunto de dados gerados artificialmente a partir de uma série temporal a qual deseja-se analisar. Tendo em mente qual característica deve ser investigada, como por exemplo não-linearidade, determinismo, pode-se definir qual o método deve ser usado para gerar os dados sub-rogados (Theiler et al., 1992; Schreiber & Schumitz, 2000; Small et al., 2001; Nakamura & Small, 2005). Dessa forma, determina-se também quais são as propriedades da série temporal que serão preservadas nas séries sub-rogadas, e qual é o fator discriminante que será usado.

O teste de hipótese baseado em séries sub-rogadas consiste em gerar um conjunto de séries artificiais que preservam algumas características estatísticas da série temporal analisada, e destroem outras. Além disso, tais séries artificiais são consistentes com a hipótese nula (H_0) de interesse.¹⁶ Em seguida, aplica-se algum

¹⁶ H_0 define a natureza do processo que deu origem à série sob análise, sendo que este processo

fator estatístico discriminante à série original e às séries artificiais. Se o valor do fator discriminante da série temporal for significativamente diferente do obtido para o conjunto de séries sub-rogadas, então pode-se rejeitar a hipótese nula como sendo uma provável origem da série temporal. Porém, se o fator discriminante da série temporal não for significativamente diferente do obtido para o conjunto de dados sub-rogados, não se pode rejeitar a H_0 .

3.6.1 Algoritmos para séries temporais sem ciclos

Alguns algoritmos foram propostos na literatura para gerar séries sub-rogadas a partir de séries temporais sem ciclos, sendo os algoritmos *Shuffled*, FT e AAFT, os três mais comumente usados (Theiler et al., 1992) e serão aqui descritos.

Algoritmo baseado em embaralhamento temporal

Nesse algoritmo, também conhecido como *random Shuffled*, os dados sub-rogados são gerados pelo embaralhamento, de maneira aleatória, da ordem temporal da série temporal de interesse. Dessa forma, pretende-se que toda a correlação linear e não-linear da série temporal seja destruída mas, a distribuição de probabilidade seja preservada. Uma das hipóteses nulas, H_0 , é que a série temporal é um ruído aleatório i.i.d. (*independent and identically distributed*).¹⁷ Sendo possível testar, a partir deste algoritmo, se há indício de alguma dinâmica determinística na série temporal. No algoritmo proposto por Theiler e colegas (1992), o embaralhamento temporal ocorre sem reposição.

Algoritmo FT

No algoritmo FT ou *phase randomized Fourier transformed*, aplica-se a transformada discreta de Fourier (DFT) à série temporal, $y(k)$, em que $0 \leq k \leq N - 1$. As fases resultantes dos pares complexos conjugados são embaralhadas, mantendo-se

pode ou não ser adequadamente explicado pelos dados disponíveis.

¹⁷(Papoulis & Pillai, 2002, cap. 7, pag. 245).

a amplitude espectral, de modo a garantir o mesmo espectro de potência da série temporal e destruir qualquer correlação não-linear. Dessa forma, obtém-se o sinal $Y(n)$, sendo que $0 \leq n \leq N - 1$. Posteriormente, a seqüência gerada é apresentada no domínio do tempo aplicando a transformada inversa de Fourier:

$$y_s(k) = \frac{1}{N} \sum_{n=0}^{N-1} |Y(n)| e^{j(2\pi + \alpha_n)k/N}, \quad (3.30)$$

sendo que $0 \leq \alpha_n \leq 2\pi$ são números aleatórios independentes com distribuição uniforme. A nova série artificial, $y_s(k)$, é aleatória e tem o mesmo espectro de potência da série original e outras estruturas não-lineares da série original destruídas.

Uma das possíveis H_0 , nesse caso, é que a série temporal é um sinal com distribuição gaussiana linearmente autocorrelacionado. Dessa forma, fica claro a dificuldade em se usar este algoritmo, uma vez que a rejeição de H_0 pode estar associada à não admissão de que a série temporal: 1) seja linear ou, 2) seja Gaussiana ou, 3) seja aleatória.¹⁸

Este método para gerar dados sub-rogados não é adequado para séries com distribuição não gaussiana, uma vez que as séries artificiais assim geradas terão distribuição diferente da série temporal, o que poderá induzir uma falsa rejeição da hipótese nula, H_0 .

Algoritmo AAFT

O algoritmo AAFT ou *amplitude adjusted Fourier transformed*, tenta contornar as limitações do algoritmo FT com relação à diferença entre a distribuição da série temporal e a série artificial. Considerando a série temporal como sendo $y(k)$, gera-se um conjunto de números aleatórios com distribuição gaussiana, $x(k)$, sendo $0 \leq k \leq N - 1$. Cada elemento é gerado independentemente a partir de um gerador de números pseudo-aleatórios gaussianos. Posteriormente, a série temporal $y(k)$ é

¹⁸As séries artificiais geradas pelo algoritmo FT são, por construção: lineares, aleatórias e com distribuição gaussiana. Logo, ao se comparar um índice estatístico entre as séries artificiais geradas pelo algoritmo FT com o índice estatístico de uma série temporal não-linear, determinística e não gaussiana, a H_0 poderá ser rejeitada.

reescalada de acordo com esta distribuição gaussiana. Calcula-se a FT desta nova seqüência, submetendo-a a um processo de embaralhamento de fases, como mencionado no algoritmo anterior. A seqüência final é gerada reescalando-se a última seqüência (com fases embaralhadas), por meio de um histograma da série temporal. Portanto, os dados sub-rogados obtidos via algoritmo AAFT conservam o espectro de potência e a distribuição da série temporal.

A hipótese nula a ser testada, neste caso, é que a série é uma transformação estática não-linear monotônica de um processo gaussiano linearmente autocorrelacionado.

A partir destes três foram propostos outros algoritmos, tais como: IAAFT (AAFT iterativo) (Schreiber & Schmitz, 1996); WFT, o qual utiliza o mesmo procedimento adotado pelo algoritmo FT, porém, faz-se o janelamento dos dados antes do cálculo da FT, para contornar o problema de altas frequências espúrias (vazamento espectral). Este janelamento também pode ser adotado no algoritmo AAFT originando o algoritmo AAWFT. O algoritmo CAAFT adota o mesmo procedimento do AAFT, mas preserva a autocorrelação real dos dados originais.

Neste trabalho, para gerar a série artificial por meio do embaralhamento temporal da série temporal (*shuffled*) foi usado o algoritmo IID, que testa a H_0 que a série temporal foi gerada por um processo gerador de ruídos do tipo i.i.d, com qualquer distribuição. Para o embaralhamento dos dados pode-se usar também o algoritmo BOOT em que embaralhamento é feito usando o método *bootstrap*, ou seja, o embaralhamento temporal da série temporal é feito com reposição, diferentemente do proposto em (Theiler et al., 1992).

Os métodos para gerar dados sub-rogados propostos por Theiler e colegas (1992) têm como principal limitação o fato de terem sido concebidos para séries estacionárias. Para contornar este problema, pode-se: dividir os dados em janelas consideradas estacionárias; ou, transformar os dados em estacionários; ou ainda, aplicar o teste de hipótese em que a não-estacionariedade é levada em consideração ao se gerar os dados sub-rogados.

3.6.2 Algoritmos para séries periódicas

Freqüentemente, para séries reais que apresentam características periódicas evidentes, a H_0 testada usando os algoritmos propostos por Theiler e colegas (1992) é trivialmente falsa. Para este tipo de série, é comum indagar se há alguma componente determinística além da componente periódica. Para isso, Theiler (1995) propôs um algoritmo alternativo para o teste de hipótese, que consiste em gerar os dados sub-rogados pelo embaralhamento individual dos ciclos dentro da série, tal algoritmo é conhecido na literatura como *cycle shuffled surrogate*.

Nesse procedimento, as séries sub-rogadas são geradas embaralhando os ciclos dentro da série. Dessa forma, a dinâmica dentro dos ciclos (intra-ciclos) é preservada, mas a dinâmica entre os ciclos (entre-ciclos), não é preservada.¹⁹ Entretanto, Theiler notou que os dados sub-rogados, assim gerados, não eram aleatórios entre os ciclos. Além disso, se os ciclos não ocorrem precisamente no mesmo valor, períodos cíclicos diferentes, os dados sub-rogados não são capazes de preservar a estacionariedade e a continuidade necessárias para o teste de hipóteses (Small et al., 2001).

Algoritmo PPS

Mais recentemente, Small e colegas 2001 propuseram um algoritmo similar ao de Theiler (1995), porém, este preserva a estacionariedade, quando houver, enquanto testa determinismo entre-ciclos. Esse algoritmo é conhecido como PPS (*Pseudo-Periodic Surrogate*).

No que se segue, Small e colaboradores 2001 utilizaram a técnica de imersão de séries temporais em coordenadas de atrasos para manter a característica periódica da série quase-periódica. Esse método procura gerar dados sub-rogados, que preservam o comportamento dos dados em larga escala (a estrutura periódica), mas destroem qualquer estrutura adicional em menor escala (determinismo caótico, linear ou não-linear).

¹⁹Apesar de Theiler 1995 denominar corretamente o método de embaralhamento da série por períodos como *period-shuffled surrogate* tal método é mais conhecido na literatura por *cycle-shuffled surrogate* (Small, 2005).

Os dados sub-rogados são gerados inferindo a dinâmica padrão segundo o modelo local e contaminando as trajetórias no espaço de estados com ruído dinâmico. Com uma escolha apropriada do nível de ruído, a dinâmica intra-ciclos é preservada, mas a dinâmica inter-ciclos não é preservada.²⁰ Logo, a hipótese nula testada usando as séries sub-rogadas, assim geradas, é de uma órbita periódica com ruído descorrelacionado. O algoritmo para gerar as séries artificiais quase-periódicas é descrito a seguir.

Assumindo como conhecidos a dimensão de imersão (d_i) e o atraso de tempo (τ) da série temporal, $y(k)$:

1. gere os vetores de atrasos para a imersão da série temporal, da seguinte maneira:

$$\mathbf{z}(k) = [y(k), y(k + \tau), y(k + 2\tau), \dots, y(k + d_i\tau)], \quad (3.31)$$

sendo a janela de imersão $d_w = d_i\tau - 1$.

2. Denotando por $A = z(k) | k = 1, \dots, N - d_w$ o atrator reconstruído, escolha uma condição inicial aleatória, $s(1) \in A$.
3. Tome $i = 1$.
4. Escolha do próximo vizinho de $s(i)$, $z(j) \in A$ que segue a seguinte relação de probabilidade:

$$\text{Prob}(z(j) = z(k)) \propto \exp \frac{-\|z(k) - s(i)\|}{\rho}. \quad (3.32)$$

em que ρ é nível do ruído adicionado às séries artificiais.

5. Tome $s(i + 1) = z(j + 1)$ como sendo o próximo $s(i)$.
6. Incremente i .

²⁰Se o nível de ruído na série sub-rogada for muito grande, as séries artificiais serão um ruído do tipo i.i.d.. Por outro lado, se o nível de ruído nos dados sub-rogados for muito pequeno, as séries sub-rogadas serão muito semelhantes à série temporal.

7. Se $i < N$, retorne ao passo 4.
8. A série sub-rogada é dada por:

$$\{(s(k))_1\} := \{(s(1))_1, (s(2))_1, \dots, s(k)_1, \dots, (s(N))_1\}$$

sendo que o número subscrito denota a primeira componente de $\{(s(k))_k\}$.

Este algoritmo necessita da escolha adequada de três parâmetros de imersão (d_i , τ e ρ), que como mencionado anteriormente na Seção 3.4.3, não é uma tarefa simples.

Algoritmo SSS

Séries sub-rogadas adequadas para testar determinismo em séries temporais quase-periódicas devem ser consistentes com a hipótese nula de nenhum determinismo intra-ciclo, nem entre-ciclos, além do comportamento periódico. Nakamura e Small (2005) propuseram um método para gerar dados sub-rogados quebrando o determinismo intra-ciclo e mantendo o determinismo entre-ciclos.

Esse método é conhecido como SSS (*Small-Shuffled Surrogate*) e é adequado para testar determinismo de curto prazo em séries quase-periódicas, que também é o objetivo deste trabalho.

Seja $i(k)$ o índice da série temporal $y(k)$, ou seja, $i(k) = k$. Considerando $g(k)$ uma série aleatória Gaussiana, A um valor qualquer, o procedimento para gerar a série sub-rogada, $s(k)$, é descrito como (Nakamura & Small, 2005):

- 1- Faça $i'(k) = i(k) + Ag(k)$;
- 2- Coloque em ordem crescente $i'(k)$ e considere o índices de $i'(k)$ como sendo $\hat{i}(k)$;
- 3- Obtenha $s(k) = y(\hat{i}(k))$

Como pode ser observado, esse algoritmo faz um embaralhamento de curto-prazo na série $y(k)$. Segundo Nakamura & Small (2005) um valor adequado para

$A = 1$, visto que para valores de $A < 1$, as séries sub-rogadas tendem a ser mais parecidas com a série temporal. Para valores de $A > 1$, as séries sub-rogadas tendem a ser mais diferentes da série temporal. Logo, neste trabalho, adotou-se $A = 1$. A Tabela 3.1 mostra um exemplo para elucidar o procedimento proposto por Nakamura e Small.

Tabela 3.1: Exemplo de uma série sub-rogada $s(k)$ gerada pelo algoritmo SSS.

$y(k)$	$i(k)$	$v'(k)$	ordem crescente de $v'(k)$	$\hat{i}(k)$	$s(k)$
15	1	1,29	0,66	2	25
25	2	0,66	1,29	1	15
12	3	3,71	3,71	3	12
47	4	5,62	4,30	5	32
32	5	4,30	5,62	4	47
50	6	6,85	6,40	8	37
12	7	8,25	6,85	6	50
37	8	6,40	7,55	9	44
44	9	7,55	8,25	7	12
29	10	10,57	10,57	10	29

3.7 Comentários finais

Este capítulo apresentou algumas das principais ferramentas para a análise de séries temporais. Procurou-se mostrar a principal diferença entre uma série temporal periódica e uma série temporal quase-periódica. Ademais, destacou-se os principais métodos usados na literatura para a redução da componente cíclica e de tendência, bem como os principais algoritmos para se gerar séries sub-rogadas para a execução do teste de determinismo, proposto neste trabalho, para séries quase-periódicas.

ANÁLISE DE SÉRIES TEMPORAIS

BASEADA EM INFERÊNCIA ESTATÍSTICA

Neste trabalho procura-se compreender o processo gerador das séries temporais quase-periódicas a partir da análise da série temporal medida. Por questões naturais, econômicas e até mesmo físicas, nem sempre é possível fazer diversas realizações para a obtenção de várias séries temporais. Logo, procura-se inferir sobre a característica determinística do processo gerador da série quase-periódica por meio de uma única amostra, a série temporal medida, usando para isso, o teste de hipóteses baseado em dados sub-rogados.

O procedimento proposto para a análise de determinismo em séries temporais consiste em gerar um conjunto de séries sub-rogadas com características estatísticas semelhantes as da série original testada (tal como: média, distribuição, etc). Porém, tais séries sub-rogadas devem ser aleatórias por construção.¹ Dessa forma, o teste de hipóteses consiste em verificar se a série original pode ser ou não oriunda de um processo aleatório, tal como o processo aleatório que gerou as séries sub-rogadas. Logo, a hipótese nula (H_0) a ser testada é que a série original é aleatória. Se H_0 for rejeitada, aceita-se a hipótese alternativa (H_1), que neste caso, é que a série original tem traços de determinismo.

Este capítulo mostra de maneira detalhada a escolha do índice discriminante

¹Como o teste de hipóteses é para analisar determinismo as séries artificiais devem ser consistentes com a H_0 , ou seja, devem que ser aleatórias.

a ser usado no teste de determinismo, bem como o cálculo do intervalo de confiança. Define, também, as hipóteses nula e alternativas baseado no índice de erro mRSE.

4.1 Estabelecimento do fator estatístico discriminante

A análise de determinismo, aqui usada, baseia-se nas informações sobre a previsibilidade da série temporal a ser investigada e das séries sub-rogadas. Se for possível conseguir um desempenho na previsão da série residual melhor do que o desempenho na previsão das séries sub-rogadas, tal série original é considerada previsível, e conseqüentemente, com tendo algum vínculo determinístico. Enfatizando que o fato de ter alguma assinatura determinística não implica que a série seja totalmente previsível, como mencionado na Seção 3.1.1.

A previsão é obtida ajustando um modelo às séries temporais, residual e sub-rogadas. Neste trabalho, utiliza-se o modelo NARMA polinomial baseada na experiência do grupo MACSIN em seu estudo, desenvolvimento, aplicação, além da disponibilidade de pacotes computacionais implementados para a construção desse tipo de modelo.

Para quantificar o desempenho da previsão obtida via modelos NARMA polinomiais utiliza-se os seguintes índices quantitativos de medida de erro de previsão, calculados em j janelas de dados deslizantes para cada i passo de previsão. Um índice simples de medida de erro é a média da raiz do erro quadrático² normalizado, dado por:

$$\text{mRSE}_N(i) = \frac{1}{n_j} \sum_{j=1}^{n_j} \frac{\sqrt{(y_j(i) - \hat{y}_j(i))^2}}{\max(y_j)}, \quad (4.1)$$

em que $y_j(i)$ é a série a ser testada na janela de dados j e, $\hat{y}_j(i)$ é a série prevista, na janela j , pelo modelo NARMA polinomial. Sendo que o índice $\text{mRSE}_N(i)$ é calculado em $j = 1, 2, \dots, n_j$ janelas de dados nas quais são realizadas as previsões de $i = 1 \dots h$ passos à frente, em que h é o horizonte máximo de previsão. Aqui, o erro é normalizado pelo valor máximo da série temporal medida na janela j , $\max(y_j)$.

²mRSE (*mean Root Square Error*).

Além do índice $\text{mRSE}_N(i)$, optou-se por usar também índices que comparam a previsão feita pelo modelo NARMA polinomial com a previsão feita por um modelo “benchmark” (padrão). Nesse caso, tem-se o índice $\text{mRSE}(i)$ e a estatística $U(i)$ de Theil, dados respectivamente por (Makridakis et al., 1998):

$$\text{mRSE}(i) = \frac{1}{n_j} \sum_{j=1}^{n_j} \frac{\sqrt{(y_j(i) - \hat{y}_j(i))^2}}{\sqrt{(y_j(i) - \bar{y}_j)^2}}, \quad (4.2)$$

$$U(i) = \frac{1}{n_j} \sum_{j=1}^{n_j} \frac{\sqrt{(y_j(i) - \hat{y}_j(i))^2}}{\sqrt{(y_j(i) - y_j(i-1))^2}}, \quad (4.3)$$

em que $y_j(i)$ é a série na janela de dados j , $\hat{y}_j(i)$ é a série prevista pelo modelo NARMA polinomial na janela de dados j , \bar{y}_j é o valor médio da série $y_j(i)$ e, $y_j(i-1)$ é o valor passado da série $y_j(i)$. Ressaltando, que tais erros são calculados em n_j janelas de dados nas quais são realizadas as previsões de $i = 1 \dots h$, em que h é o horizonte máximo de previsão.

Observa-se pela equação (4.2) que o índice $\text{mRSE}(i)$ compara o erro de previsão do modelo NARMA polinomial ($y_j(i) - \hat{y}_j(i)$) com o erro de previsão obtido pelo modelo de médias ($y_j(i) - \bar{y}_j(i)$). O índice $U(i)$ de Theil compara o erro de previsão do modelo NARMA polinomial ($y_j(i) - \hat{y}_j(i)$) com o erro de previsão obtido pelo modelo trivial (naive) ($y_j(i) - y_j(i-1)$).³

Para escolher qual índice de erro de previsão é o mais adequado para ser usado no teste de determinismo proposto neste trabalho, os três índices de erro foram empregados em uma série simulada com assinatura determinística por construção. Os resultados obtidos para a previsão de 1-10 passos à frente estão apresentados na Tabela 4.1. Espera-se que os índices de erro de previsão da série com vínculo determinístico sejam menores que os índices de erro médio das séries sub-rogadas, uma vez que tais séries são aleatórias por construção, conseqüentemente, imprevisíveis.

³Serão considerados dois tipos de modelos triviais: o modelo ingênuo *naive* e o modelo de médias. O modelo *naive* prevê o presente como sendo o valor passado ($\hat{y}(i) = y(i-1)$). O modelo de médias prevê o valor atual usando a média de todos os valores passados ($\hat{y}(i) = \bar{y}$).

Pela Tabela 4.1 pode-se verificar que o desempenho da previsão da série simulada com assinatura determinística é melhor do que o desempenho médio da previsão das séries sub-rogadas, uma vez que o índice $mRSE(i)$ da série simulada é menor do que o índice $mRSE$ médio das 1.000 séries sub-rogadas para um horizonte de 1 até 8 passos à frente. O mesmo ocorre para o índice $mRSE_N(i)$, porém, para todo o horizonte de previsão que é de 1 até 10 passos à frente.

Porém, quando se utiliza o índice de erro $U(i)$ para os passos de previsão 1-4, 6-7 e 9-10, os índices de erro para a série simulada são menores do que os índices médios das séries sub-rogadas. Esse resultado não é consistente com o que se sabe da série simulada, ou seja, esperava-se índices $U(i)$ menores para a série simulada do que para as séries artificiais.

Tabela 4.1: Índices de erro de previsão: $mRSE(i)$, $U(i)$ e $mRSE_N(i)$, obtidos para a série simulada com assinatura determinística e os erros médios para o conjunto de 1.000 séries sub-rogadas.

i	$mRSE(i)$		$U(i)$		$mRSE_N(i)$	
	série determinística	média dados sub-rogados	série determinística	média dados sub-rogados	série determinística	média dados sub-rogados
1	0,348	1,291	1,352	2,391	0,150	0,292
2	0,394	1,089	1,593	2,538	0,200	0,282
3	0,360	1,025	1,396	2,480	0,204	0,280
4	0,391	1,009	1,794	2,505	0,202	0,282
5	0,421	1,003	3,227	2,364	0,205	0,287
6	0,498	1,003	2,162	2,439	0,214	0,289
7	0,615	1,002	1,724	2,612	0,215	0,295
8	0,809	1,000	3,049	2,678	0,266	0,303
9	1,298	1,001	2,181	2,504	0,261	0,313
10	1,226	1,000	1,623	2,904	0,226	0,325

Diante dos resultados apresentados na Tabela 4.1, percebe-se que os índices de erro de previsão que foram consistentes com a série simulada com assinatura determinística são os índices $mRSE(i)$ e $mRSE_N(i)$. Uma vez que esses índices de erro foram menores para a série simulada, mostrando assim, que tem-se uma melhor previsão para a série com assinatura determinística do que para as séries sub-rogadas

(aleatórias). No entanto, como o índice $\text{mRSE}(i)$ compara as previsões feitas pelo modelo NARMA polinomial com um modelo padrão (modelo de médias), optou-se pelo uso deste como o fator estatístico do teste de determinismo aqui empregado.

Portanto, ao comparar o erro de previsão obtido pelo modelo NARMA polinomial $(y(i) - \hat{y}(i))$, com o erro de previsão obtido pelo modelo de médias $(y(i) - \bar{y})$, tem-se que, para um horizonte de previsão $i = 1, 2, \dots, h$:

- $\text{mRSE}(i) < 1$ significa que o modelo NARMA polinomial ajustado consegue prever melhor do que um modelo de média;
- $\text{mRSE}(i) \approx 1$ significa que o modelo NARMA polinomial tem um desempenho de previsão comparável ao modelo de média;
- $\text{mRSE}(i) > 1$ significa que o modelo NARMA polinomial tem um desempenho pior do que o modelo de média.

Empiricamente, sabe-se que o modelo NARMA polinomial, quando estável, é capaz de prever o valor médio de uma série temporal estacionária. Como as séries sub-rogadas ao serem geradas devem ter o mesmo valor médio da série residual, pode-se assumir que as séries sub-rogadas são estacionárias na média, por construção, além de serem aleatórias. Portanto, os modelos NARMA polinomiais ajustados a essas séries sub-rogadas, quando estáveis, são capazes de prever somente o valor médio dessas séries artificiais aleatórias.

Ressalta-se que, neste trabalho, as séries sub-rogadas são aleatórias, e por definição, sabe-se que séries aleatórias não têm uma lei que as governe, conseqüentemente, não são previsíveis. Por conseguinte, de acordo com o mRSE na equação (4.2), é provável que o erro obtido pela previsão do modelo NARMA polinomial seja igual ao erro obtido pelo modelo de médias, visto que o modelo NARMA polinomial será capaz de prever somente o valor médio da série artificial. Portanto, o índice $\text{mRSE}(i)$ para as séries sub-rogadas deve ser em torno da unidade.

Tomando como base o fator estatístico discriminante, o índice $\text{mRSE}(i)$, testa-se a H_0 de aleatoriedade na série original. Como mencionado, se o índice mRSE

da série original for igual ao índice mRSE médio das séries sub-rogadas, aceita-se H_0 o que significa que a série original é aleatória, caso contrário, rejeita-se H_0 o que implica que a série original tem algum vínculo determinístico.

4.2 Definição das hipóteses alternativas

Para definir se o teste de hipóteses, aqui empregado, deve ser unilateral ou bilateral, e assim estabelecer a(s) hipótese(s) alternativa(s), considere os seguintes exemplos usando cinco séries temporais periódicas simuladas. Foi acrescentado em todas as séries a componente cíclica com período $d = 15$ dada por:

$$c(k) = 2\text{sen}(\omega k), \quad (4.4)$$

contendo $N = 1.980$ observações, sendo $k = 1 \dots N$ e $\omega = 2\pi/15$.

Dessa forma, tem-se as séries: $s_A(k)$, $s_B(k)$, $s_C(k)$, $s_D(k)$ e $s_E(k)$ com características distintas, a saber:

- $s_A(k) \Rightarrow$ série com assinatura determinística e estacionária:

$$\begin{aligned} z(k) &= 0,8z(k-1) + 0,1z(k-2) + 0,05e(k), \\ s_A(k) &= z(k) + c(k), \end{aligned} \quad (4.5)$$

em que $e(k)$ é um sinal aleatório com média zero e variância 1.

- $s_B(k) \Rightarrow$ série sem assinatura determinística e, estacionária: gerada pela função *randn.m* do MATLAB somada à componente cíclica $c(k)$ dada em (4.4).

- $s_C(k) \Rightarrow$ série com assinatura determinística e não estacionária na média:

$$\begin{aligned} x(k) &= 0,8x(k-1) + 0,1x(k-2) + 0,5u(k), & (\text{série determinística ARX}) \\ m(k) &= \nu(k) + 0,2, & \text{tendência, em que: } -3,2 \leq \nu(k) \leq 0,87; \\ s_C(k) &= x(k) + m(k) + c(k). \end{aligned} \quad (4.6)$$

em que $u(k)$ é um sinal aleatório com média zero e variância 1.

- $s_D(k) \Rightarrow$ série determinística e não modelável por um polinômio: o mapa de Ikeda, que para os parâmetros $a = 0,9$ e $b = 6$, é dado por:

$$\begin{aligned} x(k+1) &= 1 + 0,9 \left(x(k) \cos(t(k)) - y(k) \sin(t(k)) \right), \\ y(k+1) &= 0,9 \left(x(k) \sin(t(k)) + y(k) \cos(t(k)) \right), \\ t(k) &= 0,4 - \frac{6}{1 + x(k)^2 + y(k)^2}, \end{aligned} \quad (4.7)$$

em que $s_D(k) = x(k) + c(k)$.

- $s_E(k) \Rightarrow$ série determinística caótica: o mapa de Hénon, que para os parâmetros $a = 1,4$ e $b = 0,3$, é dado por:

$$\begin{aligned} x(k+1) &= 1 - 1,4x^2(k) + y(k), \\ y(k+1) &= 0,3x(k), \end{aligned} \quad (4.8)$$

em que $s_E(k) = y(k) + c(k)$.

Tais séries periódicas, com comportamentos subjacentes distintos, estão ilustradas nos gráficos à esquerda na Figura 4.1. Usando o método FMMC descrito na Seção 3.3.4 para reduzir as componentes de tendência e cíclica, obtém-se as séries residuais correspondentes a $s_A(k)$, $s_B(k)$, $s_C(k)$, $s_D(k)$ e $s_E(k)$. As séries residuais são mostradas nos gráficos intermediários na Figura 4.1. A análise feita nesta seção

para a definição da(s) hipótese(s) alternativa(s) é feita usando tais séries residuais. Salientando-se que, o teste de determinismo é realizado nas séries residuais. Logo, toda a análise é realizada assumindo que a componente cíclica foi extraída.

Para cada série residual gerou-se um conjunto de 40 séries sub-rogadas usando o algoritmo de embaralhamento temporal (*Shuffled*), mencionado na Seção 3.6.1.

Nos gráficos à direita têm-se os índices $\text{mRSE}(i)$ para a série residual correspondente e o índice $\text{mRSE}(i)$ médio para as 40 séries sub-rogadas. Pode-se verificar pela Figura 4.1(a) que para a série residual com alguma assinatura determinística e estacionária, o índice $\text{mRSE}(i)$ para a série residual é menor do que a unidade e, conseqüentemente, menor do que o índice $\text{mRSE}(i)$ médio das séries sub-rogadas. Isso significa que a série residual com assinatura determinística é mais previsível do que as séries sub-rogadas. Logo, rejeita-se H_0 , ou seja a hipótese de que a série residual é aleatória. Portanto, aceita-se a hipótese alternativa de determinismo de curto-prazo na série $s_A(k)$ dada em (4.5), o que confere com o conhecimento *a priori*.

Para o caso da série residual sem vínculo determinístico, a qual é mostrada na Figura 4.1(b), o índice $\text{mRSE}(i)$ é muito próximo do índice $\text{mRSE}(i)$ médio das séries sub-rogadas, ou seja, em torno da unidade. Neste caso, é muito difícil diferenciar os índices $\text{mRSE}(i)$ da série residual com o índice $\text{mRSE}(i)$ médio das séries sub-rogadas. Logo, aceita-se H_0 de que a série residual é aleatória. Conseqüentemente, infere-se que $s_B(k)$ não apresenta vínculo determinístico intra-ciclo.

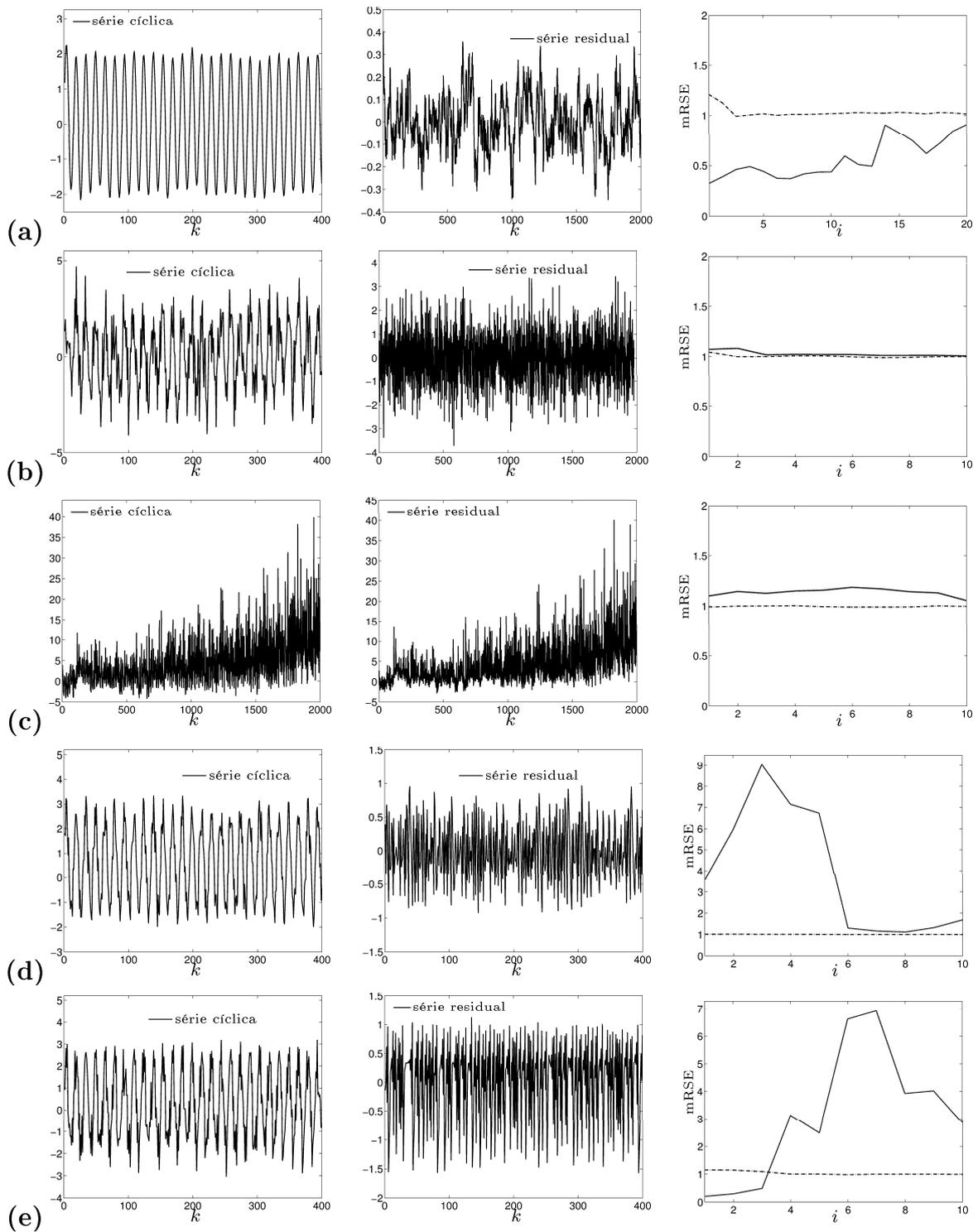


Figura 4.1: Resultado do teste de determinismo para séries cíclicas simuladas: em (a) série com assinatura determinística $s_A(k)$. Em (b) série sem vínculo determinístico $s_B(k)$. Em (c) série com assinatura determinística e não estacionária na variância $s_C(k)$. Em (d) série determinística e não representável por modelos polinomiais $s_D(k)$. Em (e) série determinística caótica $s_E(k)$. Nos gráficos da esquerda uma faixa das séries cíclicas simuladas. Nos gráficos do meio, faixa das séries residuais e nos gráficos à direita os índices $mRSE(i)$, para horizonte de $i = 1, 2, \dots$, para a série residual (—), e o índice $mRSE(i)$ médio para as séries sub-rogadas (- -).

Quando a série residual tem alguma assinatura determinística, porém, é não-estacionária na variância, o índice $\text{mRSE}(i)$ da série residual é maior do que o índice $\text{mRSE}(i)$ médio das séries sub-rogadas, como mostra a Figura 4.1(c). Isso indica que a previsão para a série residual é pior do que a previsão obtida nas séries sub-rogadas, que são aleatórias. Esse resultado poderia levar a uma conclusão errônea de que a série residual correspondente a $s_C(k)$ é aleatória, uma vez que sua previsão é pior do que a previsão das séries sub-rogadas. Porém, sabe-se que isso não é verdade visto que a série $s_C(k)$ tem vínculo determinístico por construção. Diante dessa situação não se pode inferir nada com relação à característica determinística da série residual, somente sobre a sua não estacionariedade. Dessa forma, origina-se uma segunda hipótese alternativa, H_2 , a qual será aceita caso o índice $\text{mRSE}(i)$ seja superior ao índice $\text{mRSE}(i)$ médio das séries sub-rogadas.

Salienta-se, quem em alguns casos, quando a não-estacionariedade é na média, esse resultado pode não ocorrer, visto que os modelos NARMA polinomiais são capazes de aproximar essa variação na média.

Outros dois casos interessantes acontecem quando a série periódica tem dinâmica subjacente determinística e caótica, e quando, a série periódica tem dinâmica determinística, mas, não é modelável por representações polinomiais, como as que são empregadas neste trabalho.

Se a série residual for determinística, porém, não for modelável por polinômios, como a série residual correspondente ao mapa de Ikeda, $s_D(k)$ dado em (4.7), verifica-se pela Figura 4.1(d) que o índice $\text{mRSE}(i)$ para essa série é diferente e maior do que o índice $\text{mRSE}(i)$ médio das séries sub-rogadas, nos seis primeiros passos à frente. Significando que a série residual é menos previsível do que as séries sub-rogadas. O que não implica, como no caso anterior, que a série residual seja aleatória. Esse é um segundo caso em que o índice $\text{mRSE}(i)$ da série residual com vínculo determinístico é maior do que a unidade. Completando assim a informação da hipótese alternativa H_2 : quando o índice $\text{mRSE}(i)$ da série residual for maior do que a unidade pode-se inferir, também, que a série residual não é modelável por poli-

nômios. No entanto, não se pode inferir nada sobre sua característica determinística.

Sabendo que séries caóticas são previsíveis somente para poucos passos à frente, simulou-se o mapa de Hénon via (4.8) acrescentando a componente cíclica $c(k)$ e obteve-se a série $s_E(k)$. O teste de determinismo foi aplicado à série residual correspondente à $s_E(k)$. O resultado do teste de determinismo usando o índice $mRSE(i)$ ilustrado na Figura 4.1(e) mostra que, para essa série residual, as previsões de poucos passos à frente, até aproximadamente 3 passos, apresentam erros bem menores do que a unidade, conseqüentemente, menores do que o índice $mRSE(i)$ médio das séries sub-rogadas. Logo, pode-se inferir que essa série é determinística. Percebe-se também que para um horizonte de previsão superior à 3, o índice $mRSE$ torna-se maior do que a unidade, destacando assim, a característica de séries caóticas com relação ao horizonte de previsão.

Por meio da análise dos resultados usando séries cíclicas simuladas, observou-se que o fato do índice $mRSE(i)$ obtido para a série residual ser diferente e maior do que o índice $mRSE(i)$ médio para as séries sub-rogadas não, necessariamente, significa que a série residual é aleatória ou determinística. Pode significar que não se conseguiu ajustar um modelo adequado para essa série residual, ou mesmo que a série residual é não estacionária e por isso o modelo não foi capaz de prever nem mesmo o seu valor médio, induzindo assim, a uma rejeição da hipótese nula e uma aceitação de uma outra hipótese alternativa, que não a de determinismo. Portanto, para o teste de determinismo usando o índice $mRSE(i)$ como fator estatístico discriminante tem-se duas hipóteses alternativas, a saber:

- se o índice $mRSE(i)$ para a série residual for menor do que o índice $mRSE(i)$ médio obtido para as séries sub-rogadas, a hipótese alternativa H_1 a ser aceita é que a série original apresenta uma assinatura determinística;
- se o índice $mRSE(i)$ para a série residual for maior do que o índice $mRSE(i)$ médio das séries sub-rogadas, a hipótese alternativa H_2 a ser aceita é que a série original é não-estacionária, uma vez que o modelo ajustado não foi capaz de prever nem o valor médio da série. Ou então, que a série original não pode

ser representada por modelos NARMA polinomiais.

A análise das hipóteses pode ser feita também para uma série original quase-periódica, a qual assume-se ter extraído o comportamento quase-periódico. Logo, pode-se definir as seguintes hipóteses para testar determinismo:

- H_0 : série original é procedente de um processo aleatório e é aceita quando $\text{mRSE}(i) \approx 1$.
- H_1 : série original é procedente de um processo com algum vínculo determinístico e é aceita quando $\text{mRSE}(i) < 1$.
- H_2 : série original é procedente de um processo não-estacionário ou não modelável por representações NARMA polinomiais e é aceita quando $\text{mRSE}(i) > 1$.

Logo, para efeito de ilustração, o teste de determinismo baseado em dados sub-rogados aqui usado com duas hipóteses alternativas, H_1 e H_2 , bem como as regiões de aceitação de cada hipótese é mostrado na Figura 4.2.

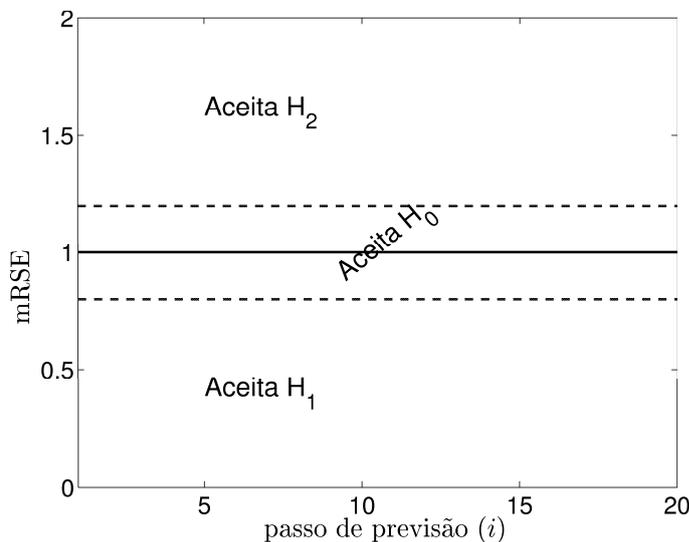


Figura 4.2: Ilustração do resultado do teste de hipóteses para a aceitação e rejeição de H_0 , H_1 e H_2 , usando como fator estatístico discriminante o índice $\text{mRSE}(i)$. Em linha contínua (—) o valor médio dos índices $\text{mRSE}(i)$ dos dados sub-rogados e em linha tracejada (- -) o intervalo de confiança para esse valor médio de erro.

Pela Figura 4.2 pode-se inferir que, se o índice $\text{mRSE}(i)$ para a série residual estiver dentro da faixa de confiança, aceita-se H_0 e sugere-se que a série original é aleatória. Se o índice $\text{mRSE}(i)$ da série residual estiver abaixo da faixa de confiança, rejeita-se H_0 e assume-se que a série original tem traço determinístico, uma vez que o desempenho da previsão da série residual foi significativamente melhor do que o desempenho médio da previsão das séries sub-rogadas. Quando o índice $\text{mRSE}(i)$ da série residual estiver acima da faixa de confiança, não se pode falar sobre a característica determinística da série e sim inferir que a série original é não estacionária, ou que talvez, o modelo NARMAX polinomial não é o mais adequado para representar tal série.

Como mencionado anteriormente, um simples valor pontual do índice discriminante pode não ser suficiente para a realização do teste de hipóteses. Para evitar conclusões errôneas, bem como atribuir um grau de confiança ao teste de determinismo, utiliza-se um intervalo de confiança para o índice mRSE médio das séries sub-rogadas. O procedimento para a construção do intervalo de confiança será descrito na próxima seção.

4.3 Intervalos de confiança

Intervalos de confiança ajudam a interpretar o teste de hipóteses. Além disso, o uso do intervalo de confiança possibilita expressar o grau de confiança associado à estimativa do índice $\text{mRSE}(i)$, conseqüentemente, o grau de confiança do teste de hipóteses.

Definindo o nível de significância α a ser usado no teste, ou seja, a probabilidade de se rejeitar H_0 quando for verdadeira (erro do Tipo I), define-se também o nível de confiança do teste de hipóteses como sendo $(1 - \alpha) \times 100\%$.

Para determinar o nível de significância (α) do teste, e conseqüentemente, o intervalo de confiança,⁴ deve-se determinar qual é o tipo de distribuição de probabi-

⁴Intervalos de confiança, normalmente, estão relacionados ao grau de confiança sobre determinada estimação. Neste trabalho, além dessa associação, o intervalo de confiança é usado para

lidade dos índices $\text{mRSE}(i)$ obtidos para as séries sub-rogadas.⁵

Considerando a distribuição dos índices $\text{mRSE}(i)$ como uma Normal, calcula-se o intervalo de confiança da seguinte maneira (Petruccelli et al., 1999):

$$IC_R = \bar{x}_R \pm z_c \frac{\sigma_R}{\sqrt{n_R}} \quad (4.9)$$

em que \bar{x}_R é estimativa da média dos índices $\text{mRSE}(i)$, σ_R é a estimativa do desvio-padrão dos índices $\text{mRSE}(i)$, n_R é o tamanho de amostras de índices $\text{mRSE}(i)$ considerados e, z_c é o valor crítico que depende do nível de confiança (α) escolhido pelo analista.

No entanto, na prática, percebe-se que o fator estatístico empregado neste trabalho, o índice mRSE , não apresenta uma distribuição normal. Para efeito de ilustração, gerou-se um conjunto de 1.000 séries sub-rogadas para testar determinismo em uma série com assinatura determinística por construção. Ajustou-se um modelo NARMA polinomial às 1.000 séries sub-rogadas e realizou-se a previsão de 1 até 6 passos à frente, calculando o índice mRSE para cada um dos passos de previsão.

O histograma do índice mRSE para os horizontes de previsão de 1-6 passos à frente está representado na Figura 4.3. Por meio de uma análise visual, verifica-se na Figura 4.3 que a distribuição dos índices $\text{mRSE}(i)$ nem sempre se apresentam na forma de uma Normal. Dependendo do horizonte de previsão a distribuição do $\text{mRSE}(i)$ é diferente de uma distribuição normal, como acontece para os horizontes 1 e 2 mostrados nas Figuras 4.3 (a) e (b), respectivamente. Verifica-se também que à medida que o horizonte de previsão aumenta, as distribuições dos índices $\text{mRSE}(i)$ aproximam-se de uma distribuição Normal, como mostram as Figuras 4.3 (c), (d), (e) e (f).

Pôde-se observar por meio de séries simuladas que o supracitado, geralmente, ocorre quando a série original tem algum traço determinístico. Conjectura-se, portanto, que ao se gerar as séries sub-rogadas a partir da série original, tais séries

auxiliar na tomada de decisão do teste de hipóteses.

⁵Os tipos de distribuições de probabilidade mais conhecidos na literatura são: Normal, qui-quadrada e *t-student*. Para maiores detalhes ver (Papoulis & Pillai, 2002).

artificiais mantêm alguma característica determinística da série original. Fazendo, então que, os modelos ajustados a tais séries artificiais tentem modelar esse determinismo herdado. O mesmo não ocorre quando a série original é aleatória, como será visto a seguir.

A partir de uma série aleatória por construção gerou-se um conjunto de 1.000 séries sub-rogadas. Um modelo NARMA polinomial foi ajustado às séries sub-rogadas e foram feitas previsões de 1-6 passos à frente. A Figura 4.4 ilustra o histograma dos índices mRSE das séries sub-rogadas geradas a partir de uma série original aleatória. Percebe-se, por meio da análise visual da Figura 4.4, que a distribuição dos índices $\text{mRSE}(i)$ obtidos para as séries sub-rogadas, geradas a partir da série original aleatória, apresentam distribuição mais próxima da Normal para todos os horizontes de previsão.

Para contornar o problema de estimar o intervalo de confiança assumindo que a distribuição dos índices $\text{mRSE}(i)$ é uma Normal, quando na verdade ela não o é, pode-se calcular o intervalo de confiança por meio de percentis ou quartis.⁶ Essas medidas são usadas, particularmente, em distribuições assimétricas, bem como distribuições simétricas como no caso da Normal (Petruccelli et al., 1999). Um percentil é uma medida da p -ésima posição de uma observação com relação às demais observações. Para calcular o p -ésimo percentil, deve-se ordenar as observações dos índices mRSE na forma crescente, e assim calcular:

$$p = 100 \times \frac{(j - 1)}{n_R - 1}, \quad (4.10)$$

em que n_R é o número de amostras de índices mRSE levados em consideração, e j é a posição da observação no vetor mRSE. Portanto, o p -ésimo termo será mRSE_p ,

⁶Os percentis de números 25, 50, 75 são chamados de 1^o quartil (Q1), 2^o quartil (Q2 ou mediana) e 3^o quartil (Q3).

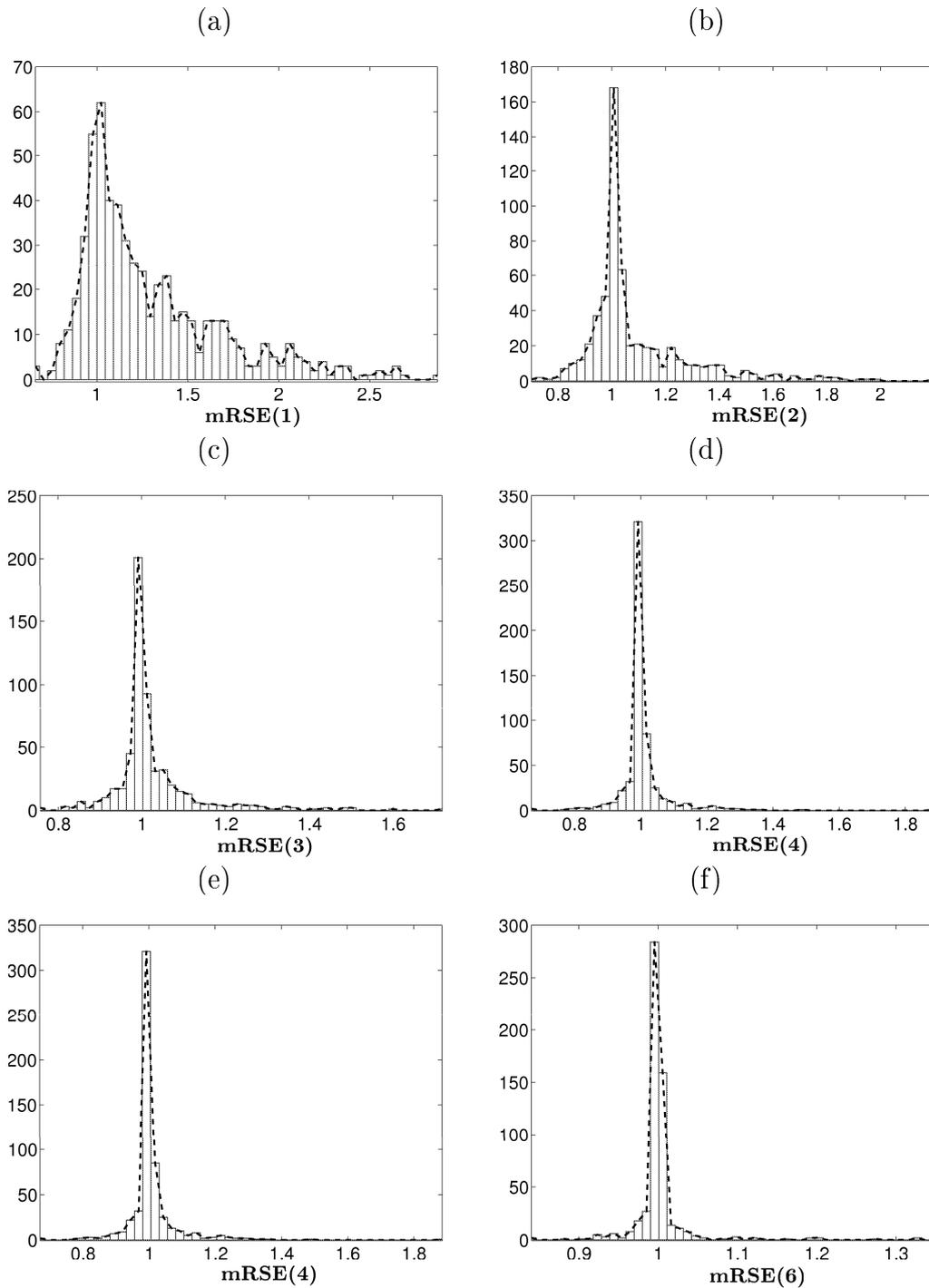


Figura 4.3: Histograma dos índices $mRSE(i)$ obtidos para a previsão de $i = 1 \cdots 6$ passos à frente para as séries sub-rogadas geradas a partir da série com vínculo determinístico. Em (a) $mRSE(i)$ para um passo à frente. Em (b) $mRSE(i)$ para dois passos à frente, e assim sucessivamente.

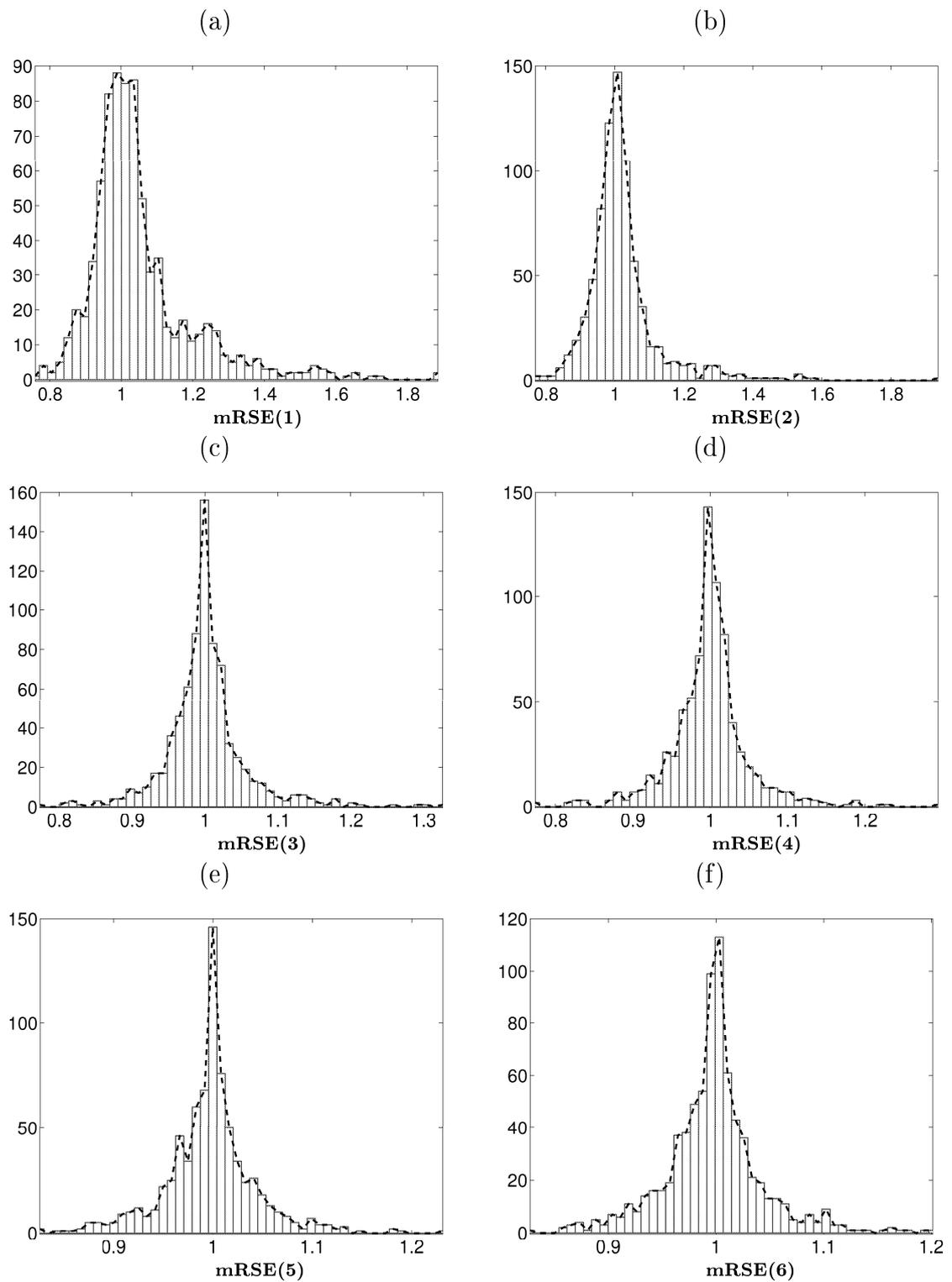


Figura 4.4: Histograma dos índices $mRSE(i)$ obtidos para a previsão de $i = 1 \dots 6$ passos à frente para as séries sub-rogadas geradas a partir de uma série aleatória. Em (a) $mRSE(i)$ para um passo à frente. Em (b) $mRSE(i)$ para dois passos à frente, e assim sucessivamente.

obtido rearranjando a equação (4.10) como:

$$\begin{aligned} j &= \frac{p(n_R - 1) + 100}{100}, \\ \text{mRSE}_p &= \text{mRSE}(j). \end{aligned} \quad (4.11)$$

Logo, $p\%$ dos valores de mRSE estarão abaixo do valor mRSE_p e $(100 - p)\%$ dos valores de mRSE estarão acima de mRSE_p . Se a posição j for um número fracionário deve-se aproximá-lo pelo próximo valor inteiro. Por exemplo, se para $p = 5^{\text{o}}$ percentil corresponde a posição no vetor $j = 3,4$, então, $\text{mRSE}_{5^{\text{o}}} = \text{mRSE}(4)$.

Por exemplo: Considere 15 amostras de índices mRSE conforme mostra a Tabela 4.3. Pela equação (4.10), tem-se os seguintes percentis:

$$\begin{aligned} p &= 100 \times \frac{(1 - 1)}{15 - 1} = 0^{\text{o}}, \\ p &= 100 \times \frac{(2 - 1)}{15 - 1} = 7,17^{\text{o}}, \\ p &= 100 \times \frac{(3 - 1)}{15 - 1} = 14,29^{\text{o}}, \\ &\quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\ p &= 100 \times \frac{(14 - 1)}{15 - 1} = 92,86^{\text{o}}, \\ p &= 100 \times \frac{(15 - 1)}{15 - 1} = 100^{\text{o}}, \end{aligned} \quad (4.12)$$

Pode-se verificar pela Tabela 4.3 que o 20^o percentil refere-se a quarta observação, ou seja, $\text{mRSE}_{20^{\text{o}}} = \text{mRSE}(4) = 1,5$. Isto significa que 20% dos índices mRSE são menores do que 1,5 e que 80% dos índices mRSE são maiores do que 1,5.

Não se deve confundir percentil com porcentagem. Um percentil está relacionado somente com a posição relativa de uma observação quando comparada com outras observações. Ao longo do trabalho, o intervalo entre os percentis delimitarão os intervalos de confiança dos índices mRSE.

posição j	índice mRSE	índice mRSE ordenado	Percentil
1	1,50	1,45	0 ^o
2	1,45	1,50	7,14 ^o
3	1,80	1,50	14,29 ^o
4	1,58	1,55	21,43^o
5	1,90	1,58	28,57 ^o
6	1,50	1,58	35,71 ^o
7	1,65	1,60	42,86 ^o
8	1,55	1,63	50 ^o
9	1,58	1,65	57,14 ^o
10	1,63	1,65	64,29 ^o
11	1,65	1,65	71,43 ^o
12	1,69	1,69	78,57 ^o
13	1,65	1,70	85,71 ^o
14	1,70	1,80	92,86 ^o
15	1,60	1,90	100 ^o

A diferença entre se calcular o intervalo de confiança do índice mRSE das séries sub-rogadas considerando uma distribuição Normal e considerando uma distribuição qualquer está ilustrada na Figura 4.3. Na Figura 4.3(a) o intervalo de confiança, destacado como a parte clara do histograma, foi calculado considerando a distribuição dos índices mRSE como uma Normal. Para um $\alpha = 0,3173$, o nível de confiança do teste é $(1 - 0,3173) \times 100\% = 68,27\%$ e $z_c = 1$. Na Figura 4.3(b) o intervalo de confiança foi calculado como um intervalo entre o 84,12^o e o 15,89^o percentis ($84,12 - 15,89 = 68,23\%$).

Pode-se notar pelas Figuras 4.3(c) e (d) que o intervalo de confiança considerando uma distribuição Normal é mais largo do que o intervalo obtido via percentis considerando uma distribuição qualquer. Para um índice mRSE = 0,90, se o teste de hipóteses for baseado na distribuição Normal aceita-se H_0 considerando a série original como sendo aleatória. Se o teste de hipóteses for baseado em uma distribuição qualquer, rejeita-se H_0 e pode-se inferir uma assinatura determinística na série original.

Logo, para evitar conclusões errôneas a partir do teste de hipóteses, neste trabalho, optou-se em considerar a distribuição dos índices mRSE(i) como sendo

uma distribuição qualquer, obtendo assim, os intervalos de confiança por meio dos intervalos entre percentis.

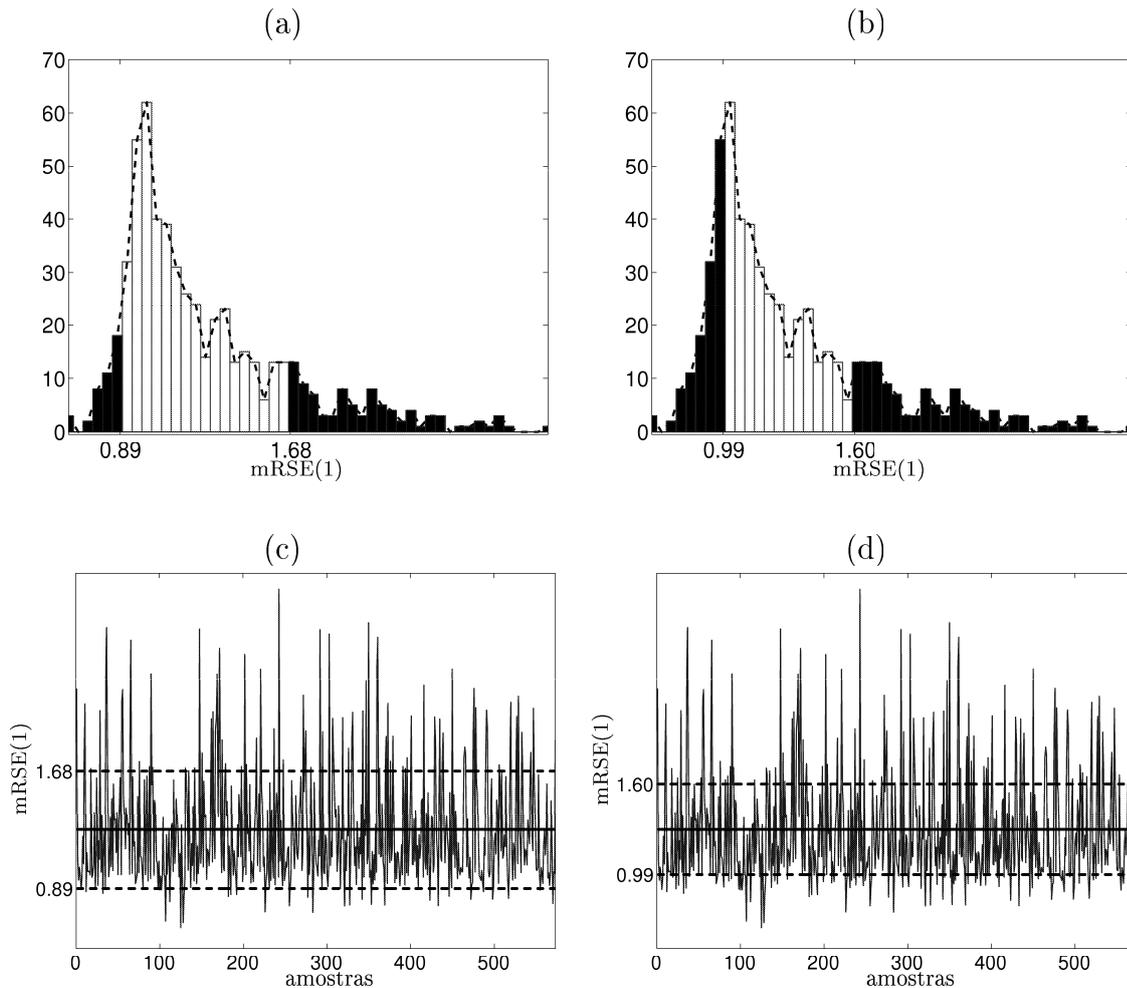


Figura 4.5: Histograma dos índices $mRSE(1)$ da previsão de 1 passo à frente das séries subrogadas, considerando a distribuição dos índices $mRSE(i)$ como: em (a) uma Normal com 68,27% de confiança representado como a parte clara do histograma. Em (b) como uma distribuição qualquer também com 68,27% de confiança representado como a parte clara do histograma. Em (c) o intervalo de confiança para o índice $mRSE(i)$ quando considerado com distribuição Normal e em (d) quando considerado com distribuição qualquer usando o intervalo entre percentis. Em (—) média dos índices $mRSE(i)$ e em (- -) valores críticos que delimitam a faixa de confiança.

4.4 Comentários Finais

Neste capítulo definiu-se que o fator discriminante do teste de hipóteses a ser usado é o índice $\text{mRSE}(i)$, o qual compara a qualidade de previsão da série residual com as das séries sub-rogadas. Mediante o índice $\text{mRSE}(i)$ estabeleceu-se as hipóteses nula H_0 e alternativas, H_1 e H_2 .

Verificou-se também, por meio de inspeção visual das distribuições dos índices mRSE , que essa distribuição é diferente de uma Normal. Conseqüentemente, optou-se por calcular o intervalo de confiança do índice mRSE como sendo o intervalo entre percentis.

ANÁLISE DE DETERMINISMO EM SÉRIES QUASE-PERIÓDICAS

Freqüentemente, deseja-se encontrar as equações que governam a dinâmica de um sistema do qual se possui um sinal medido. Para séries quase-periódicas é natural perguntar se há alguma dinâmica determinística, de curto ou longo prazo - intra-ciclo ou entre-ciclos¹ - além do comportamento quase-periódico. Porém, nem sempre isso é verificado (Weigend & Gershenfeld, 1994; Zhang et al., 1998). Este capítulo descreve um procedimento para investigar a existência de característica determinística intra-ciclo em uma série quase-periódica, além do seu padrão de comportamento cíclico. O teste de hipóteses baseado em dados sub-rogados é empregado na série quase-periódica e na série residual, a qual é obtida extraíndo/reduzindo o(s) ciclo(s) da série.

5.1 Tratamento dos dados

Na maioria das séries temporais reais, os dados medidos podem não estar completos e/ou apresentarem erros grosseiros de medição, tais como: mudanças abruptas (*outliers*), ausência de medições (*missings*) e taxas de amostragem diferentes. Para evitar uma análise e até uma modelagem incorreta dos dados, é necessário

¹Dinâmica de curto prazo (intra-ciclo) é a dinâmica existente no intervalo menor do que o período cíclico da série. Dinâmica de longo prazo (entre-ciclos) é a dinâmica existente no intervalo maior do que o período cíclico da série, ou seja, é a dinâmica que governa a relação entre os ciclos.

“limpar” os dados brutos (medidos). Dessa forma, quando detectadas tais falhas, estas podem ser contornadas por meio de interpolação polinomial.

Neste trabalho, assume-se que se tem uma série temporal quase-periódica, $y(k)$, com N observações e período cíclico d . O procedimento para reconstruir a série temporal utiliza aproximadamente 10 observações adjacentes à falha ($n_a = 10$), ou seja, tanto as observações anteriores quanto as posteriores à falha são usadas na interpolação, da seguinte forma:

- primeiramente, são detectadas as observações em que há falhas de medição. Neste trabalho, foram consideradas falhas as observações que variam cerca de 50% do desvio-padrão da série $y(k)$, ou seja, as observações que não obedecem a seguinte relação:

$$\mu_y - 1,5\sigma_y < y(k) < \mu_y + 1,5\sigma_y, \quad (5.1)$$

em que $k_f = k$ é o índice da observação em que há falha de medição, σ_y e μ_y são o desvio-padrão e a média de $y(k)$, respectivamente. Os valores de σ_y e μ_y podem ser calculados em faixas de dados, uma vez que a série $y(k)$ pode ser não-estacionária;

- posteriormente, substituiu-se as observações com falhas pela média dos valores adjacentes às falhas:

$$m_y = \frac{1}{2n_a} \left[\sum_{i_1=1}^{n_a} y(k_f - i_1) + \sum_{i_2=1}^{n_a} y(k_f + i_2) \right],$$

$$\hat{y}(k_f) = m_y;$$

em que $n_a = 10$ é o número observações adjacentes à falha;

- ao vetor $\hat{y}(k_2)$, para $k_2 = k_f - n_a \dots k_f \dots k_f + n_a$ ajusta-se um polinômio $\hat{p}(k_2)$ de ordem j , tal que:

$$\hat{p}(k_2) = a_0 + a_1 k_2 + a_2 k_2^2 + \dots + a_j k_2^j.$$

- a série temporal reconstruída é então dada por:

$$\hat{y}(k_f) = \hat{p}(k_f),$$

para k_f na janela de dados que há falha de medição.

A escolha da taxa de amostragem também é importante, pois taxas de amostragem diferentes resultam em modelos diferentes. Pode-se classificar os dados com relação à taxa de amostragem em três categorias: subamostrados, bem amostrados e superamostrados. Dados subamostrados tipicamente não contêm informação dinâmica suficiente, impossibilitando assim, a determinação de um modelo para o processo. Um sinal superamostrado, em que sucessivas amostras tendem a estar fortemente correlacionadas, pode proporcionar mau condicionamento numérico e dificultar a determinação da estrutura do modelo (Billings & Aguirre, 1995), porém, essa dificuldade pode ser contornada decimando os sinais, possibilitando assim, a utilização deles na estimação e validação do modelo.

Deve-se destacar que, em algumas séries reais usadas neste trabalho, o tempo de amostragem já é definido. Além disso, pode haver trechos ao longo da série cujo tempo de amostragem é diferente do restante. Nesse caso, pode-se utilizar também a interpolação polinomial (se o trecho tiver maior tempo de amostragem) ou a decimação (se o trecho tiver menor tempo de amostragem).

Uma vez reconstruída a série temporal, é interessante extrair o máximo de informação da mesma antes de aplicar qualquer ferramenta mais sofisticada de análise de séries temporais.

Em alguns casos, somente uma inspeção visual da série temporal possibilita a detecção da presença da componente quase-periódica. Porém, em outros casos, as componentes cíclicas da série temporal podem ser detectadas somente por meio da análise da função de autocorrelação e do espectro de potência da série.

Portanto, na análise das séries temporais quase-periódicas, aqui investigadas,

emprega-se a função de autocorrelação linear (FAC) e o espectro de potência da série, para a detecção de quase-periodicidades e tendência. Além disso, é possível estimar, por meio destas ferramentas, os períodos cíclicos (d).

Geralmente, séries quase-periódicas apresentam frequências dominantes, as quais revelam-se na forma de picos bem definidos no espectrograma gerado pela FT (transformada de Fourier) da série.

Sabe-se também que por meio da análise dos coeficientes da FAC da série temporal pode-se ter uma noção da medida da dependência temporal entre as observações. Quando o gráfico da FAC apresenta um comportamento periódico, com altos valores de correlação em espaços d entre as observações, isso indica a presença de comportamento periódico ou quase-periódico com período d . Além disso, a FAC pode revelar se a série temporal é estacionária.

5.1.1 Teste de estacionariedade

Como mencionado na Seção 3.6.1, os algoritmos para gerar dados sub-rogados propostos por Theiler e colegas (1992) têm como principal limitação o fato de terem sido concebidos para séries estacionárias.

O conceito de estacionariedade está, de certa forma, relacionado ao conceito de invariância temporal. Ser invariante no tempo significa que a natureza do processo não muda ao longo do tempo (Haykin & Van Veen, 1999). Dessa forma, um único modelo (quando houver) será capaz de representar de forma adequada o processo gerador da série temporal em qualquer instante de tempo.

Na análise de séries temporais lineares, a estacionariedade é geralmente testada a partir da análise de parâmetros estatísticos (momentos) de primeira e de segunda ordem, tais como a média, a variância e a densidade de potência espectral (estacionariedade fraca) (Priestley, 1988). Deve-se destacar que existem outras técnicas para detectar estacionariedade em séries temporais, para maiores detalhes ver Duarte (2006).

Uma outra maneira de se avaliar se uma série temporal é ou não estacionária

é por meio da função de autocorrelação. A FAC de uma série temporal estacionária não depende da janela de dados analisada, desprezando-se os efeitos das bordas. Dessa forma, pela FAC de janelas de dados distintas, i e j , da série, pode-se inferir que a série temporal é estacionária se seus respectivos coeficientes de autocorrelação, r_i e r_j , obedecerem a seguinte condição:

$$r_i(k) = r_j(k), \quad \forall k. \quad (5.2)$$

Métodos para tratamento de séries temporais não-estacionários ainda são escassos e não existe um método considerado padrão. Processos reais podem apresentar estacionariedade localizada no tempo, sob a forma de janelas. Logo, uma mesma série temporal, globalmente não-estacionária, pode conter várias janelas de dados estacionárias com dinâmica similar dispostas de forma não-adjacente no tempo. Então, é necessário organizar essas janelas similares em grupos, a partir das quais se pode fazer uma análise utilizando os métodos adequados.

Uma das maneiras de se verificar qual o tamanho da janela de dados da série temporal a qual pode ser considerada estacionária, baseia-se no teste usado por Guimarães (1996) em séries de variabilidade de frequência cardíaca. Tal teste consiste na obtenção de modelos paramétricos para faixas de dados da série temporal. Nesse caso, cada janela de dados é dividida em dois segmentos para os quais modelos polinomiais AR de mesma ordem dinâmica são estimados. Se os pólos referentes ao modelo do segundo segmento localizam-se em regiões de confiança dos pólos do primeiro segmento, pode-se considerar esta janela de dados como estacionária, uma vez que há um modelo AR, com características dinâmicas similares, que pode representar a série em segmentos distintos. Deve-se notar que esse teste é incapaz de distinguir não-estacionariedade de não-linearidade (Duarte, 2006).

Outra forma de se verificar o tamanho e quais janelas de dados podem ser consideradas estacionárias é por meio da FAC. Para isto, determina-se uma janela de dados com amostra inicial k_1 e final k_j . Para esta janela de dados, $y(k_1, \dots, k_j)$, calcula-se a FAC, $r(k_1, \dots, k_j)$. Aumenta-se o tamanho da janela mantendo a amostra inicial,

ou seja, $y(k_1, \dots, k_{j+1})$ e calcula-se novamente a FAC, $r(k_1, \dots, k_{j+1})$. Isto é feito sucessivamente, até atingir um valor $j + c$, para o qual a FAC ainda se mantém “igual” à FAC da janela de dados inicial.

5.2 Teste de determinismo proposto

O teste de determinismo proposto neste trabalho testa hipóteses baseado em dados sub-rogados para detectar determinismo intra-ciclo em séries temporais quase-periódicas.

A partir da série residual são geradas seqüências artificiais, denominadas dados sub-rogados (*surrogates*), que preservam algumas características estatísticas da série residual e destroem outras. Além disso, tais séries artificiais devem ser consistentes com a hipótese nula (H_0) a ser testada. Neste trabalho, a H_0 que se deseja testar é que a série quase-periódica tem comportamento aleatório, além do comportamento determinístico quase-periódico. Portanto, H_0 deverá ser rejeitada se os dados que estão sendo testados possuírem alguma característica determinística, além do comportamento quase-periódico e das flutuações aleatórias.

A fim de se avaliar o aspecto determinístico dos dados, deve-se utilizar um fator estatístico discriminante endereçado a esta característica. O índice empregado neste trabalho é o mRSE dado pela equação (4.2), conforme descrito na Seção 4.1. Esse índice discriminante é obtido após a modelagem não-linear e predição de 1 até h passos à frente da série residual e das séries artificiais.

Dessa forma, obtêm-se modelos NARMA polinomiais para a série residual, que está sob investigação, e para as séries artificiais. A partir desses modelos são obtidas as predições de 1 até h passos à frente e calculados os índices mRSE (4.2) para cada passo de previsão de ambas as séries. Os valores dos índices mRSE são colocados em um gráfico à semelhança de Gomes (2001). Os algoritmos, aqui usados, para gerar as séries sub-rogadas foram descritos na Seção 3.6.

Ressalta-se que há outros trabalhos disponíveis na literatura cujo fator discriminante é diferente do aqui utilizado, como por exemplo, o coeficiente de correlação

utilizado por Small e colaboradores em (2001) e a FAC e a informação mútua usada por Nakamura e Small em (2005). Para detalhes, ver Seção 3.4.3.

5.2.1 Algoritmo proposto para a detecção de determinismo

O algoritmo utilizado para testar determinismo intra-ciclo em séries temporais quase-periódicas, por meio de dados sub-rogados, é semelhante ao proposto por Gomes (2001), e pode ser descrito pelos seguintes passos principais:

- 1^o - estimar e extrair/reduzir a componente quase-periódica da série temporal, obtendo assim, a série residual;
- 2^o - gerar as séries sub-rogadas a partir da série residual;
- 3^o - estimar modelos NARMA polinomiais para a série residual e para as séries sub-rogadas;
- 4^o - realizar a previsão livre de 1 até h passos à frente e calcular os índices mRSE para cada horizonte;
- 5^o - testar a hipótese nula.

Tendo em mente que a componente quase-periódica tende a mascarar qualquer comportamento subjacente a ela na série temporal, o primeiro passo no teste de determinismo é estimar e extrair essa componente quase-periódica ($\hat{c}(k)$) e de tendência ($\hat{m}(k)$) da série temporal quase-periódica $y(k)$, obtendo assim a série residual $\hat{x}(k) = y(k) - \hat{c}(k) - \hat{m}(k)$. Conhecendo *a priori* os ciclos que compõem a série, a extração da componente quase-periódica é realizada por meio do método FMMC, o qual foi descrito na Seção 3.3.4. É importante destacar que *todos os passos e análises subsequentes do teste de determinismo são feitos assumindo que a componente quase-periódica foi corretamente extraída*. Porém, nem sempre é possível extraí-la completamente, havendo casos em que a componente quase-periódica é somente reduzida.

Diante da série residual, $\hat{x}(k)$, gera-se um conjunto de n_{surr} séries sub-rogadas, $(s_i(k))$, usando os algoritmos baseados nos métodos propostos por Theiler e colaboradores em (1992). Deve-se atentar que, um outro motivo para a extração da componente quase-periódica é que os algoritmos propostos em (Theiler et al., 1992) não foram desenvolvidos para série com comportamento quase-periódico.²

Posteriormente, divide-se as séries, residual e artificiais, em conjuntos para a identificação e validação dos modelos NARMA polinomiais e escolhe-se empiricamente os valores máximos de n_y , n_e , ℓ e o número máximo de termos, n_t , que comporão os modelos. Apesar de ser uma escolha empírica, deve-se levar em consideração as informações *a priori* da série original. Se é conhecido que a componente quase-periódica devida ao período cíclico d_j não foi completamente extraída, deve-se escolher o máximo atraso $n_y < d_j$. Dessa forma, é possível garantir que o padrão determinístico quase-periódico com período d_j não comprometa o resultado do teste de determinismo.

Uma vez divididas as janelas e definido os limites máximos das estruturas, obtém-se uma família de modelos NARMA, conforme foi descrito na Seção 2.1.2. Os termos de cada modelo são escolhidos entre todos os termos candidatos, de acordo com o critério ERR (Seção 2.2). O modelo NARMA polinomial escolhido para a série residual é o melhor modelo dentre todos os possíveis segundo o menor mRSE (4.2) para o primeiro passo-à-frente calculado em aproximadamente 10 janelas diferentes dos dados de validação.

Por meio dos modelos NARMA polinomiais identificados, faz-se a previsão livre de $1-h$ passos à frente em, no máximo, 10 janelas de dados diferentes. Calculam-se os índices mRSE, referentes a cada passo de previsão ao longo das diferentes janelas. Os índices mRSE de cada passo de previsão para a série residual são então computados.

²No entanto, os algoritmos PPS e CSS propostos por Small et al. (2001) e por (Nakamura & Small, 2005), respectivamente, foram desenvolvidos para séries com padrões quase-periódicos. Sendo que o PPS serve para testar determinismo entre-ciclos e o CSS para testar determinismo intra-ciclo.

O mesmo procedimento é feito para as séries sub-rogadas. Levando em consideração que somente o índice mRSE das séries sub-rogadas com modelos estáveis devem ser usados, a partir dos quais calcula-se o intervalo de confiança obtido pelas previsões das séries artificiais por meio do cálculo do intervalo entre percentis, conforme descrito na Seção 4.3. Ressaltando que o nível de confiança é estabelecido gradualmente aumentando-se o nível de confiança do teste de determinismo começando com um nível inicial de 60% até um nível de confiança para o qual o resultado do teste de determinismo (aceitação ou rejeição da hipótese nula) não se altere. Em seguida, coloca-se em um gráfico os índices mRSE assim calculados e, compara-se o desempenho qualitativo obtidos pelas previsões das séries residual e das séries sub-rogadas. Dessa forma, testa-se a hipótese nula tendo como base tais gráficos.

A definição da hipótese nula (H_0) e das hipóteses alternativas (H_1 e H_2) foi feita baseada no índice mRSE, conforme mostrado na Seção 4.1, sendo portanto, atribuído à:

- H_0 : série residual é procedente de um processo aleatório e é aceita quando $mRSE \approx 1$.
- H_1 : série residual tem um vínculo determinístico e é aceita quando $mRSE < 1$.
- H_2 : série residual é procedente de um processo não-estacionário ou não modelável por representações NARMA polinomiais e é aceita quando $mRSE > 1$.

Logo, a hipótese nula H_0 , de que a série residual ($\hat{x}(k)$) provém de um sistema gerador de séries aleatórias, será rejeitada se o desempenho do modelo para a série residual com relação à predição for estatisticamente diferente e melhor do que o desempenho médio obtido pelas séries sub-rogadas, ou seja, $mRSE < 1$. Nesse caso, aceita-se H_1 e pode-se considerar que a série residual ($\hat{x}(k)$) apresenta um traço determinístico de curto-prazo e, conseqüentemente, que a série temporal quase-periódica, $y(k)$, apresenta assinatura determinística intra-ciclo além da quase-periodicidade.

Quando o desempenho das previsões para a série residual for estatisticamente igual ao desempenho médio das séries artificiais, ou seja, os índices mRSE para

a série residual estiverem dentro da faixa de confiança, não se pode rejeitar H_0 . Logo, conjectura-se que a série $\hat{x}(k)$ é aleatória, conseqüentemente, infere-se que a série quase-periódica $y(k)$ não tem traço determinístico intra-ciclo. Por outro lado, quando o desempenho das previsões para a série residual for pior do que o das séries artificiais, com $\text{RMSE} > 1$, rejeita-se H_0 e aceita-se H_2 , podendo-se inferir que a série quase-periódica $y(k)$ apresenta um comportamento não-estacionário ou que a mesma não pode ser representada por modelos NARMA polinomiais.

Quando se rejeita H_0 em favor de H_2 , é aconselhável refazer o teste de determinismo mudando-se os valores da estrutura máxima do modelo, para assegurar que o resultado do teste não seja mascarado pela escolha inadequada da estrutura do modelo NARMA polinomial.

5.3 Algoritmos para gerar dados sub-rogados

Como descrito na Seção 3.6, há vários algoritmos para a geração de dados sub-rogados. Porém, deve-se verificar quais são os mais apropriados para testar uma dada hipótese. Neste trabalho, a hipótese nula está relacionada à característica determinística intra-ciclo da série temporal quase-periódica, a qual é testada tomando como base a sua previsibilidade.

Para verificar quais algoritmos apresentados na Seção 3.6 são os mais indicados para a análise de determinismo, serão apresentados a seguir os resultados das análises de determinismo intra-ciclo em duas séries quase-periódicas simuladas.

Inicialmente, considere as duas séries quase-periódicas simuladas mostradas no Capítulo 1 na Figura 1.1. A série $y_1(k)$ apresenta um vínculo determinístico de curto prazo além do padrão quase-periódico. A série $y_2(k)$ tem um comportamento aleatório subjacente ao padrão quase-periódico.

A componente quase-periódica é gerada usando os períodos cíclicos, $d_1 = 13$ e $d_2 = 10\pi \approx 314,159 \dots$, cuja a relação $f_1/f_2 = 24,166097 \dots$, tal que:

$$c(k) = 5\text{sen}(\omega_1 k) - 1\text{sen}(\omega_2 k). \quad (5.3)$$

A componente dinâmica com assinatura determinística de curto-prazo é obtida simulando o modelo do tipo ARX(3):

$$x_1(k) = 0,9x(k-1) + 0,02x(k-2) + 0,03x(k-3) + 0,2u(k); \quad (5.4)$$

em que $u(k)$ é um ruído branco com variância 0,04.

Somando-se à série dinâmica determinística, $x_1(k)$, a componente quase-periódica, $c(k)$ dada em (5.3), obtém-se a série quase-periódica com vínculo determinístico de curto-prazo, $y_1(k)$, da seguinte forma:

$$y_1(k) = c(k) + x_1(k). \quad (5.5)$$

Para gerar a série quase-periódica com comportamento aleatório subjacente, foi adicionado à componente quase-periódica, $c(k)$ dada em (5.3), um ruído branco com variância 1, $x_2(k)$, tal que:

$$y_2(k) = c(k) + x_2(k). \quad (5.6)$$

Primeiramente, o teste de determinismo usado por Nakamura e Small (2005) é aplicado à série quase-periódica com determinismo de curto-prazo $y_1(k)$ em (5.5) mostrada na Figura 5.1 (a). Destaca-se que o algoritmo SSS é aplicado diretamente à série quase-periódica. Gera-se um conjunto de 400 séries sub-rogadas a partir da série quase-periódica $y_1(k)$ dada em (5.5) usando o algoritmo SSS, conforme os passos descritos na Seção 3.6.2. Uma faixa de dados de uma das 200 séries sub-rogadas está na Figura 5.1 (b). Pode-se perceber que a série sub-rogada é bem similar a série quase-periódica $y_1(k)$ dada em (5.5).

No trabalho de Nakamura e Small (2005), os autores empregam como índices discriminantes: a função de autocorrelação (para detectar determinismo linear) e a informação mútua (para detectar determinismo não-linear).

No presente trabalho, utiliza-se somente a FAC da série quase-periódica com

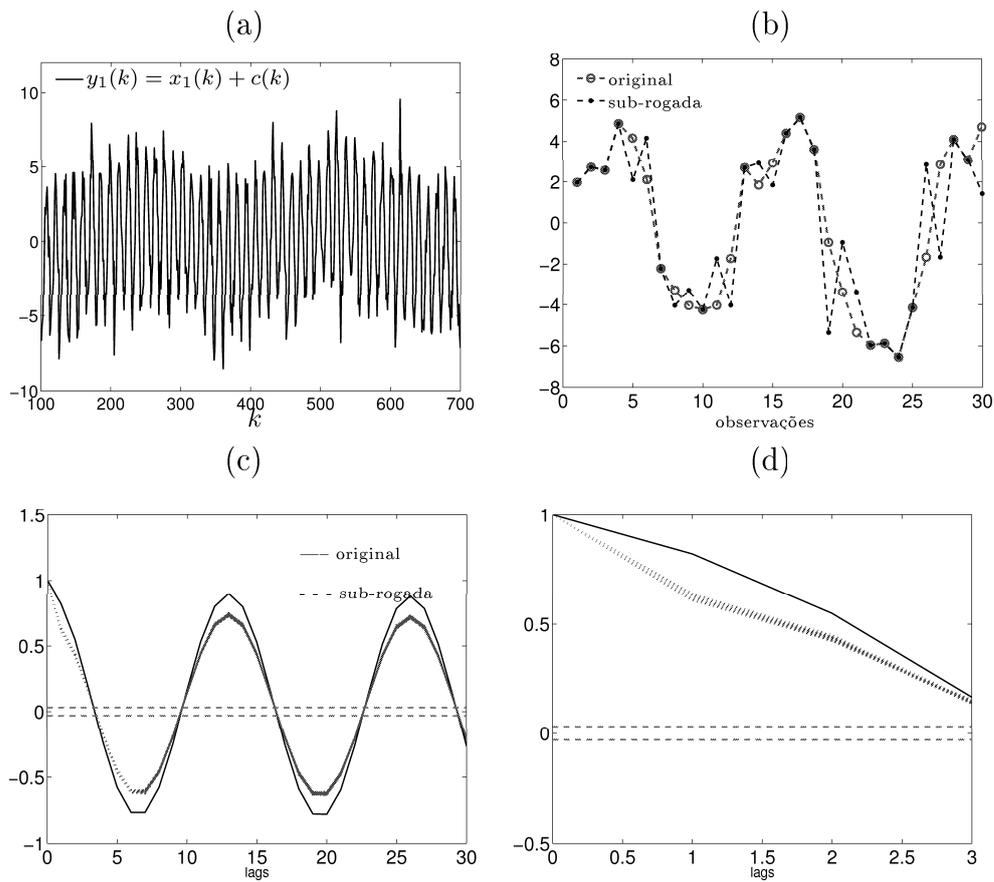


Figura 5.1: Em (a) série quase-periódica com vínculo determinístico de curto-prazo, $y_1(k)$ dada em (5.5). Em (b) faixa da série quase-periódica $y_1(k)$ em (-o-) e, faixa de uma das 200 séries sub-rogadas (-.-) geradas segundo o método SSS (Nakamura & Small, 2005). Em (c) FAC da série $y_1(k)$ em (—) e das séries sub-rogadas em (...). Em (d) destaque da FAC mostrada em (c).

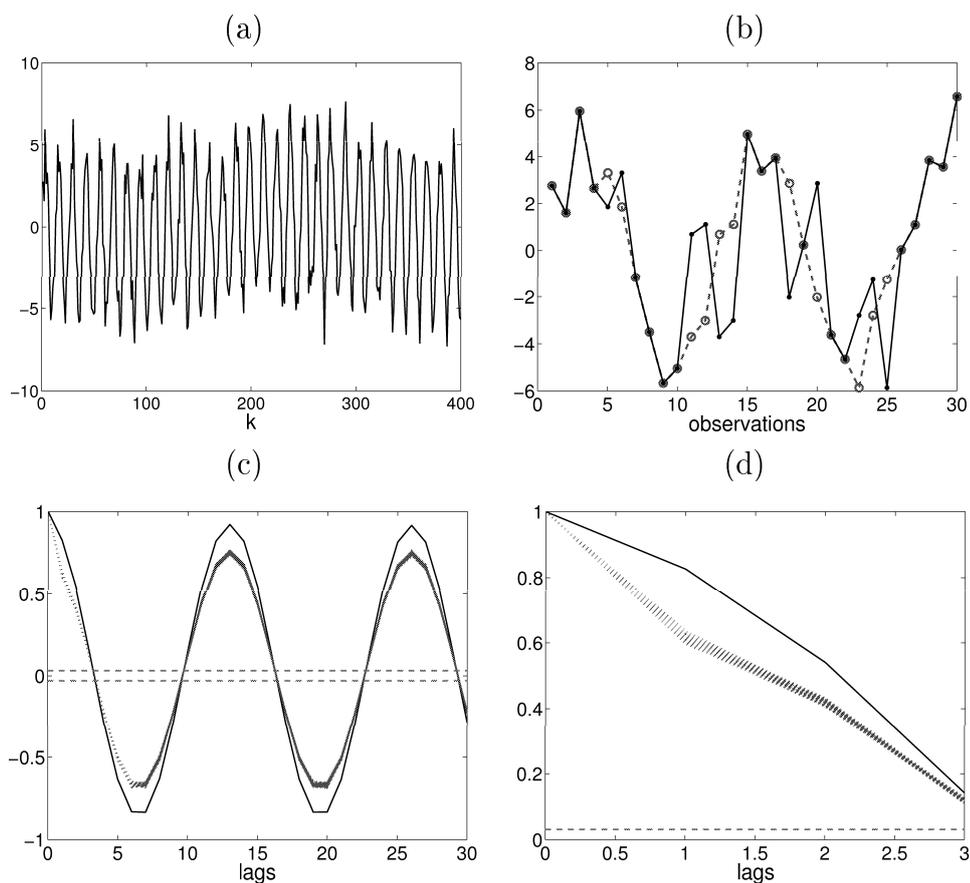


Figura 5.2: Em (a) série quase-periódica com comportamento aleatório subjacente, $y_2(k)$ dada em (5.6). Em (b) faixa da série quase-periódica $y_2(k)$ em (-o-) e, faixa de uma das 200 séries sub-rogadas em (-.-) geradas segundo o método SSS (Nakamura & Small, 2005). Em (c) FAC da série original $y_2(k)$ em (—) e das séries sub-rogadas em (...). Em (d) destaque da FAC mostrada em (c).

vínculo determinístico, $y_1(k)$ dada em (5.5), e das 200 séries sub-rogadas, a qual é mostrada na Figura 5.1(c), com destaque na Figura 5.1 (d). A análise da FAC mostra o resultado esperado, ou seja, a FAC da série quase-periódica com assinatura determinística subjacente, $y_1(k)$, é diferente da FAC das séries sub-rogadas. Logo, pode-se inferir determinismo de curto-prazo na série $y_1(k)$.

Porém, quando a série quase-periódica tem um comportamento aleatório subjacente aos ciclos, como $y_2(k)$ dada em (5.6), o resultado do teste de determinismo proposto por Nakamura & Small (2005) pode não ser o adequado, como mostra a Figura 5.2(c). Aqui, também foi gerado um conjunto de 200 séries sub-rogadas, sendo que uma delas pode ser comparada à série original $y_2(k)$ na Figura 5.2(b).

Nota-se pela Figura 5.2(c) que a FAC da série quase-periódica, $y_2(k)$, é diferente da FAC das séries sub-rogadas. Segundo os autores, essa diferença é suficiente para inferir que a série $y_2(k)$ tem determinismo de curto-prazo. O que não confere com o que se sabe *a priori* de $y_2(k)$ dada em (5.6).

Portanto, neste trabalho, optou-se pelo uso de um dos algoritmos propostos por Theiler e colaboradores em (1992). Para avaliar quais são os mais adequados para a análise de determinismo intra-ciclo, realizou-se o teste de determinismo nas duas séries simuladas quase-periódicas: $y_1(k)$ (com vínculo determinístico de curto-prazo dada em (5.5)) e, $y_2(k)$ (com comportamento aleatório subjacente aos ciclos dada em (5.6)).

A princípio, deve-se retirar a componente quase-periódica das séries visto que tais algoritmos foram propostos para serem empregados em séries não-periódicas (Theiler et al., 1992). Para isto é necessário aplicar uma técnica para a redução e, idealmente, a extração, da(s) componente(s) cíclica(s).

Uma das técnicas mais usadas na área econômica para extrair as componentes sazonal e de tendência é conhecida como X-11-ARIMA (Shiskin et al., 1967) e, sua versão adaptada o X-12-ARIMA (Findley et al., 1998). Esse método foi desenvolvido tendo em vista o ajustamento sazonal em séries econômicas que, freqüentemente, apresentam sazonalidades mensais, trimestrais e semestrais, com respectivos períodos

sazonais $d = 12, 4, 2$. No entanto, em aplicações empíricas constata-se que o método X-11-ARIMA não é aplicável em séries físicas, biológicas, as quais podem apresentar períodos cíclicos que sejam maiores do que 12.

Outro método bastante difundido para a redução de ciclos, sazonais ou não, é o filtro de média móvel para componentes cíclicas (FMMC), o qual é empregado neste trabalho e está descrito na Seção 3.3.4.

5.3.1 FMMC para redução da componente quase-periódica

Um passo importante no teste de determinismo, aqui proposto, é com relação a extração da componente quase-periódica. Para extrair a quase-periodicidade das séries optou-se pelo uso do método I descrito na Seção 3.3.4, o qual emprega o filtro de média móvel para componentes cíclicas (FMMC) para estimar a componente quase-periódica. Deve-se destacar que, em algumas áreas, séries quase-periódicas são aquelas que apresentam mais que um período cíclico e a relação entre as frequências correspondentes aos períodos cíclicos é um número irracional. Neste trabalho, porém, quase-periódico será usado para designar sinais cíclicos que não são matematicamente periódicos.

Para se extrair duas componentes cíclicas por meio do FMMC, primeiro deve-se retirar a componente correspondente a um dos ciclos, obtendo assim, uma série residual. Posteriormente, deve-se aplicar o FMMC nessa série residual e extrair a componente referente ao outro ciclo que restou. Esse procedimento pode ser aplicado quando a série quase-periódica apresentar mais de dois ciclos (duas frequências). Quando um dos períodos dos ciclos não for um valor inteiro este deve ser aproximado pelo valor inteiro mais próximo. Salienta-se que o método é bastante sensível a escolha do período do ciclo.

Quando não se conhece os períodos dos ciclos *a priori* pode-se detectá-los por meio da função de autocorrelação e do periodograma da série temporal quase-periódica, sendo este último mais recomendado.

Como mencionado no Capítulo 3, neste trabalho, lida-se com as componentes

da série temporal $y(k)$ como se essas estivessem dispostas em um modelo de decomposição clássico aditivo:

$$y(k) = s(k) + c(k),$$

em que a $y(k)$ é a série temporal quase-periódica, $c(k)$ é a componente cíclica com períodos d_j , para $j = 1, 2, \dots$

O filtro de média móvel para componentes cíclicas (FMMC) deve ser capaz de estimar a componente quase-periódica $\hat{c}(k)$ da melhor forma possível. Dessa forma, é possível estimar a componente residual subtraindo a componente cíclica estimada, $\hat{c}(k)$, da série quase-periódica, da seguinte forma:

$$\hat{x}(k) = y(k) - \hat{c}(k).$$

Com a componente residual $\hat{x}(k)$ sem o comportamento quase-periódico, aplica-se os métodos convencionais para gerar os dados sub-rogados propostos por Theiler e colaboradores (1992) e assim, testa-se determinismo de curto-prazo na série residual. Caso haja determinismo de curto-prazo na série residual, conjectura-se que a série temporal quase-periódica $y(k)$ apresenta vínculo determinístico intra-ciclo.

Retornando as séries anteriores, tem-se que ao aplicar o método FMMC em $y_1(k)$ dada por (5.5), com $d_1 = 13$ e $d_2 = 314$, esse deve ser capaz de estimar $\hat{c}(k)$ adequadamente, de modo que, a série residual assim estimada:

$$\hat{x}_1(k) = y_1(k) - \hat{c}(k) \tag{5.7}$$

seja o mais próximo possível de $x_1(k)$ dada em (5.4), ou seja, com assinatura determinística subjacente ao comportamento quase-periódico. As extrações das componentes cíclicas são mostradas nas Figuras 5.3(a) e (b), as quais mostram, respectivamente, a série residual estimada $\hat{x}_1(k)$ e uma comparação entre a FAC da séries residual e da série quase-periódica $y_1(k)$. Uma comparação entre a FAC de $y_1(k)$ e $\hat{x}_1(k)$ na Figura 5.3 (b) destaca que a componente quase-periódica com períodos $d_1 = 13$ e $d_2 = 314,1593\dots$ foi significativamente reduzida em $\hat{x}_1(k)$. Também é

possível perceber a adequabilidade do filtro por meio dos índices de erro médio, que é igual a 0,29 e, o de correlação entre $\hat{x}_1(k)$ e $x_1(k)$ em (5.4) igual a 0,96.

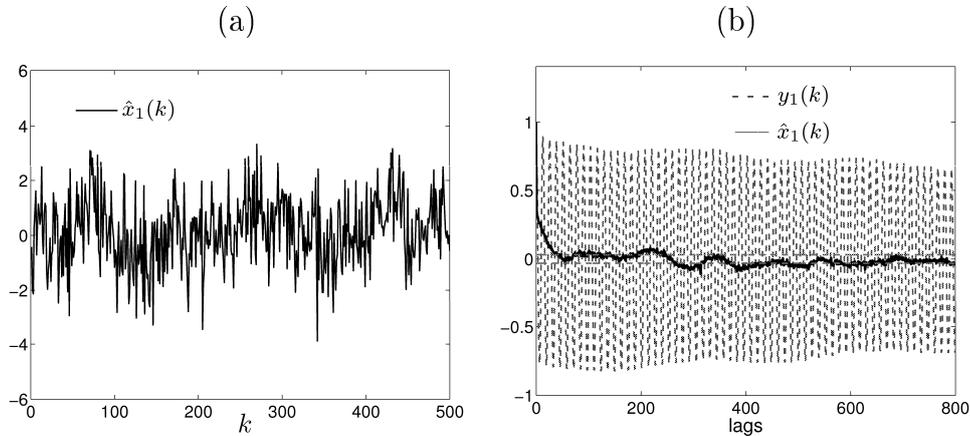


Figura 5.3: Em (a) a série residual com assinatura determinística $\hat{x}_1(k)$. Em (b) uma comparação da FAC de $y_1(k)$ em (- - -) e de $\hat{x}_1(k)$ em (—). O erro médio entre $x_1(k)$ em (5.4) e $\hat{x}_1(k)$ é igual a 0,29.

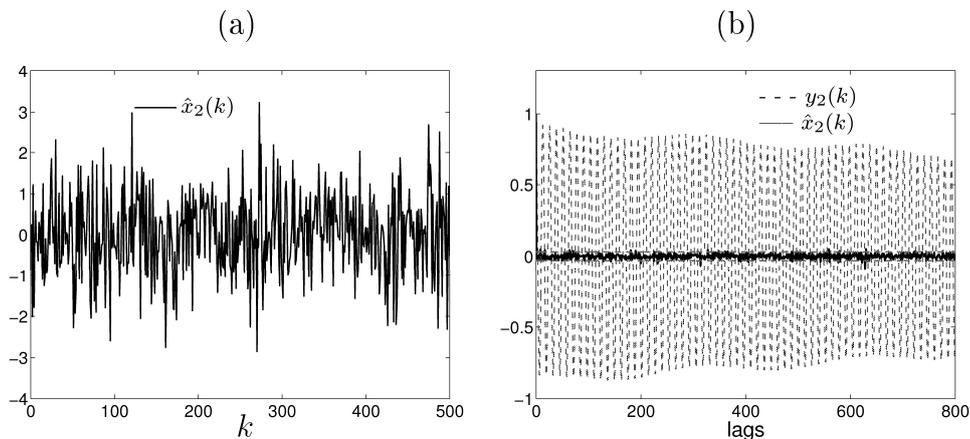


Figura 5.4: Em (a) a série residual com comportamento aleatório $\hat{x}_2(k)$. Em (b) FAC de $y_2(k)$ em (- - -) e de $\hat{x}_2(k)$ em (—). O erro médio entre $x_2(k)$ e $\hat{x}_2(k)$ é igual a 0,23.

Para a série quase-periódica com comportamento aleatório subjacente aos ciclos, as componentes cíclicas são estimadas para $d_1 = 13$ e $d_2 = 314$, aplicando o FMMC e, obtendo assim, a série residual:

$$\hat{x}_2(k) = y_2(k) - \hat{c}(k). \quad (5.8)$$

A série residual aleatória estimada está representada na Figura 5.4(a). Uma comparação entre a FAC da série $y_2(k)$ e da série residual estimada $\hat{x}_2(k)$ mostra que houve uma redução significativa da componente quase-periódica. Sendo o erro médio igual a 0,23 e, a correlação entre $x_2(k)$ e $\hat{x}_2(k)$ igual a 0,96.

Apesar desses resultados, o método FMMC parece não extrair completamente as componentes cíclicas de algumas das séries quase-periódicas reais utilizadas neste trabalho. Esses resultados poderão ser vistos, posteriormente, no Capítulo 6. Resultados similares são encontrados quando as séries cíclicas são periódicas.

Vale ressaltar que, além do método baseado em filtros de médias móveis, também foi testado o método de agrupamento de padrões usando o k -médias para reduzir a quase-periodicidade das séries, no entanto, os resultados obtidos por esse último foram similares ao obtido pelo FMMC. Uma vez obtidas as séries residuais, $\hat{x}_1(k)$ e $\hat{x}_2(k)$, aplica-se o teste de determinismo em tais séries.

A princípio, gera-se um conjunto de 100 séries sub-rogadas para cada um dos algoritmos descritos na seção 3.6: AAFT, AAWFT, BOOT, CAAFT, FT, GAUSS, IID e WFT, por meio da série residual $\hat{x}_1(k)$ em (5.5). Foram usados modelos NARMA polinomiais para a previsão das séries residual e sub-rogadas com os seguintes valores máximos de estrutura: $\ell = 2$, $n_y = 10$, $n_e = 5$ e $n_t = 10$. Os resultados dos testes de determinismo usando os algoritmos AAFT, AAWFT, BOOT, CAAFT, FT, GAUSS, IID e WFT, para gerar os dados sub-rogados a partir da série residual, $\hat{x}_1(k)$, estão ilustrados nas Figuras 5.5 e 5.6.

A Figura 5.5 ilustra, em cada gráfico superior, a série residual a qual se testa determinismo de curto-prazo; no intermediário, uma das séries sub-rogadas, e, em cada gráfico inferior o desempenho das previsões da série residual e o desempenho médio das séries artificiais, ou seja, seus respectivos índices mRSE.

A hipótese nula (H_0) de que a série residual provém de um processo aleatório, será rejeitada quando o desempenho da previsão da série residual $\hat{x}_1(k)$ for significativamente diferente ao desempenho médio das séries artificiais. Como ocorreu no caso em que as séries sub-rogadas foram geradas pelos algoritmos BOOT, GAUSS e

IID ilustrados nas Figuras 5.5 e 5.6. Por outro lado, no teste de determinismo usando os algoritmos AAWFT, CAAFT, FT, e WFT para gerar as séries sub-rogadas, as previsões da série residual $\hat{x}_1(k)$ tiveram um desempenho similar ao desempenho médio das previsões das séries artificiais. Dessa forma, não se pode rejeitar a H_0 . Contestando o que se sabe sobre $\hat{x}_1(k)$, a qual é possui assinatura determinística por construção (ver equação (5.4)).

Para o teste de determinismo na série com comportamento aleatório $\hat{x}_2(k)$, Figura 5.4(a), gerou-se um conjunto de 100 dados artificiais usando cada um dos algoritmos AAFT, AAWFT, BOOT, CAAFT, FT, GAUSS, IID e WFT. Os resultados do teste de determinismo estão ilustrados nas Figuras 5.7 e 5.8. Todos os resultados apontam para a não rejeição da H_0 , visto que o desempenho das previsões para a série residual $\hat{x}_2(k)$ são estatisticamente iguais ao desempenho médio obtidos para as séries artificiais, o que condiz com o que se conhece de $x_2(k)$, a qual é aleatória por construção.

Conjectura-se que os algoritmos baseados na transformada de Fourier da série residual (FT, AAFT) são mais indicados para testar linearidade em séries temporais, uma vez que as séries artificiais geradas por estes algoritmos tendem a manter a correlação linear e destruir a dinâmica não-linear da série.

Por meio dos testes de determinismo realizados nas duas séries simuladas, com vínculo determinístico e com comportamento aleatório subjacentes aos ciclos, $y_1(k)$ e $y_2(k)$ respectivamente, observou-se que os algoritmos mais adequados para investigar determinismo em séries temporais são os baseados no embaralhamento temporal da série (*Shuffled*). Uma vez que, na presença de determinismo os demais algoritmos não conseguem distinguir significativamente tal característica, como pode ser observado nos resultados apresentados nas Figuras 5.5 e 5.6. Portanto, neste trabalho, será usado o algoritmo IID para gerar, individualmente, as seqüências de dados sub-rogados a partir das séries residuais obtidas via filtro FMMC.

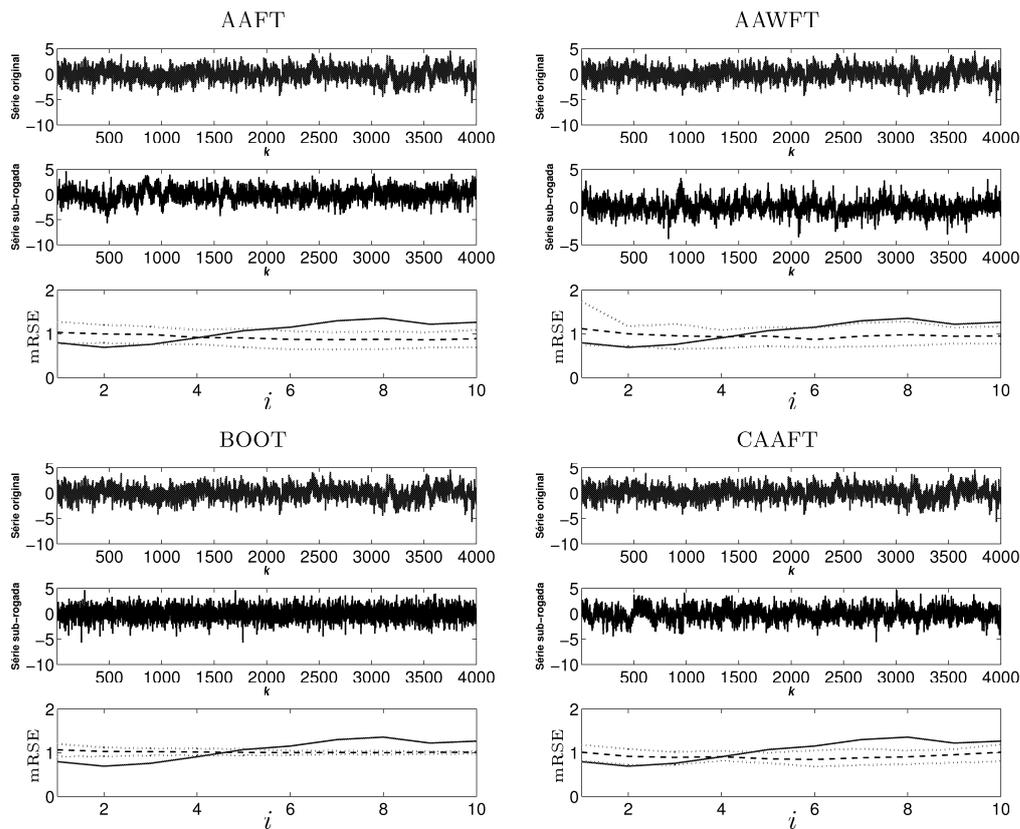


Figura 5.5: Resultados do teste de determinismo usando dados sub-rogados obtidos a partir da série residual com assinatura determinística $\hat{x}_1(k)$, usando os algoritmos AAFT, AAWFT, BOOT e CAAFT. O gráfico superior ilustra a série residual $\hat{x}_1(k)$, o intermediário uma das séries sub-rogadas e o inferior o resultado das previsões para a série $\hat{x}_1(k)$ e para as séries artificiais. O índice mRSE, calculado para um horizonte de previsão de $i = 1, 2, \dots, 10$, da série residual $\hat{x}_1(k)$ em (—) e, o índice mRSE médio das séries sub-rogadas em (- - -). Sendo a faixa de confiança de 60% representada em (...).

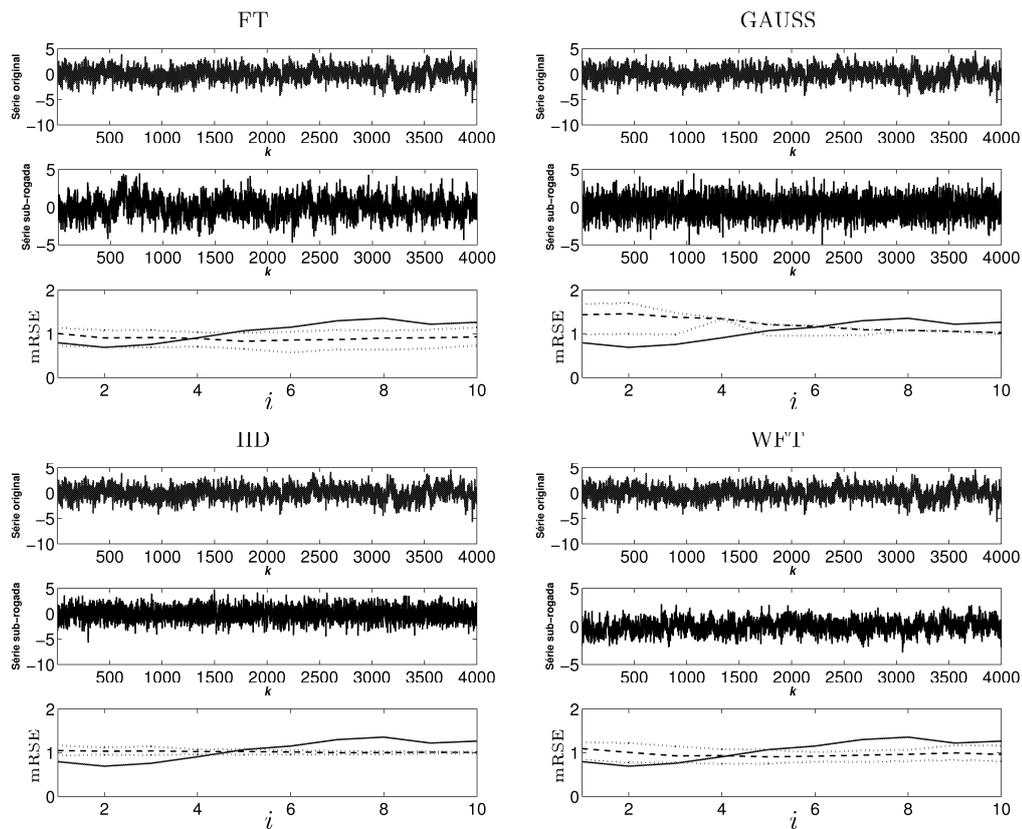


Figura 5.6: Resultados do teste de determinismo baseado em dados sub-rogados para a série residual com assinatura determinística $\hat{x}_1(k)$, usando os algoritmos FT, GAUSS, IID e WFT. O gráfico superior ilustra a série $\hat{x}_1(k)$, o intermediário uma das séries sub-rogadas e o inferior o resultado das previsões para a série $\hat{x}_1(k)$ e para as séries artificiais. O índice mRSE, calculado para um horizonte de previsão de $i = 1, 2, \dots, 10$, da série $\hat{x}_1(k)$ em (—) e, o índice mRSE médio das séries sub-rogadas em (- -). A faixa de confiança de 60% representada em (...).

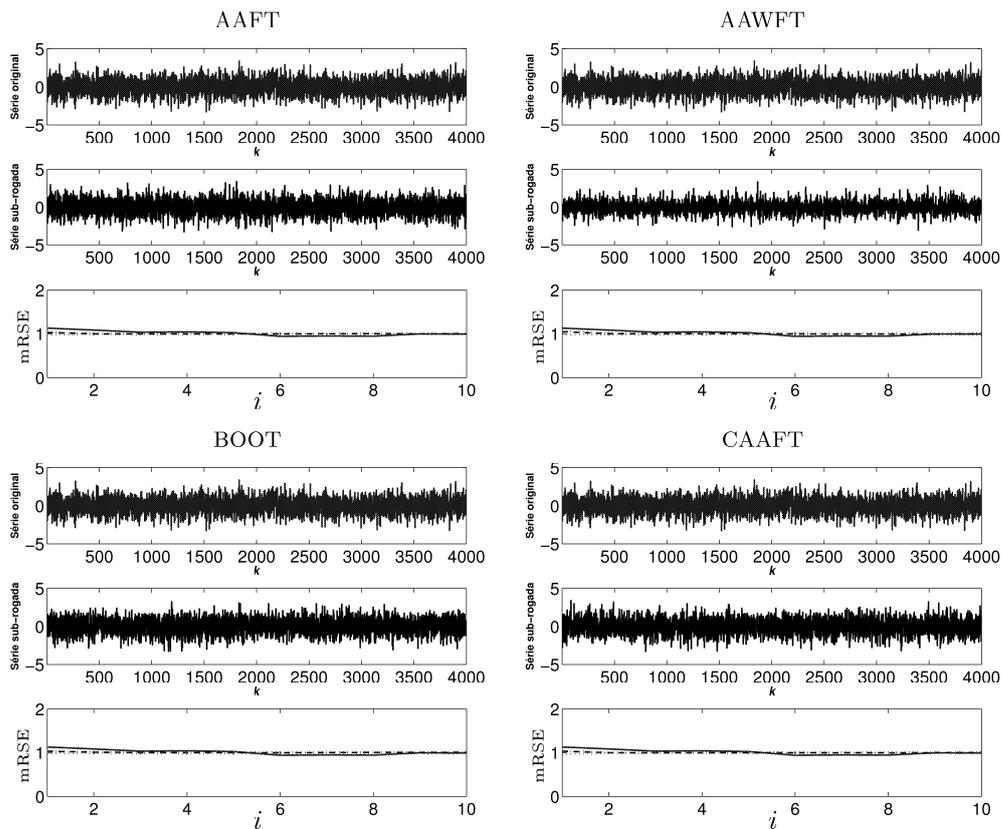


Figura 5.7: Resultados do teste de determinismo baseado nos dados sub-rogados para uma série residual aleatória $\hat{x}_2(k)$, usando os algoritmos AAF, AAW, BOOT e CAAFT. O gráfico superior ilustra a série residual $\hat{x}_2(k)$, o intermediário uma das séries sub-rogadas e o inferior o resultado das previsões para a série $\hat{x}_2(k)$ e para as séries artificiais. Sendo o índice mRSE, calculado para um horizonte de previsão de $i = 1, 2, \dots, 10$, da série original em (—) e, o índice mRSE médio das séries sub-rogadas em (- - -). A faixa de confiança representada em (...) é de 60%.

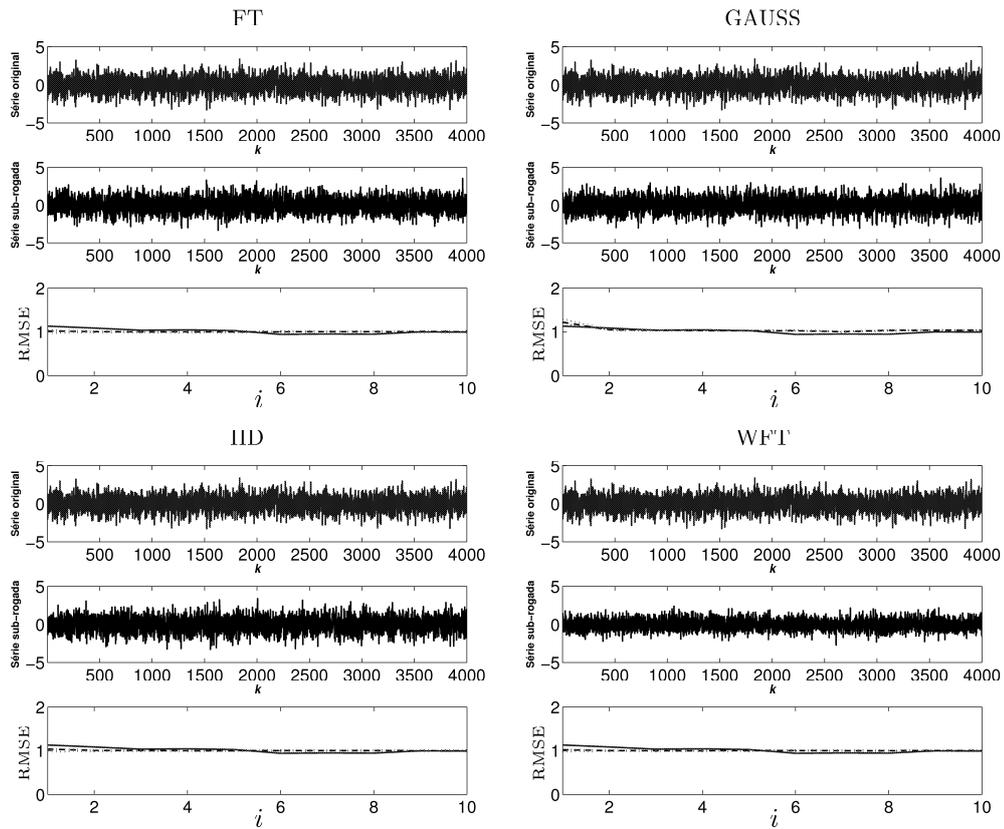


Figura 5.8: Resultados do teste de determinismo baseado nos dados sub-rogados para uma série residual aleatória $\hat{x}_2(k)$, usando os algoritmos FT, GAUSS, IID e WFT. O gráfico superior ilustra a série residual $\hat{x}_2(k)$, o intermediário uma das séries sub-rogadas e o inferior o resultado das previsões para a série $\hat{x}_2(k)$ e para as séries artificiais. O índice mRSE, calculado para um horizonte de previsão de $i = 1, 2, \dots, 10$, da série $\hat{x}_2(k)$ em (—) e, o índice mRSE médio das séries sub-rogadas em (- -). A faixa de confiança representada em (...) é de 60%.

Por meio destes resultados percebe-se que os algoritmos usados para gerar os dados sub-rogados dependem do que se deseja investigar, como por exemplo, determinismo ou linearidade. Como foi verificado ao longo desta seção, os algoritmos baseado no embaralhamento temporal da série residual (BOOT, GAUSS e IID) são mais indicados para a análise de determinismo, uma vez que estes tendem a destruir a correlação linear da série.³ Lembrando que para séries quase-periódicas, estes algoritmos devem ser aplicados na série residual, pois eles foram desenvolvidos para séries não periódicas (Theiler et al., 1992).

Como mencionado no início desta seção, o teste de determinismo baseado em dados sub-rogados, aqui proposto, baseou-se no método proposto por Gomes (2001), no entanto, com algumas diferenças a serem destacadas a seguir.

Em Gomes (2001) o índice RMSE é calculado em janelas de dados deslizantes ao longo da série temporal, como ilustra a Figura 5.9 (a) para uma janela deslizante de 3 passos. No entanto, no algoritmo usado neste trabalho, o índice de erro usado é o mRSE, o qual é calculado como a média de cada horizonte de previsão em no máximo 10 janelas diferentes, como ilustra a Figura 5.9 (b) em que o horizonte vai de 1 até 4 passos.

O cálculo do índice RMSE para uma janela deslizante segundo (Gomes, 2001) tem como principal vantagem o fato de poder encontrar, ao longo dos dados de validação da série, uma faixa de dados que seja previsível. Por outro lado, o cálculo do índice mRSE para cada horizonte de previsão, como feito neste trabalho, pode proporcionar uma idéia de até qual passo à frente a série pode ser considerada previsível.

5.4 Comentários

Um passo importante no teste de determinismo, aqui proposto, é com relação à extração da componente quase-periódica. Para extrair a quase-periodicidade das séries optou-se pelo uso do método FMMC para estimar e reduzir a componente

³Conjectura-se que a correlação não-linear da série também é destruída.

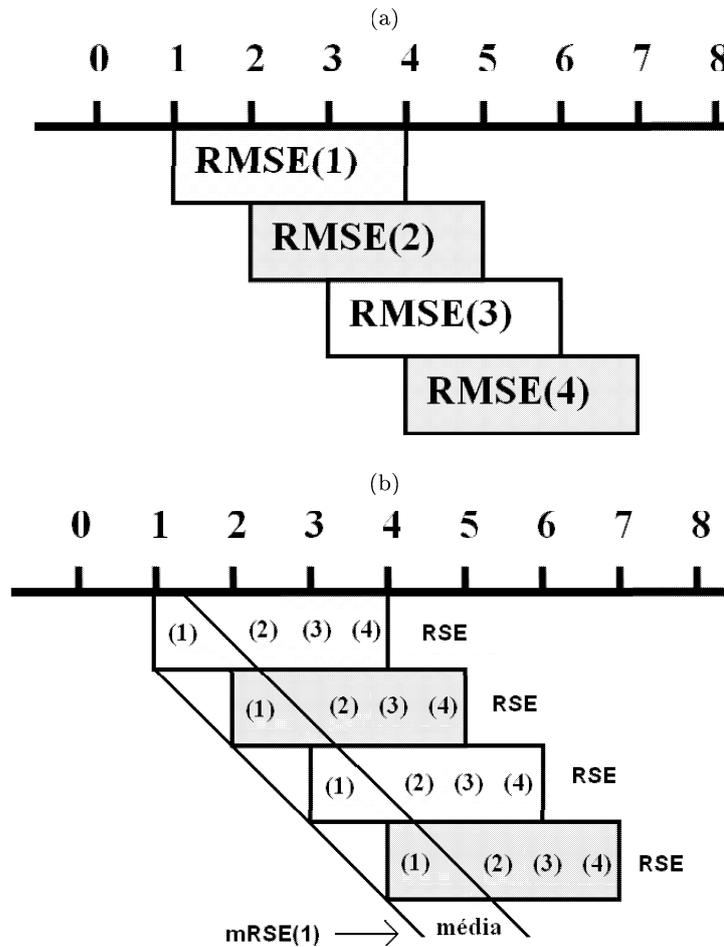


Figura 5.9: Em (a) o cálculo do índice RMSE feito em (Gomes, 2001) para uma janela deslizando de 3 passos. Em (b) o cálculo do índice mRSE para um horizonte que varia de 1 até 4 passos à frente.

quase-periódica. Deve-se destacar que, séries quase-periódicas apresentam mais que um ciclo e que, na teoria, a relação entre as frequências correspondentes aos ciclos é um número irracional.

No entanto, não se sabe ao certo o que a extração dos ciclos e tendência dos dados pode acarretar na série temporal. No sentido de que esta pode extrair não somente os ciclos mas também outro comportamento dinâmico da série temporal.

Uma análise empírica dos resultados obtidos com séries reais, ao longo deste estudo, mostra que os períodos dos ciclos não são sempre fixos, ou seja, esses podem

variar ao longo da série. Além disso, pôde-se perceber que as componentes da série - tendência, ciclos e resíduo - nem sempre são separáveis.

DETERMINISMO INTRA-CICLO EM SÉRIES QUASE-PERIÓDICAS REAIS

O interesse em se conhecer a dinâmica subjacente a uma série quase-periódica medida é comum em diversas áreas. Neste capítulo, a análise de determinismo é feita em séries quase-periódicas reais das ciências biológicas, físicas, econômicas e das engenharias. O teste de determinismo descrito na Seção 5.2.1 é aplicado às respectivas séries residuais das séries quase-periódicas. Para a redução da quase-periodicidade utiliza-se o método FMCC, descrito na Seção 3.3.4.

As séries quase-periódicas serão analisadas segundo a ordem:

- $y_Q(k)$: série diária de consumo de energia elétrica;
- $y_C(k)$: série mensal de casos de caxumba registrados em Nova Iorque.
- $y_L(k)$: série mensal de produção de leite.
- $y_{s_a}(k)$: série anual de número de sunspots (manchas solares).
- $y_{s_D}(k)$: série diária de número de sunspots (manchas solares).

6.1 Série de carga

A previsão de demanda de carga é uma tarefa difícil, uma vez que a série de carga é complexa e exhibe comportamentos quase-periódicos (ciclos). Além disso,

a carga é influenciada por diversos fatores exógenos, tais como fatores temporais, climáticos e aleatórios. Logo, não se pode partir da premissa de que há determinismo intra-ciclo na série de carga além do seu comportamento quase-periódico.

A série de consumo de energia elétrica ou série de carga $y_Q(k)$, aqui estudado, é referente a 22 de abril a 11 de dezembro de 1998.¹ Os dados disponíveis apresentam falhas nas medições, as quais foram contornadas por meio da interpolação polinomial, descrita na Seção 5.1, usando 10 valores adjacentes a estas observações. Realizou-se a decimação dos dados, para se ter acesso à previsão de curto-prazo em diante,² resultando em uma série temporal com tempo de amostragem de 1 hora. Dessa forma, tem-se uma série com 5.616 observações horárias (234 dias). A Figura 6.1 (a) ilustra a série de carga e a Figura 6.1 (b) uma faixa dos dados, destacando os ciclos diários e semanais.

Por inspeção visual da série temporal mostrada nas Figuras 6.1 (a) e (b), verificou-se a presença da componente cíclica referentes ao:

- padrão de demanda de carga semanal;
- padrão de demanda de carga diária.

Como mencionado ao longo do texto, as componentes cíclicas e de tendência de uma série temporal podem ser detectadas por meio da análise da função de autocorrelação e do periodograma da série. Para a série de carga, $y_Q(k)$, os ciclos podem ser explicitados por meio da FAC e do periodograma da série, como mostram as Figuras 6.2(a) e (b). As figuras destacam a FAC da série original, $y_Q(k)$, para uma janela de 5.616 observações (234 dias) e 720 observações (1 mês), respectivamente. Pode-se verificar que há um padrão que se repete a cada 168 observações (uma semana). Esse comportamento da FAC ilustra a presença do ciclo semanal da

¹A série de carga é a potência ativa em kW consumida na região de Queensland, na Austrália.

²A classificação da previsão de demanda de carga com relação ao horizonte de previsão desejado é dada por: previsão a curtíssimo prazo (o horizonte de interesse é de poucos minutos a uma hora à frente); previsão a curto prazo (refere-se à previsão horária da carga, ou seja, das próximas 24 horas a uma semana a frente); previsão a médio prazo (referente a previsão horária ou de pico de carga para um horizonte de um a vários meses); e, previsão a longo prazo (refere-se a previsão feita para períodos superiores a um ano).

série de carga. Também, pode-se verificar pela Figura 6.2(b) que há um padrão de comportamento que se apresenta a cada 24 horas, denotando assim, o ciclo diário da série $y_Q(k)$.

O periodograma da série original, $y_Q(k)$, é mostrado na Figura 6.2(c). Pela análise da figura verifica-se que os períodos cíclicos, $d_1 = 24$ e $d_2 = 168$, encontrados via análise da FAC de $y_Q(k)$ mostrados na Figura 6.2 (b), estão coerentes. Além da componente diária e semanal, a figura destaca também a presença de componentes cíclicas referentes aos períodos de 12 e 84 horas, $d_3 = 84$ e $d_4 = 12$.

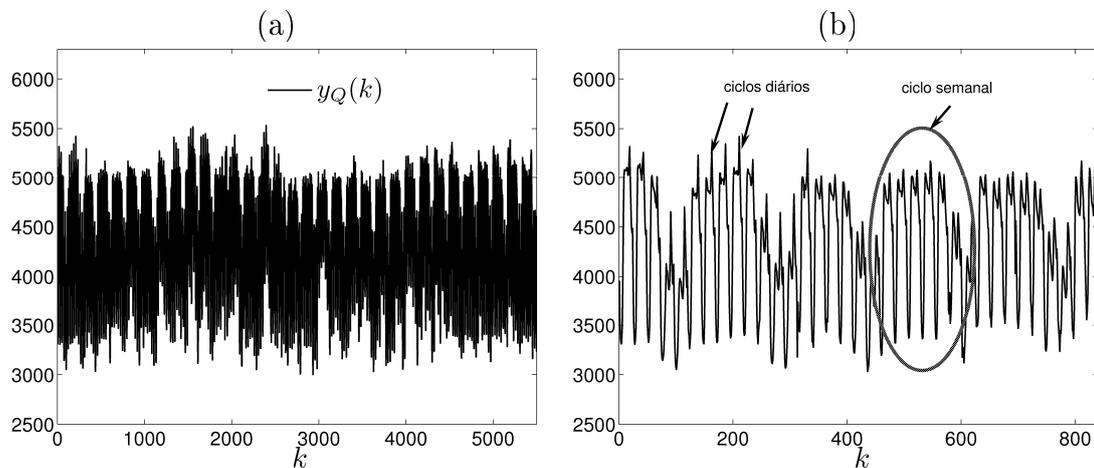


Figura 6.1: Em (a) série de demanda de carga da região de Queensland, na Austrália. No eixo das abscissas as observações k dada em horas e no eixo das ordenadas a amplitude do sinal em kW. Em (b) é mostrado uma faixa dos dados.

6.1.1 Teste de estacionariedade

Como a FAC mede a dependência temporal linear entre as observações, pode-se considerar que a FAC de uma série temporal estacionária não dependerá da janela de dados analisada. Dessa forma, considere as três janelas de dados da série de carga, contendo 168 observações cada, destacadas na Figura 6.3 (a). Pela FAC das *janelas* 1, 2 e 3, pode-se inferir que a série de carga, aqui estudada, é não-estacionária em

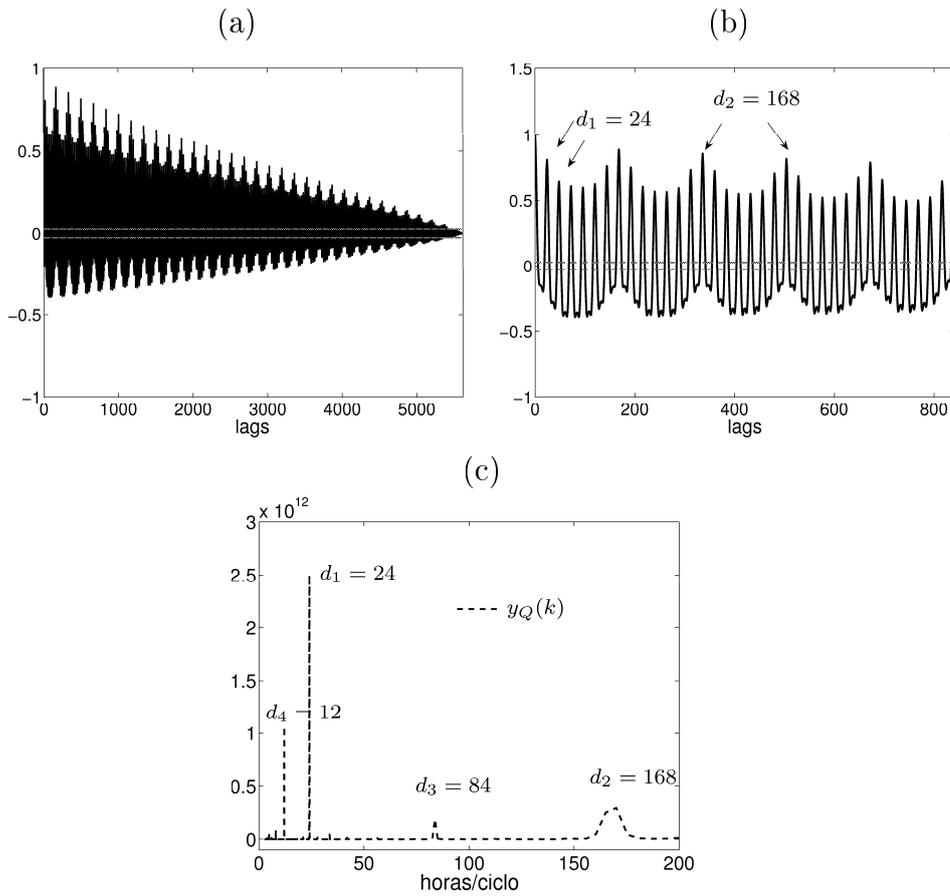


Figura 6.2: Em (a) FAC da série de carga para um atraso (*lag*) de 5.616 observações (234 dias). Em (b) FAC da série de carga para uma janela de 5 semanas (840 horas). Em (c) periodograma da série de carga. As componentes de frequência, dadas em Hz/segundo, são: 1 semana ($1/168 \times 3.600 = 0,17 \times 10^{-5}$ Hz), 1 dia ($1/24 \times 3.600 = 1,16 \times 10^{-5}$ Hz), 12 horas ($1/12 \times 3.600 = 2,32 \times 10^{-5}$ Hz) e 84 horas ($1/84 \times 3.600 = 3,31 \times 10^{-6}$ Hz).

alguma medida, uma vez que:

$$r_1(k) \neq r_2(k) \neq r_3(k), \quad \forall k, \quad (6.1)$$

em que r_1 , r_2 e r_3 são, respectivamente, os coeficientes de autocorrelação das *janelas 1, 2 e 3*.

Como pode ser observado, nem todas as janelas de dados analisadas apresentam a mesma FAC e o mesmo periodograma, como mostram as Figuras 6.3 (b) e (c). Dessa forma, conclui-se que a série de carga, aqui estudada, não é estacionária. O que implicará em uma dificuldade maior em se analisar e modelar adequadamente a série. Logo, deve-se encontrar faixas de dados ao longo da série temporal para as quais a série possa ser considerada estacionária.

Para determinar o tamanho da janela de dados que pode ser considerada estacionária, aplicou-se o teste desenvolvido em (Guimarães, 1996) e usado por Gomes (2001) em séries de variabilidade de frequência cardíaca. Tal teste consiste na obtenção de modelos paramétricos lineares para faixas de dados da série temporal. Nesse caso, cada janela é dividida em dois segmentos para os quais modelos polinomiais AR de mesma ordem dinâmica são estimados.

Considere a *janela 1* de dados da Figura 6.3 (a). Dividiu-se a janela em dois segmentos e estimou-se modelos polinomiais AR, com ordem dinâmica igual a 4 ($n_y = 4$), para ambos os segmentos. Os pólos dos modelos AR do segundo segmento da *janela 1* estão localizados dentro das regiões de confiança associadas aos pólos estimados para o primeiro segmento, conforme pode ser visto na Figura 6.4 (a). Isto sugere que a *janela 1* é estacionária, uma vez que há um modelo AR com características dinâmicas similares que pode representar a série em faixas distintas.³ O mesmo ocorre na *janela 3*, Figura 6.3 (b). Dividiu-se a janela em dois segmentos e estimou-se modelos polinomiais AR, com ordem dinâmica igual a 4, para ambos os segmentos. Os pólos dos modelos AR do segundo segmento estão localizados dentro

³Para a janela de dados 1, o resultado foi o mesmo para a ordem dinâmica do modelo AR variando de 3 a 8.

das regiões de confiança associadas aos pólos estimados para o primeiro segmento, conforme pode ser visto na Figura 6.3 (b). Sugerindo que a *janela 3* é estacionária.⁴ Por outro lado, na *janela 2* o resultado do teste apresentado na Figura 6.3 (c) mostra que para um modelo AR com ordem dinâmica igual a 6, a maioria dos pólos do segundo segmento não estão dentro da faixa de confiança associadas aos pólos do primeiro segmento, o que sugere que essa janela não é estacionária.

Diante desses resultados, as janelas de dados da série de carga, aqui estudada, conterão aproximadamente 240 observações, visto que as ferramentas de modelagem NARMA que serão utilizadas foram desenvolvidas para séries temporais estacionárias.

⁴Para a janela de dados 3, o resultado foi o mesmo para a ordem dinâmica do modelo AR variando de 3 a 7.

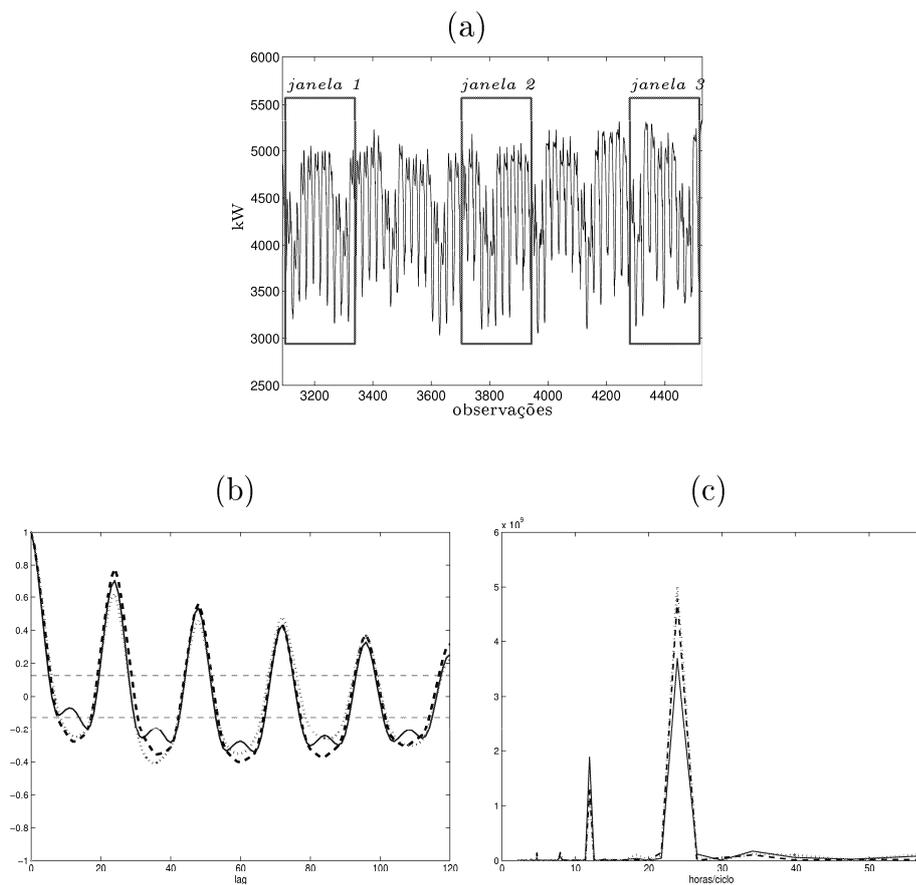


Figura 6.3: Teste de estacionariedade para a série de carga. Em (a) série de carga (`q04a1298a.dat`) com destaques das *janelas 1, 2 e 3*. Em (b) FAC das *janelas 1* (—), *2* (- - -) e *3* (...). Em (c) periodograma para as *janelas 1* (—), *2* (- - -) e *3* (...). Sendo a média e o desvio-padrão: para a *janela 1*, $\mu_1 = 4259.4$ e $\sigma_1 = 540$; para a *janela 2*, $\mu_2 = 4336$ e $\sigma = 550$; e, para a *janela 3*, $\mu_3 = 4.378$ e $\sigma = 585$.

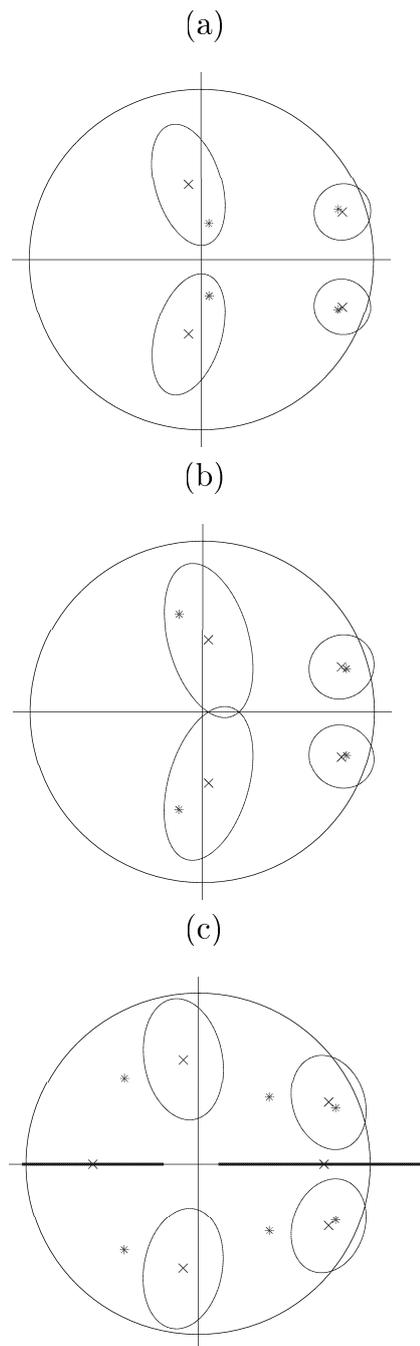


Figura 6.4: Testes de estacionariedade para as *janelas 1, 2 e 3* da série de carga. Em (a) pólos do modelo AR do segundo segmento (*) da *janela 1*, em (b) os pólos do modelo AR do segundo segmento (*) da *janela 3* estão dentro das regiões de confiança associadas aos pólos obtidos no primeiro segmento (x) de ambas as janelas. Em (c) nem todos os pólos do modelo AR do segundo segmento (*) da *janela 2* estão dentro das regiões de confiança associadas aos pólos obtidos no primeiro segmento (x) da janela.

6.1.2 Teste de determinismo na série quase-periódica

Inicialmente, aplicou-se o teste de determinismo proposto por Nakamura & Small (2005). O teste foi aplicado diretamente à série quase-periódica.

Foi gerado a partir da série $y_Q(k)$ um conjunto de 100 séries sub-rogadas usando o algoritmo SSS, o qual foi descrito na Seção 3.6.2, para o parâmetro $A = 1$. Em seguida, dividiu-se a série quase-periódica $y_Q(k)$ e as séries sub-rogadas em seis janelas distintas, conforme Tabela 6.1.

A faixa de uma das séries artificiais comparada à série quase-periódica é mostrada na Figura 6.5. Percebe-se que a série sub-rogada é bastante similar à série quase-periódica $y_Q(k)$. O fator discriminante, aqui usado, é a FAC à semelhança de Nakamura & Small (2005).

A FAC das 100 séries sub-rogadas e da série quase-periódica $y_Q(k)$ para as quatro primeiras janelas está ilustrada na Figura 6.5. Por meio das figuras verificou-se que a FAC de $y_Q(k)$ apresenta uma leve diferença das FAC das séries artificiais para as quatro janelas de dados. Segundo Nakamura & Small (2005), isso é um indício de que a série quase-periódica $y_Q(k)$ apresenta um traço determinístico intra-ciclo ao longo das quatro janelas.

Tabela 6.1: Teste de determinismo usando o método SSS (Seção 3.6.2) nas janelas de dados da série de carga $y_Q(k)$.

janela	faixa de dados	determinismo via SSS	determinismo via método proposto
1	4.000 - 4.167 - 4.239	sim	sim
2	4.239 - 4.406 - 4.478	sim	—
3	4.478 - 4.645 - 4.717	sim	sim
4	4.717 - 4.884 - 4.956	sim	—
5	4.956 - 5.123 - 5.195	sim	sim
6	5.195 - 5.362 - 5.434	sim	sim

6.1.3 Redução da quase-periodicidade da série $y_Q(k)$

O primeiro passo para se aplicar o teste de determinismo, proposto neste trabalho, consiste em extrair ou reduzir a componente quase-periódica da série $y_Q(k)$.

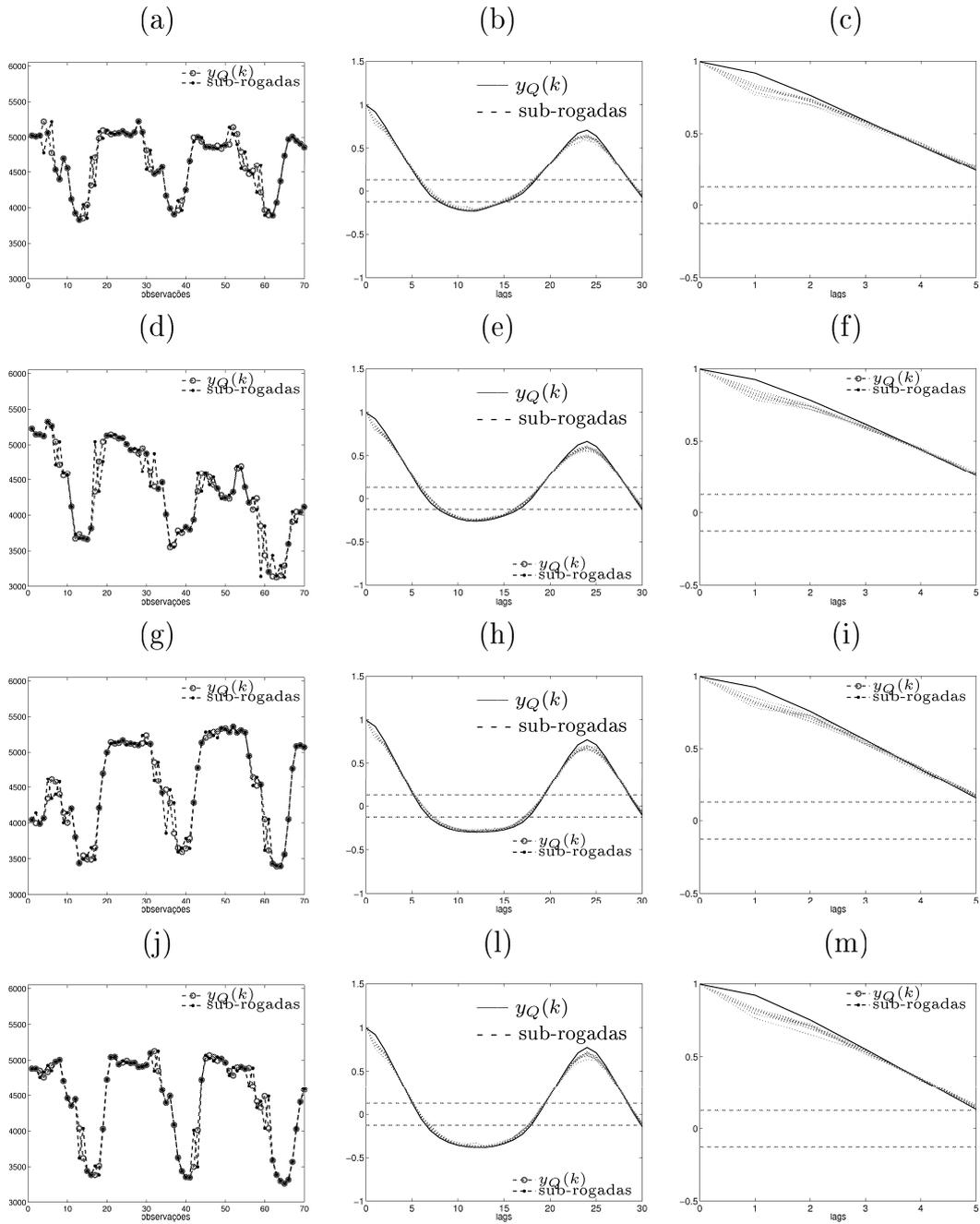


Figura 6.5: Teste de determinismo usando o método SSS proposto por Nakamura & Small (2005) aplicado à série $y_Q(k)$. Em (a) faixa de dados de $y_Q(k)$ (-o-) e uma das 100 séries sub-rogadas (- -) para a janela 1. Em (b) FAC de $y_Q(k)$ (—) e das 100 séries sub-rogadas (- -) para a janela 1. Em (c) detalhe de (b) para a janela 1. Em (d), (e) e (f) para a janela 2. Em (g), (h) e (i) para a janela 3. Em (j), (l) e (m) para a janela 4.

Para isso emprega-se o método FMMC descrito na Seção 3.3.4. Assim, a série de carga, $y_Q(k)$, é descrita como uma composição de três componentes:

$$y_Q(k) = c_Q(k) + m_Q(k) + x_Q(k),$$

sendo $c_Q(k)$ a componente cíclica, $m_Q(k)$ a componente de tendência e $x_Q(k)$ a componente residual. $c_Q(k)$ e $m_Q(k)$ serão estimadas e retiradas da série original, $y_Q(k)$, para a obtenção da série residual $x_Q(k)$, tal que:

$$\hat{x}(k) = y(k) - \hat{c}(k) - \hat{m}(k).$$

Redução do ciclo diário

Sabendo previamente que a série de carga, $y_Q(k)$, possui duas componentes cíclicas referentes ao ciclo diário ($d_1 = 24$ hs) e ao ciclo semanal ($d_2 = 168$ hs) e, considerando as 5.616 observações da série, a componente de tendência de $y_Q(k)$ foi estimada aplicando-se um filtro de média móvel escolhido desconsiderando as componentes $x_Q(k)$ e $c_Q(k)$. Para o ciclo com $d = 24$, tem-se $q = 24/2 = 12$, logo:

$$\hat{m}_{0_{24}}(k) = (0,5y_Q(k - 12) + y_Q(k - 11) + \dots + y_Q(k + 11) + 0,5y_Q(k + 12))/24.$$

O passo seguinte consiste em estimar a componente cíclica diária, $\hat{c}_{Q_{24}}(k)$. Para cada $k_0 = 1, \dots, 24$, estimou-se a média $\hat{w}_{24}(k_0)$ da diferença dada por:

$$\hat{w}_{24}(k_0) = \{ (y(k_0 + 24j) - \hat{m}_{0_{24}}(k_0 + 24j)), \quad 24 < k_0 + 24j \leq (5.616 - 12) \}.$$

sendo $j = 1, \dots, n_c$ e n_c o número de dias (ciclo diário) na série quase-periódica $y_Q(k)$. Dessa forma, a componente cíclica diária, $c_{Q_{24}}(k)$, pôde ser estimada da seguinte maneira:

$$\hat{c}_{Q_{24}}(k) = \hat{w}_{24}(k) - \frac{1}{24} \sum_{i=1}^{24} \hat{w}_{24}(i),$$

$k = 1, \dots, 24$ e $\hat{c}_{24}(k) = \hat{c}_{24}(k - 24)$, $k > 24$.

A série resultante sem o ciclo diário, $\hat{D}(k)$, foi definida extraindo-se a componente cíclica estimada, $\hat{c}_{Q_{24}}(k)$, da série original, $y_Q(k)$:

$$\hat{D}_{24}(k) = y_Q(k) - \hat{c}_{Q_{24}}(k), \quad k = 1, \dots, 5.616. \quad (6.2)$$

A tendência da série, $\hat{D}_{24}(k)$, foi estimada ajustando-se um polinômio de ordem dois, conforme descrito na Seção 3.3.3. Dessa forma, tem-se:

$$\hat{m}_Q(k) = 1,5 \times 10^{-6}k^2 + 0,006k + 4.348,90.$$

Em termos das componentes cíclica e de tendência estimadas, $\hat{c}_{Q_{24}}(k)$ e $\hat{m}_Q(k)$, respectivamente, a série residual, $x_{24}(k)$, foi obtida como:

$$\hat{x}_{Q_{24}}(k) = y_Q(k) - \hat{m}_Q(k) - \hat{s}_{Q_{24}}(k), \quad k = 1, \dots, 5.616. \quad (6.3)$$

A Figura 6.6(a) ilustra a série residual $\hat{x}_{Q_{24}}(k)$ sem o ciclo diário, porém, com o ciclo semanal. Isso pode ser confirmado observando a FAC, ilustrada na Figura 6.6(b), a qual compara a FAC da série quase-periódica $y_Q(k)$ com a FAC da série residual $\hat{x}_{Q_{24}}(k)$. O mesmo pode ser visto pelo periodograma representado na Figura 6.6(c). Percebe-se que, apesar da significativa redução dos ciclos referentes a $d_1 = 24$ e $d_4 = 12$, os ciclos com $d_2 = 168$ e $d_3 = 84$ permanecem na série residual $\hat{x}_{Q_{24}}(k)$. O que pode mascarar os resultados do teste de determinismo.

Redução do ciclo diário e semanal

Considerando o o ciclo semanal com $d = 168$, tem-se $q = 168/2 = 84$, portanto:

$$\hat{m}_{0_{168}}(k) = (0,5y_Q(k - 84) + y_Q(k - 83) + \dots + y_Q(k + 83) + 0,5y_Q(k + 84))/168.$$

A componente cíclica, $\hat{c}_{Q_{168}}(k)$ foi estimada para cada $k_0 = 1, \dots, 168$, calculando-

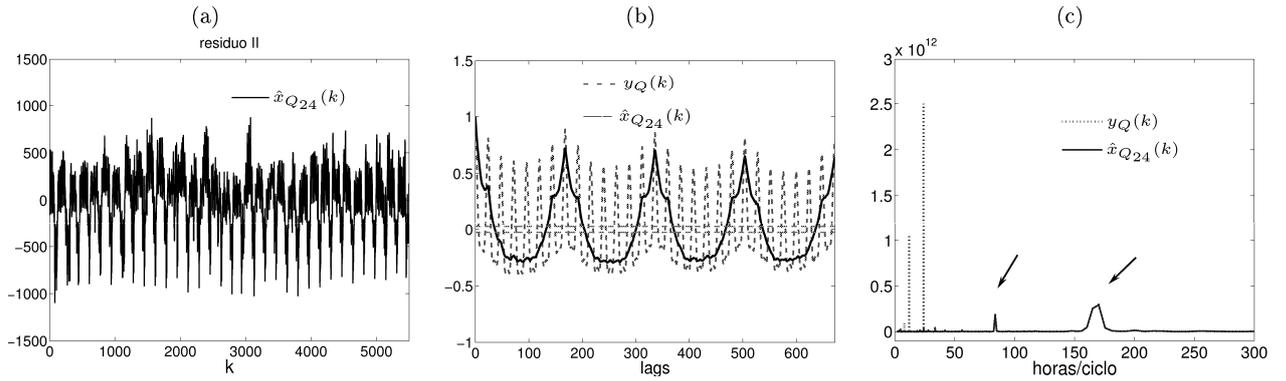


Figura 6.6: Em (a) série residual $\hat{x}_{Q_{24}}(k)$. Em (b) FAC e em (c) periodograma da série quase-periódica $y_Q(k)$ em (- -) e da série residual $\hat{x}_{Q_{24}}(k)$ em (—).

se a média $\hat{w}_{168}(k_0)$ da diferença dada por:

$$\hat{w}_{168}(k_0) = \left\{ (y(k_0 + 168j) - \hat{m}_{0_{168}}(k_0 + 168j)), \quad 168 < k_0 + 168j \leq (5.616 - 84) \right\}.$$

em que $j = 1, \dots, n_c$ e n_c o número de semanas (ciclo semanal) presentes na série quase-periódica $y_Q(k)$. Dessa forma, a componente cíclica, $c_{Q_{168}}(k)$, foi estimada da seguinte maneira:

$$\hat{c}_{Q_{168}}(k) = \hat{w}_{168}(k) - \frac{1}{168} \sum_{i=1}^{168} \hat{w}_{168}(i),$$

$$k = 1, \dots, 168 \text{ e } \hat{c}_{Q_{168}}(168) = \hat{c}_{Q_{168}}(336), \dots$$

Os dados sem o ciclo semanal $\hat{D}_{168}(k)$ foi definido extraíndo-se a componente cíclica semanal estimada, $\hat{c}_{Q_{168}}(k)$, da série original $y_Q(k)$:

$$\hat{D}_{168}(k) = y_Q(k) - \hat{c}_{Q_{168}}(k), \quad k = 1, \dots, 5.616. \quad (6.4)$$

A componente de tendência da série $\hat{D}_{168}(k)$ foi estimada usando-se um polinômio de ordem dois:

$$\hat{m}_Q(k) = -5,27 \times 10^{-7} k^2 + 0,0016k + 4.339,44.$$

Enfim, em termos das componentes cíclicas e de tendência estimadas, $\hat{c}_{Q_{168}}(k)$ e $\hat{m}_Q(k)$, respectivamente, a série residual foi obtida como:

$$\hat{x}_{Q_{168}}(k) = y_Q(k) - \hat{m}_Q(k) - \hat{c}_{Q_{168}}(k), \quad k = 1, \dots, n. \quad (6.5)$$

A Figura 6.7(a) ilustra a série residual $\hat{x}_{Q_{168}}(k)$. Percebe-se que ao se extrair o ciclo semanal o ciclo diário foi também reduzido. Pela FAC e o periodograma na Figura 6.7(b) e (c), respectivamente, percebe-se que as componentes cíclicas com períodos $d_2 = 168$, $d_3 = 84$ e $d_4 = 12$ foram completamente extraídas e o ciclo diário com $d_1 = 24$ foi apenas reduzido.

Vale ressaltar que os outros métodos para a redução de ciclos e sazonalidades, apresentados na Seção 3.3.4, também não foram capazes de eliminar a componente cíclica diária.

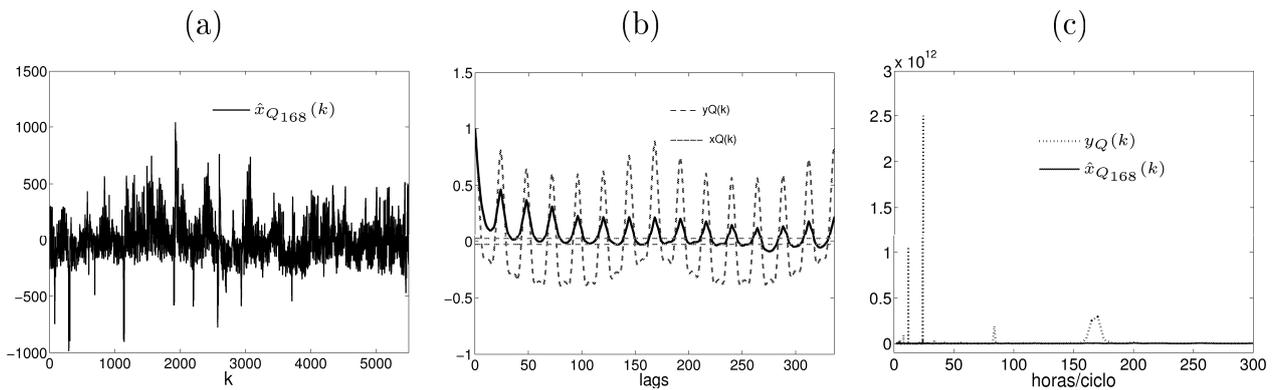


Figura 6.7: Em (a) série residual $\hat{x}_{Q_{168}}(k)$. Em (b) FAC e em (c) periodograma da série quase-periódica $y_Q(k)$ (- -) e da série residual $\hat{x}_{Q_{168}}(k)$ (—).

Testando determinismo em $\hat{x}_{Q_{168}}(k)$

O teste de determinismo foi, então, aplicado à série residual $\hat{x}_{Q_{168}}(k)$ mostrada na Figura 6.7(a). Foram geradas 100 séries sub-rogadas a partir do algoritmo de embaralhamento temporal (Theiler et al., 1992), descrito na Seção 3.6.1. Uma das 100 séries subrogadas é mostrada nos gráficos intermediários da Figura 6.8.

Como analisado na Seção 6.1.1, a série quase-periódica $y_Q(k)$ pode ser considerada estacionária para janelas de dados de aproximadamente 240 observações. Portanto, a série residual correspondente, $\hat{x}_{Q_{168}}(k)$, e as séries artificiais foram divididas em seis janelas, contendo 240 observações cada, como mostrado na Tabela 6.1. O teste de determinismo proposto foi, então, aplicado às seis janelas de dados.

Para cada janela de dados, utilizaram-se as 168 observações iniciais para estimar os modelos NARMA polinomiais e, as 73 observações finais, para a validação dos modelos. Os modelos NARMA, para cada a série original $\hat{x}_{Q_{168}}(k)$ e para cada uma das 100 séries sub-rogadas, foram escolhidos conforme o menor índice mRSE para o primeiro passo de previsão. A estrutura dos modelos foi definida segundo os valores máximo de $\ell = 2$, $n_y = 11$, $n_e = 5$ e $n_t = 20$.

Por meio dos modelos NARMA escolhidos, foram feitas as previsões de 1 até 10 passos à frente e foram calculados os índices mRSE, tanto para a série residual $\hat{x}_{Q_{168}}(k)$ quanto para as séries artificiais, para cada uma das seis janelas de dados. Os índices mRSE, para cada uma das janelas, estão representados nos gráficos inferiores da Figura 6.8.

Finalmente, foi testada a hipótese nula, H_0 , para a janela 1 (4.000-4.239). Verifica-se, por meio da Figura 6.8, que o desempenho da previsão da série residual $\hat{x}_{Q_{168}}(k)$ foi melhor do que o desempenho médio das previsões das 100 séries sub-rogadas para os dois primeiros passos de previsão, portanto, pode-se rejeitar H_0 . Assim, aceita-se H_1 e infere-se que ao longo da janela 1 da série residual $\hat{x}_{Q_{168}}(k)$ tem alguma característica determinística de curto-prazo, além dos ciclos que foram removidos. Conseqüentemente, conjectura-se que a série de carga $y_Q(k)$, referente à janela 1, tem traço determinístico intra-ciclo.

A janela 2 (4.239-4.478) tem um desempenho de previsão da série $\hat{x}_{Q_{168}}(k)$ pior que o desempenho médio das previsões das 100 séries sub-rogadas, conforme o índice mRSE mostrado a Figura 6.8. Porém, como definido anteriormente, pode-se rejeitar H_0 , e aceitar H_2 . Assim, não se pode inferir sobre sua característica determinística da série residual $\hat{x}_{Q_{168}}(k)$ ao longo da janela 2. Conseqüentemente,

conjectura-se que a série de carga $y_Q(k)$, ao longo da janela 2, pode ser não-estacionária ou não modelável por NARMA polinomais.

Para a janela 3 (4.478-4.717) verifica-se, pela Figura 6.8, que o desempenho da série residual $\hat{x}_{Q_{168}}(k)$ é melhor para todos os horizontes de previsão. Logo, pode-se rejeitar H_0 e aceita-se H_1 , inferindo assim, que há determinismo de curto-prazo em $\hat{x}_{Q_{168}}(k)$, além do ciclo que foi removido. Portanto, conjectura-se que a série quase-periódica $y_Q(k)$, referente à janela 3, apresenta característica determinística intra-ciclo.

Pelo teste de determinismo para a janela 4 (4.717-4.956), apresentado na Figura 6.8, pode-se inferir que a mesma apresenta um comportamento não-estacionário, ou seja, rejeita-se H_0 e aceita-se H_2 . Visto que, o desempenho da previsão da série residual $\hat{x}_{Q_{168}}(k)$ nessa janela ora apresenta-se pior, ora similar (passos 5-7), ora melhor (passos 9-10) do que o desempenho médio das previsões das séries subrogadas, Portanto, conjectura-se que a série de carga $y_Q(k)$, referente à janela 4, é não estacionária.

A análise do índice mRSE na Figura 6.8 mostra que, para a janela 5, a série residual apresenta um desempenho melhor para os 2 primeiros passos de previsão, e para a janela 6, um melhor desempenho para os 10 primeiros passos de previsão. Portanto, pode-se inferir determinismo de curto-prazo na série de carga $y_Q(k)$, ao longo das janelas 5 e 6.

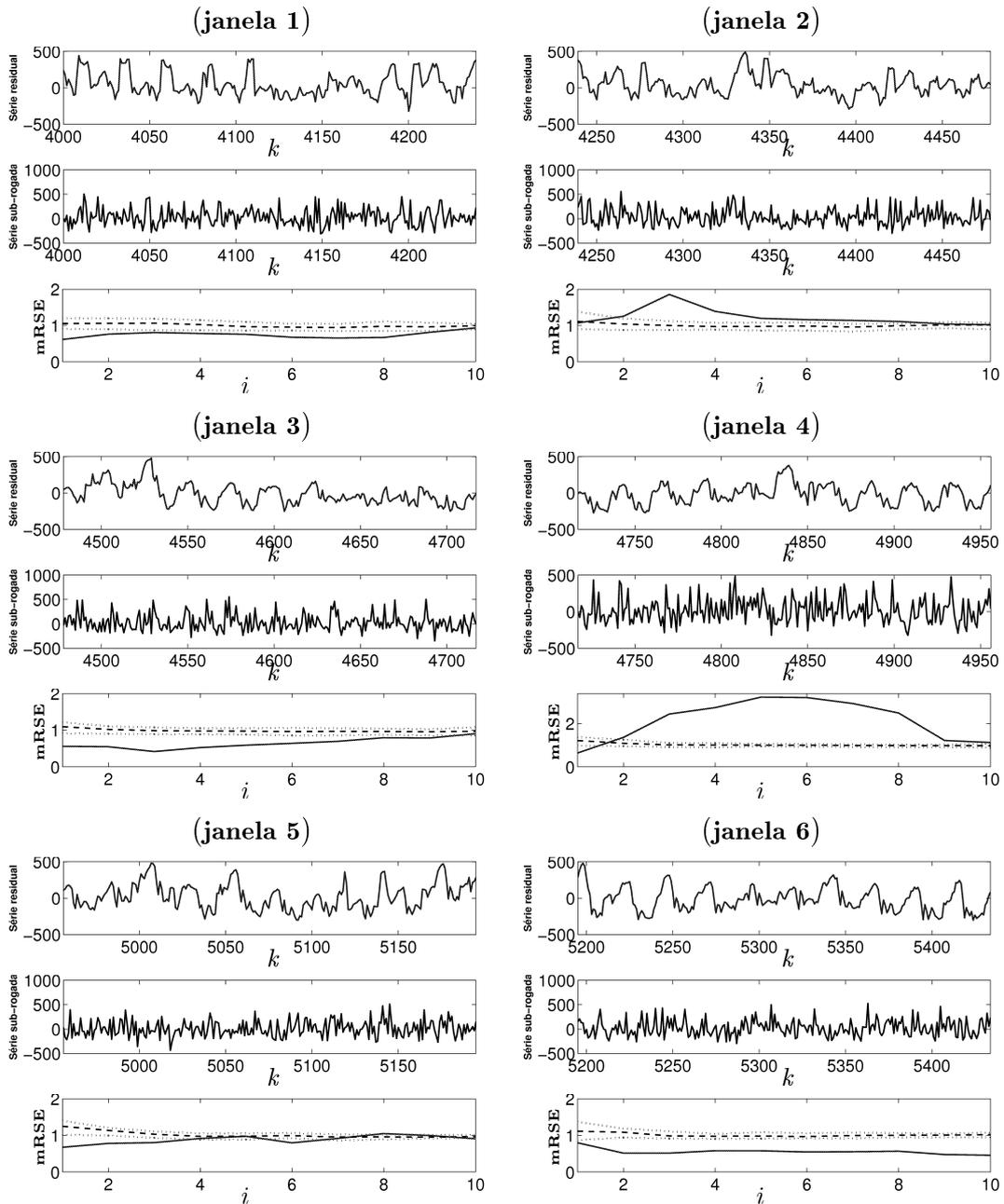


Figura 6.8: Teste de determinismo intra-ciclo da série de carga $y_{Q_{168}}(k)$, realizado a partir de sua respectiva série residual $\hat{x}_{Q_{168}}(k)$. Nos gráficos superiores janelas de dados da série residual $\hat{x}_{Q_{168}}(k)$. Nos gráficos intermediários, uma das séries artificiais. E nos gráficos inferiores os índices mRSE para a série residual $\hat{x}_{Q_{168}}(k)$ (—), o índice mRSE médio para as séries sub-rogadas (- -), além do intervalo de confiança de 60%, para as janelas dados 1, 2, 3, 4, 5 e 6.

6.2 Índices mensais de caxumba

Segundo a Figura 6.9(a), a série temporal quase-periódica de casos mensais de caxumba $y_C(k)$, registrados em Nova Iorque entre 1928-1972, apresenta um padrão acentuado que se repete a cada 12 meses (padrão sazonal).⁵ Aparentemente, os altos índices de caxumba ocorrem entre os meses de março-junho. Às vezes, pode haver um interesse em verificar como é a evolução subjacente da doença ao longo dos meses. Para isso, aplica-se o teste de determinismo na série quase-periódica $y_c(k)$ com 528 observações mensais.

Pela FAC e pelo periodograma mostrados nas Figuras 6.9(c) e (d), percebe-se que a série quase-periódica $y_c(k)$ apresenta uma componente sazonal, com período $d_1 = 12$, e um ciclo com $d_2 = 31$. Esses padrões cíclicos foram estimados e extraídos de $y_c(k)$ pelo método FMCM descrito na Seção 3.3.4, o qual abrange qualquer valor inteiro de período. Logo, assumindo que $y_C(k)$ pode ser representado como um modelo aditivo, tem-se:

$$y_c(k) = s_{c_1}(k) + c_{c_2}(k) + m_c(k) + x_c(k), \quad (6.6)$$

sendo $m_c(k)$ a componente de tendência, $s_{c_1}(k)$ a componente sazonal com $d_1 = 12$, $c_{c_2}(k)$ a componente cíclica com $d_2 = 31$ e, $x_c(k)$ a componente residual.

O FMCM é então aplicado à $y_c(k)$ com $N = 528$ observações para $d_1 = 12$ e $q_1 = 6$, obtendo assim:

$$\hat{m}_0(k) = \left(0,5y_c(k-6) + y_c(k-5) + \dots + y_c(k+5) + 0,5y_c(k+6) \right) / 12.$$

O passo seguinte consiste em estimar a componente sazonal com $d_1 = 12$, $\hat{s}_{c_1}(k)$. Para cada $k_0 = 1, \dots, 12$; estima-se a média da diferença, $\hat{w}(k_0)$, dada por:

$$\hat{w}(k_0) = y_c(k_0 + 12j) - \hat{m}_0(k_0 + 12j), \quad 12 < k_0 + 12j < 528 - 6,$$

⁵A série de casos de caxumba (`mumps.dat`) encontra-se no sítio abaixo na subpasta `Health`: <http://www-personal.buseco.monash.edu.au/%7Ehyndman/TSDL/>

sendo $j = 1 \dots, n_c$, em que $n_c = 528/12 = 44$ é o número de ciclos. Dessa forma, tem-se a primeira componente sazonal como:

$$\hat{s}_{c_1}(k) = \hat{w}(k) - \frac{1}{12} \sum_{i=1}^{12} \hat{w}(i), \quad (6.7)$$

para $k = 1, \dots, 12$, ressaltando que $\hat{s}_{c_1}(k) = \hat{s}_{c_1}(k+6) = \hat{s}_{c_1}(k+12) = \dots$

A componente cíclica, $\hat{c}_{c_2}(k)$ com $d_2 = 31$, foi estimada conforme procedimento anterior, portanto, tem-se para $d_2 = 31$ e $q_2 = 15$:

$$\hat{m}_0(k) = \left(y_c(k-15) + y_c(k-14) + \dots + y_c(k+14) + y_c(k+15) \right) / 31.$$

Para cada $k_0 = 1, \dots, 31$; estima-se a nova média da diferença dada por:

$$\hat{w}(k_0) = y_c(k_0 + 31j) - \hat{m}_0(k_0 + 31j), \quad 31 < k_0 + 31j < 528 - 15,$$

sendo $j = 1 \dots, n_c$, em que $n_c = 528/31 = 17$ é o número de ciclos que ocorrem ao longo da série. Dessa forma, tem-se:

$$\hat{c}_{c_2}(k) = \hat{w}(k) - \frac{1}{31} \sum_{i=1}^{31} \hat{w}(i), \quad (6.8)$$

para $k = 1, \dots, 31$; ressaltando que, $\hat{s}_{c_2}(k) = \hat{s}_{c_2}(k+31) = \hat{s}_{c_2}(k+62) = \dots$

O próximo passo consiste em estimar a componente de tendência $\hat{m}_c(k)$ por meio do ajuste polinomial de ordem 2, assim:

$$\hat{m}_c(k) = 297,09k^2 + 2,47k - 0,01. \quad (6.9)$$

Dessa forma, subtraindo-se (6.7), (6.8) e (6.9) de $y_c(k)$, estima-se a série residual:

$$\hat{x}_c(k) = y_c(k) - \hat{s}_{c_1}(k) - \hat{c}_{c_2}(k) - \hat{m}_c(k). \quad (6.10)$$

A série residual $\hat{x}_c(k)$ está representada na Figura 6.9(b). Comparando a FAC e o periodograma da série de caxumba, $y_c(k)$, com a sua respectiva série residual, $\hat{x}_c(k)$, nas Figuras 6.9(c) e (d), verifica-se uma redução significativa das componentes cíclicas. O teste de determinismo foi então aplicado à série residual $\hat{x}_c(k)$ em (6.10).

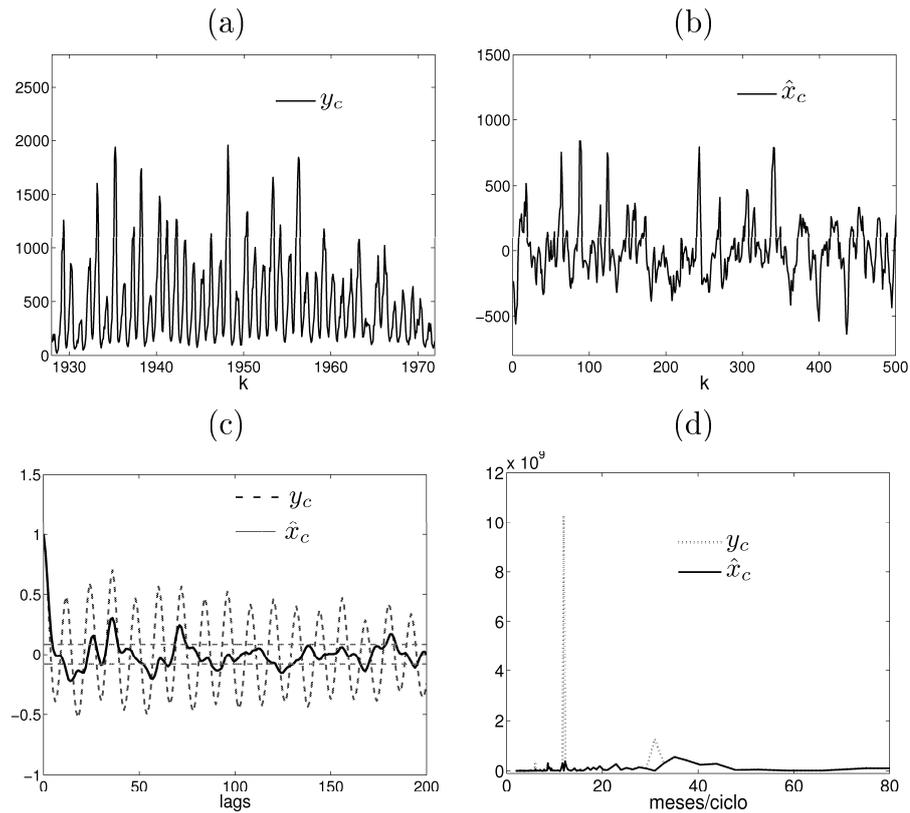


Figura 6.9: Em (a) série temporal quase-periódica de casos mensais de caxumba $y_c(k)$, registrados em Nova Iorque entre 1928-1972. Em (b) série residual $\hat{x}_c(k)$. Em (c) e (d), respectivamente, FAC e periodograma da série $y_c(k)$ (- - -) e da série $\hat{x}_c(k)$ (—).

Detectando determinismo de curto-prazo na série residual $\hat{x}_c(k)$

O primeiro passo do teste de determinismo descrito na Seção 5.2.1 consiste em extrair a componente quase-periódica, o qual foi executado anteriormente. Diante da série residual $\hat{x}_c(k)$ em (6.10), representada na Figura 6.9(b), gerou-se um conjunto de $n_{surr} = 100$ séries sub-rogadas usando o algoritmo de embaralhamento

temporal (Theiler et al., 1992). Uma das 100 séries artificiais está ilustrada no gráfico intermediário da Figura 6.10.

O próximo passo no teste de determinismo consiste em estimar os modelos NARMA polinomiais para a série residual $\hat{x}_c(k)$ e para as 100 séries sub-rogadas. A série $\hat{x}_c(k)$ foi dividida em 335 observações (≈ 28 anos) para a identificação e as 194 observações restantes (≈ 16 anos) para a validação dos modelos. Para uma estrutura definida pelos valores máximos $\ell = 2$, $n_y = 10$, $n_e = 5$, $n_t = 20$, o modelo escolhido para a série residual, $\hat{x}_c(k)$, é dado por:

$$\begin{aligned}
\hat{x}_c(k) = & 1,23\hat{x}_c(k-1) - 0,37\hat{x}_c(k-2) - 0,0012\hat{x}_c(k-3)\hat{x}_c(k-1) - 0,25\hat{x}_c(k-10) + \\
& + 0,20\hat{x}_c(k-9) - 0,0003\hat{x}_c(k-9)\hat{x}_c(k-7) + 0,0003\hat{x}_c(k-1)\hat{x}_c(k-1) + \\
& + 0,0014\hat{x}_c(k-3)\hat{x}_c(k-2) - 0,0005\hat{x}_c(k-3)\hat{x}_c(k-3) + \\
& + 0,0003\hat{x}_c(k-10)\hat{x}_c(k-7) - 0,0016\hat{x}_c(k-10)\hat{x}_c(k-2) + \\
& + 0,0010\hat{x}_c(k-9)\hat{x}_c(k-1) - 0,0004\hat{x}_c(k-2)\hat{x}_c(k-2) + \\
& + 0,0007\hat{x}_c(k-10)\hat{x}_c(k-3) + 16,70.
\end{aligned} \tag{6.11}$$

Os modelos NARMA para as 100 séries sub-rogadas foram estimados usando a mesma estrutura de (6.11), ou seja, os modelos permaneceram com os mesmos termos e, estimou-se novos parâmetros para cada termo. Por meio dos modelos escolhidos, realizou-se a previsão de 1-10 passos à frente.

A Figura 6.10, em seu gráfico inferior, ilustra o desempenho das previsões para a série residual $\hat{x}_c(k)$ e para as séries sub-rogadas por meio dos índices mRSE calculados para cada passo de previsão. A partir desta figura testa-se a H_0 . Verifica-se que o índice mRSE para a série residual, $\hat{x}_c(k)$, é menor do que o índice mRSE médio para as séries sub-rogadas para um horizonte de no máximo 2 passos à frente. Dessa forma, rejeita-se H_0 e aceita-se H_1 , em que se infere, então, que a série $\hat{x}_c(k)$ tem vínculo determinístico de curto-prazo, além dos ciclos que foram removidos. Conseqüentemente, conjectura-se que a série quase-periódica de casos de caxumba $y_c(k)$ apresenta traço de determinismo intra-ciclo.

Pedir a opiniao do Eriva e citar sua tese.

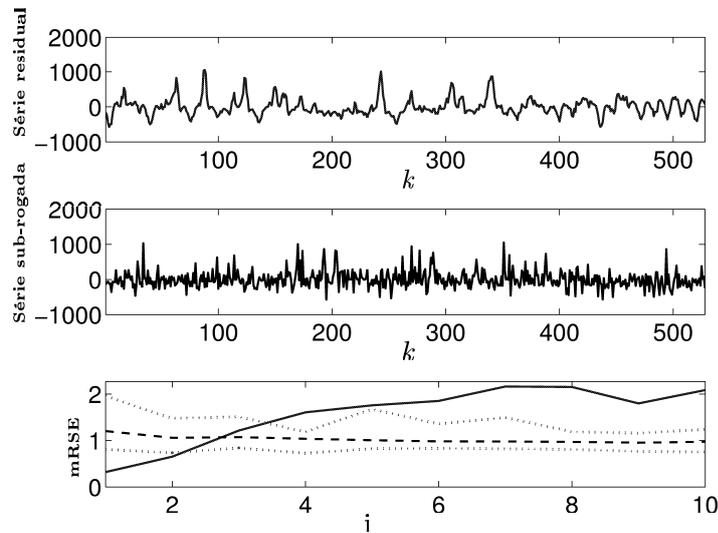


Figura 6.10: Teste de determinismo para a série de índices mensais de caxumba $y_c(k)$, a partir de sua respectivamente série de resíduos $\hat{x}_c(k)$ dada em (6.10). No gráfico superior a série residual $\hat{x}_c(k)$. No gráfico intermediário uma das 100 séries sub-rogadas. No gráfico inferior índices mRSE para $\hat{x}_c(k)$ (—) e, o índice mRSE médio para as séries sub-rogadas (- - -), além da faixa de confiança de 95% em (...).

6.3 Produção mensal de leite

A série quase-periódica de produção mensal de leite, $y_L(k)$, registrada durante os anos de 1962-1975 é mostrada na Figura 6.11(a).⁶

Percebe-se pela figura que a série $y_L(k)$ apresenta uma componente de tendência além de ciclos, os quais são salientados pela FAC e pelo periodograma da série $y_L(k)$ nas Figuras 6.11 (c) e (d), respectivamente. Por meio da FAC e do periodograma verifica-se que os períodos cíclicos de $y_L(k)$ são $d_1 = 6$ e $d_2 = 12$, os quais referem-se à ciclos sazonais.

Aplicando o FMMS à série quase-periódica $y_L(k)$, obteve-se a série residual correspondente $\hat{x}_L(k)$ ilustrada na Figura 6.11 (b), por meio da qual, percebe-se que as componentes cíclicas e de tendência foram significativamente reduzidas. A

⁶A série de produção de leite (milk.dat) pode ser encontrada no na subpasta Agriculture no sítio: <http://www-personal.buseco.monash.edu.au/%7Ehyndman/TSDL/>

comparação da FAC e do periodograma, entre a série quase-periódica $y_L(k)$ e a série residual $\hat{x}_L(k)$, nas Figuras 6.11 (c) e (d), confirmam esse resultado.

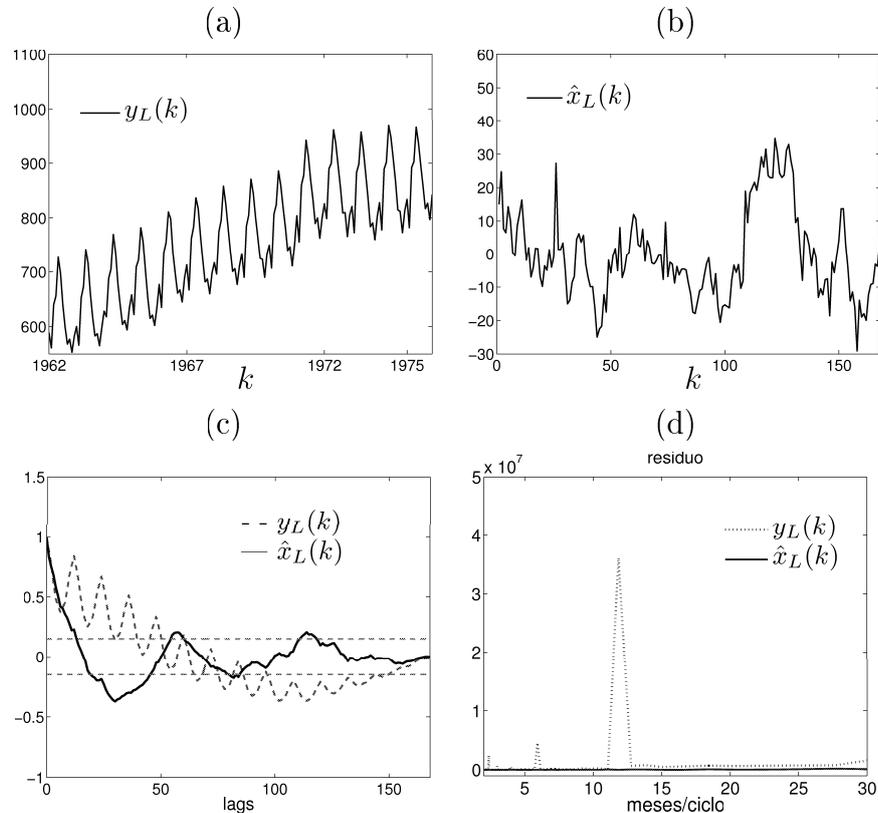


Figura 6.11: Em (a) série temporal periódica de produção mensal de leite $y_L(k)$ com ciclos em $d_1 = 6$ e em $d_2 = 12$. Em (b) série residual $\hat{x}_L(k)$ correspondente. Em (c) e (d), respectivamente, FAC e periodograma de $y_L(k)$ (- -) e de $\hat{x}_L(k)$ (—).

A partir da série residual $\hat{x}_L(k)$ foram geradas 100 séries sub-rogadas usando o algoritmo de embaralhamento temporal (Theiler et al., 1992), sendo que uma das 100 séries artificiais está representada no gráfico intermediário da Figura 6.12.

A série $\hat{x}_L(k)$ com 168 observações, na Figura 6.11 (b) foi, então, dividida em dois conjuntos: um contendo as 118 observações iniciais para a identificação e as 51 observações restantes para a validação dos modelos NARMA polinomiais.

Para uma estrutura com limites máximos definidos por: $\ell = 2$, $n_y = 8$, $n_e = 5$ e $n_t = 4$, foram estimados vários modelos NARMA polinomiais. O modelo que apresentou o menor índice mRSE médio para o primeiro passo de previsão foi o

escolhido. Para a série residual $\hat{x}_L(k)$, o modelo AR polinomial escolhido foi:

$$\hat{x}_L(k) = 0,74\hat{x}_L(k-1) + 0,029\hat{x}_L(k-2). \quad (6.12)$$

Para as 100 séries sub-rogadas utilizou-se a mesma estrutura do modelo (6.12), no entanto, com parâmetros diferentes, que foram reestimados para cada uma das séries sub-rogadas.

A partir dos modelos encontrados para $\hat{x}_L(k)$ e para as 100 séries artificiais, realizou-se a previsão de 1-10 passos à frente e calculou-se os índices mRSE correspondentes, conforme mostra o gráfico inferior da Figura 6.12. Por meio deste gráfico testa-se a hipótese nula (H_0).

Verifica-se que o índice mRSE para $\hat{x}_L(k)$ é menor e está fora da faixa de confiança para todo o horizonte de previsão, indicando que o desempenho da previsão de $\hat{x}_L(k)$ é melhor do que o desempenho médio das previsões para as séries artificiais. Dessa forma, pode-se rejeitar H_0 e aceitar H_1 , inferindo assim, que $\hat{x}_L(k)$ apresenta vínculo determinístico de curto-prazo, além dos ciclos que foram removidos. Conseqüentemente, a série quase-periódica $y_L(k)$ pode apresentar assinatura determinística intra-ciclo.

Deve-se destacar que a FAC da série residual $\hat{x}_L(k)$ na Figura 6.11(c) apresenta uma correlação entre as primeiras 20 observações o que poderia ser um indício de vínculo determinístico na série $\hat{x}_L(k)$.

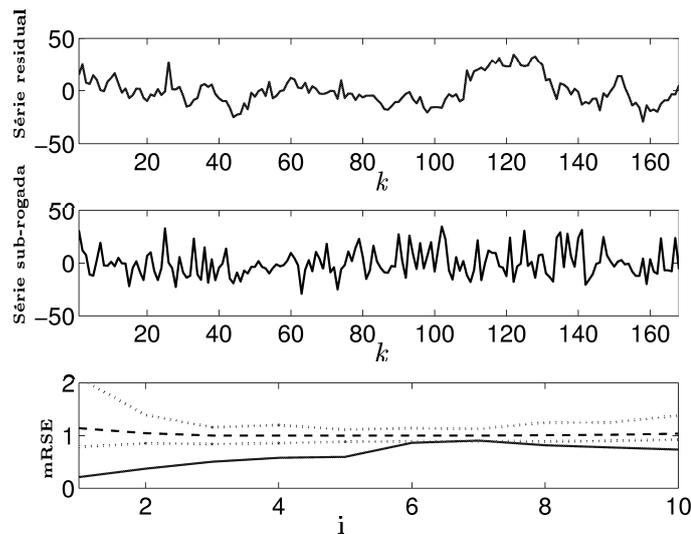


Figura 6.12: Teste de determinismo para a série de produção mensal de leite, $y_L(k)$, realizado a partir de sua correspondente série de resíduos $\hat{x}_L(k)$. No gráfico superior série residual $\hat{x}_L(k)$. No gráfico intermediário uma das 100 séries sub-rogadas. No gráfico inferior, os índices mRSE para $\hat{x}_L(k)$ (—) e os índices mRSE médio para as séries sub-rogadas (- - -), além da faixa de confiança de 95% representada em (...).

6.4 Série de números anuais de manchas solares

O teste de determinismo foi aplicado à série quase-periódica, $y_{s_a}(k)$, de números anuais de manchas solares (*sunspots*), representada na Figura 6.13(a).⁷ Essa série apresenta ciclos com $d_1 \approx 95$, $d_2 = 11$ e $d_3 = 57$ evidenciados na FAC e no periodograma de $y_{s_a}(k)$ nas Figuras 6.13(b) e (c), respectivamente.⁸

De maneira similar às séries anteriores, aplicou-se o FMMC à série quase-periódica $y_{s_a}(k)$, com $N = 289$ observações, para valores de $d_1 = 96$ e $d_2 = 11$. No entanto, essa redução dos ciclos foi realizada em duas etapas descritas a seguir.

⁷A série quase-periódica de números anuais de manchas solares ou *sunspots* (`sunspot.dat`) pode ser encontrada na subpasta `Physics` no sítio: <http://www-personal.buseco.monash.edu.au/%7Ehyndman/TSDL/>

⁸A componente cíclica com período $d_3 = 57$ foi desconsiderada neste estudo, uma vez que, por meio de resultados obtidos, percebeu-se que a redução do ciclo com $d_3 = 57$ implica em uma pior redução da componente cíclica com $d_2 = 11$.

1ª etapa: redução dos ciclos usando FMCC com $d_1 = 96$

Descrevendo a série quase-periódica de números anuais de manchas solares, $y_{s_a}(k)$, como uma composição somente das componentes residual e de ciclo com $d_1 = 96$, sem considerar, a princípio, a componente de tendência, tem-se:

$$y_{s_a}(k) = c_{1_{s_a}}(k) + x_{1_{s_a}}(k).$$

Para o ciclo com $d_1 = 96$, tem-se $q_1 = 96/2 = 48$, logo:

$$\hat{m}_{0_{96}}(k) = (0,5y_{s_a}(k - 48) + y_{s_a}(k - 47) + \dots + y_{s_a}(k + 47) + 0,5y_{s_a}(k + 48))/96.$$

Para cada $k_0 = 1, \dots, 96$, estimou-se a média $\hat{w}_{96}(k_0)$ da diferença dada por:

$$\hat{w}_{96}(k_0) = \left\{ (y(k_0 + 96j) - \hat{m}_{0_{96}}(k_0 + 96j)), \quad 96 < k_0 + 96j \leq (N - 48) \right\}.$$

sendo $N = 289$, $j = 1, \dots, n_c$ e n_c o número de ciclos com $d_1 = 96$ existentes na série quase-periódica $y_{s_a}(k)$. Dessa forma, a primeira componente cíclica, $\hat{c}_{1_{s_a}}(k)$, pôde ser estimada:

$$\hat{c}_{1_{s_a}}(k) = \hat{w}_{96}(k) - \frac{1}{96} \sum_{i=1}^{96} \hat{w}_{96}(i),$$

$k = 1, \dots, 96$ e $\hat{c}_{1_{s_a}}(k) = \hat{c}_{1_{s_a}}(k - 96)$, $k > 96$.

A série residual, $\hat{x}_{1_{s_a}}(k)$ foi, então definida, extraindo-se a componente cíclica estimada, $\hat{c}_{1_{s_a}}(k)$, da série quase-periódica, $y_{s_a}(k)$:

$$\hat{x}_{1_{s_a}}(k) = y_{s_a}(k) - \hat{c}_{1_{s_a}}(k), \quad k = 1, \dots, N. \quad (6.13)$$

A série residual, $\hat{x}_{1_{s_a}}(k)$, assim obtida, está representada na Figura 6.13(d). Percebe-se pela sua FAC e pelo seu periodograma - mostrados, respectivamente, nas Figuras 6.13(e) e (f) - que houve uma redução significativa da componente referente ao ciclo com período $d_1 = 96$, e que, o segundo ciclo com período $d_2 = 11$, não foi

reduzido. Dessa forma, repete-se a aplicação do FMMC à série residual $\hat{x}_{1_{s_a}}(k)$ dada em (6.13) usando-se $d_2 = 11$.

2ª etapa: redução dos ciclos usando FMMC com $d_2 = 11$

A série residual estimada, $\hat{x}_{1_{s_a}}(k)$, pode ser considerada como sendo composta pelas seguintes componentes:

$$\hat{x}_{1_{s_a}}(k) = c_{2_{s_a}}(k) + m_{2_{s_a}}(k) + x_{2_{s_a}}(k).$$

Percebe-se que, nesta segunda etapa, foi considerada a componente de tendência $m_{2_{s_a}}(k)$, um segundo ciclo $c_{2_{s_a}}(k)$ com $d_2 = 11$ e uma segunda componente residual $x_{2_{s_a}}(k)$.

Para o ciclo com período $d_2 = 11$, tem-se $q_2 = (11 - 1)/2 = 5$, logo:

$$\hat{m}_{0_{11}}(k) = (0,5y_{s_a}(k - 5) + y_{s_a}(k - 4) + \dots + y_{s_a}(k + 4) + 0,5y_{s_a}(k + 5))/11.$$

Para cada $k_0 = 1, \dots, 11$, estimou-se a média $\hat{w}_{11}(k_0)$ da diferença dada por:

$$\hat{w}_{11}(k_0) = \{ (y(k_0 + 11j) - \hat{m}_{0_{11}}(k_0 + 11j)), \quad 11 < k_0 + 11j \leq (N - 5) \}.$$

em que $N = 289$, $j = 1, \dots, n_c$, sendo n_c o número de ciclos com $d_2 = 11$ existentes ao longo da série $\hat{x}_{1_{s_a}}(k)$ dada em (6.13). Dessa forma, a componente cíclica, $\hat{c}_{2_{s_a}}(k)$, pôde ser estimada da seguinte maneira:

$$\hat{c}_{2_{s_a}}(k) = \hat{w}_{11}(k) - \frac{1}{11} \sum_{i=1}^{11} \hat{w}_{11}(i),$$

$k = 1, \dots, 11$ e $\hat{c}_{2_{s_a}}(k) = \hat{c}_{1_{s_a}}(k - 11)$, $k > 11$.

A série sem o ciclo com $d_2 = 11$, $\hat{D}_2(k)$, foi definida extraindo-se a componente cíclica estimada, $\hat{c}_{2_{s_a}}(k)$, da série $\hat{x}_{1_{s_a}}(k)$ dada em (6.13):

$$\hat{D}_2(k) = \hat{x}_{1_{s_a}}(k) - \hat{c}_{2_{s_a}}(k), \quad k = 1, \dots, N. \quad (6.14)$$

A componente de tendência da série, $\hat{D}_2(k)$, foi estimada ajustando-se um polinômio de ordem dois, conforme descrito na Seção 3.3.3. Dessa forma, tem-se:

$$\hat{m}_{2_{s_a}}(k) = 58,56k^2 - 40,90k + 49,76.$$

Em termos das componentes cíclica e de tendência estimadas, $\hat{c}_{2_{s_a}}(k)$ e $\hat{m}_{2_{s_a}}(k)$, respectivamente, a nova série residual, $\hat{x}_{2_{s_a}}(k)$, foi obtida como:

$$\hat{x}_{2_{s_a}}(k) = \hat{x}_{1_{s_a}}(k) - \hat{m}_{2_{s_a}}(k) - \hat{c}_{2_{s_a}}(k), \quad k = 1, \dots, N. \quad (6.15)$$

A série residual, $\hat{x}_{2_{s_a}}(k)$, obtida nessa segunda etapa de aplicação do FMMC está representada na Figura 6.13(h). Percebe-se pela comparação entre a FAC e o periodograma das séries: quase-periódica ($y_{s_a}(k)$) e residual ($\hat{x}_{2_{s_a}}(k)$), que o ciclo com $d_1 = 11$ foi apenas reduzido e não eliminado. Este fato deve ser considerado na ocasião de se escolher a máxima estrutura dos modelos NARMA polinomiais, para que estes não incorporem a informação devido ao ciclo com $d_1 = 11$.

Portanto, os valores máximos para as estruturas dos modelos são: $\ell = 2$, $n_y = 8$, $n_e = 5$ e $n_t = 10$. O modelo AR polinomial escolhido para a série residual, $\hat{x}_{2_{s_a}}(k)$, foi:

$$\hat{x}_{2_{s_a}}(k) = 1,27\hat{x}_{2_{s_a}}(k-1) - 0,708\hat{x}_{2_{s_a}}(k-2). \quad (6.16)$$

Os modelos AR polinomial para as 100 séries sub-rogadas têm a mesma estrutura do modelo (6.16), porém, os parâmetros foram reestimados para cada uma das séries artificiais.

Os resultados das previsões para a série residual, $\hat{x}_{2_{s_a}}$, e para as séries sub-rogadas estão representados na Figura 6.14, por meio da qual, verifica-se que o desempenho das previsões de $\hat{x}_{2_{s_a}}(k)$ é melhor do que o desempenho médio das previsões das séries sub-rogadas para os dois primeiros passos de previsão. Por conseguinte, rejeita-se H_0 a favor de H_1 , inferindo assim, vínculo determinístico de

curto-prazo na série residual $\hat{x}_{2_{s_a}}(k)$, além dos ciclos que foram removidos. Assim, conjectura-se que a série quase-periódica de números anuais de manchas solares, $y_{s_a}(k)$, apresenta assinatura determinística intra-ciclo.

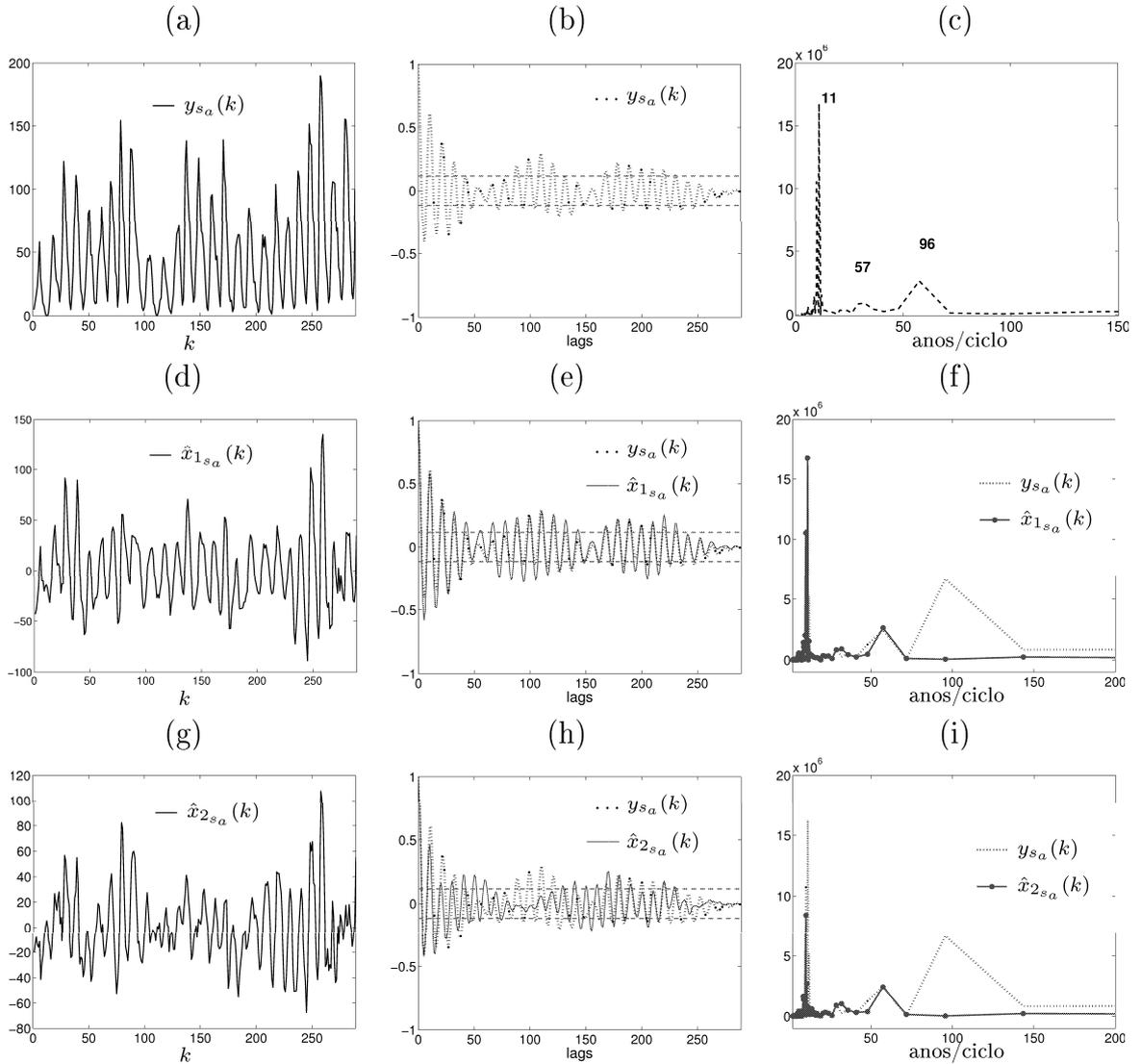


Figura 6.13: Em (a) série quase-periódica de números anuais de manchas solares (*sunspots*), $y_{s_a}(k)$, com ciclos em períodos $d_1 = 96$ e $d_2 = 11$. Em (b) FAC de $y_{s_a}(k)$. Em (c) periodograma de $y_{s_a}(k)$. Em (d) série residual correspondente, $\hat{x}_{1_{s_a}}(k)$, com o ciclo em $d_1 = 96$ reduzido. Em (e) e (f) uma comparação entre a FAC e o periodograma de $y_{s_a}(k)$ (- -) e de $\hat{x}_{1_{s_a}}(k)$ (—), respectivamente. Em (g) série residual correspondente, $\hat{x}_{2_{s_a}}(k)$, com os ciclos em $d_1 = 96$ e $d_2 = 11$ reduzidos. Em (h) e (i) uma comparação entre a FAC e o periodograma de $y_{s_a}(k)$ (- -) e de $\hat{x}_{2_{s_a}}(k)$ (—), respectivamente.

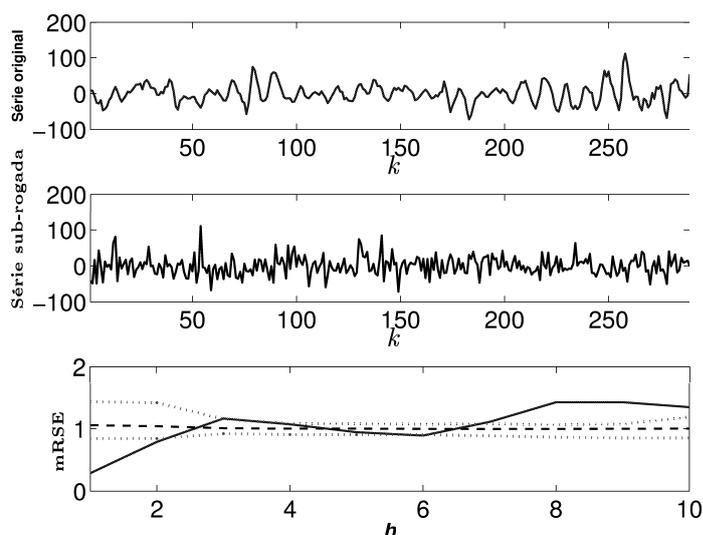


Figura 6.14: Teste de determinismo para a série quase-periódica, $y_{s_a}(k)$, de números anuais de manchas solares (*sunspots*) realizado a partir de sua respectiva série residual $\hat{x}_{2s_a}(k)$. No gráfico superior a série residual $\hat{x}_{2s_a}(k)$. No gráfico intermediário, uma das 100 séries sub-rogadas. No gráfico inferior, os índices mRSE para $\hat{x}_{2s_a}(k)$ (—) e os índices mRSE médio para as séries sub-rogadas (- - -). Além da faixa de confiança de 95% representada em (...).

6.5 Série de números diários de manchas solares

A série de número diários de manchas solares (*sunspots*), $y_{s_D}(k)$, usada nesta seção, tem um período de amostragem diferente da anterior, ou seja, diário. A princípio, sabe-se que séries de manchas solares apresentam um padrão cíclico de 11 anos. Como a série quase-periódica $y_{s_D}(k)$ foi amostrada em dias, essa apresentará a cada 11 anos quantidades diferentes de anos bissextos, portanto, quantidades diferentes de observações em cada ciclo. Isso implica em períodos que variam ao longo da série.⁹

O método FMDC foi aplicado à série quase-periódica, $y_{s_D}(k)$, para o período $d_1 = 4.017$. Esse período está destacado no periodograma da série $y_{s_D}(k)$ na Figura 6.15 (d). Assim, obtém-se a série residual correspondente $\hat{x}_{s_D}(k)$ ilustrada na Figura 6.15(b). Uma comparação visual da FAC e do periodograma, nas Figuras 6.15(c) e (d), mostra que a série residual $\hat{x}_{s_D}(k)$ não apresenta o mesmo ciclo com período

⁹Os ciclos associados à 11 anos, ora terá 4.017 dias ($11\text{anos} \times 365\text{dias} + 2\text{dias}$) devido aos dois anos bissextos ocorridos ao longo desses 11 anos. Ora, terá 4.018 dias ($11\text{anos} \times 365\text{dias} + 3\text{dias}$) devido aos três anos bissextos ocorridos ao longo de outros 11 anos.

$d_1 = 4.017$ de $y_{s_D}(k)$.

Foram geradas 100 séries sub-rogadas a partir da série residual $\hat{x}_{s_D}(k)$, sendo que uma das 100 está ilustrada nos gráficos intermediários da Figura 6.16. A série residual e as séries sub-rogadas foram divididas em quatro janelas de dados abrangendo faixas de dados conforme a Tabela 6.2.

Para cada uma das janelas de dados, procurou-se por modelos com estruturas máximas definidas como: $\ell = 2$, $n_y = 20$, $n_e = 10$ e $n_t = 5$. Em cada janela de dados foram usadas as 400 primeiras observações para a identificação dos modelos e as 101 restantes na validação dos modelos. O modelo escolhido, para cada janela, foi aquele que apresentou o menor índice mRSE médio para o primeiro passo de previsão, conforme mostra a Tabela 6.2.

A partir dos modelos escolhidos, em cada janela, para a série residual $\hat{x}_{s_D}(k)$ e para as séries artificiais, fez-se a previsão de 1-10 passos a frente e calculou-se os índices mRSE, os quais estão ilustrados nos gráficos inferiores da Figura 6.16.

Para as janelas 1 e 2 o desempenho da previsão da série residual $\hat{x}_{s_D}(k)$ foi pior do que o desempenho médio das séries sub-rogadas, conforme Figura 6.16. Porém, como definido anteriormente, nesse caso, rejeita-se H_0 em favor de H_2 . Portanto, conjectura-se que a série de números diários de manchas solares $y_{s_D}(k)$, ao longo das janelas 1 e 2, é não-estacionária ou não modelável por NARMA polinomial.

No entanto, infere-se que há traços determinísticos de curto-prazo na série residual $\hat{x}_{s_D}(k)$, ao longo das janelas 3 e 4. Uma vez que o desempenho das previsões da série residual $\hat{x}_{s_D}(k)$ no decorrer dessas janelas é melhor do que a previsão média das séries sub-rogadas, veja a Figura 6.16. Dessa forma, rejeita-se H_0 e aceita-se H_1 . Portanto, infere-se que há determinismo intra-ciclo na série de números diários de manchas solares $y_{s_D}(k)$, ao longo das janelas 3 e 4.

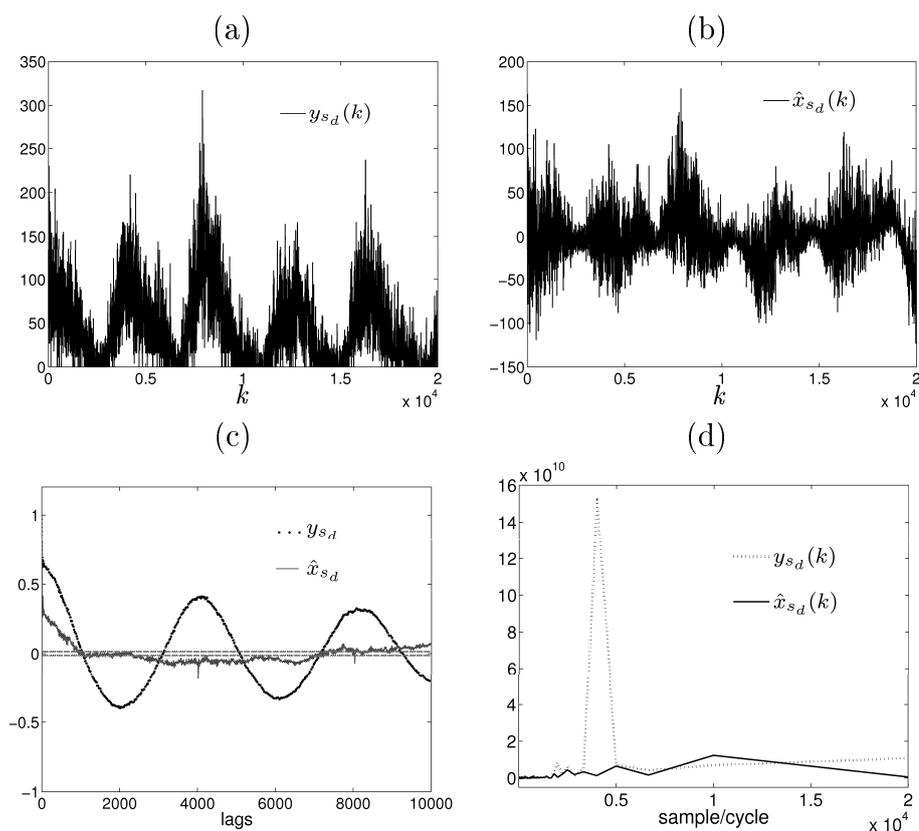


Figura 6.15: Em (a) série de números diários de manchas solares (*sunspots*), $y_{s_D}(k)$. Em (b) série residual correspondente, $\hat{x}_{s_D}(k)$. Em (c) FAC e em (d) periodograma de $y_{s_D}(k)$ em (- -) e de $\hat{x}_{s_D}(k)$ em (—), respectivamente.

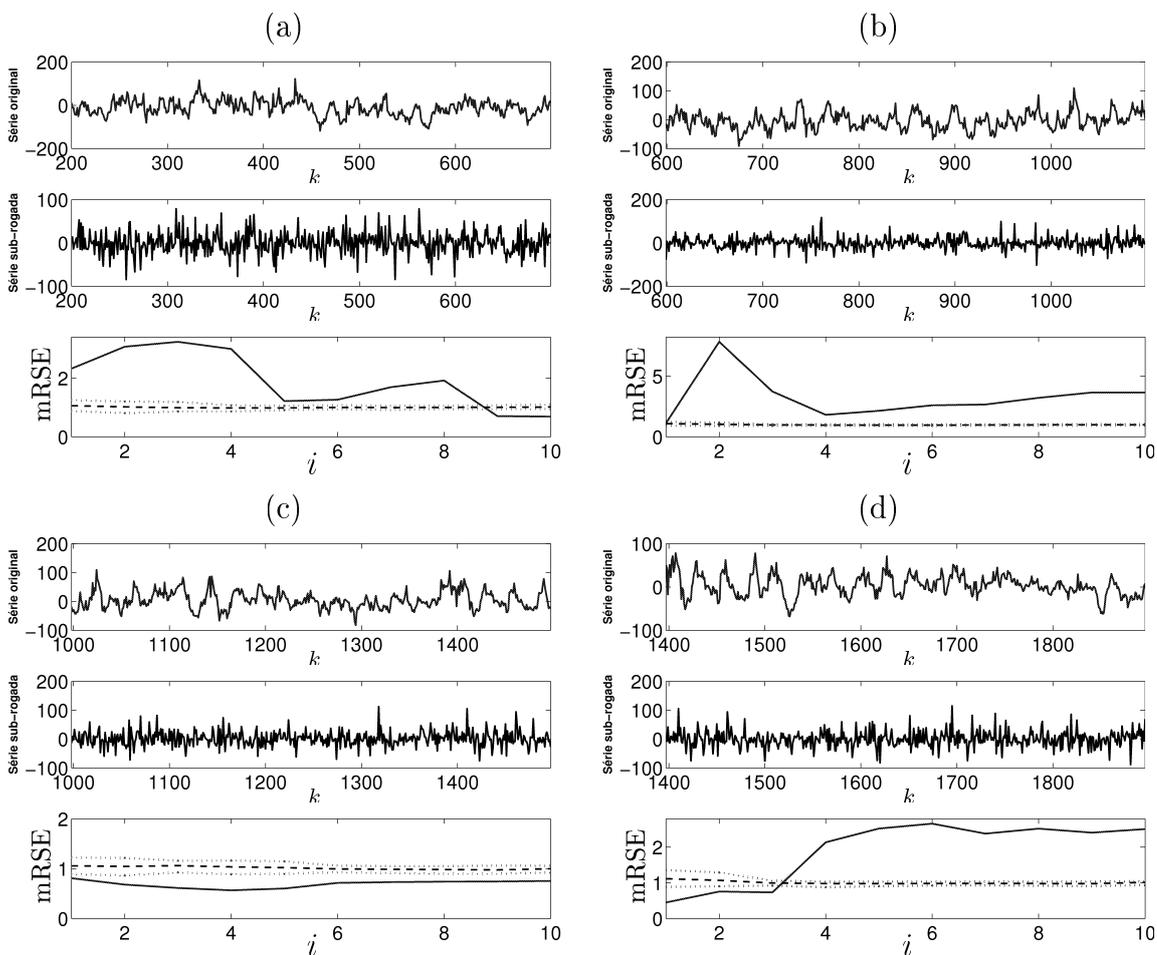


Figura 6.16: Teste de determinismo para as quatro janelas distintas ao longo da série de resíduos $\hat{x}_{s_D}(k)$. Para cada janela, no gráfico superior a série residual $\hat{x}_{s_D}(k)$. No gráfico intermediário uma das 100 séries sub-rogadas. No gráfico inferior os índices mRSE para $\hat{x}_{s_D}(k)$ em (—) e o índice mRSE médio para as séries sub-rogadas em (- -). Além da faixa de confiança de 70% em (...).

Tabela 6.2: Modelos NARMA polinomiais usados na previsão e os resultados do teste de determinismo para cada uma das quatro janelas de dados da série residual $\hat{x}_{s_D}(k)$ testadas.

janela	faixa de dados	modelo NARMA polinomial para a série residual	indício de vínculo determinístico
1	200 - 599 - 699	$\hat{x}_{s_D}(k) = 0,73\hat{x}_{s_D}(k-1) - 0,0029\hat{x}_{s_D}(k-1)^2$ $- 0,0044\hat{x}_{s_D}(k-14)\hat{x}_{s_D}(k-20)$ $+ 0,004\hat{x}_{s_D}(k-13)\hat{x}_{s_D}(k-20)$ $- 0,0015\hat{x}_{s_D}(k-2)\hat{x}_{s_D}(k-12)$	—
2	599 - 998 - .098	$\hat{x}_{s_D}(k) = 0,75\hat{x}_{s_D}(k-1) + 0,5$ $- 0,0031\hat{x}_{s_D}(k-1)\hat{x}_{s_D}(k-10)$ $0,0053\hat{x}_{s_D}(k-9)\hat{x}_{s_D}(k-17)$ $0,0015\hat{x}_{s_D}(k-8)\hat{x}_{s_D}(k-10)$	—
3	998 - 1.397 - 1.497	$\hat{x}_{s_D}(k) = 0,74\hat{x}_{s_D}(k-1) - 0,16\hat{x}_{s_D}(k-10)$ $+ 0,0028\hat{x}_{s_D}(k-10)\hat{x}_{s_D}(k-15)$ $+ 0,096\hat{x}_{s_D}(k-2)$	sim
4	1.397 - 1.796 - 1.896	$\hat{x}_{s_D}(k) = 0,86\hat{x}_{s_D}(k-1) + 2,2$ $- 0,1301\hat{x}_{s_D}(k-9) - 0,0024\hat{x}_{s_D}(k-1)^2$ $- 0,0023\hat{x}_{s_D}(k-1)\hat{x}_{s_D}(k-6)$	sim

6.6 Comentários Finais

Neste capítulo, o teste de determinismo intra-ciclo para séries temporais quase-periódicas, descrito no capítulo 5, foi aplicado a séries reais de carga, de casos de caxumba, produção de leite, números anuais e diários de manchas solares, ou número de *sunspots*. A primeira série a ser analisada foi a série de carga $y_Q(k)$ na Seção 6.1.

Série de carga $y_Q(k)$

Para a série de carga $y_Q(k)$ foi feita uma análise de estacionariedade de modo a dividir a série em janelas estacionárias. Segundo os testes usados neste capítulo, pode-se considerar janelas com 240 observações como janelas estacionárias. Dessa forma, utilizou-se seis janelas de dados contendo 240 observações, conforme mostra a Tabela 6.1.

O teste foi realizado sob duas perspectivas: a primeira usando a série de carga (com os padrões quase-periódicos) e a segunda, usando a série residual (sem os padrões quase-periódicos), empregando os algoritmos SSS (Nakamura & Small,

2005) e de embaralhamento temporal (Theiler et al., 1992), respectivamente, para gerar as séries artificiais.

Para a série quase-periódica, o vínculo determinístico foi testado analisando as FAC da série quase-periódica e das séries artificiais à semelhança de Nakamura & Small (2005). Para as seis janelas de dados analisadas, o teste indicou vínculo determinístico de curto-prazo em todas, conforme mostra a Figura 6.5.

Para aplicar o teste de determinismo intra-ciclo, proposto neste trabalho, foi empregado o método FMMC, o qual foi capaz de reduzir significativamente ciclos com períodos $d_2 = 168$, $d_3 = 84$ e $d_4 = 12$. Porém, ciclo diário com período $d_1 = 24$ foi apenas reduzido. Outros métodos de redução de ciclos obtiveram resultados similares ao método FMMC. Conjectura-se que esse ciclo com $d_1 = 24$ presente na série de residual não seja separável.

A escolha do modelo NARMA polinomial levou esse fato em consideração. Dessa forma, o modelo NARMA foi escolhido entre um conjunto de modelos com atraso máximo $n_y = 11 < 24$. O teste detectou indício de vínculo determinístico intra-ciclo nas janelas 1, 3, 5 e 6, além dos ciclos removidos. A Tabela 6.1 resume os resultados encontrados.

Destaca-se ainda, que na série de resíduos algumas janelas de dados apresentaram um desempenho de previsão da série residual, $\hat{x}_{Q_{168}}(k)$, estatisticamente diferente, porém, inferior ao desempenho médio das previsões obtido para as séries sub-rogadas. Em outras palavras, os índices mRSE para a série residual eram maiores do que os índices mRSE médios das séries artificiais, o que parece indicar problemas na modelagem, apesar de ter sido buscado o melhor modelo dentre a estrutura máxima.

Série de casos mensais de caxumba $y_C(k)$

Foi possível extrair a componente quase-periódica com períodos $d_1 = 12$ e $d_2 = 31$ da série de casos mensais de caxumba $y_C(k)$ usando o método FMMC. O teste de determinismo proposto aponta a existência de vínculo determinístico de

curto-prazo na série residual ($\hat{y}_C(k)$), como mostra a Figura 6.10. Conseqüentemente, conjectura-se que a série quase-periódica $y_C(k)$ apresenta vínculo determinístico intra-ciclo além dos ciclos que foram removidos.

Série de produção mensal de leite $y_L(k)$

Na série de produção de leite $y_L(k)$, também, foi possível extrair as componentes cíclicas com períodos $d_1 = 6$ e $d_2 = 12$. Detectou-se vínculo determinístico de curto-prazo na série residual ($\hat{x}_L(k)$) como ilustrado na Figura 6.12. Dessa forma, infere-se que a série de produção mensal de leite $y_L(k)$ apresenta traço determinístico de curto-prazo subjacente aos ciclos que foram removidos.

Série de números anuais de manchas solares ou de *sunspots*, $y_{s_a}(k)$

O método FMCC não conseguiu reduzir significativamente a componente cíclica com período $d_1 = 11$ da série de números anuais de manchas solares (*sunspots*), $y_{s_a}(k)$. Assim, escolheu-se o máximo atraso dos possíveis modelos NARMA polinomiais como sendo $n_y = 10 < 11$. O teste apontou para a existência de traço determinístico de curto-prazo na série residual correspondente, $\hat{x}_{2_{s_a}}(k) y_{s_a}(k)$. Portanto, infere-se que a série de números anuais de manchas solares, $y_{s_a}(k)$, tem vínculo determinístico de curto-prazo além dos ciclos que foram removidos. O que é um resultado consistente com o trabalho (Letellier, Aguirre, Maquet & Gilmore, 2005).

Da mesma forma que aconteceu com a série de carga $y_Q(k)$, não se sabe ao certo qual a influência na detecção de determinismo na série $y_{s_a}(k)$ devido à periodicidade restante na série residual.

Série de números diários de manchas solares ou de *sunspots*, $y_{s_D}(k)$

A quase-periodicidade da série de números diários de manchas solares, $y_{s_D}(k)$, foi significativamente reduzida usando o método FMCC, como mostra a Figura 6.15.

A série $y_{s_D}(k)$ foi dividida em quatro janelas com 500 observações cada e o teste de determinismo foi aplicado a cada uma das janelas e constatou-se traço

determinístico de curto-prazo na série residual, $\hat{x}_{s_D}(k)$, ao longo das janelas 3 e 4. Portanto, infere-se que há vínculo determinístico na série de números diários de manchas solares $y_{s_D}(k)$, além dos ciclos que foram removidos.

DISCUSSÃO E SUGESTÕES PARA A CONTINUIDADE DO TRABALHO

Prever e entender, por meio de valores passados, a dinâmica subjacente a uma dada série temporal é de grande interesse em diversas áreas. No entanto, esta não é uma tarefa fácil, principalmente quando se trata de séries temporais com comportamento quase-periódico. Este comportamento tende a mascarar qualquer outro comportamento determinístico, que possivelmente exista, além dele. Tal fato pode comprometer a qualidade da análise e da previsão de tais séries.

Baseado na hipótese de que previsibilidade implica em algum determinismo, este trabalho apresentou uma versão adaptada do teste de determinismo proposto por Gomes, Souza, Guimarães & Aguirre (2000). O novo procedimento difere no objetivo inicialmente proposto em (Gomes et al., 2000). O novo procedimento procura testar determinismo intra-ciclo em séries quase-periódicas. Assim, seguem-se algumas diferenças entre os métodos:

- a série temporal deve ser pré-processada com o intuito de reduzir/extrair as componentes quase-periódicas;
- em (Gomes et al., 2000) emprego-se os algoritmos FT e o AAFT. O algoritmo, aqui usado, para gerar dados sub-rogados é o baseado em embaralhamento temporal *Shuffled*.
- em (Gomes et al., 2000), o fator estatístico RMSE proporciona uma distinção

de quais janelas ao longo das séries pode-se inferir a existência de traços determinísticos. O fator estatístico discriminante usado neste trabalho, o índice mRSE, proporciona uma visão do horizonte de previsão das séries temporais, visto que se procura por determinismo de curto-prazo nas séries.

7.1 Discussões e contribuições

Até a década passada os pesquisadores procuravam por determinismo em séries temporais não periódicas (Sugihara & May, 1990; Theiler et al., 1992; Kaplan & Glass, 1992; Schreiber, 2000; Gomes et al., 2000). No entanto, sabe-se que diversas séries reais em áreas distintas apresentam comportamento quase-periódico. Para lidar com esses tipos de séries, destacam-se alguns trabalhos como os algoritmos CSS¹ (Theiler, 1995) e PPS² (Small et al., 2001), os quais procuram por determinismo entre-ciclos de séries quase-periódicas.

Recentemente, Nakamura e Small (2005) apresentaram o método SSS para detectar determinismo intra-ciclo em séries quase-periódicas. O teste de determinismo proposto por Nakamura e Small (2005) tem como principal vantagem o fato de poder ser aplicado diretamente à série quase-periódica. No entanto, como visto na Seção 5.3, o teste parece apresentar resultados espúrios ao inferir determinismo de curto-prazo na série quase-periódica com comportamento aleatório subjacente (veja a Figura 5.2). Esse fato pode ter ocorrido devido à escolha empírica do parâmetro A usado no algoritmo SSS, descrito na Seção 3.6.2 para gerar as séries sub-rogadas. Segundo Nakamura & Small (2005) $A = 1$ é um valor adequado e, esse foi o valor adotado neste trabalho.

Como pode ser observado, os testes de determinismo em séries quase-periódicas, até então propostos, sugerem uma nova maneira de se gerar dados sub-rogados. O teste proposto neste trabalho, procura conciliar o que já se tem bem estabelecido na literatura. Este trabalho procurou verificar a presença de vínculo

¹*Cycle Shuffled Surrogate*

²*Pseudo-Periodic Surrogate*

determinístico intra-ciclo em séries temporais quase-periódicas, utilizando o teste de hipóteses baseado em dados sub-rogados (Theiler et al., 1992; Gomes, 2001).

Nesse contexto, procurou-se investigar a capacidade dos algoritmos apresentados por Theiler e colegas (1992), bastante usados na literatura, para gerar dados sub-rogados adequados ao teste de determinismo. Foi verificado, por meio de séries simuladas quase-periódicas, que o algoritmo mais indicado para o teste de determinismo é o baseado em embaralhamento temporal da série (*shuffled*), conforme mostra os resultados na Seção 5.3. Porém, antes de empregá-lo, deve-se extrair ou reduzir o comportamento quase-periódico da série temporal, uma vez que esse algoritmo foi desenvolvido para séries que não apresentam tal característica. Por meio dos resultados mostrados no Capítulo 5, verificou-se que os algoritmos que empregam a transformada de Fourier da série original para gerar dados sub-rogados mantêm algum traço de determinismo nas séries sub-rogadas.

Dentre os métodos usados para extrair os ciclos e tendências optou-se pelo método que emprega o filtro de média móvel para componentes cíclicas (FMMC). O método FMMC apresenta um bom desempenho em séries quase-periódicas simuladas, conforme mostram os Capítulos 3 e 5. Porém, no Capítulo 6 foi mostrado que em séries reais quase-periódicas o método FMMC nem sempre consegue reduzir significativamente as componentes cíclicas.

Uma análise empírica dos resultados obtidos com séries reais quase-periódicas, ao longo deste estudo, mostrou que os períodos dos ciclos não são sempre fixos, ou seja, esses podem variar ao longo da série. Tal fato pode ter comprometido os resultados da extração dos ciclos, em séries com períodos variantes, usando o método FMMC. Além disso, pôde-se perceber que as componentes da série - tendência, ciclos, sazonalidade e resíduo - nem sempre são separáveis, como no caso das séries de carga e da série de números anuais de manchas solares. Outro método testado para retirar os ciclos da série foi o método de agrupamento de padrões usando o k -médias. Contudo, os resultados obtidos foram semelhantes aos alcançados pelo método FMMC.

Não se sabe ao certo o que uma extração de ciclos pode acarretar na série temporal. No sentido de que este pode extrair não somente os ciclos, mas também, outro comportamento dinâmico da série temporal.

O procedimento para testar determinismo intra-ciclo, aqui proposto, baseia-se na hipótese de que alguma previsibilidade implica em algum determinismo. Logo, é necessário escolher um modelo adequado e índices de erro que quantifiquem a qualidade da previsão. A classe de modelos escolhida para o teste foi NARMA polinomial, a qual é bem estudada pelo grupo MACSIN³, além de existir várias rotinas implementadas no grupo.

No capítulo 4 fez-se um estudo diante de resultados simulados e verificou-se que o índice mRSE é apropriado para quantificar a qualidade da previsão. Além de quantificar o erro, o índice mRSE proporciona uma idéia da qualidade da previsão do modelo NARMA polinomial quando comparado a um modelo trivial de médias.

Baseado no índice mRSE calculado em séries quase-periódicas simuladas, estabeleceu-se a hipótese nula a ser testada, bem como as hipóteses alternativas.

Ainda no capítulo 4, foi feito um estudo sobre a distribuição aproximada dos índices mRSE e observou-se que, nem sempre, estes índices se apresentam na forma de uma distribuição Normal. Diante dos resultados encontrados, conjecturou-se que, quando a série residual apresentar algum traço determinístico, as distribuições dos índices mRSE para os primeiros passos de previsão não se apresentam na forma de uma distribuição Normal. Esse fato pode ocorrer devido às séries sub-rogadas manterem alguma característica determinística da série original e os modelos ajustados às séries sub-rogadas procurarem modelar ou aproximar essa característica determinística. Esse é um assunto que merece ser melhor investigado. No entanto, à medida que se aumenta o horizonte de previsão as distribuições do mRSE assemelham-se a uma normal. O cálculo do intervalo de confiança foi feito como o intervalo entre percentis, o qual pode ser empregado em qualquer tipo de distribuição.

Diante do intervalo entre percentis dos índices mRSE atribui-se um grau

³<http://www.cpdee.ufmg.br/macsin/>

de confiança ao teste de determinismo. Neste trabalho, o intervalo de confiança foi aumentado progressivamente a partir de 50%, até que o resultado do teste de determinismo não se altere do resultado obtido inicialmente.

O procedimento proposto para testar determinismo intra-ciclo baseado em dados sub-rogados foi avaliado em séries reais quase-periódicas de diversas áreas: série de carga, série de números de manchas solares, série de produção de leite e série de casos de caxumba.

Em suma, tem-se como principais contribuições deste trabalho:

- A escolha do algoritmo apropriado para gerar dados sub-rogados quando se objetiva testar determinismo baseado em previsibilidade.
- A definição do índice de erro apropriado ao teste, no caso o índice mRSE.
- A partir do índice mRSE estabelecer as hipóteses nula e alternativas para testar determinismo.
- A atribuição de um grau de confiança ao teste de determinismo proposto, usando o intervalo entre percentis, o qual independe do tipo de distribuição dos índices de erro usado.

7.2 Sugestões de Trabalhos Futuros

Como foi observado, o procedimento proposto para o teste de determinismo intra-ciclo, ainda permanece com algumas questões a serem melhor investigadas, tais como:

- verificar se as componentes cíclicas, de tendência e residual das séries estão associadas de forma aditiva ou multiplicativa;
- procurar um método capaz de reduzir completamente as componentes quase-periódicas;

- investigar a distribuição dos índices de erro de modo a compreender os diferentes tipos de distribuição ocorridas em diferentes horizontes de previsão;
- expandir o teste de hipótese baseado em dados sub-rogados para a análise de linearidade.

BIBLIOGRAFIA

- Abarbanel, H. D. (1993). The analysis observed chaotic data in physical systems. *Reviews of Modern Physics*, 65, 1331–1392.
- Aguirre, L. A. (1994). Term clustering and the order selection of linear continuous systems. *Journal of the Franklin Institute*, 331B(4), 403–415.
- Aguirre, L. A. (2004). *Introdução à Identificação de Sistemas - Técnicas Lineares e Não-Lineares Aplicadas a Sistemas Reais*. Belo Horizonte, MG: Editora da UFMG, 2ª edição.
- Aguirre, L. A. & Billings, S. A. (1995a). Dynamical effects of overparametrization in nonlinear models. *Physica D*, 80(1,2), 26–40.
- Aguirre, L. A. & Billings, S. A. (1995b). Improved structure selection for nonlinear models based on term clustering. *International Journal of Control*, 62(3), 569–587.
- Aguirre, L. A. & Mendes, E. M. A. M. (1996). Global nonlinear polynomial models: Structure, term cluster and fixedpoints. *Int. J. Bif. Chaos*, 6(2), 279–294.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.
- Amaral, G. F. V. (2006). *Síntese de Dinâmica Não-linear por meio de modelos afins por partes: Um método baseado em topologia*. PhD thesis, PPGEE - UFMG, Belo Horizonte.

- Bagarinao, E., Pakdaman, K., Nomura, T., & Sato, S. (1999). Reconstructing bifurcation diagrams from noisy time series using nonlinear autoregressive models. *Physical Review E*, 60(1), 1073–1076.
- Barroso, M. F. d. S. (2001). Métodos de otimização mono-objetivo aplicadas a Identificação caixa-cinzade sistemas não-lineares. *Dissertação de Mestrado PPGEE. Universidade Federal de Minas Gerais.*
- Barroso, M. F. S. (2006). *Otimização Bi-Objetivo Aplicada à Estimação de parâmetros de modelos NARX polinomiais: Caracterização e tomada de decisão.* PhD thesis, PPGEE - UFMG, Belo Horizonte, MG.
- Billings, S. A. & Aguirre, L. A. (1995). Effects of the Sampling Time and Identification of Nonlinear Models. *Int. J. of Bifurcation and Chaos*, 5(6), 1541–1556.
- Billings, S. A., Chen, S., & Korenberg, M. J. (1989). Identification of MIMO nonlinear systems using a forward-regression orthogonal estimator. *International Journal of Control*, 49(6), 2157–2189.
- Box, G. E. P. & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control.* Holden Day, Oakland, California, USA.
- Brockwell, P. J. & Davis, R. A. (1987). *Time Series: Theory and Methods* (2^a ed.). Springer Series in Statistics.
- Brockwell, P. J. & Davis, R. A. (2002). *Introduction to Time Series and Forecasting* (2^a ed.). Springer Text in Statistics.
- Broomhead, D. S. & Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2, 321–355.
- Caminhas, W. & Gomide, F. (2000). A fast learning algorithm for neofuzzy networks. *The 8th Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Madrid, Espanha.*

- Cao, L. (1997). Practical method for determining the minimum embedding dimension of a scalar time series. *Physica D: Nonlinear Phenomena*, 110, 43–50.
- Cao, L. & Soofi, A. S. (1999). Nonlinear deterministic forecasting of daily dollar exchange rates. *International Journal of Forecasting*, 15, 421–430.
- Casdagli, M. (1989). Nonlinear prediction of chaotic time series. *Physica D: Nonlinear Phenomena*, 35, 335–356.
- Casella, G. & Berger, R. L. (2001). *Statistical Inference* (2th ed.). DUXBURY Press.
- Cassini, C. (1999). Estimaco Recursiva de no-linearidades estticas usando modelos NARMAX polinomiais. Master's thesis, PPGEE, Universidade Federal de Minas Gerias, Belo Horizonte, MG, Brasil.
- Chang, F.-J., Liang, J.-M., & Chen, Y.-C. (2001). Flood forecastin using radial basis function neural networks. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews*, 31(4), 530–535.
- Chatfield, C. (2003). *The Analysis of time Series: An introduction* (6th ed.). Chapman & Hall/CRC.
- Chen, S. & Billings, S. A. (1989). Representations of non-linear systems the NARMAX model. *Int. J. Control*, 49(3), 1013–1032.
- Chen, S., Billings, S. A., Cowan, C., & Grant, P. (1990a). Non-linear systems identification using radial basis functions. *Int. J. Systems Sci.*, 21(12).
- Chen, S., Billings, S. A., Cowan, C., & Grant, P. (1990b). Practical identification of NARMAX models using radial basis function. *Int. J. Control*, 52(6).
- Chen, S., Billings, S. A., & Luo, W. (1989). Orthogonal least-squares methods and their applications to non-linear system identification. *Tinternational Journal of Control*, 50(5), 1873–1896.

- Chen, S., Grant, P., & Cowan, C. (1991). Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Trans. on Neural Network*, 2, 302–309.
- Coelho, M. C. S., Oliveira, D. J., & Mendes, E. M. A. M. (2004). Teste de determinismo em séries temporais usando dados sub-rogados. *CBA 2004 - XV Congresso Brasileiro de Automática, Gramado*.
- Coelho, M. C. S., Silva, V., Fonseca, A. R., & Aguirre, L. A. (2003). Metodologia para a obtenção de modelos narma polinomiais para previsão de carga a curto prazo. *Anais do II CITENEL : II Congresso de Inovação Tecnológica em Engenharia Elétrica, ANEEL, Brasília, 2*, 782–784.
- Coelho, M. C. S., Silva, V. C., Martinez, C. B., & Aguirre, L. A. (2003). Modelo NARMA polinomial para a previsão de curto prazo de consumo de energia elétrica. *X Encontro Latino Americano e do Caribe sobre Pequeno Aproveitamentos Hidroenergéticos- X ELPAH*, 184–189.
- Corrêa, M. (2001). *Identificação caixa-cinza de sistemas não-lineares utilizando representações NARMAX racionais e polinomiais*. PhD thesis, PPGEE, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.
- Darbellay, G. A. & Slama, M. (2000). Forecasting the short-term demand for electricity: Do neural networks stand a better chance? *International Journal of Forecasting*, 16, 71–83.
- Daubechies, I. (1990). The Wavelet transform, time-frequency localization and signal analysis. *IEEE Trans. Inform. Theory*, 36(5), 961–1005.
- Dickey, D. A. & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unitroot. *Journal of the American Statistical Association*, 74(366), 427–431.
- Dolan, K. & Spano, M. L. (2001). Surrogate for nonlinear time series analysis. *Physical Review E*, 64, 046128.

- Doveh, E., Feigin, P., Greig, D., & Hyams, L. (1999). Experience with fnn models for medium term power demand predictions. *IEEE Transactions on Power Systems*, *14*(2), 538–546.
- Duarte, M. M. (2006). Modelagem paramétrica de sistemas dinâmicos não-estacionários. Master's thesis, Programa de Pós-Graduação em Engenharia Elétrica, UFMG, Belo Horizonte, MG, Brasil.
- Eckmann, J. P., Kamphorst, S. O., Ruelle, D., & Ciliberto, S. (1986). Liapunov exponents from time-series. *Physical Review A*, *34*(6), 4971–4979.
- Eckmann, J. P. & Ruelle, D. (1985). Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics*, *57*(3), 617–656.
- Fiedler-Ferrara, N. & Prado, C. (1994). *Caos - Uma Introdução*. Editora Edgard Bücher.
- Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C., & Chen, B. C. (1998). New capabilities and methods of the X-12-ARIMA seasonal-adjustment program. *Journal of Business & economic statistics*, *16*(2), 127–152.
- Fraser, A. M. & Swinney, H. (1986). Independent coordinates in strange attractors from mutual information. *Phys. Rev. A*, *33*, 1134–1140.
- Garcia, C. (2006). *Modelagem e Simulação de Processos Industriais e de Sistemas Eletromecânicos* (2^a ed.). Coleção Acadêmica, EDUSP.
- Goldberger, A. L. (1996). Non-linear dynamics for clinicians: Chaos theory, fractals, and complexity at the bedside. *LANCET*, *347*(9011), 1312–1314.
- Gomes, M. (2001). Técnicas de sistemas dinâmicos não-lineares na análise da modulação autonômica da variabilidade da frequência cardíaca. *Tese de Doutorado PPGEE, Universidade Federal de Minas Gerais, Belo Horizonte*.
- Gomes, M. E. D., Souza, Á. V. P., Guimarães, H. N., & Aguirre, L. A. (2000). Investigation of determinism in heart rate variability. *Chaos*, *10*(2), 398–410.

- Gontar, Z., Sideratos, G., & Hatziaargyriou, N. (2004). Short-term load forecasting with radial basis function networks. *Methods and Applications of Artificial Intelligence, Proceedings Lecture Notes in Computer Science, 3025*, 432–438.
- Grassberger, P. (1986). Do climatic attractors exist. *Nature*, *323*(6089), 609–612.
- Grassberger, P. & Procaccia, I. (1983a). Characterization of strange attractors. *Physical Review Letters*, *50*(5), 346–349.
- Grassberger, P. & Procaccia, I. (1983b). Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, *9*(1-2), 189–208.
- Guimarães, H. N. (1996). Análise da variabilidade da frequência cardíaca - método e implicações fisiológicas. *Tese de Doutorado, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte*.
- Guo, J.-J. & Luh, B. (2003). Selecting input factors for clusters of gaussian radial basis function networks to improve market clearing price prediction. *IEEE Transactions on Power Systems*, *18*(2), 665–672.
- Haykin, S. (2001). *Redes Neurais. Princípios e prática*. Porto Alegre: Bookman, 2ª edição.
- Haykin, S. & Van Veen, B. (1999). *Signal and Systems*. J. Wiley & Sons.
- Hénon, M. (1976). A Two-dimensional Mapping with Strange Attractor. *Commun. Math. Phys.*, *50*, 69–77.
- Hippert, H. S., Pedreira, C. E., & CastroSouza, R. (2001). Neural network for short-term load forecasting: A review and evaluation. *IEEE Transactions on Power Systems*, *16*(1), 44–55.
- Hsieh, D. A. (1991). Chaos and nonlinear dynamics - application to financial-markets. *Journal of Finance*, *46*(5), 1839–1877.

- Hwang, H. B. & Ang, H. T. (2001). A simple neural network for ARMA(p,q) time series. *The International Journal of Management Science*, 29, 319–333.
- Jácome, C. (1996). Uso de Conhecimento Prévio na Identificação de Modelos Polinomiais NARMAX. Master's thesis, PPGEE, Universidade Federal de Minas Gerais, Belo Horizonte, Brasil.
- Kanters, J. K., Holsteinrathlou, N. H., & Agner, E. (1994). Lack of evidence for low-dimensional chaos in heart-rate-variability. *Journal Of Cardiovascular Electrophysiology*, 5(7), 591–601.
- Kaplan, D. T. & Cohen, R. J. (1990). Is fibrillation chaos? *Circulation Research*, 67(4), 886–892.
- Kaplan, D. T. & Glass, L. (1992). Direct test for determinism in a time series. *Physical Review of Letters*, 68(4), 427–430.
- Kaplan, D. T. & Glass, L. (1993). Coarse-grained embeddings of time-series - random-walks, Gaussian random-processes, and deterministic chaos. *Physica D: Nonlinear Phenomena : Nonlinear Phenomena*, 64(4), 431–454.
- Kennel, M. B., B. R. & Abarbanel, H. D. I. (1992). Determining embedding dimensions for phase-space reconstruction using geometrical construction. *Physical Review A*, 45(6), 3403–3411.
- Kennel, M. & Isabelle, S. (1992). Method to distinguish possible chaos from colored noise and to determine embedding parameters. *Physical Review A*, 46(6), 3111–3118.
- Korenberg, M., Billings, S., Liu, Y., & Mcilroy, P. (1988). Orthogonal parameter estimation algorithm for non-linear stochastic systems. *Int. J. Control*, 48(1), 193–210.

- Lapedes, A. & Farber, R. (1987). Nonlinear signal processing using neural networks: prediction and system modeling. LA-UR-87-2662, LA-UR-87-2662, Los Alamos National Laboratory, Los Alamos, New Mexico.
- Laplace, S. P. (1812). Introduction à la théorie analytique des probabilités. *s.n.*
- Leontaritis, I. J. & Billings, S. A. (1985a). Input-output parametric models for non-linear systems - part I: deterministic non-linear systems. *Int. J. Control*, 41(2), 303–328.
- Leontaritis, I. J. & Billings, S. A. (1985b). Input-output parametric models for non-linear systems - part II: stochastic non-linear systems. *Int. J. Control*, 41(2), 329–344.
- Letellier, C., Aguirre, L. A., Maquet, J., & Gilmore, R. (2005). Evidence for low dimensional chaos in sunspot cycles. *Astronomy & Astrophysics*, 449(1), 379–387.
- Lima, S. T., Rodrigues, D. D., & Aguirre, L. A. (2002). Nonlinear analysis and forecasting of electrical load time series. *XIV Congresso Brasileiro de Automática, Natal, RN.*
- Ljung, L. (1987). *System Identification: Theory for the User*. Prentice-Hall, London.
- Lorenz, E. N. (1963). Deterministic nonperiodic flow. *J. Atmos. Sci.*, 20, 283–293.
- Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation time-series methods - results of a forecasting competition. *Journal of Forecasting*, 1(2), 111–153.
- Makridakis, S. & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, 16, 451–476.
- Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: Methods and applications* (3^a ed.). John Wiley & Sons. Inc.

- Mason, J. & Kambhampati, C. (1996). Predictive control of a mixing tank using radial basis function networks. *IEEE - Proceedings of the 35th Conference on Decision and Control, Kobe, Japan*, 478–479.
- Mendes, E. M. A. M. & Billings, S. A. (1998). On overparametrization of nonlinear discrete systems. *Int. J. Bif. Chaos*, 8(3), 535–556.
- Mingoti, S. A. & Neves, O. F. (1999). A metodologia de geoestatística como alternativa na análise de séries temporais. *Revista Escola de Minas (REM), Ouro Preto*, 52(3), 182–187.
- Monteiro, L. H. (2002). *Sistemas Dinâmicos*. Editora Livraria da Física.
- Moon, T. K. & Stirling, W. C. (2000). *Mathematical Methods and Algorithms for Signal Processing*. Prentice Hall, New Jersey.
- Morettin, P. A. & Toloí, C. M. (2004). *Análise de séries temporais*. São Paulo: Edgard Blücher LTDA.
- Naguib, R. N. G., Wayman, J., Bennett, M. K., Raimes, S. A., & Griffin, S. (1998). Pre- and Post-operative predictions of recurrence in patients with cancer of the oesophago-gastric junction using radial basis function artificial neural networks. *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 20(6), 3335–3338.
- Nakamura, T. & Small, M. (2005). Small-shuffle surrogate data: Testing for dynamics in fluctuating data with trends. *Physical Review E*, 72(056216), 1–5.
- Nicolis, C. & Nicolis, G. (1984). Is there a climatic attractor? *Nature*, 311(5986), 529–532.
- Nicolis, C. & Nicolis, G. (1987). Evidence for climatic attractors. *Nature*, 326(6112), 523–523.
- Ogata, K. (1998). *Engenharia de Controle Moderno*. Rio de Janeiro, RJ: Prentice-Hall do Brasil.

- Oliveira, C., Machado, L. J. B., & Wazlawick, R. S. (2000). Modelo de rede neural artificial para previsão automática da carga ativa nodal em sistemas de energia elétrica. *XIII Congresso Brasileiro de Automática. Florianópolis, SC*, 586–591.
- Ortega, G. J. (2000). Using topological statistics to detect determinism in time series. *Physical Review E*, *62*(3), 3419–3428.
- Ortega, G. J. & Louis, E. (1998). Smoothness implies determinism in time series: A measure based approach. *Physical Review Letters*, *81*(20), 4345–4348.
- Osborne, A. R. & Provenzale, A. (1989). Finite correlation dimension for stochastic systems with power-law spectra. *Physica D*, *35*(3), 357–381.
- Packard, N. H., Crutchfield, J. P., Farmer, J. D., & Shaw, R. S. (1980). Geometry from a time series. *Physical Review Letters*, *45*, 712–715.
- Palus, M. (1996). Nonlinearity in normal human eeg: Cycles, temporal asymmetry, nonstationarity and randomness, not chaos. *Biological Cybernetics*, *75*(5), 389–396.
- Papoulis, A. (1991). *Probability, random variables, and stochastic process*. New York: MacGraw-Hill.
- Papoulis, A. & Pillai, S. U. (2002). *Probability, random variables, and stochastic process* (4^a ed.). New York: MacGraw-Hill.
- Park, Y. R., Murray, T. J., & Chen, C. (1996). Predicting sun spots using a layered perceptron neural network. *IEEE Transactions on Neural Networks*, *7*(2), 501–505.
- Petrucci, J. D., Nandram, B., & Chen, M. (1999). *Applied Statistic for Engineers and Scientists*. Upper Saddle River - NJ, USA: Prentice Hall, Inc.
- Priestley, M. (1988). *Non-linear and non-stationary time series analysis*. Academic Press, London.

- Ranaweera, D. K., Hubele, N. F., & Papalexopoulos, A. D. (1995). Application of radial basis function neural network model for short-termload forecasting. *IEEE Proceedings Generation, Transmission and Distribution*, 142(1), 45–50.
- Reis, A. J. R. (1999). Pré-processamento de séries temporais para um previsor de carga elétrica a curto prazo via redes neurais artificiais. Master's thesis, Escola Federal de Engenharia de Itajubá, Itajubá, MG.
- Rodrigues, D. D. (2002). Previsão de fornecimento de energia elétrica: Um estudo de caso do ponto de vista do barramento. Master's thesis, Universidade Federal de Minas Gerais - PPGEE.
- Salvino, L. W. & Cawley, R. (1994). Smoothness implies determinism: A method to detect it in time series. *Physical Review Letters*, 73(8), 1091–1094.
- Salvino, L. W., Cawley, R., Grebogi, C., & Yorke, J. A. (1995). Predictability in time series. *Physics Letters A*, 209, 327–332.
- Sauer, T., Yorke, J. A., & Casdagli, M. (1991). Embedology. *Journal of Statistical Physics*, 65(3-4), 579–616.
- Schreiber (1999). Interdisciplinary application of nonlinear time series methods. *Physics Reports*, 308, 1–64.
- Schreiber, T. (2000). Surrogate time series. *Physica D: Nonlinear Phenomena*, 346–382.
- Schreiber, T. & Schmitz, A. (1996). Improved surrogate data for nonlinearity tests. *Physical Review Letters*, 77(4), 635–638.
- Schreiber, T. & Schumitz, A. (2000). Review paper: Surrogate time series. *Physica D: Nonlinear Phenomena*, 142, 346–382.
- Shiskin, J., Young, A. H., & Musgrave, J. C. (1967). The X-11 variant of the census method ii seasonal adjustment program. Technical paper no. 15, Department of Commerce, Bureau of Economic Analysis.

- Silva, A. P. A. & Moulin, L. S. (2000). Confidence intervals for neural network based short-term load forecasting. *IEEE Transactions on Power Systems*, 15(4), 1191–1196.
- Silva, A. P. A., Rodrigues, U. P., RochaReis, A. J., Moulin, L. S., & Neto, C. A. S. (21 a 26 de outubro de 2001). Oráculo - Uma ferramenta para previsão de carga. *XVI Seminário Nacional de Produção e Transmissão de Energia Elétrica, Campinas-São Paulo - Brasil*.
- Sinclair, M. J., Musavi, M. T., & Qiao, M. (1995). *Radial Basis Function Neural Network as Predictive Process Control Model*.
- Small, M. (2005). *Applied nonlinear time series analysis: Applications in physics, physiology and finance*, volume 52 of *A. World Scientific Series on Nonlinear Science*.
- Small, M. & Tse, C. (2003). Evidence for deterministic nonlinear dynamics in financial time series data. *Proceedings of the IEEE International Conference on Computational Intelligence for Financial Engineering*, 339–346.
- Small, M. & Tse, C. K. (2002). Applying the method of surrogate data to cyclic time series. *Physica D: Nonlinear Phenomena*, 164, 187–201.
- Small, M., Yu, D., & Harrison, R. G. (2001). Surrogate test for pseudoperiodic time series data. *Physical Review Letters*, 87(4), 188101.
- Soofi, A. S. & Cao, L. (2002). *Modelling and Forecasting Financial Data: Techniques of Nonlinear Dynamics*. Studies in Computational Finance.
- Stam, C. J., Pijn, J. P. M., & Pritchard, W. S. (1998). Reliable detection of non-linearity in experimental time series with strong periodic components. *Physica D: Nonlinear Phenomena*, 112(3-4), 361–380.
- Stewart, I. (1991). *Será que Deus joga dados? A nova matemática do caos*. Ciência e Cultura, Rio de Janeiro.

- Sugihara, G. & May, R. M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, 344(6268), 734–741.
- Sze, T. L. (1995). *System Identification using radial basis function networks*. PhD thesis, Department of Automatic Control and Systems Engineering University of Sheffield, Sheffield, U.K.
- Takens, F. (1980). Detecting strange attractors in turbulence. In Rand, D. & Young, L. (Eds.), *Dynamical systems and turbulence - Lecture Notes in Mathematics*, volume 898, (pp. 366–381). Springer-Verlag, Berlin.
- Theiler, J. (1991). Some comments on the correlation dimension of $1/f^\alpha$ noise. *Physica Letters A*, 155, 480–493.
- Theiler, J. (1995). On the evidence for low-dimensional chaos in an epileptic electroencephalogram. *Phys. Lett. A*, 196, 335–341.
- Theiler, J., Eubank, S., Longtin, A., & Farmer, B. G. J. D. (1992). Testing for nonlinearity in time series: the method of surrogate data. *Physica D: Nonlinear Phenomena*, 58, 77–94.
- Theiler, J. & Rapp, P. E. (1996). Re-examination of the evidence for low-dimensional, nonlinear structure in the human electroencephalogram. *Electroencephalography and Clinical Neurophysiology*, 98(13), 213–222.
- Theiler, J. P. (1988). *Quantifying Chaos: Practical Estimation of the correlation dimension*. PhD thesis, Caltech (California Institute of Technology).
- Tong, H. (1990). *Non-linear Time Series - A Dynamical System Approach*. Oxford Science Publications.
- Wayland, R., Bromley, D., Pickett, D., & Passamante, A. (1993). Recognizing determinism in a time series. *Physical Review Letters*, 70(5), 580–582.
- Weigend, A. S. & Gershenfeld, N. A. (1994). *Time Series Prediction: Forecasting the Future and Understanding the Past*. Addison Wesley Longman.

- Wold, H. (1954). *A study in the analysis of stationary time series*. 2nd edição
Estocolmo: Almqvist & Wiksel.
- Wolf, A., Swinney, J. B. S. H. L., & Vastano, J. A. (1985). Determining lyapunov
exponents from a time-series. *Physica D: Nonlinear Phenomena*, 16(3), 285–317.
- Yang, H.-T. & Huang, C.-M. (1998). A new short-term load forecasting approach
using self-organizing fuzzy ARMAXmodels. *IEEE Transactions on Power Systems*,
13(1), 217–225.
- Yule, G. U. (1927). On a method of investigating periodicities in disturbed series
with specialreference to Wolfer’s sunspot numbers. *Philos. Trans. R. Soc., A226*,
267–298.
- Zhang, G. P., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural
networks: The state of the art. *International Journal of Forecasting*, 14, 35–62.
- Zhang, G. P. & Qi, M. (2005). Neural network forecasting for seasonal and trend
time series. *European Journal of Operational Research*, 160(2), 501–514.