

Laboratório de Modelagem, Análise e Controle de Sistemas Não-Lineares

Departamento de Engenharia Eletrônica

Universidade Federal de Minas Gerais

Av. Antônio Carlos 6627, 31270-901 Belo Horizonte, MG Brasil

Fone: +55 3499-4866 - Fax: +55 3499-4850



Detecção de Determinismo e Modelagem Preditiva de Séries Temporais de Consumo de Energia Elétrica

Tales Argolo Jesus

Dissertação 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 Mestre em Engenharia Elétrica.

Orientador: Prof. Eduardo Mazoni Andrade Marçal Mendes, PhD

Belo Horizonte, outubro de 2008

Dedicatória

*Aos meus pais, pelo amor e pela dedicação.
A minha irmã, pela amizade e pelo zelo.
A minha namorada, pelo carinho e pela paciência.*

Agradecimentos

Não existe nenhum trabalho exclusivamente individual. A lei de cooperação determina que toda atividade humana se processe no âmbito do auxílio mútuo e da divisão de tarefas.

Dessa forma, não é possível trazer este texto à lume sem registrar os seguintes agradecimentos:

A Deus, pela imerecida oportunidade de continuar os meus estudos na capital mineira.

Aos colegas de pós-graduação e amigos Gledson e Lucymara. A convivência com vocês com certeza suavizou o muitas vezes árido ambiente acadêmico.

Ao Professor Luis Aguirre, pelo voto de confiança ao me admitir como bolsista de mestrado no projeto de pesquisa de previsão de consumo de energia a longo prazo.

Ao Professor Leonardo Tôrres, que não foi meu orientador e nem participou deste projeto de pesquisa, mas pelo exemplo de dedicação e competência enquanto professor e pesquisador¹.

Ao Professor Eduardo Mazoni, por aceitar ser meu orientador, pela paciência de estar sempre disponível para me atender em sua sala, pela oportunidade da primeira experiência em docência e pelo companheirismo que faz com que a nossa relação enquanto pesquisadores transcenda o âmbito profissional e atinja o campo da amizade. Espero que possamos continuar a trabalhar juntos em um futuro não muito distante.

À CEMIG, pelo apoio financeiro que sem dúvida foi indispensável à realização deste trabalho.

¹Sem contar sua atuação como conselheiro em termos de decisão de rumos na academia!

Epígrafe

*“Fui átomo, vibrando entre as forças do Espaço,
Devorando amplidões, em longa e ansiosa espera...
Partícula, pousei... Encarcerado, eu era
Infusório do mar em montões de sargaço.*

*Por séculos fui planta em movimento escasso,
Sofri no inverno rude e amei na primavera;
Depois, fui animal, e no instinto da fera
Achei a inteligência e avancei passo a passo...*

*Guardei por muito tempo a expressão dos gorilas,
Pondo mais fé nas mãos e mais luz nas pupilas,
A lutar e chorar para, então, compreendê-las!...*

*Agora, homem que sou, pelo Foro Divino,
Vivo de corpo em corpo a forjar o destino
Que me leve a transpor o clarão das estrelas!...”*

Soneto intitulado *Jornada*, de autoria de Adelino da Fontoura Chaves. Extraído do livro *Antologia dos Imortais*, psicografado por Francisco Cândido Xavier.

Sumário

Resumo	xiii
Abstract	xv
Lista de Figuras	xix
Lista de Tabelas	xxii
Lista de Abreviações	xxiii
1 Introdução	1
1.1 Definição do Problema	1
1.2 Motivação	2
1.3 Objetivos do Projeto	5
1.4 Organização do Trabalho	5
2 Uma Breve Revisão de Modelagem e Previsão de Séries Temporais	7
2.1 Modelagem e Previsão de Séries Temporais	7
2.2 Modelagem e Previsão de Séries Temporais de Consumo de Energia Elétrica	11
3 Descrição e Análise de Dados	17
3.1 Introdução	17
3.2 Séries Temporais de Consumo de Energia Elétrica	18
3.2.1 Séries da CEMIG	18
3.2.2 Série de New England	24
3.3 Decomposição de Séries Temporais	25
3.3.1 Análise Espectral Singular	26

3.3.2	Decomposição de séries temporais de consumo de energia elétrica	28
3.4	Detecção de Determinismo	31
3.4.1	Teste de Dados Sub-Rogados	31
3.4.2	Aplicação do Teste de Dados Sub-rogados SSS	37
3.4.3	Aplicação do Teste de Dados Sub-rogados CSS	43
3.5	Conclusão	45
4	Metodologias de Previsão	47
4.1	Introdução	47
4.2	Medidas de Erro	47
4.3	Definição de um preditor trivial	48
4.4	Modelos Auto-Regressivos Múltiplos	49
4.5	Rede Neural Artificial MLP	53
4.6	Rede Neuro-Fuzzy Linear Local	56
4.7	Conclusão	61
5	Resultados de Previsão	63
5.1	Introdução	63
5.2	Preditor trivial	63
5.2.1	Série de Consumo Mensal de New England	63
5.2.2	Série de Consumo Mensal de Baixa Tensão da CEMIG	64
5.2.3	Série de Consumo Mensal Total da CEMIG	65
5.3	Modelos Auto-Regressivos Múltiplos	66
5.3.1	Série de Consumo Horário de New England	66
5.3.2	Série de Consumo Mensal Total da CEMIG	70
5.4	Rede Neural Artificial MLP	71
5.4.1	Série de Consumo Mensal de New England	71
5.4.2	Série de Consumo Mensal de Baixa Tensão da CEMIG	73
5.5	Rede Neuro-Fuzzy Linear Local	75
5.5.1	Série de Consumo Mensal de New England	75
5.5.2	Série de Consumo Mensal de Baixa Tensão da CEMIG	78
5.6	Síntese comparativa dos resultados de previsão	79
6	Conclusões	83

Bibliografia	92
A Teste de Grubbs	93

Resumo

A previsão de séries temporais de consumo de energia elétrica a longo prazo é tarefa de grande relevância num contexto que envolve questões relacionadas ao planejamento do mercado energético de um país ou região. Estimativas confiáveis do consumo de energia possibilitam a definição de estratégias que venham a facilitar a ampliação e a capacitação do sistema energético que atende a uma determinada população dentro de um determinado contexto climático, cultural, político, socio-econômico, etc.

O presente trabalho de mestrado trata da aplicação da metodologia conhecida como Análise de Dados Sub-Rogados em uma tentativa de detectar determinismo nas séries de consumo de energia elétrica do Estado de Minas Gerais (Brasil) e do Estado de New England (Estados Unidos da América), bem como da utilização de ferramentas de modelagem matemática existentes em Identificação de Sistemas com a finalidade de se realizarem previsões de tais séries a longo prazo, num horizonte de previsão máximo de 60 meses. Ainda na etapa de análise de dados é apresentada uma metodologia para decomposição de séries temporais chamada de Análise Espectral Singular, a qual se mostrou eficiente na extração de características das séries como as suas componentes de tendência.

Três foram as ferramentas matemáticas escolhidas para a construção dos modelos preditivos: (i) modelos auto-regressivos múltiplos; (ii) redes neurais artificiais do tipo perceptron multicamada; (iii) e redes neuro-fuzzy do tipo linear local. Além delas, definiu-se um preditor trivial a partir do conhecimento prévio dos comportamentos de tendência e sazonalidade das séries a fim de se ter em mãos uma espécie de modelo mais simples porém útil para a validação das referidas ferramentas utilizadas para lidar com o problema de previsão.

No que diz respeito à detecção de determinismo, a aplicação do teste de dados sub-rogados revelou a existência de informação dinâmica intra-ciclo e entre ciclos na série de consumo de New England, enquanto que no caso da série da CEMIG tal detecção só se patentou no caso das séries de alguns setores de baixo peso no consumo geral e nas séries agrupadas por unidades de tempo menores do que um mês. Tal verificação se refletiu na etapa de construção de modelos, em que a similitude entre as previsões e os valores reais de consumo foi mais evidente para série de New England do que para as séries da CEMIG. Em termos de medidas de erro, para as séries de New England e de baixa tensão da CEMIG, o melhor modelo estimado em cada uma das três categorias de modelos aplicados neste trabalho conduziram a medidas de erro percentual absoluto médio (MAPE) sempre abaixo de 3%, considerando-se horizontes de previsão de 60 e 24 meses, respectivamente. Na série de consumo total da CEMIG, o MAPE obtido em um horizonte de 24 meses com os modelos ARM foi inferior a 5%. O único caso em que o preditor trivial superou as outras estratégias preditivas foi no caso da série de consumo de baixa tensão da CEMIG, justamente uma das séries para as quais não foi verificada a existência de determinismo intra-ciclo e entre-ciclos via análise de dados sub-rogados,

de sorte que um mero ajuste de tendência e sazonalidade à guisa de mapeamento de padrões foi suficiente para a obtenção de um resultado de previsão bastante acurado (MAPE de 1,99% para um horizonte de 24 meses).

Mais do que chegar a resultados satisfatórios em termos de minimização das medidas de erro de previsão, este trabalho ratificou o conhecido fato de que a complexidade dos modelos nem sempre está relacionada com a sua capacidade em reproduzir com fidelidade os comportamentos apresentados pelos sistemas que existem no mundo real.

Abstract

Long-term load forecasting is a task of great relevance in the context of energy market planning of a country or a region. Reliable forecasts of energy consumption make possible the definition of strategies that can promote the expansion of the electrical system that supplies a population into an specific climatic, cultural, political and social-economic context.

The present work deals with the application of Surrogate Data Analysis in an attempt to detect determinism in the energy consumption time series of Minas Gerais State (Brazil) and New England State (United States of America), and also with the use of mathematical modelling tools typical in Systems Identification to perform long-term load forecasting of such time series, considering a maximum prediction horizon of 60 months. In the stage of data analysis it is also presented a methodology to decompose time series called Singular Spectral Analysis, which has been shown to be efficient in the extraction of time series characteristics, such as the trend components.

Three were the mathematical tools chosen to construct the predictive models: (i) multiple auto-regressive-like models; (ii) multi-layer perceptron artificial neural networks; and (iii) local linear neuro-fuzzy networks. Besides them, it was also defined a trivial predictor, based on the previous knowledge of the series trend and seasonality behavior, with the objective of using it as a tool to validate the methods used to deal with the forecasting problem.

As far as determinism detection is concerned, the application of the surrogate data test has revealed the existence of inter-cycle and intra-cycle dynamics in *i)* the time series of the electricity consumption of the state of New England, *ii)* in those time series of specific sectors of little impact in the total electricity consumption of the state of Minas Gerais and *iii)* in the aggregated time series of the same state when the sampling time is smaller than one month. Such results have reflected in the modelling stage by revealing that the similarity among the predictions and the real consumption values was more evident when working with the New England time series than with CEMIG time series. In terms of error measures, for both the New England and the low voltage CEMIG time series, the best model in each of the three categories of models used in this work resulted in mean absolute percentual errors (MAPE) smaller than 3%, when considering 60 and 24 month prediction horizon. In the case of CEMIG total consumption time series, the MAPE was smaller than 5% for a 24 month prediction horizon when ARM models are used. The only situation in which the trivial predictor was superior than the other prediction strategies was for CEMIG low voltage time series, in which the existence of intra-cycle and inter-cycle determinism via surrogate data analysis was not verified. In such a case a simple trend and seasonality fitting was sufficient to obtain very accurate prediction results (MAPE smaller than 1,99% to a 24 months prediction horizon).

More than achieving satisfactory results in terms of minimizing the forecasting error, this work ratifies the already known fact that the model complexity is not always related to its ability in reproducing the dynamic behavior exhibited by systems in the real world.

Lista de Figuras

3.1	Séries dos consumidores de baixa tensão da Cemig divididos em setores - i) comercial; ii) consumo próprio; iii) industrial; iv) iluminação pública; v) residencial; vi) rural; vii) poder público; viii) serviço público; ix) concessionárias.	19
3.2	Consumo Faturado Total - O gráfico à esquerda inclui todos os setores, enquanto que gráfico à direita inclui quase todos os setores, exceção feita à categoria 'concessionárias'.	19
3.3	Consumo Faturado - Industrial. São apresentadas a série original e a série sem <i>outliers</i> já interpolada, cada uma ladeada por seu respectivo histograma.	20
3.4	Consumo Faturado - Serviço Público. São apresentadas a série original e a série sem <i>outliers</i> já interpolada, cada uma ladeada por seu respectivo histograma.	20
3.5	Séries de consumo total (exceto pela categoria 'concessionárias') após as retirada dos <i>outliers</i>	22
3.6	Carga por hora referente ao ano 2002. Série original e série sem <i>outliers</i> já interpolada, cada uma ladeada por seu respectivo histograma.	22
3.7	Carga por hora referente ao ano 2003. Série original e série sem <i>outliers</i> já interpolada, cada uma ladeada por seu respectivo histograma.	23
3.8	Carga por hora referente ao ano 2004. Série original e série sem <i>outliers</i> já interpolada, cada uma ladeada por seu respectivo histograma.	23
3.9	Carga por hora referente ao ano 2005. Série original e série sem <i>outliers</i> já interpolada, cada uma ladeada por seu respectivo histograma.	23
3.10	À esquerda, carga diária total referente ao período que se estende de janeiro de 2000 a dezembro de 2006; à direita, carga mensal total referente ao mesmo período.	24
3.11	À esquerda, série de consumo horário de New England no ano de 1994; à direita, um <i>zoom</i> que evidencia o formato da curva ao longo de duas semanas.	25
3.12	À esquerda, a série de consumo mensal de New England no intervalo de tempo entre Janeiro de 1980 e Dezembro de 1998; à direita, um <i>zoom</i> no período correspondente ao ano de 1994.	26
3.13	Série $y(t)$ à esquerda e suas componentes de tendência $m(t)$, sazonalidade $s(t)$ e resíduos $r(t)$ à direita.	26
3.14	As três componentes da série da Figura 3.13 obtidas via SSA juntamente com os seus respectivos valores singulares.	28

3.15	Análise Espectral Singular - Comparação entre as componentes de maior peso e a série $\{y(t) - r(t)\}$ para $M \in \{5, 10, 30, 50, 100, 500\}$, juntamente com o respectivo espectro singular das diferentes decomposições realizadas.	29
3.16	À esquerda, série mensal de consumidores de baixa tensão da Cemig, série de consumo total da Cemig e série de consumo mensal de New England juntamente com as componentes principais obtidas via SSA para $M = 30$; à direita, os valores singulares associados a cada componente da série.	30
3.17	Série original aleatória e quatro séries de dados sub-rogados.	33
3.18	Função de autocorrelação e informação mútua média da série original aleatória e das séries artificiais	33
3.19	Série original com dependência temporal e quatro séries de dados sub-rogados.	34
3.20	Função de autocorrelação e informação mútua média da série original com dependência temporal e das séries artificiais	34
3.21	À esquerda, série original e componente de tendência estimada; à direita, série resultante após a extração da componente de tendência.	38
3.22	FAC e IMM das séries de consumo mensal de baixa tensão da CEMIG e das respectivas séries sub-rogadas. A significância define a distância entre a série original e as séries artificiais.	38
3.23	Dados de consumo mensal de baixa tensão da CEMIG - À esquerda, série original e componente de tendência estimada; à direita, série resultante após a extração da componente de tendência.	40
3.24	Dados de consumo mensal de New England - À esquerda, série original e componente de tendência estimada; à direita, série resultante após a extração da componente de tendência.	42
3.25	FAC e IMM das séries de consumo mensal de New England e das respectivas séries sub-rogadas. A significância define a distância entre a série original e as séries artificiais.	42
4.1	Resultados da previsão de 1993 a 1997 a partir do consumo de 1992 com o preditor ingênuo.	49
4.2	Séries de consumo evidenciando correlações entre semanas do ano existentes nos dados de consumo dos anos de 2004 e 2005.	50
4.3	Séries de consumo evidenciando correlações entre horas do dia existentes nos dados de consumo dos anos de 2004 e 2005.	51
4.4	Diagrama esquemático de um neurônio artificial.	53
4.5	Diagrama esquemático de uma rede neural artificial MLP com uma camada escondida.	54
4.6	Diagrama esquemático de uma RNFL, em que $\vec{z}$ é o vetor de entradas das funções de validade e $\vec{x}$ é o vetor de entradas dos MLLs (Hoffman and Nelles, 2001). No caso geral considerado nessa dissertação, $\vec{u} = \vec{z} = \vec{x}$.	57
4.7	Diagrama esquemático do particionamento do espaço de entrada bidimensional ($p = 2$) de uma rede RNFL realizado pelo algoritmo LOLI-MOT (Hoffman and Nelles, 2001).	61

4.8	Funcionamento do algoritmo LOLIMOT para ajuste dos parâmetros da rede neuro-fuzzy linear local. Da esquerda para a direita, de cima para baixo, são exibidas as seis primeiras iterações do algoritmo para aproximar uma função não-linear estática.	62
5.1	Resultados da previsão de consumo em New England de 1994 a 1998 utilizando-se o preditor trivial.	64
5.2	Resultados da previsão de consumo de baixa tensão da CEMIG de 2005 e 2006 utilizando-se o preditor trivial.	65
5.3	Resultados da previsão de consumo total da CEMIG de 2005 e 2006 utilizando-se o preditor trivial.	65
5.4	Resultados da previsão de 1994 a partir da matriz de 1993. (a) Todo o ano de 1994 e (b) as primeiras semanas de 1994.	66
5.5	Resultados da previsão de 1994 a partir da matriz de 1993 sem o ajuste de correção anual. (a) Todo o ano de 1994 e (b) as primeiras semanas de 1994.	67
5.6	Resultados da previsão de 1994 a partir da matriz de 1993 com o algoritmo modificado. No gráfico à direita estão as séries real e prevista de consumo mensal.	69
5.7	Resultados da previsão de 1994 a 1998 a partir da matriz de 1993 com o algoritmo modificado.	70
5.8	Resultados da previsão de 2005 e 2006 a partir da matriz de 2004 com o algoritmo modificado.	71
5.9	À esquerda, série de tendência de consumo mensal total da CEMIG; à direita, a primeira diferença da série de tendência de consumo mensal total da CEMIG. O trecho em destaque corresponde ao ano de 2004. . . .	71
5.10	Módulo da transformada de Fourier da série de consumo mensal de New England.	73
5.11	Resultados da previsão de 1994 a 1998 utilizando rede neural artificial MLP.	73
5.12	Módulo da transformada de Fourier da série de consumo mensal de baixa tensão da CEMIG.	74
5.13	Resultados da previsão de 2005 e 2006 utilizando rede neural artificial MLP.	75
5.14	À esquerda, série de tendência de consumo mensal total de baixa tensão da CEMIG; à direita, a primeira diferença da série de tendência de consumo mensal total de baixa tensão da CEMIG. O trecho em destaque corresponde aos ano de 2004, 2005 e 2006.	76
5.15	Resultados da previsão de 1994 a 1998 utilizando a rede neuro-fuzzy. . .	76
5.16	Resultados da previsão de 1994 a 1998 utilizando a rede neuro-fuzzy com inserção arbitrária de não-estacionariedade na série.	77
5.17	Resultados da previsão de 2005 e 2006 utilizando a rede neuro-fuzzy. . .	78

Lista de Tabelas

3.1	Consumo percentual correspondente a cada setor de consumidores de baixa tensão.	21
3.2	Evidência de rejeição de H_0 em função do p -valor.	32
3.3	Séries de consumo total da CEMIG - Significância associada à FAC e à IMM para atrasos de 1 a 4 amostras.	39
3.4	Séries de consumo total da CEMIG - p -valores associados à FAC e à IMM para atrasos de 1 a 4 amostras.	39
3.5	Séries de consumo mensal de baixa tensão da CEMIG - Significância associada à FAC e à IMM para atrasos de 1 a 4 amostras.	41
3.6	Séries de consumo mensal de baixa tensão da CEMIG - p -valores associados à FAC e à IMM para atrasos de 1 a 4 amostras.	41
3.7	Aplicação do SSS às séries de consumo mensal de New England - Significância associada à FAC e à IMM para atrasos de 1 a 4 amostras.	43
3.8	Aplicação do SSS às séries de consumo mensal de New England - p -valores associados à FAC e à IMM para atrasos de 1 a 4 amostras.	43
3.9	Aplicação do CSS às séries de consumo total da CEMIG - Significância associada à FAC e à IMM para atrasos de 12, 24, 36 e 48 amostras.	43
3.10	Aplicação do CSS às séries de consumo total da CEMIG - p -valores associados à FAC e à IMM para atrasos de 12, 24, 36 e 48 amostras.	43
3.11	Séries de consumo mensal de baixa tensão da CEMIG - Significância associada à FAC e à IMM para atrasos de 12, 24, 36 e 48 amostras.	44
3.12	Séries de consumo mensal de baixa tensão da CEMIG - p -valores associados à FAC e à IMM para atrasos de 12, 24, 36 e 48 amostras.	44
3.13	Aplicação do CSS às séries de consumo mensal de New England - Significância associada à FAC e à IMM para atrasos de 1 a 4 amostras.	45
3.14	Aplicação do CSS às séries de consumo mensal de New England - p -valores associados à FAC e à IMM para atrasos de 1 a 4 amostras.	45
4.1	Erros da previsão de consumo de 1993 a 1997 a partir do consumo de 1992 com o preditor ingênuo.	49
5.1	Erros da previsão de consumo em New England de 1994 a 1998 utilizando-se o preditor trivial.	64
5.2	Erros da previsão de consumo de baixa tensão da CEMIG de 2005 e 2006 utilizando-se o preditor trivial.	64
5.3	Erros da previsão de consumo total da CEMIG de 2005 e 2006 utilizando-se o preditor trivial.	66
5.4	Erros da previsão de consumo de 1994 a partir do consumo de 1993 . . .	67

5.5	Erros da previsão de consumo de 1994 a partir do consumo de 1993 sem o ajuste de correção anual	68
5.6	Erros da previsão de consumo de 1994 a partir do consumo de 1993 com o algoritmo modificado	68
5.7	Erros da previsão de consumo de 1994 a 1998 a partir dos dados de consumo de 1993 com o algoritmo modificado.	70
5.8	Erros da previsão de consumo de 2005 e 2006 a partir dos dados de consumo de 2004 com o modelos AR múltiplos.	70
5.9	Erros da previsão de consumo de 1994 a 1998 a utilizando rede MLP (regressores 6 e 12).	74
5.10	Erros da previsão de consumo de 2005 a 2006 a utilizando rede MLP (regressores 4, 6 e 12).	75
5.11	Erros da previsão de consumo de 1994 a 1998 a utilizando a rede neuro-fuzzy (regressor 12).	77
5.12	Erros da previsão de consumo de 1994 a 1998 a utilizando a rede neuro-fuzzy com inserção arbitrária de não-estacionariedade na série (regressores 6 e 12).	78
5.13	Erros da previsão de consumo de 2005 a 2006 utilizando a rede neuro-fuzzy (regressor 12).	79
5.14	Medidas de erro percentuais ao longo dos respectivos horizontes de previsão para cada um dos modelos e cada uma das séries.	79
5.15	Média e desvio padrão das medidas de erro percentuais ao longo dos respectivos horizontes de previsão para os modelos RNA, RNFL e RN-FLLE e para as séries de New england e de consumo de baixa tensão da CEMIG.	80

Lista de Abreviações

AR	Auto-Regressivos;
ARIMA	<i>Auto-Regressive Integrated Moving Average</i> ;
ARM	Auto-Regressivos Múltiplos;
CEMIG	Companhia Energética de Minas Gerais;
CSS	<i>Cycle-Shuffle Surrogate</i> ;
ED	Estatística Discriminante;
MQ	Estimador de Mínimos Quadrados;
MQP	Estimador de Mínimos Quadrados Ponderado;
ERR	Taxa de Redução de Erro (<i>Error Reduction Rate</i>);
FAC	Função de Autocorrelação;
FV	Função de Validade;
IMM	Informação Mútua Média;
LM	Levenberg-Marquardt;
LOLIMOT	<i>Local Linear Model Tree</i> ;
MAPE	<i>Mean Absolute Percentual Error</i> ;
MLL	Modelo Linear Local;
MLP	Perceptron Multicamada (<i>Multi-Layer Perceptron</i>);
MPE	<i>Mean Percentual Error</i> ;
PSO	<i>Particle Swarm Optimization</i> ;
RNA	Rede Neural Atifical;
RNFL	Rede Neuro-Fuzzy Linear Local;
S	Significância;
SSA	Análise Espectral Singular (<i>Singular Spectral Analysis</i>);
SSS	<i>Small-Shuffle Surrogate</i> ;
TPE	<i>Total Percentual Error</i> ;

Introdução

“Se queres conversar comigo, define primeiro os termos que usas.”

Voltaire (1694-1778)

Este capítulo consiste na exposição preliminar do problema de previsão de consumo de energia a longo prazo, cuja pesquisa e análise foram assumidas por uma equipe da Universidade Federal de Minas Gerais (UFMG), constituída de alunos de graduação, de pós-graduação e de professores, conforme solicitação da Companhia Energética de Minas Gerais (CEMIG). O projeto intitula-se “Pesquisa de Modelo de Previsão de Consumo de Energia, para Curto e Médio Prazo, utilizando Inteligência Computacional”, dentro do qual está inserido o escopo dessa dissertação de mestrado, que tratará, especificamente, da detecção de características nas séries temporais de consumo de energia (por exemplo, determinismo) a partir do método de análise de dados subrogados, e da proposição de metodologias de previsão de consumo de energia a longo prazo. A seguir, define-se o problema e são apresentados a motivação e os objetivos deste trabalho.

1.1 Definição do Problema

Desde que Galileu Galilei (1564-1642) estabeleceu como base do método científico a observação dos fatos com vistas à elaboração de teorias consentâneas com a realidade mensurável do Universo¹, a Ciência tem utilizado a experimentação, dentro de cabíveis e rigorosos critérios de controle, como uma verdadeira pedra de toque para a validação das teorias que surgem nos mais diversos campos do conhecimento humano, seja na Biologia, na Química, na Física, na Astronomia ou em qualquer dos outros ramos da Ciência.

Hoje, com o avanço tecnológico difundido em todo o planeta e a era da informatização já muito bem estabelecida, é relativamente fácil se ter acesso a medições de inúmeras grandezas físicas (ou variáveis quaisquer) ao longo do tempo, a exemplo a temperatura de um forno em uma indústria (Teixeira, 2004), o consumo energético mensal de um determinado país (Jesus, 2006), a intensidade de radiação eletromagnética emitida por um pulsar localizado a centenas ou milhares de anos-luz da Terra (Johnston et al., 2006), ou até mesmo a vazão de ar nos canais respiratórios de uma criança (Small et al., 1999). Esses conjuntos de dados são conhecidos como séries temporais, ou seja, sequências

¹“Aquilo que não pode ser medido e quantificado não é científico”(Galileu Galilei).

de valores de uma determinada grandeza distribuídos ao longo do tempo. De posse desses dados, algumas questões interessantes podem ser levantadas: (i) Essas séries são majoritariamente caracterizadas por aleatoriedade ou determinismo²? (ii) Apresentando elas algum determinismo, é possível quantificá-lo? (iii) Como realizar previsões para essas séries, isto é, a partir de dados presentes e passados, é possível a estimação de seus valores futuros?

Para responder a primeira questão proposta acima (e talvez a segunda), existe um procedimento desenvolvido no início da década de 90, inspirado na clássica metodologia estatística de teste de hipóteses, que consiste na execução de uma bateria de testes a fim de se realizar inferências quanto à natureza da dinâmica subjacente a uma determinada série temporal. Essa abordagem de caracterização dinâmica é conhecida como Análise de Dados Sub-Rogados, e o seu trabalho seminal foi publicado por Theiler et al. (1992). Desde então, inúmeros trabalhos começaram a vir a lume abordando uma diversidade de problemas, sejam eles a detecção de gaussianidade e reversibilidade temporal em séries (Hinich et al., 2005), a detecção de determinismo em séries temporais quase-periódicas (Coelho, 2006) e em séries de variabilidade da frequência cardíaca (Gomes, 2001), a identificação de caos de baixa dimensão em dados eletroencefalográficos (Theiler and Rapp, 1996), a caracterização de dependência ou correlação temporal entre ciclos de um conjunto de dados que apresenta uma forte componente periódica (Theiler, 1995), a distinção entre determinismo e aleatoriedade nas flutuações irregulares de uma série temporal (Nakamura and Small, 2005; Jesus et al., 2007), dentre tantos outros. No que tange ao problema abordado nesta pesquisa, os algoritmos apresentados nas duas últimas citações foram aplicados a séries de consumo de energia elétrica para identificação de determinismo entre ciclos e intra-ciclo, respectivamente. Um estudo comparativo de diversas técnicas de análise de dados sub-rogados já propostas pode ser encontrado em (Costa, 2008).

A terceira pergunta, que traz em seu bojo a problemática da previsão de séries temporais, já foi abordada na literatura sob os mais diversos aspectos e diferentes enfoques, conforme está registrado mais adiante no Capítulo 2 deste trabalho.

O presente trabalho se propõe, dentro do espírito galileano de observação dos fatos (medições) para se tentar explicar com matemática (modelos preditivos) o fenômeno em análise, a estudar técnicas variadas de previsão de séries temporais que possam ser aplicadas a dados reais, em especial a séries de consumo de energia elétrica, bem como a tratar do problema de detecção de determinismo nas referidas séries via análise de dados sub-rogados.

1.2 Motivação

Mudanças viscerais no cenário do mercado de energia elétrica têm se revelado uma realidade em muitos países nas últimas décadas, à guisa de um processo de

²Diz-se que um sistema dinâmico é determinístico se existe uma regra que pode ser utilizada para se prever os seus estados futuros dado que haja conhecimento suficiente acerca do seu estado atual (Small, 2005).

transformação econômica que assume proporções em nível mundial. Países como a Argentina, o Canadá, a Noruega, a França e os Estados Unidos submeteram-se, nos últimos anos, a mudanças significativas em seu modelo de setor elétrico (Aguirre et al., 2006).

Na Argentina, por exemplo, houve uma reforma do modelo na década de 1980, por ocasião do governo de Carlos Menen e em atendimento a recomendações do Banco Mundial, com os objetivos de resolução do problema do déficit de energia e de melhoramento da eficiência econômica do setor.

No Canadá, onde se registra o maior consumo de energia elétrica *per capita* do mundo (produção majoritariamente hidrelétrica), a separação dos elos da cadeia produtiva da indústria elétrica se processou em meados da década de 1990. Cumpre ressaltar que tal reforma “(...) *se preocupa em construir um sistema competitivo e confiável e em manter um setor gerador de renda e de emprego para a sociedade*” (Aguirre et al., 2006).

Outro país marcado por mudanças é a Noruega, em que o antigo modelo misto público-privado, originado nas concessões tanto de direito de propriedade sobre as quedas d’água como da geração e distribuição local de energia, submeteu-se a um processo de reestruturação que orientou o setor elétrico na direção de um mercado de energia competitivo. Desdobrando-se esse processo sem a intervenção de iniciativas de privatização, definiu-se uma clara separação entre o Estado e as agências operadoras do sistema de energia. Interessante frisar que o novo modelo noruguês nasceu nas universidades por solicitação governamental e se concretizou a partir de 1991.

O modelo francês caracteriza-se pela forte influência do Estado como consequência natural do desenrolar histórico do país. A falência do modelo liberal antes da crise de 1929, as tradições republicanas e napoleônicas e o fato de os industriais terem assumido uma postura ambígua frente aos invasores no período pós-guerra contribuíram para o recrudescimento do modelo de Estado centralizador, o que implica o estabelecimento de um Estado diretor da economia “*por meio do planejamento indicativo dos instrumentos de crédito e de um conjunto de grandes empresas ditas estratégicas*”. (Aguirre et al., 2006). Atuando nesse cenário, a Eletricité de France (EDF) é uma empresa que se estruturou sobre tal ideologia, tornando-se assim detentora da maior capacidade instalada da Europa Ocidental, atendendo ao mercado interno e exportando parte da produção excedente.

Os Estados Unidos também passaram por um processo de desregulamentação em vários setores de sua economia, dentre eles o setor elétrico. Em sua nova formatação de mercado de energia, os consumidores tem liberdade de escolha entre as empresas geradoras, os preços de tarifa devem ser definidos pelo próprio mercado, transmissão e distribuição continuam a ser atividades reguladas, enquanto que os fornecedores de energia não são regulados, e controles de ordem federal e estadual são definidos como profilaxia para que as empresas que prestam serviços não criem barreiras para novos entrantes.

No Brasil, o ano de 2004 foi marcado por mudanças profundas no setor elétrico nacional, com o processo de desverticalização das empresas que ocorreu visando atender à Lei nº 10.848, de 15 de março do mesmo ano. No caso da CEMIG, a empresa desmembrou-se em duas: a CEMIG Distribuição de Energia S.A. e a CEMIG Geração e

Transmissão S.A. Essa Lei estabeleceu os termos em que seria processada a comercialização da energia elétrica em ambiente de contratação livre ou regulado. De acordo com a própria Lei, "a contratação livre dar-se-á nos termos do art. 10 da Lei nº 9.648, de 27 de maio de 1998, mediante operações de compra e venda de energia elétrica envolvendo os agentes concessionários e autorizados de geração, comercializadores e importadores de energia elétrica e os consumidores que atendam às condições previstas nos arts. 15 e 16 da Lei nº 9.074, de 7 de julho de 1995, com a redação dada por esta Lei." (Lei, 2004). Dentre as referidas condições, a do artigo 16 é de que: "é de livre escolha dos novos consumidores, cuja carga seja igual ou maior que 3.000 kW, atendidos em qualquer tensão, o fornecedor com quem contratará sua compra de energia elétrica." (Lei, 1995). Isso é o que caracteriza o chamado consumidor livre, o qual pode adquirir energia junto ao fornecedor que lhe aprouver, diferentemente do consumidor que não satisfaz essa condição de carga mínima, o qual é chamado de consumidor cativo. Este acaba tendo que recorrer obrigatoriamente ao distribuidor ou concessionária local.

A comercialização a que se refere a citada Lei, "será realizada nos termos da Convenção de Comercialização, a ser instituída pela Agência Nacional de Energia Elétrica - ANEEL (...)" (Lei, 1995). Será papel da ANEEL prever: "I - as obrigações e os direitos dos agentes do setor elétrico; II - as garantias financeiras; III - as penalidades; e IV - as regras e procedimentos de comercialização, inclusive os relativos ao intercâmbio internacional de energia elétrica." (Lei, 1995). Além disso, a mesma Lei ainda autoriza a criação da Câmara de Comercialização de Energia Elétrica - CCEE, a qual, sob autorização do Poder Concedente e regulação e fiscalização pela Agência Nacional de Energia Elétrica - ANEEL, tem como função viabilizar a comercialização de energia elétrica.

Em 30 de julho de 2004, foi assinado o Decreto nº 5.163, o qual, dentre as providências a que se destina, regulamenta a comercialização de energia elétrica e o processo de outorga de concessões de autorizações de geração de energia elétrica, dentro do novo modelo do setor elétrico instaurado com a Lei nº 10.848. De acordo com esse Decreto, caberá à ANEEL expedir a convenção de comercialização, as regras de comercialização e os procedimentos de comercialização.

Dentre as importantes definições estabelecidas, ressaltem-se as seguintes: "(...) VI - ano-base "A" o ano de previsão para o início do suprimento da energia elétrica adquirida pelos agentes de distribuição por meio dos leilões de que trata este Decreto; VII - ano "A - 1" o ano anterior ao ano-base "A" em que se realizam os leilões de compra de energia elétrica; VIII - ano "A - 3" o terceiro ano anterior ao ano-base "A" em que se realizam os leilões de compra de energia elétrica; IX - ano "A - 5" o quinto ano anterior ao ano-base "A" em que se realizam os leilões de compra de energia elétrica; (...)" (Decreto (2004)).

Esses referidos leilões são realizados no ACR (Ambiente de Contratação Regulada), e comercializam a energia elétrica proveniente de empreendimentos de geração já existentes (empreendimentos que até a data de publicação do respectivo edital de leilão: I - não sejam detentores de concessão, permissão ou autorização; ou II - sejam parte de empreendimento existente que venha a ser objeto de ampliação, restrito ao acréscimo da sua capacidade instalada) (Decreto, 2004) e proveniente de novos empreendimentos de geração. De acordo com o Decreto: "1º Os leilões para compra de energia elétrica

de que trata o caput serão promovidos, observado disposto nos arts. 60 a 64, nos: I - anos "A - 5" e "A - 3" para energia elétrica proveniente de novo empreendimento de geração; II - ano "A - 1" para energia elétrica proveniente de empreendimento de geração existente."(Decreto, 2004).

Em outras palavras, as concessionárias deverão efetuar a contratação da energia junto às geradoras com um, três ou cinco anos de antecedência, por meio do leilão de energia realizado pela ANEEL, conforme as diretrizes do Ministério de Minas e Energia.

Além disso, a partir de 2005, todos os agentes de distribuição, vendedores, autoprodutores e os consumidores livres deverão informar ao Ministério de Minas e Energia, até 1º de agosto de cada ano, as previsões de seus mercados ou cargas para os cinco anos subseqüentes.

Dentro dessa estrutura, após realizadas as contratações de energia, existe uma possibilidade de ajuste dentro de uma dada margem até a data de entrega da energia às distribuidoras. Qualquer ajuste além dessa faixa implica multas para as concessionárias de energia. Por essa razão, a necessidade da CEMIG Distribuição em ter em mãos um preditor que seja capaz de estimar o consumo energético com uma precisão muito grande, de preferência com um horizonte de previsão de cinco anos, em que a margem de erro aceita é de +3%. O descumprimento dessa exigência implica multas para as empresas distribuidoras de energia.

1.3 Objetivos do Projeto

Dentro do contexto apresentado acima, o presente trabalho tem o objetivo de, no primeiro momento, analisar os dados em uma tentativa de se detectar determinismo nas séries temporais de consumo de energia elétrica. Para tanto será utilizada a análise de dados sub-rogados, que nada mais é do que um teste de hipóteses aplicado a séries temporais.

Num segundo momento, serão apresentados os resultados atingidos com as ferramentas matemática aplicadas ao problema de previsão de séries a longo prazo. Para a realização dessas previsões, foram utilizadas três ferramentas diferentes, uma delas inspirada nos já tão conhecidos modelos auto-regressivos (AR) e as outras duas oriundas da área de pesquisa em inteligência artificial: a rede neural artificial (RNA) e a rede neuro-fuzzy linear local (RNFL). Esses modelos serão descritos em detalhes no Capítulo 4.

O grande desafio desse trabalho, conforme se verá no seu decorrer, reside no fato de que as séries estudadas, e em especial as séries da CEMIG, são não-estacionárias, com flutuações irregulares e constituídas de um número bastante reduzido de amostras.

1.4 Organização do Trabalho

No Capítulo 2 desta dissertação, é apresentada uma revisão bibliográfica sobre as diversas abordagens utilizadas em todo o mundo para lidar com o problema de

previsão de séries temporais a longo prazo, e em especial no que diz respeito a séries temporais de consumo ou demanda de energia. O Capítulo 3 consiste na descrição e análise dos dados, em que são realizadas: (i) a retirada dos valores espúrios das séries; (ii) a decomposição das séries em componentes constituintes; e (iii) a análise de dados-sub-rogados, utilizada para identificação de determinismo em séries temporais. No Capítulo 4, são descritas as metodologias de previsão que serão utilizadas nesse trabalho. No Capítulo 5, são apresentados os resultados obtidos com os modelos preditivos apresentados no capítulo 4. Por fim, no Capítulo 6, tem-se a conclusão da dissertação.

Uma Breve Revisão de Modelagem e Previsão de Séries Temporais

“Os físicos acham que tudo o que temos de fazer é dizer: estas são as condições, o que acontece em seguida?”

Richard Phillips Feynman (1918-1988)

2.1 Modelagem e Previsão de Séries Temporais

Considere a seguinte definição de série temporal dada por Tong (1996):

Definição 2.1.1 *Uma série temporal, $\{Y_t\}$, é uma sequência de valores reais indexados por $t \in \mathbb{Z}$, em que $\mathbb{Z}$ representa o conjunto dos números inteiros.*

Num contexto geral, séries temporais podem ser contínuas ($t \in \mathbb{R}$) ou discretas ($t \in \mathbb{Z}$). Portanto a Definição 2.1.1 seria a de uma série temporal de tempo discreto. Todavia, como nesse trabalho não serão analisadas séries de tempo contínuo, sempre que utilizarmos a expressão “série temporal”, estaremos nos referindo ao caso discreto.

Vasta é a literatura científica em que se podem encontrar bem estabelecidas as bases da modelagem de séries temporais, em especial no que diz respeito ao caso linear e estacionário, ou seja, a séries que podem ser representadas pelo seguinte modelo dinâmico geral:

$$Y_t = \sum_{i=1}^{n_Y} a_i Y_{t-i} + \sum_{j=0}^{n_U} b_j U_{t-j} + \sum_{k=0}^{n_\xi} c_k \xi_{t-k} \quad (2.1)$$

em que n_Y , n_U e n_ξ são respectivamente o máximo valor de atraso da série (Y_t), da entrada (U_t) e dos termos de ruído (ξ_t).

Com relação a séries não-lineares, muito já se produziu e se continua a produzir em termos de conhecimento e de ferramentas de análise, modelagem e previsão, mas em comparação com o que já existe de bem definido no mundo linear, a abordagem não-linear ainda carece de um amadurecimento mais profundo. De acordo com (Gooijer and Hyndman, 2006), diante dos fundamentos teóricos do estudo de séries temporais lineares, a análise e a previsão de séries temporais não-lineares ainda se encontram em um estado incipiente de desenvolvimento.

Segundo Tong (1996), talvez o objetivo mais importante no estudo de uma série temporal consista na busca da descoberta da lei dinâmica que lhe é subjacente. Em

outras palavras, qual é a lei que governa a geração da série? Para um entendimento completo dessa lei, necessário se faz um conhecimento igualmente profundo das leis físicas, químicas ou biológicas que descrevem o tipo de processo em estudo. Como em muitas situações esse conhecimento prévio não é possível, o que normalmente se dá pela complexidade do sistema que gerou a série temporal ou pelo desconhecimento da teoria que rege o tipo de fenômeno analisado, o que resta ao pesquisador são unicamente os dados. Diante disso, um possível procedimento a ser adotado seria (Tong, 1996):

1. o reconhecimento de características importantes dos dados disponíveis;
2. a construção de um modelo empírico¹ que seja, tanto quanto possível, teoricamente embasado;
3. a constatação de que o modelo é capaz de capturar as características reconhecidas no primeiro estágio e a verificação da possibilidade de se incorporar alguma melhoria ao modelo proposto.

Como a especificação de um modelo para séries temporais depende muitas vezes de um julgamento subjetivo, não é absurdo dizer que essa tarefa é tanto uma arte quanto é uma ciência. É até natural pensar assim, à maneira do que acontece em modelagem matemática de sistemas dinâmicos (Aguirre, 2007), uma vez que não existe uma solução ou metodologia aplicável a todos os tipos de problemas em séries temporais que conduza sempre a resultados satisfatórios. A eficácia da metodologia aplicada está intimamente ligada, pelo menos, à natureza dos dados (determinismo, aleatoriedade, linearidade, não-linearidade, estacionariedade) e à adequação da ferramenta à tarefa que se deseja realizar (previsão, controle, detecção de faltas), etc.

Um trabalho que pode ser considerado um marco na história da previsão de séries temporais é o livro de Box and Jenkins (1976). De acordo com Gooijer and Hyndman (2006), essa obra integrou todo o conhecimento existente até então (meados da década de 70 do século XX) na área de séries temporais, no que dizia respeito a estimação de parâmetros, identificação e previsão². Ainda no seu prefácio, os autores deixam bem claro que se propuseram a abordar a construção de modelos estocásticos no domínio do tempo para séries temporais discretas que sejam os mais simples possíveis, com número mínimo de parâmetros, satisfazendo portanto um critério de parcimônia. Dentre as razões que eles elencam para a escolha de tal critério, enumeram-se: (i) o fato de esses modelos serem capazes de revelar algo sobre a natureza dos processos que geram as séries temporais³; e (ii) a possibilidade de esses modelos serem utilizados para a obtenção de previsões ótimas de valores futuros das séries.

¹Um modelo que representa uma hipótese relacionada à probabilidade de transição ao longo do tempo, isto é, um modelo dinâmico (Tong, 1996).

²Vale dizer que uma grande contribuição para o desenvolvimento de ferramentas e métodos para a análise de séries temporais deve-se ao trabalho de Yule (1927), em que ele propôs a idéia de se olhar para uma série temporal simplesmente como a realização de um processo estocástico.

³E certamente modelos mais simples são mais fáceis de ser interpretados, em termos de como cada parâmetro influencia a sua resposta, por exemplo.

Ainda nesse livro, é proposta uma metodologia para a modelagem de séries temporais, a qual ficou conhecida como abordagem de Box-Jenkins. Constitui-se ela de quatro estágios:

1. a definição de uma classe de modelos que satisfaça os objetivos da modelagem;
2. dentro dessa classe, que pode ser extensa, *identificar* uma subclasse mais adequada a partir de conhecimentos prévios acerca do sistema e dos dados. Nesse etapa, podem até ser utilizadas estimativas mais “grosseiras” dos parâmetros para se definir essa subclasse;
3. a *estimação* de parâmetros do modelo selecionado. As estimativas preliminares que por ventura tenham sido realizadas no estágio 2 podem ser utilizadas como valores iniciais de métodos iterativos de estimação de parâmetros;
4. realização de um *diagnóstico*, que nada mais é do que a verificação da qualidade do modelo obtido. Caso ele não se ajuste bem aos dados, um ciclo iterativo de *identificação*, *estimação* e *diagnóstico* pode ser iniciado em busca de uma representação matemática que melhor se adeque ao problema em questão.

Uma outra questão relevante ainda abordada por Box and Jenkins (1976) é a do estabelecimento dos limites de confiança para as previsões obtidas com os modelos estimados considerando-se diferentes horizontes de previsão. Em outras palavras, a definição de faixas em tornos dos valores previstos dentro das quais o valor real da série assuma algum valor com uma determinada probabilidade associada. Dessa maneira, tem-se um cálculo dos riscos decorrentes das decisões que venham a ser tomadas com base nas previsões obtidas.

Em um outro livro interessante, Weigend and Gershenfeld (1996) apresentam os resultados de uma competição organizada no Instituto de Santa Fé (USA) em torno da temática de previsão de séries temporais. Em suma, os autores disponibilizaram na internet diferentes bases de dados das mais diversas procedências (a exemplo conjuntos de dados biomédicos de um paciente com apnéia durante o sono, dados astrofísicos de uma anã-branca, e até mesmo uma fuga de Bach que o renomado compositor alemão não foi capaz de terminar antes de sua morte), de modo que diferentes pesquisadores tivessem acesso a eles e tentassem aplicar-lhes técnicas de extração de informação visando caracterização, modelagem e/ou previsão das séries. Trata-se, portanto, de uma compilação de artigos produzidos pelos diferentes autores da comunidade científica internacional que se dedicaram à competição. Um dos objetivos de tal publicação era demonstrar que técnicas de análise de séries temporais, então recém-desenvolvidas no âmbito dos estudos de sistemas caóticos, seriam aplicáveis a diferentes áreas, extrapolando assim os limites de suas origens, bem como evidenciar que muitas vezes teorias consideradas abrangentes se mostrariam inadequadas quando aplicadas a problemas que envolvessem dados reais contaminados com ruído e não poucas vezes gerados por sistemas fortemente não-lineares.

Uma exaustiva revisão de literatura quanto ao que se tem realizado na área de previsão de séries temporais nos últimos 25 anos em todo o mundo pode ser encontrada

em (Gooijer and Hyndman, 2006). Em suas trinta páginas, o artigo apresenta discussões em torno de modelos das mais diversas naturezas, sejam de suavização exponencial (Holt, 2004), modelos ARIMA (Auto-Regressive Integrated Moving Average) (Box and Jenkins, 1976; Campos, 2008), modelos em espaço de estados (como o filtro de Kalman) (Al-Hamadi and Soliman, 2006), modelos neurais (Braga et al., 2000), etc.

Em (Ghoulipour et al., 2007), é apresentada uma abordagem bastante interessante ao problema de previsão de longo prazo⁴. Em primeiro lugar, os autores fazem a extração das principais características ou componentes da série temporal via análise espectral singular (SSA). Essas componentes principais (PCs) identificadas apresentam tendências lineares ou não-lineares e também padrões periódicos, o que supostamente as torna previsíveis a longo prazo, segundo os autores. Em seguida à etapa de extração de características, modela-se cada uma das PCs com uma rede neuro-fuzzy linear local (RNFL), que nada mais é do que uma espécie de modelo Takagi-Sugeno (Takagi and Sugeno, 1985; Nelles, 2001), e faz-se a previsão em separado de cada uma delas. Para se obter a previsão total da série, basta que se combinem aditivamente as previsões de cada uma de suas componentes. A metodologia foi validada utilizando-se séries temporais de sistemas não-lineares caóticos amplamente estudados, como o mapa de Ikeda e o sistema de Lorenz, e aplicada a dados reais de atividade solar.

Em (Perepelitsa et al., 2006), encontra-se uma abordagem diferenciada para previsão de séries utilizando autômatos celulares, métodos de classificação e teoria de conjuntos nebulosos. Os autores afirmam ser essa uma forma mais interessante de se tratar o problema em situações nas quais métodos clássicos de previsão não conduzem a resultados satisfatórios em termos de redução das medidas de erro.

Na mesma linha de trabalhos não-convencionais, por assim dizer, uma outra abordagem pouco usual é apresentada em (Nunn and White, 2005), em que os autores mesclam dois métodos oriundos da Computação Evolucionária: um método de busca antigênica, inspirado no comportamento dos sistemas imunológicos naturais, e o PSO (do inglês *Particle Swarm Optimization*), inspirado na existência de indivíduos que vivem em grupo, sendo levadas em conta, portanto, as experiências individuais de cada um de seus membros e a experiência do grupo como um todo, a exemplo os enxames de abelhas, colônias de formigas, revoadas de pássaros, dentre outros. Apesar de ser uma abordagem aparentemente sofisticada, é questionável a sua real eficácia uma vez que, no dizer dos próprios autores, ela só se presta a tratar do problema de previsão de um passo à frente. Além disso, os resultados são apresentados no artigo de forma muito confusa, o que dificulta uma avaliação mais concreta de sua qualidade.

No que diz respeito ao melhoramento ou entendimento mais profundo dos resultados das previsões, uma abordagem que parece ser de grande utilidade seria a aplicação de *wavelets*, conforme se verifica em (Ramsey, 1999). Em verdade, a própria modelagem matemática de sistemas dinâmicos não-lineares pode ser realizada utilizando-se as chamadas *wavelet networks*, à maneira do que se encontra em (Postalcioğlu and Becerikli, 2007) e (Ozkurt and Savaci, 2006). Outras metodologias baseadas em *wavelets* e com finalidades de previsão de séries temporais são apresentadas em (Yousefi et al., 2005) e

⁴Terminologia definida na página seguinte.

(Frizlewics et al., 2003).

2.2 Modelagem e Previsão de Séries Temporais de Consumo de Energia Elétrica

De acordo com (Ghiassi and Saidane, 2006), o problema de previsão de consumo ou demanda de energia, do ponto de vista do horizonte de previsão, pode ser dividido em três grandes subcategorias:

- Previsão de curto prazo, em que horizonte de previsão máximo considerado é de dois dias. Convém ressaltar que essa é a subcategoria mais estudada na literatura científica;
- Previsão de médio prazo, em que o horizonte de previsão se estende até um limite superior de dois anos;
- Previsão de longo prazo, em que o horizonte de previsão é maior do que dois anos. De todas as três subcategorias, é a que menos recebe considerações na literatura.

Devido às recentes mudanças no setor energético mundial, o problema de previsão de consumo ou demanda a longo prazo tem começado a receber uma maior atenção na literatura científica, ainda que de forma um tanto tímida. É fundamental reconhecer que se trata de um problema bastante difícil, uma vez que os fatores que influenciam o comportamento do setor elétrico são os mais diversos (econômicos, políticos, climáticos, etc). Além desses aspectos, é justo ressaltar que a grande maioria dos métodos de previsão existentes se adequam bem a série temporais estacionárias, ao passo que as séries temporais de consumo de energia são comumente não-estacionárias⁵. Não é difícil perceber, portanto, que a complexidade do problema recrudescer à medida que se aumenta o horizonte de previsão desejado.

A seguir são apresentadas algumas das técnicas recentemente propostas na literatura para a solução do problema de previsão a longo prazo de consumo ou demanda de energia elétrica.

Em (Kermanshahi and Iwamiya, 2002), é apresentada a previsão de picos de demanda para as nove companhias de energia no Japão, até o ano de 2020, utilizando-se redes neurais artificiais (na verdade, dois tipos: rede back-propagation com três camadas e rede neural recorrente). Segundo os autores, as previsões de longo prazo são mais afetadas por fatores econômicos do que por condições climáticas, ao contrário das previsões de curto prazo. Após uma investigação cautelosa, eles elegeram como entradas dos modelos neurais: (i) o produto nacional bruto (PNB); (ii) o produto interno bruto (PIB); (iii) a população; (iv) o número de casas; (v) o número de aparelhos

⁵De acordo com Aguirre (2007), "Um processo é estacionário se as leis de probabilidade que o regem não variam com o tempo. Pode-se dizer que um processo estacionário está em equilíbrio estatístico. O conceito de estacionariedade está intimamente ligado ao de *invariância*." Há definições mais formais de não-estacionariedade, conforme pode ser verificado em (Papoulis and Pillai, 2002), por exemplo.

de ar condicionado; (vi) a quantidade de CO₂ proveniente de poluição; (vii) o índice de produção industrial; (viii) o preço do petróleo; (ix) o consumo de energia; e (x) o preço da energia elétrica. Ambas as redes foram treinadas com dados de 1975 a 1995, com tempo de amostragem de um ano o que implica num conjunto de dados de treinamento com 21 valores. As previsões foram realizadas para os anos de 1999, 2000, 2005, 2010, 2015 e 2020. Os resultados de previsão para o ano de 1999 apresentaram erro de apenas 3%, o que é bastante satisfatório, uma vez que os erros de predição aceitáveis nas companhias japonesas são da ordem de 10%, segundo os autores.

De um modo geral, de acordo com (Jia et al., 2001), os algoritmos utilizados para previsão de carga a longo prazo são de dois tipos: métodos paramétricos e métodos baseados em inteligência artificial (IA). Os primeiros, por expressarem as relações entre a carga e os fatores que a afetam através de fórmulas explícitas, têm a capacidade de lidar com informação lógica. Por outro lado, os métodos baseados em IA lidam com informação intuitiva, uma vez que tentam mimetizar o comportamento do sistema nervoso humano, no caso específico das redes neurais artificiais. Estes últimos, sendo modelos do tipo caixa-preta, dificultam a compreensão da relação entre as entradas e saídas do sistema. Na tentativa de superar essas limitações, Jia e colegas propuseram um modelo para previsão a longo prazo baseado na Teoria da Simulação Geral (GSIM), a qual lida com informações lógicas e intuitivas e consideram as relações entre as variáveis que afetam a carga de forma simultânea e integrada.

Os fatores que afetam a carga considerados nessa abordagem foram: (i) energia para os aparelhos de ar condicionado; (ii) produto interno bruto (PIB); (iii) demanda de energia; (iv) preço do petróleo; (v) número de empregados; (vi) abastecimento e fornecimento de energia; (vii) potência elétrica gerada; (viii) população; (ix) investimento em instalações; (x) índice de produção industrial; e (xi) preço da energia elétrica. Aplicando o modelo GSIM para o caso do sistema de distribuição de energia de Tokyo, Jia e colegas utilizaram um histórico de dados anual de 25 valores que vai de 1972 a 1996 (os primeiros 20 para treinamento e os outros 5 para previsão). Para um horizonte de previsão de cinco anos, eles encontraram um erro máximo de -2,83% e um erro médio absoluto de 0,89% no consumo médio anual. O método GSIM mostrou-se superior em termos de minimização do erro cometido na previsão quando comparado com o método tradicional de regressão linear múltipla.

Em (Kandil et al., 2002), é mostrado um aspecto relevante na escolha dos modelos utilizados em problemas de previsão de demanda a longo prazo. Comparando-se a previsão de pico de demanda para um cenário de desenvolvimento normal, de característica mais estável, com um cenário de crescimento mais rápido (variância temporal mais intensa, devido a fatores como a introdução de competição no ambiente de geração de energia com possíveis mudanças nas estruturas tarifárias), verifica-se que os métodos clássicos de previsão conduzem a bons resultados no primeiro caso, mas o erro associado ao segundo é muito expressivo, o que torna necessária uma busca da melhoria dos modelos disponíveis atualmente. A solução proposta pelos autores é a de desenvolvimento de um sistema especialista que, com base em regras de decisão que relacionem as variáveis que afetam a previsão, escolha, em sua base de dados, o algoritmo preditivo mais apropriado para a conjuntura do cenário energético naquele

momento. Para o caso de um cenário de rápido crescimento, com um conjunto de treinamento que vai de 1981 a 1993 (13 valores, uma para cada ano), os resultados obtidos para previsão num horizonte de 5 anos (1994 a 1998) com o sistema especialista são muito mais eficazes, segundo os autores, do que utilizando-se métodos clássicos baseados em séries temporais ou métodos econométricos.

Uma situação semelhante ao que acontece no mercado energético brasileiro atualmente encontra-se em (Nakashima, 2000). Neste trabalho, o autor faz uma análise da conjuntura do ponto de vista de três segmentos: o operador do sistema, que tem o papel de agir como entidade neutra, evitando o conflito de interesses na relação entre os fornecedores (geração) e os participantes da compra (distribuição); o regulador, que estabelece as regras de competição para reduzir o preço da energia elétrica; e os participantes da compra da energia, correspondentes às empresas distribuidoras. Deste último ponto de vista, é necessário que haja uma estimativa do preço da energia e da sua demanda para fins de planejamento econômico e tomada de decisões efetivas. O modelo levantado para a previsão de ambas as variáveis (preço e demanda) é baseado em teoria de possibilidades e auto-regressão nebulosa (fuzzy), e é aplicado a dados reais do ano de 1999 no estado da Califórnia (EUA) num horizonte de previsão de um mês (o conjunto de dados de treinamento foi de 52 valores correspondentes ao período entre 8 de janeiro e 28 de fevereiro, e os valores previstos foram os de cada dia do mês de março do referido ano). A principal razão para a escolha deste tipo de representação é o fato de que na modelagem nebulosa assume-se que os preços e o consumo de energia refletem mais a atitude de tomada de decisão humana do que um processo estocástico, como costuma-se assumir em modelos convencionais. Para ilustrar isso, o autor cita o caso de transações de energia elétrica baseadas no mercado, as quais tendem a tornar a demanda mais volátil em altos volumes do que em baixo volumes, o que faz com que os dados, nesse caso, não sejam explicados satisfatoriamente por uma distribuição normal.

Nessa abordagem, a estrutura do modelo é identificada por meio de análise de autocorrelação e os parâmetros de regressão são dados por números nebulosos, ou seja, os parâmetros não apresentam um valor pontual exato, determinístico, mas uma distribuição de valores dada por uma função de pertinência triangular. Esses parâmetros são calculados via programação linear, tendo como critério de otimização a minimização da imprecisão do modelo.

Uma outra vantagem do modelo nebuloso escolhido é que a previsão realizada inclui faixas de possibilidade (incerteza), que podem auxiliar em tomadas de decisão, ao passo que os modelos convencionais convergem para um único ponto mais provável de ocorrência em um dado instante de tempo.

Em (Cullen, 1999), são realizadas previsões de picos de demanda residencial durante o verão no sul de West Virgínia (EUA) por meio de modelos de regressão e via simulação de Monte Carlo. A principal limitação inerente ao projeto foi a insuficiência do conjunto de dados disponíveis, que consiste nos picos de demanda anuais de 1980 a 1995. O modelo de regressão apresenta como resposta o valor esperado com sua incerteza associada, a qual indica a precisão da previsão. Já a simulação de Monte Carlo gera uma distribuição de valores da demanda futura a partir da distribuição de valores

das variáveis de entrada assumidas (temperatura, preço da energia elétrica, preço do gás natural, e alguns indicadores econômicos). O horizonte de predição considerado é de dez anos, o que enquadra o problema na categoria de previsão a longo prazo. Como o trabalho foi realizado durante o período em que estava em desenvolvimento o plano de desregulamentação do setor de geração de energia elétrica em West Virgínia, a necessidade de se obter previsões a longo prazo mais precisas se tornou uma questão premente, devido ao elemento de incerteza adicionado pela mudança na estrutura do mercado energético com o estabelecimento do novo modelo proposto.

A conclusão a que a autora chegou é de que nenhum dos dois modelos conduziu a resultados aceitáveis. A implicação mais notória, portanto, é que um conjunto de dados insuficiente impossibilita a obtenção de bons resultados com os modelos de predição, conduzindo a uma sub ou superestimação dos picos de demanda.

Uma abordagem diferente e interessante do problema de predição em um ambiente de mudanças rápidas é encontrada em (Punjabi, 2005), ainda que não necessariamente no contexto do consumo de energia elétrica. O caso estudado neste trabalho é o da indústria de grãos de feijão nos Estados Unidos. Como decorrência das mudanças, nos últimos anos, no campo da biotecnologia, da industrialização da agricultura e das preferências dos consumidores, a indústria alimentícia passou por alterações em sua estrutura e organização nunca dantes vistas. Para realização de planejamentos e predições, tipicamente são realizadas análises históricas com fins de extrapolação para a dedução dos comportamentos futuros. Todavia, em ambientes de mudança estrutural, a análise baseada apenas em dados históricos não parece ser a mais adequada. Dessa forma, o que o autor propõe é que, juntamente com a abordagem econométrica tradicional, seja realizada uma análise de cenários, que é prática comum na área de administração estratégica para planejamento de ambientes em rápidas mudanças, a qual consiste na identificação das forças orientadoras (sociais, econômicas, políticas, ambientais, etc), identificação das tendências básicas e das principais incertezas dentre as forças orientadoras, e no estabelecimento de alguns diferentes cenários com base na interação entre as tendências básicas e as principais incertezas. Esse estudo de cenários permitiu que se identificassem a demanda de grãos de feijão e o impacto da globalização como as mais prováveis forças orientadoras que afetam a indústria. Os cenários futuros foram obtidos incorporando-se ao modelo econométrico os resultados da análise de cenários. Três foram os cenários delineados para a indústria de grãos de feijão: i) “More of the Same” (mais do mesmo), que seria uma continuidade do comportamento corrente, sem mudanças significativas; ii) “Challenged in a Global Market” (desafiado em um mercado global), em que a indústria norte-americana enfrentaria a competitividade dos produtores globais; iii) “Saved by Consumer Demand” (salvo pela demanda dos consumidores), que seria o cenário mais otimista, resultante de uma elevação considerável da demanda de feijão.

A principal vantagem dessa abordagem é que ela permite vislumbrar um conjunto de possibilidades futuras, auxiliando no processo de tomada de decisão em um ambiente de alta variabilidade comportamental, o que não seria possível com a abordagem tradicional baseada unicamente nos dados históricos.

Na China, conforme encontra-se em (Liao, 2001), foi realizado um estudo do impacto das tecnologias utilizadas para o melhoramento da eficiência energética no setor residencial, a partir de predições de demanda energética, uma vez que a quantidade de recursos energéticos *per capita* no país se encontra abaixo da média mundial. Também foram realizadas previsões de demanda residencial sob diferentes cenários de crescimento econômico e preço da energia. A abordagem ao problema foi feita de acordo com um modelo de previsão de vários níveis (MLFM - Multiple-Level of Forecasting Model). No presente caso, foram considerados quatro níveis: RHEM (Residential Household Energy Model), que representa o consumo em uma única casa, RCEM (Residential Community Energy Model), que representa o consumo em comunidades residenciais, RREM (Regional Residential Energy Model), que integra o consumo em uma região que compreende um conjunto de comunidades, e NREM (National Residential Energy Model), que engloba o consumo total da China. Essa maneira de tratar o problema permite o isolamento das variáveis mais relevantes em cada um dos níveis tratados. Os modelos gerados para a previsão foram de natureza neuro-fuzzy. Para treinamento, foram utilizados conjuntos de dados históricos numa faixa de 5 a 10 anos e o horizonte de previsão considerado foi de 10 anos.

Um estudo de previsão de temperatura, que é uma das variáveis climáticas que mais afetam a previsão de carga, é realizado em (Narasimha, 2003). O modelo utilizado é baseado em redes neurais artificiais (RNAs). Segundo o autor, as redes costumam ser de elevada complexidade para problemas de previsão, o que torna os tempos de treinamento muito longos. Para solucionar essa questão, propõe uma rede mais simples que realiza previsão de temperatura com a mesma eficiência que as redes maiores. Para tanto, o que se fez foi um pré-processamento das entradas da rede de modo a se eliminar redundâncias existentes nos dados. O horizonte de predição considerado é de um dia, o que caracteriza o problema como de previsão a curto prazo. Os dados para treinamento apresentados à rede foram amostrados por hora.

A forte relação entre as variáveis climáticas e a demanda de carga é bem evidenciada em (Tribble, 2003). Neste trabalho, são usadas informações de carga de quatro sub-estações no estado de Oklahoma (EUA) e observações climáticas para determinar o impacto das variáveis climáticas no consumo de eletricidade por três categorias de clientes: residencial, industrial e comercial. Fica demonstrado pelo autor que o entendimento correto da relação entre variáveis climáticas e demanda de eletricidade melhoram a precisão dos modelos de previsão de demanda a curto prazo (horizonte de previsão de um dia). Dois foram os tipos de modelos levantados: redes neurais artificiais (RNA's) e modelo de regressão múltipla. A grande diferença entre os modelos está no pré-processamento e na capacidade computacional requerida. No caso das RNA's, o esforço computacional é maior; no modelo de regressão múltipla, o pré-processamento é maior.

Uma abordagem ao problema de previsão de demanda a médio e longo prazo, diferente das apresentadas anteriormente, é encontrada em (Al-Hamadi and Soliman, 2005). Os autores propuseram um modelo que combina o comportamento diário (24h) e anual (52 semanas) da demanda de energia para realizar a previsão num horizonte de um ano. Os dados utilizados foram os de uma companhia de energia elétrica do

Canadá no ano de 1994, sendo a previsão realizada para 1995. O erro percentual médio absoluto obtido foi menor do que 3,8%.

Em um outro trabalho publicado no ano seguinte (Al-Hamadi and Soliman, 2006), os autores acrescentaram como uma nova etapa ao algoritmo a utilização do filtro de Kalman com o objetivo de melhorar a acurácia da previsão obtida com os múltiplos modelos auto-regressivos. Dessa forma, o erro médio absoluto obtido foi reduzido de 3,8% para 2,3%. Nesse trabalho, os autores frisam que diferentes conjuntos de dados procedentes de diferentes companhias energéticas podem conduzir a diferentes acurácias, o que é uma constatação bastante razoável⁶. Resultados obtidos para outras séries temporais de consumo de energia elétrica por meio de uma versão modificada dessa metodologia foram publicados em (Jesus and Mendes, 2008).

Mais uma metodologia em que se utilizam redes neurais para a previsão de séries de consumo de energia elétrica é apresentada em (González-Romera et al., 2007). Nesse artigo, os autores trabalham com a série de consumo de energia elétrica mensal da Espanha. Visando quebrar o problema em partes menores, a série é dividida em duas partes principais: tendência e flutuações mensais. Cada uma dessas componentes é modelada com uma rede MLP (Multi-Layer Perceptron) e são realizadas previsões para um horizonte de 60 meses, atingindo-se erros sempre abaixo de 7%.

Dentro do projeto em que também se situa esta dissertação de mestrado, já foram produzidos três trabalhos acadêmicos, sendo eles duas monografias de conclusão de curso de graduação em Engenharia de Controle e Automação da UFMG (Jesus, 2006; Mauro, 2007) e uma dissertação de mestrado pelo Programa de Pós-Graduação em Engenharia Elétrica da UFMG (Campos, 2008). No primeiro desses trabalhos (Jesus, 2006), apresentam-se os resultados obtidos com o algoritmo proposto em (Al-Hamadi and Soliman, 2005) modificado. Trabalhando-se com a série de consumo mensal de New England (EUA), o autor atingiu um MAPE de 3,24% e um MPE de 0,48% para um horizonte de previsão de 60 amostras (janeiro de 1993 a dezembro de 1997). Já em (Mauro, 2007), o autor tratou da modelagem preditiva de séries temporais mensais da CEMIG utilizando redes neurais do tipo MLP. Ajustando-se uma rede para cada uma das oito categorias de consumidores e tomando-se como previsão total a soma de cada uma das previsões individuais, em um horizonte de 24 meses, atingiu-se um MAPE 1,78% de e um MPE de -0,15%. Por fim, Campos (2008) aplicou quatro ferramentas diferentes ao problema de previsão de consumo de energia: modelos ARIMA, modelos NARIMA, redes MLP e redes neuro-fuzzy. Das quatro abordagens, a que conduziu a melhores resultados em termos de minimização das medidas de erro para o caso da previsão de 60 meses da série de New England foi o modelo ARIMA em termos do MPE (0,13%) e a rede neuro-fuzzy em termos do MAPE (2,67%).

⁶Isso ficará evidente no capítulo 5, em que são apresentados os resultados de previsão aplicando-se uma mesma metodologia a conjuntos de dados de diferentes procedências.

Descrição e Análise de Dados

3.1 Introdução

Em ciência, a uma porção do Universo que pode ser isolada para fins de estudo, sendo portanto conhecidas (ou assumidas conhecidas) as influências internas e externas que atuam sobre ela, dá-se o nome de *sistema*. Os sistemas que evoluem no tempo são especialmente classificados como *sistemas dinâmicos* (Tôrres, 2001). Um sistema dinâmico, de forma geral, pode estar sob a ação de p sinais de entrada $\{u_1(t), u_2(t), \dots, u_p(t)\}$ e revelar seu comportamento por meio de q sinais de saída $\{y_1(t), y_2(t), \dots, y_q(t)\}$, em que a variável independente t representa o tempo. É importante enfatizar que os sinais de entrada e saída de um sistema dinâmico podem ser interpretados como séries temporais, isto é, valores de uma determinada grandeza em função do tempo (Tong, 1996).

Em qualquer estudo que envolva a manipulação de dados cuja fonte geradora (que pode ser um sistema dinâmico) seja suficientemente complexa ao ponto de não serem conhecidas as leis que a regem em termos de comportamento, sempre é de bom alvitre a realização de uma análise ou pré-processamento de tais dados antes de se tirar qualquer conclusão a priori acerca da natureza do sistema em análise.

No contexto do processo de consumo de energia elétrica, que se trata de um sistema dinâmico com uma saída $y(t)$ (consumo de energia) e múltiplas entradas $u_1(t), u_2(t), \dots$ (população, PIB, renda per capita, preço do petróleo, etc), a quantidade de fatores que influenciam o comportamento das séries temporais de consumo é tão significativa que chega a ser imperiosa uma análise cuidadosa de tais dados, em reforço à filosofia de “deixar os dados falarem por si mesmos”. Assumindo-se que o que se tem em mãos é unicamente a série temporal $y(t)$ em kWh, MWh ou GWh, com $t \in \mathbb{N}^+$ medido em horas, dias ou meses, algumas questões podem ser levantadas, como por exemplo: (i) Essa série é estacionária ou não-estacionária? (ii) Ela é determinística ou estocástica? (iii) Ela apresenta padrões bem definidos, como sazonalidade ou tendência de crescimento? (iv) É possível decompor tal série em componentes?

Diante de tais apontamentos, definimos o mérito deste capítulo, o qual consiste em tratar basicamente de dois aspectos relevantes no contexto do presente projeto de pesquisa: o primeiro relacionado a abordagens para a decomposição de séries temporais, tendo por inspiração a clássica mas não menos eficiente estratégia de dividir para conquistar, a qual será utilizada na etapa de caracterização dinâmica e na definição do preditor trivial; e o segundo referente a ferramentas existentes para a caracterização dinâmica dessas séries temporais, etapa em que será enfocada especialmente a análise de dados sub-rogados. Antes de mais nada, porém, serão apresentadas as séries

estudadas e analisadas ao longo deste trabalho.

3.2 Séries Temporais de Consumo de Energia Elétrica

3.2.1 Séries da CEMIG

Duas foram as bases de dados fornecidas pela CEMIG à equipe da UFMG. A primeira consiste em 10 séries de consumo mensal de energia elétrica que compreendem o período entre Dezembro de 1996 e Setembro de 2007, concernentes às categorias: i) residencial; ii) industrial; iii) comercial; iv) rural; v) poder público; vi) iluminação pública; vii) serviço público; viii) consumo próprio; ix) concessionárias; e x) total. Os dados da segunda base correspondem à carga horária total na área coberta pela CEMIG nos anos de 2000 a 2006. A primeira base analisada foi a de dados mensais.

Base de dados mensais de consumo de baixa tensão - Dezembro de 1996 a Setembro de 2007

Todas as séries dos nove setores que formam a classe dos consumidores de baixa tensão da CEMIG são apresentadas na Figura 3.1. A Figura 3.2 apresenta duas curvas de consumo mensal¹, sendo que a primeira inclui todas as nove categorias citadas anteriormente, enquanto que a segunda é o somatório de oito daquelas categorias, exceção feita à categoria 'concessionárias', cuja série estava incompleta, compreendendo somente o período que vai de dezembro de 1996 a novembro de 1998. Uma vez que o consumo da referida categoria corresponde a um percentual considerável do consumo total, é razoável então que haja uma discrepância significativa entre as duas curvas apresentadas, especificamente no trecho correspondente aos dois primeiros anos do histórico de dados.

Para a identificação de *outliers*², foi aplicado o teste de Grubbs (Apêndice A). Nesse algoritmo, considera-se que os dados são descritos razoavelmente por uma distribuição normal, de modo que os pontos que pareçam não seguir esse padrão estatístico são caracterizados como *outliers*. Retirados os valores espúrios, uma simples interpolação polinomial de terceira ordem foi suficiente para a reconstrução da série.

Na primeira série da Figura 3.3, os valores de consumo em outubro de 1997, janeiro de 1998 e fevereiro de 1998 parecem não corresponder à realidade por não seguirem o padrão estatístico da série. É importante observar que há claramente uma mudança de comportamento por volta de maio 2001 (tempo $\approx$ 55 meses nas Figuras 3.2 e 3.3), com uma queda brusca no consumo em praticamente todos os setores (ver Figura 3.1). Esse período foi caracterizado pelo racionamento, efeito da crise energética por que passou o Brasil durante o governo Fernando Henrique Cardoso, a qual gerou cortes forçados de consumo de energia em todo o país. Observa-se, todavia, que houve um aumento do consumo nos últimos meses, como uma tentativa do sistema de retomar o seu antigo comportamento no período pré-acionamento. De qualquer

¹Por questões de sigilo, todas as séries da CEMIG apresentadas nessa dissertação foram normalizadas.

²Valores espúrios que não devem ser considerados válidos ou confiáveis.

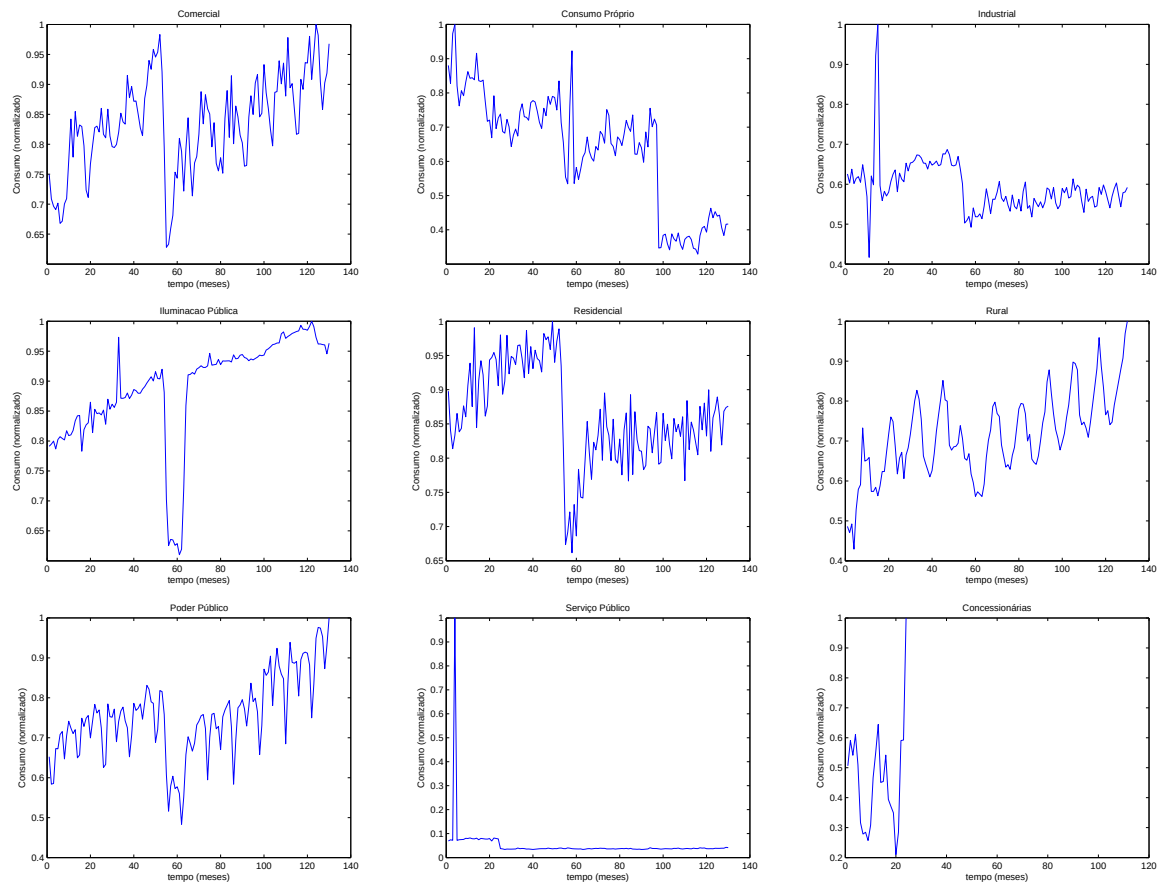


Figura 3.1: Séries dos consumidores de baixa tensão da Cemig divididos em setores - i) comercial; ii) consumo próprio; iii) industrial; iv) iluminação pública; v) residencial; vi) rural; vii) poder público; viii) serviço público; ix) concessionárias.

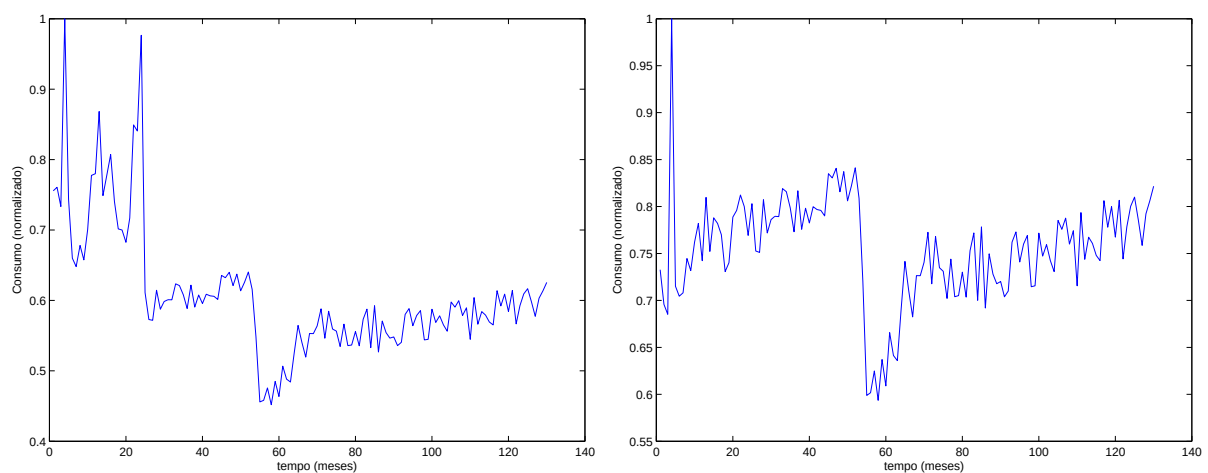


Figura 3.2: Consumo Faturado Total - O gráfico à esquerda inclui todos os setores, enquanto que gráfico à direita inclui quase todos os setores, exceção feita à categoria 'concessionárias'.

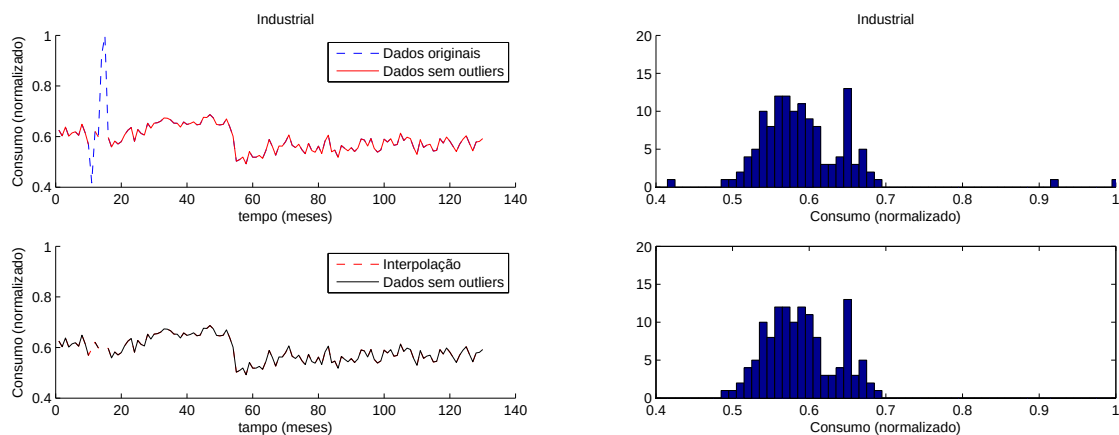


Figura 3.3: Consumo Faturado - Industrial. São apresentadas a série original e a série sem *outliers* já interpolada, cada uma ladeada por seu respectivo histograma.

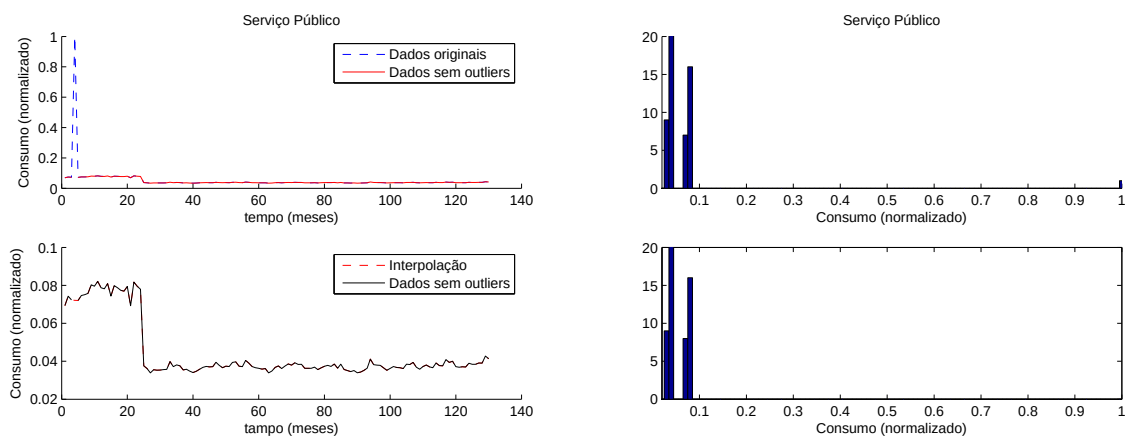


Figura 3.4: Consumo Faturado - Serviço Público. São apresentadas a série original e a série sem *outliers* já interpolada, cada uma ladeada por seu respectivo histograma.

forma, ainda assim é possível perceber, a partir da inspeção visual da Figura 3.2, que a taxa média de crescimento no período pós-acionamento ainda é menor do que no período pré-acionamento. É razoável considerar que isso se deve a uma mudança de cultura nacional em relação à problemática da economia de energia elétrica. Ainda na Figura 3.3, os histogramas mostram que após a remoção dos *outliers*, seguida da devida interpolação, a série passou a ser estatisticamente mais ‘bem comportada’.

Conforme se vê na Figura 3.4, o valor de consumo no mês de março de 1997 foi identificado como *outlier* (uma ordem de grandeza acima dos outros valores da série). É curioso observar que a queda brusca no consumo desse setor se dá num momento diferente do que se verifica nas Figuras 3.2 e 3.3, fato esse a que não foi possível atribuir uma causa. Vê-se no segundo histograma da Figura 3.4 que a série não tem um comportamento parecido com o de uma distribuição normal, mas ainda assim, usando-se o algoritmo de Grubbs, foi possível detectar a inconsistência existente, uma vez que o seu valor é notoriamente discrepante em relação aos outros valores da série³.

Após a retirada de todos os *outliers*, temos finalmente a série de consumo total pré-processada na Figura 3.5. É interessante observar que esse gráfico em muito se assemelha às séries dos setores ‘residencial’ e ‘comercial’ (ver Figura 3.1), o que não é nada surpreendente ao se olhar para a Tabela 3.1, a qual revela que juntos esses dois setores respondem por quase 3/4 (três quartos) de todo o consumo em nível de baixa tensão no Estado de Minas Gerais. Um outro aspecto digno de nota é que a série de consumo total não revela um padrão de comportamento anual bem evidente como o que se observa na série de consumo rural, por exemplo, ou na série de consumo comercial. Percebe-se, portanto, a prevalência da série do setor residencial, a qual, à maneira da série total, não apresenta ciclos anuais bem definidos, pelo menos não identificáveis por uma mera e descuidada inspeção visual dos dados.

Tabela 3.1: Consumo percentual correspondente a cada setor de consumidores de baixa tensão.

Setor	Consumo Percentual
Residencial	54,60%
Comercial	18,48%
Rural	10,20%
Iluminação pública	7,61%
Industrial	4,32%
Poder público	2,60%
Serviço público	2,00%
Consumo próprio	0,19%

³A realização do teste de Grubbs sobre as séries de baixa tensão da CEMIG após a extração das componentes de tendência, procedimento que seria mais razoável de se realizar tendo em vista a notória não-gaussianidade das séries a partir da observação dos histogramas apresentados nas Figuras 3.3 e 3.4, conduziu a resultados quase idênticos aos apresentados neste trabalho.

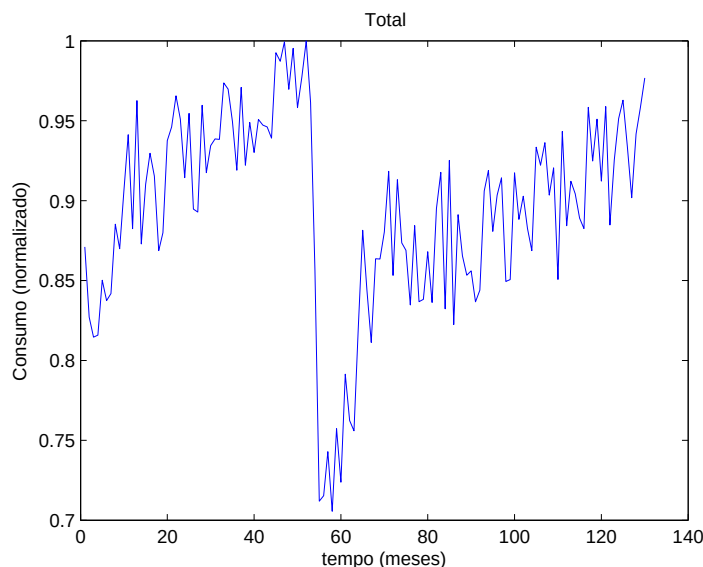


Figura 3.5: Séries de consumo total (exceto pela categoria ‘concessionárias’) após a retirada dos *outliers*.

Base de dados distribuídos por hora - Janeiro de 2000 a Dezembro de 2006

A segunda base de dados, ainda que não dividida em setores, permite uma análise do consumo energético no estado de Minas Gerais com maior nível de granulação ou detalhamento, uma vez que esses dados estão distribuídos por hora.

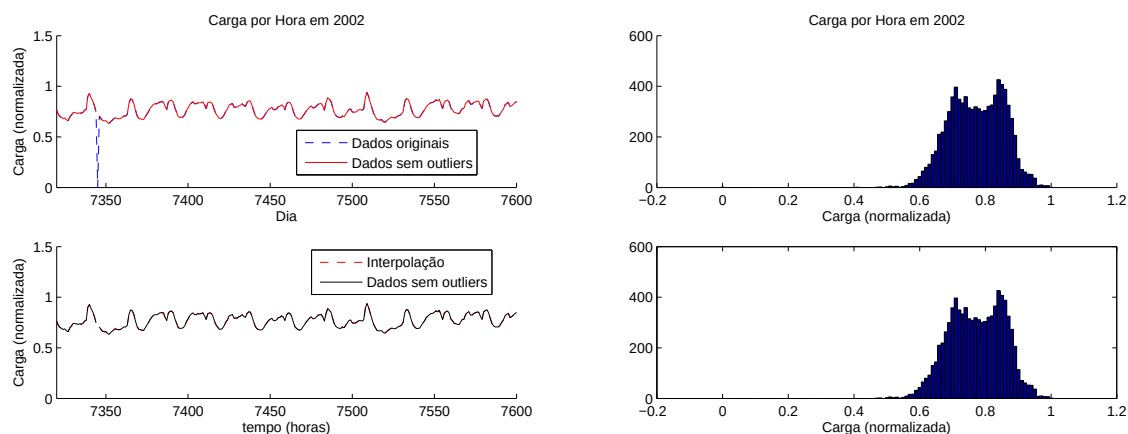


Figura 3.6: Carga por hora referente ao ano 2002. Série original e série sem *outliers* já interpolada, cada uma ladeada por seu respectivo histograma.

As Figuras 3.6, 3.7, 3.8 e 3.9 apresentam os resultados do mesmo teste de Grubbs para esse segundo conjunto de dados, bem como da interpolação das séries após a retirada dos *outliers* detectados. Os histogramas mostram que, em relação à primeira base de dados, esses dados possuem uma distribuição mais próxima da normal, conforme é requerido pelo teste de Grubbs.

Note-se uma questão curiosa: na Figura 3.6, o valor retirado corresponde à primeira

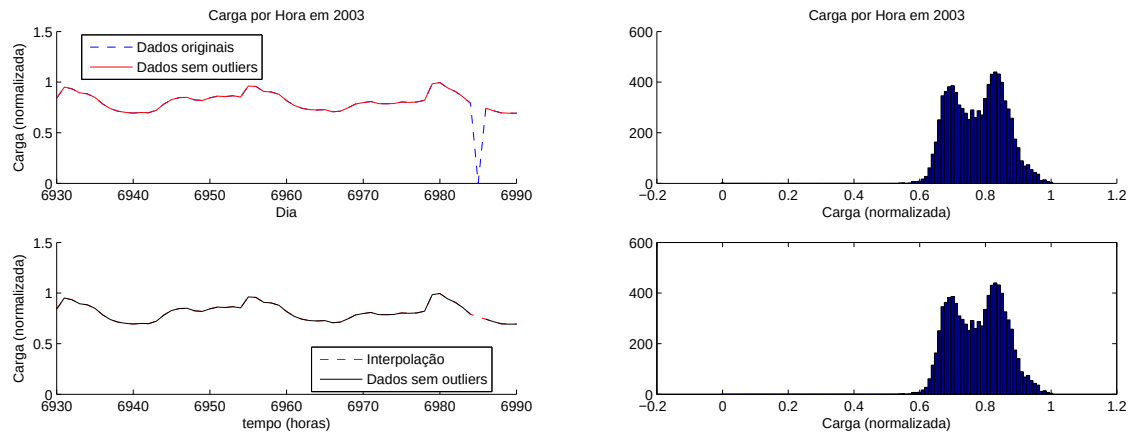


Figura 3.7: Carga por hora referente ao ano 2003. Série original e série sem *outliers* já interpolada, cada uma ladeada por seu respectivo histograma.

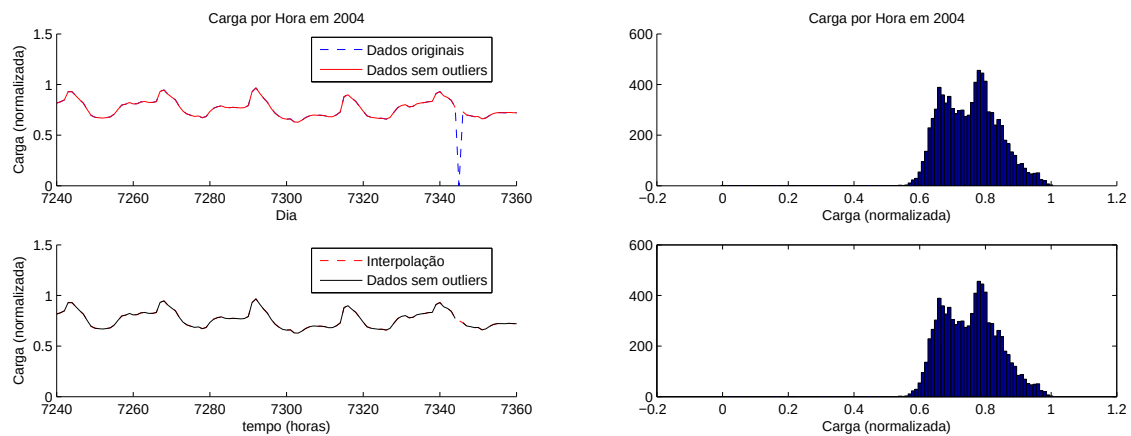


Figura 3.8: Carga por hora referente ao ano 2004. Série original e série sem *outliers* já interpolada, cada uma ladeada por seu respectivo histograma.

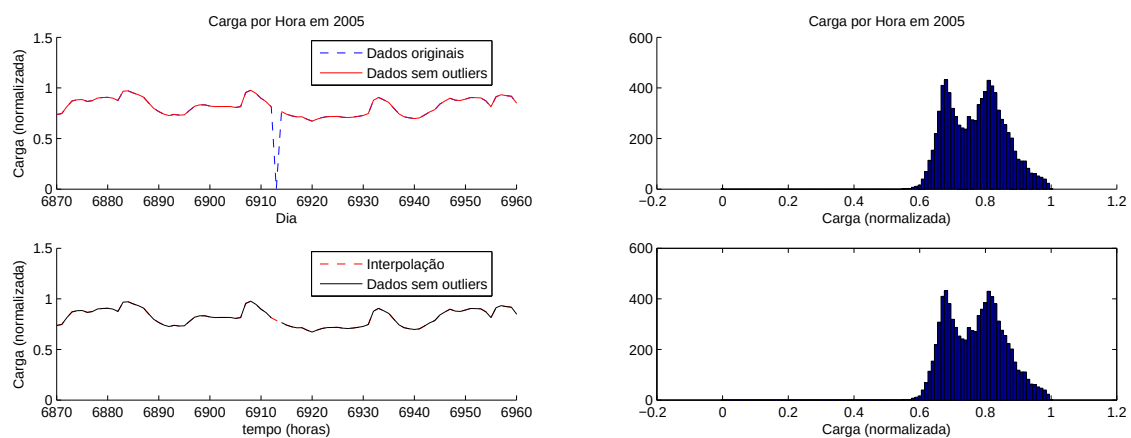


Figura 3.9: Carga por hora referente ao ano 2005. Série original e série sem *outliers* já interpolada, cada uma ladeada por seu respectivo histograma.

hora de 3 de novembro de 2002; na Figura 3.7, à primeira hora de 19 de outubro de 2003; na Figura 3.8, o valor retirado novamente corresponde à primeira hora de 3 de novembro de 2004; por fim, na Figura 3.9, o valor retirado corresponde à primeira hora de 16 de outubro de 2005. É muita coincidência que os quatro *outliers* identificados correspondam à primeira hora do dia e que dois deles tenham ocorrido no mesmo dia em anos diferentes. Esses valores revelam, na verdade, um acontecimento periódico que se repetiu ao longo dos quatro anos: o início do chamado *horário de verão*. Ainda assim, é válido que sejam retirados da série como valores espúrios, já que eles destoam completamente dos valores normais de consumo.

Como os dados de carga estão agrupados por hora, o valor de demanda (potência requerida) e de consumo (energia utilizada) de cada ponto é numericamente o mesmo. Dessa forma, é possível agrupá-los em consumo por dia e por mês, conforme se verifica na Figura 3.10. É interessante perceber que os perfis de ambas as curvas são muito semelhantes entre si. Mais uma vez fica evidenciada a mudança de comportamento das séries no período pós-acionamento (tempo ≈ 500 dias no gráfico de consumo diário; tempo ≈ 18 meses no gráfico de consumo mensal). Outro aspecto interessante de se perceber em ambas as curvas é a sazonalidade anual. No caso mensal, é bastante evidente a existência de um ciclo de 12 meses combinado aditivamente com uma tendência polinomial e uma componente residual. No caso diário, a sazonalidade é de 365 (ou 366) dias, também relativa ao período de um ano.

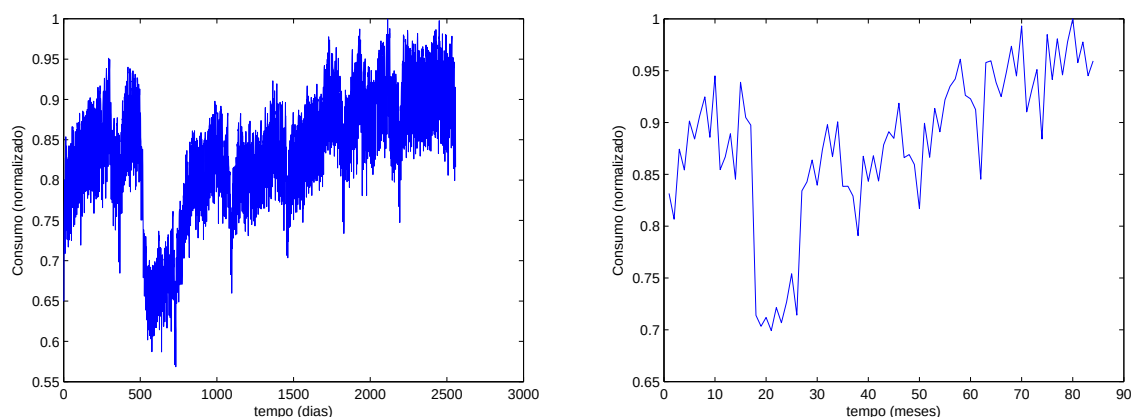


Figura 3.10: À esquerda, carga diária total referente ao período que se estende de janeiro de 2000 a dezembro de 2006; à direita, carga mensal total referente ao mesmo período.

3.2.2 Série de New England

Uma outra série de consumo de energia utilizada nesse trabalho é a do Estado de New England (EUA), a qual pode ser encontrada sem maiores dificuldades em uma busca pela Internet⁴. Trata-se de uma série de consumo de energia elétrica por hora, em kWh, ao longo de 19 anos, de Janeiro de 1980 a Dezembro de 1998, num total de 166560 valores.

⁴Dados encontrados no site <http://www.iso-ne.com/>.

A Figura 3.11 apresenta a série de consumo do ano de 1994, juntamente com um detalhamento que revela o comportamento da série em um período de duas semanas. É interessante perceber que os dias de consumo mais baixo correspondem aos finais de semana, separados entre si pelos cinco dias úteis, que se situam em um patamar de consumo superior.

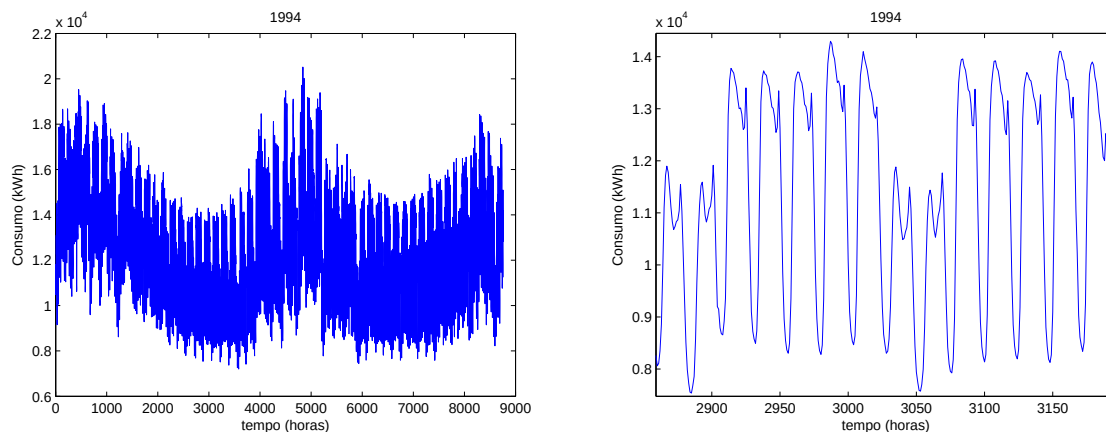


Figura 3.11: À esquerda, série de consumo horário de New England no ano de 1994; à direita, um *zoom* que evidencia o formato da curva ao longo de duas semanas.

Realizando-se uma espécie de integração numérica da série de consumo por hora (na verdade, uma soma acumulada), é possível construir a série de consumo mensal, a qual é apresentada na Figura 3.12. Note-se que diferentemente da série de consumo de baixa tensão mensal total da CEMIG (ver Figura 3.5), a série de New England apresenta um padrão de sazonalidade anual mais bem definido. Uma observação minimamente cuidadosa da Figura 3.12 revelará a existência de ciclos anuais que visualmente se assemelham à letra 'W', o que revela que o fim e o meio do ano são as ocasiões caracterizadas por valores de consumo de energia mais altos.

3.3 Decomposição de Séries Temporais

Apresentadas as séries temporais de consumo de energia elétrica, entremos no mérito nas ferramentas matemáticas que nos permitem a sua análise.

É prática comum, em se tratando de séries temporais, a decomposição de um conjunto de dados em componentes constituintes. Uma maneira clássica de se realizar tal decomposição pode ser representada pela equação

$$y(t) = m(t) + s(t) + r(t) \quad (3.1)$$

em que $m(t)$, $s(t)$ e $r(t)$ são, respectivamente a tendência, a sazonalidade e os resíduos da série $y(t)$. O sinal $y(t)$ apresentado na Figura 3.13 à esquerda foi construído somando-se as três componentes apresentadas na mesma figura à direita (uma reta como tendência, uma função seno como sazonalidade e uma realização de um processo estocástico gaussiano de media zero e variância unitária como resíduos).

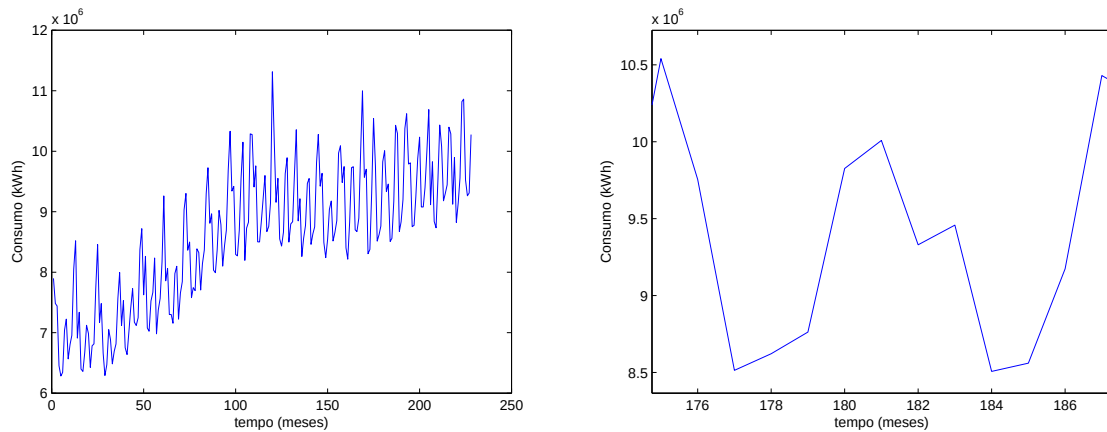


Figura 3.12: À esquerda, a série de consumo mensal de New England no intervalo de tempo entre Janeiro de 1980 e Dezembro de 1998; à direita, um *zoom* no período correspondente ao ano de 1994.

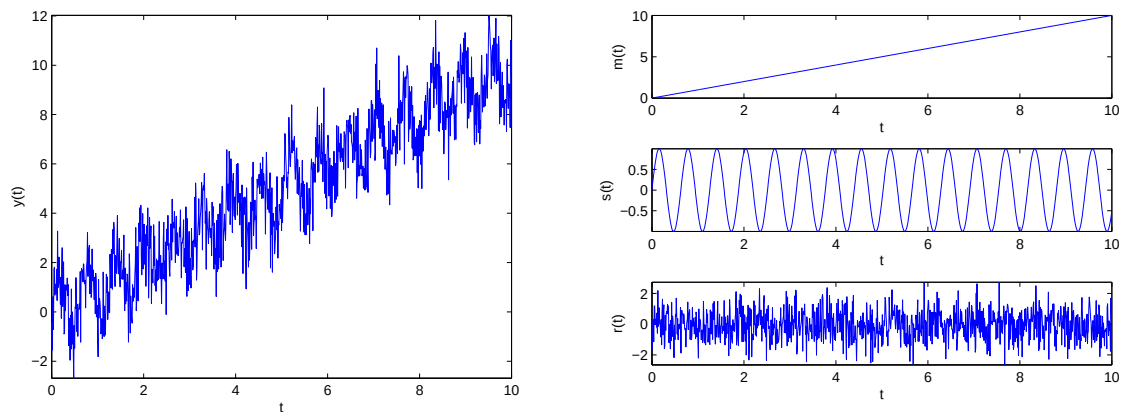


Figura 3.13: Série $y(t)$ à esquerda e suas componentes de tendência $m(t)$, sazonalidade $s(t)$ e resíduos $r(t)$ à direita.

Além dessa abordagem, existem outras decomposições que extrapolam o número de três componentes. A seguir será apresentada uma metodologia que pode ser aplicada à decomposição de uma série em quantas componentes se queira, quantidade essa que pode ser definida arbitrariamente, portanto.

3.3.1 Análise Espectral Singular

Em um trabalho relativamente antigo (Vautard et al., 1992), foi proposta uma técnica para a extração de informação de séries temporais pequenas, caóticas e ruidosas, a qual recebeu o nome de análise espectral singular (SSA - *Singular Spectrum Analysis*), que se trata basicamente de uma espécie de decomposição da série em diversas partes constituintes. Uma aplicação dessa técnica a séries temporais climáticas pode ser encontrada em (Ghil et al., 2002), que é um artigo de revisão extenso o qual explica de modo detalhado a utilização dessa ferramenta.

O que se faz nessa metodologia, em primeiro lugar, é realizar a imersão da série $\{X(t) : t = 1, \dots, N\}$ por coordenadas de atraso, de modo a se construir uma sequência $\{\tilde{X}(t)\}$ M -dimensional:

$$\tilde{X}(t) = \{X(t), X(t+1), \dots, X(t+M-1)\}, \quad t = 1, \dots, N', \quad N' = N - M + 1 \quad (3.2)$$

Considerando-se essas séries temporais deslocadas no tempo como vetores coluna, é possível concatená-las lado a lado de modo a se obter uma matriz de trajetória D de dimensão $N' \times M$ (Vautard et al., 1992). A partir de D , é calcula-se a matriz de covariância C_X de $\tilde{X}(t)$ da seguinte forma:

$$C_X = \frac{1}{N'} D^T D. \quad (3.3)$$

Da álgebra linear elementar, é possível calcular os autovalores λ_k e os autovetores v_k , $k = 1, \dots, M$ a partir da igualdade

$$C_X v_k = \lambda_k v_k. \quad (3.4)$$

Por fim, realiza-se a projeção da série original $X(t)$ na direção de cada um dos autovetores da matriz de covariância C_X :

$$A_k(t) = \sum_{j=1}^M X(t+j-1) v_k(j). \quad (3.5)$$

Para a reconstrução da série original, basta que se combinem as componentes $A_k(t)$ da seguinte maneira:

$$X(t) = \sum_{k=1}^M R_k(t), \quad R_k(t) = P_t \sum_{j=L_t}^{U_t} A_k(t-j+1) v_k(j) \quad (3.6)$$

em que P_t , L_t e U_t são o fator de normalização, o limite inferior e o limite superior do procedimento de reconstrução, respectivamente, os quais são definidos pela expressão

$$(P_t, L_t, U_t) = \begin{cases} (\frac{1}{t}, 1, t), & 1 \leq t \leq M-1 \\ (\frac{1}{M}, 1, M), & M \leq t \leq N' \\ (\frac{1}{N-t+1}, t-N+M, M), & N'+1 \leq t \leq N \end{cases} \quad (3.7)$$

Aplicando-se esse algoritmo de decomposição à série apresentada na Figura 3.13, para $M = 3$, obtém-se as componentes apresentadas na Figura 3.14.

Note-se que a componente PC1 é a que está associada ao maior dos valores singulares (duas ordens de grandeza acima dos outros valores singulares), fato que se reflete na similitude entre ela e a série original $y(t)$. Outro aspecto relevante a ser observado é o fato de as três componentes obtidas via SSA serem completamente diferentes das componentes $m(t)$, $s(t)$ e $r(t)$ utilizadas para compor o sinal $y(t)$. Isto não é surpreendente, uma vez que a formulação matemática da análise espectral singular é completamente

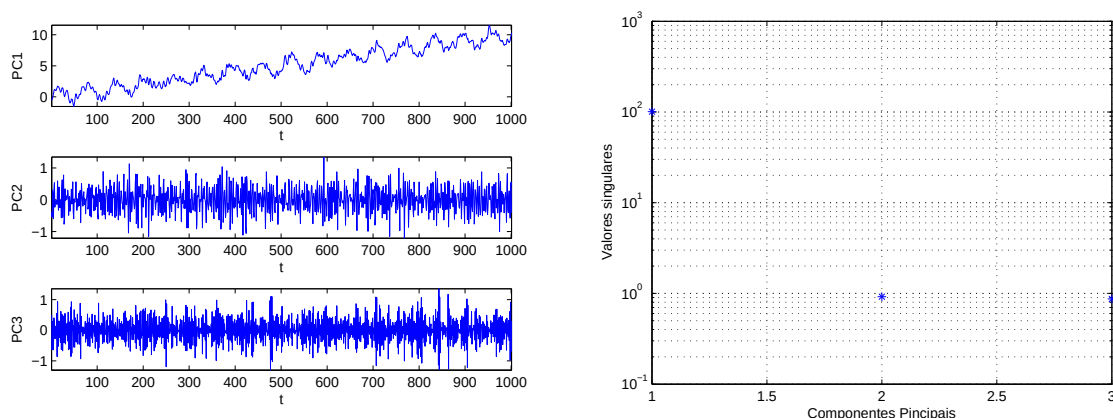


Figura 3.14: As três componentes da série da Figura 3.13 obtidas via SSA juntamente com os seus respectivos valores singulares.

independente da maneira pela qual a série original venha a ser construída ou obtida.

Considerando-se decomposições com diferentes valores de M , chega-se aos resultados apresentados na Figura 3.15. Em cada par de gráficos espectro-reconstrução, são consideradas as componentes de maior peso obtidas na decomposição e pode-se observar que essas componentes, ao serem combinadas aditivamente entre si, em muito se aproximam do formato da série original sem ruído. Tal fato se mostra mais evidente com o aumento de M , conforme se verifica claramente na Figura 3.15 (da esquerda para a direita e de cima para baixo).

3.3.2 Decomposição de séries temporais de consumo de energia elétrica

A análise espectral singular foi aplicada a algumas das séries temporais apresentadas na seção anterior a fim de se verificar a eficácia dessa técnica na identificação das componentes de tendência das referidas séries.

A Figuras 3.16 apresenta os resultados obtidos para as duas séries mensais da CEMIG (consumo de baixa tensão e consumo total) e para a série de consumo mensal de New England. Em cada um dos gráficos foram plotadas, juntamente com a série de consumo, algumas reconstruções das séries originais considerando-se até as quatro primeiras componentes principais de maior peso de acordo com a SSA.

É interessante verificar que em todas as séries, à exceção da série de consumidores de baixa tensão da CEMIG, a reconstrução composta apenas pela primeira componente principal reproduz com bastante fidelidade a tendência de comportamento das respectivas séries de consumo. A consideração de duas até quatro componentes revela um tentativa da série reconstruída em mimetizar tanto a tendência como o comportamento cíclico das séries. Salientamos que essa possibilidade de extração de tendência via SSA será utilizada em três pontos deste trabalho: nos testes de dados sub-rogados, na definição de um preditor trivial e na escolha das variáveis regressoras para as redes MLP e RNFL (ver Capítulo 4).

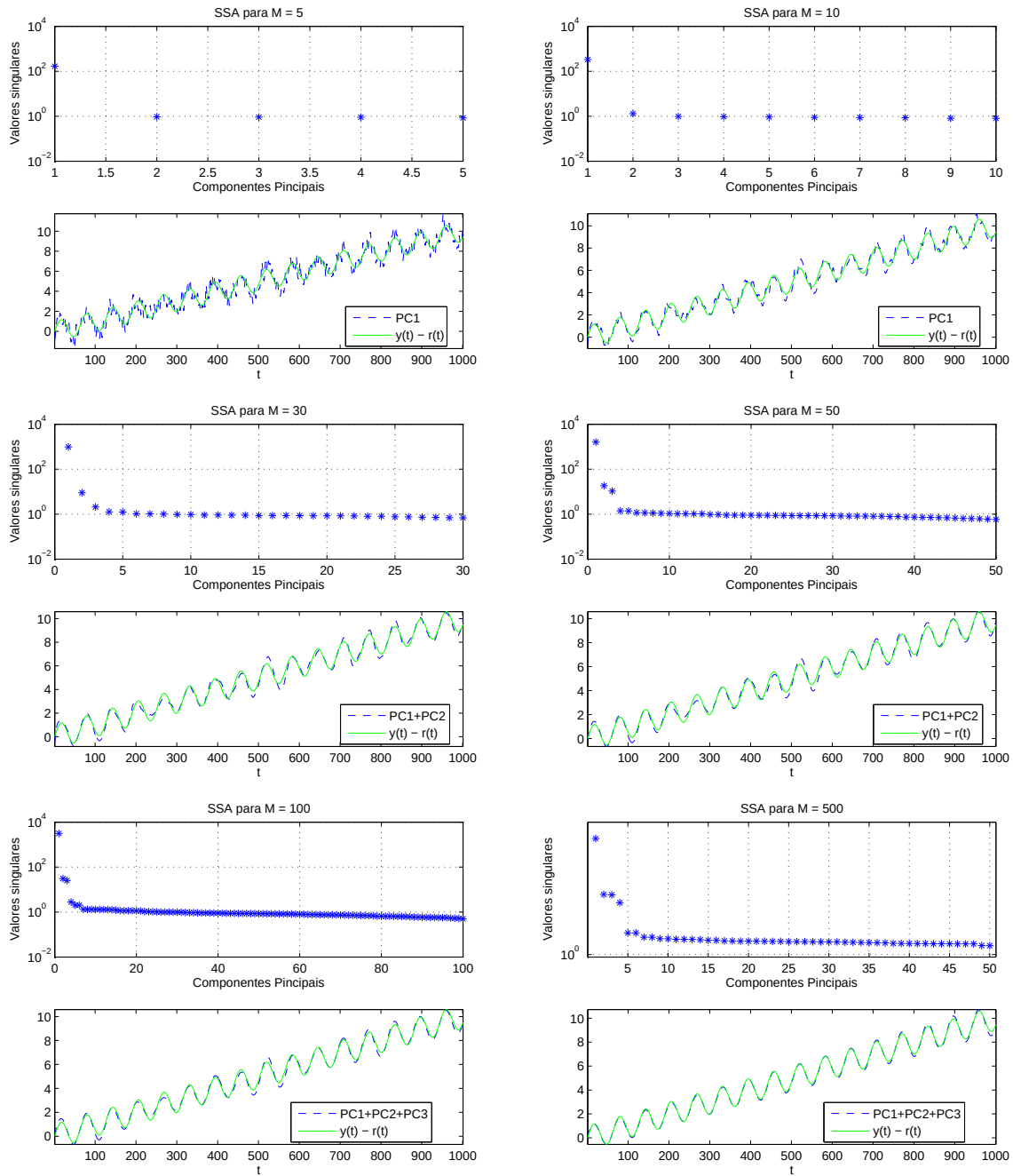


Figura 3.15: Análise Espectral Singular - Comparação entre as componentes de maior peso e a série $\{y(t) - r(t)\}$ para $M \in \{5, 10, 30, 50, 100, 500\}$, juntamente com o respectivo espectro singular das diferentes decomposições realizadas.

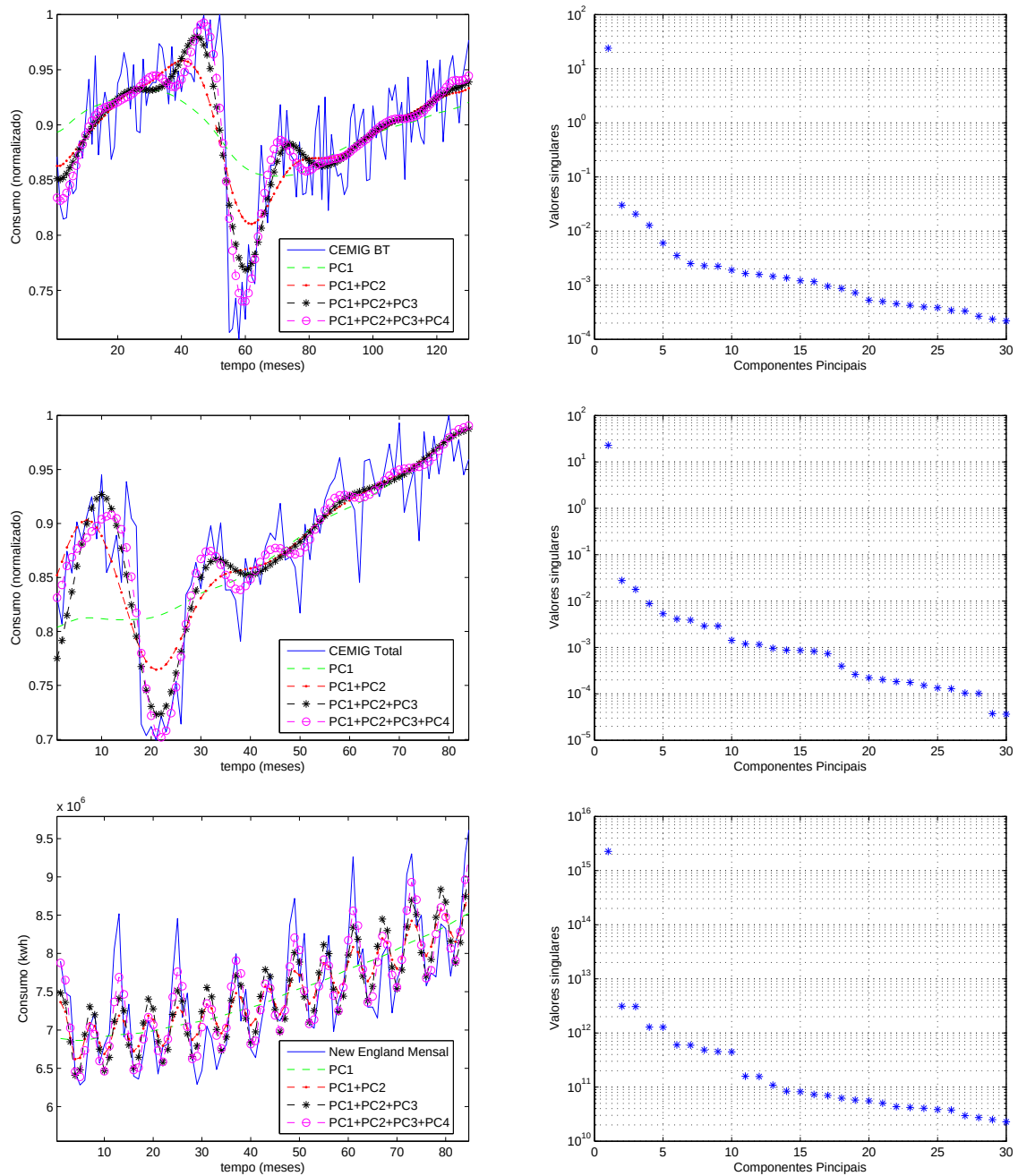


Figura 3.16: À esquerda, série mensal de consumidores de baixa tensão da Cemig, série de consumo total da Cemig e série de consumo mensal de New England juntamente com as componentes principais obtidas via SSA para $M = 30$; à direita, os valores singulares associados a cada componente da série.

3.4 Detecção de Determinismo

O objetivo da presente seção é apresentar uma metodologia desenvolvida na década de 90 para a caracterização da dinâmica em séries temporais por meio da identificação de características como determinismo e não-linearidade. Além disso, são exibidos os resultados obtidos com a aplicação de tal técnica às séries de consumo de energia elétrica anteriormente apresentadas neste trabalho (ver Seção 3.2).

3.4.1 Teste de Dados Sub-Rogados

Em diversos estudos de séries temporais, é válido que se levantem questionamentos quanto à natureza do processo subjacente àquele conjunto de dados apresentados, ou seja, qual o tipo de dinâmica do processo que originou aquelas séries. Para tal fim, existe na Estatística o chamado teste de hipóteses, que foi aplicado a problemas de dados dinâmicos não-lineares em (Theiler et al., 1992), com o nome de Análise de Dados Sub-Rogados ⁵. Nesse trabalho, Theiler e colegas propuseram três algoritmos para identificar a existência de determinismo e não-linearidade nos dados das séries temporais estudadas. O princípio básico do método consiste em gerar, a partir da série de dados originais, várias séries artificiais que satisfaçam a uma determinada hipótese nula e que se assemelhem, de uma certa forma, aos dados originais. A essa coleção de séries artificiais dá-se o nome de conjuntos de dados sub-rogados (*surrogate data sets*). Feito isso, é computado o valor de alguma estatística discriminante (ED) (função de auto-correlação (FAC), informação mútua média (IMM), dimensão de correlação, expoentes de Lyapunov etc. (Small, 2005)) para a série original e para as séries artificiais. Se o valor desse discriminante para os dados originais for significativamente diferente dos valores calculados para os conjuntos de dados sub-rogados, então a hipótese nula pode ser rejeitada. Em caso contrário, a hipótese não pode ser rejeitada (o que não significa que ela seja verdadeira). O discriminante utilizado pelos autores nos casos apresentados no artigo foi a dimensão de correlação.

Para verificar se a hipótese nula H_0 pode ou não ser rejeitada, calcula-se a chamada significância, a qual é dada pela expressão

$$S = \frac{|Q_D - \mu_H|}{\sigma_H} \quad (3.8)$$

em que Q_D é o valor da ED para a série original, μ_H e σ_H são a média e o desvio padrão, respectivamente, dos valores da ED para as séries de dados sub-rogados. A significância S trata-se portanto de uma medida da distância entre a série original e as séries de dados sub-rogados sob a consideração de que a ED segue uma distribuição gaussiana.

Como essa hipótese de gaussianidade nem sempre pode ser garantida, uma outra alternativa que se tem em mãos é o cálculo do p -valor (ou probabilidade de significância), que nada mais é do que a probabilidade de se observar uma significância

⁵Do inglês *Surrogate Data Analysis*.

maior do que S em caso de a hipótese nula H_0 ser verdadeira (Theiler et al., 1992). Em outras palavras, quanto menor o p -valor, mais evidente se torna a rejeição de H_0 . Matematicamente, o p -valor P_v pode ser expresso por

$$P_v = \begin{cases} 2i, & \text{se } i \leq 0,5 \\ 2(1 - i), & \text{se } i > 0,5 \end{cases} \quad (3.9)$$

em que i é o índice de p_i , que por sua vez é o menor i -ésimo percentil inferior da ED de todas as séries de dados sub-rogados em que se encontra o valor de ED da série original. Para fins práticos, a Tabela 3.2 apresenta uma classificação dos níveis de evidência de rejeição de H_0 de acordo com os p -valores (Wasserman, 2004).

Tabela 3.2: Evidência de rejeição de H_0 em função do p -valor.

p -valor (P_v)	Evidência
< 0.01	evidência muito forte contra H_0
$0.01 - 0.05$	evidência forte contra H_0
$0.05 - 0.1$	evidência fraca contra H_0
> 0.1	pouca ou nenhuma evidência contra H_0

No primeiro algoritmo proposto por Theiler et al. (1992), a hipótese nula é de que os dados originais são puramente aleatórios. Para gerar uma série de dados sub-rogados que obedeça à hipótese nula de aleatoriedade, o que se faz é embaralhar completamente os valores da série original de modo que qualquer dinâmica presente dos originais seja destruída. A repetição desse procedimento N vezes produz N conjuntos de dados sub-rogados. Para visualização do funcionamento do algoritmo, apresenta-se abaixo um exemplo simples.

Exemplo de aplicação do algoritmo 0 proposto por Theiler

Seja uma série temporal X_t completamente aleatória constituída de 100 valores. São geradas 100 séries ($S_t^1, S_t^2, \dots, S_t^{100}$) completamente aleatórias simplesmente embaralhando-se a ordem dos valores da série original X_t . A série original é apresentada na Figura 3.17, juntamente com quatro dessas séries artificiais (séries de dados sub-rogados). Como estatísticas discriminantes, utilizam-se a FAC e a IMM, os quais são calculados tanto para a série original como para as séries artificiais, considerando-se atrasos entre 1 e 10 amostras. Matematicamente, elas podem ser expressas por

$$FAC(\tau) = \frac{1}{N} \sum_{t=1}^{N-\tau} X_t X_{t+\tau}, \quad \tau = 0, 1, \dots, N-1, \quad (3.10)$$

e

$$IMM(\tau) = \frac{1}{N} \sum_{t=1}^N p(X_t, X_{t+\tau}) \log_2 \left[\frac{p(X_t, X_{t+\tau})}{p(X_t)p(X_{t+\tau})} \right]$$

em que N é o número de amostras da série, τ é o defasamento temporal em amostras para o qual se calculam as EDs e $p(X_t)$ é a probabilidade de se observar o valor X_t na série temporal.

A partir das EDs, calcula-se a significância S para cada um dos 10 atrasos⁶.

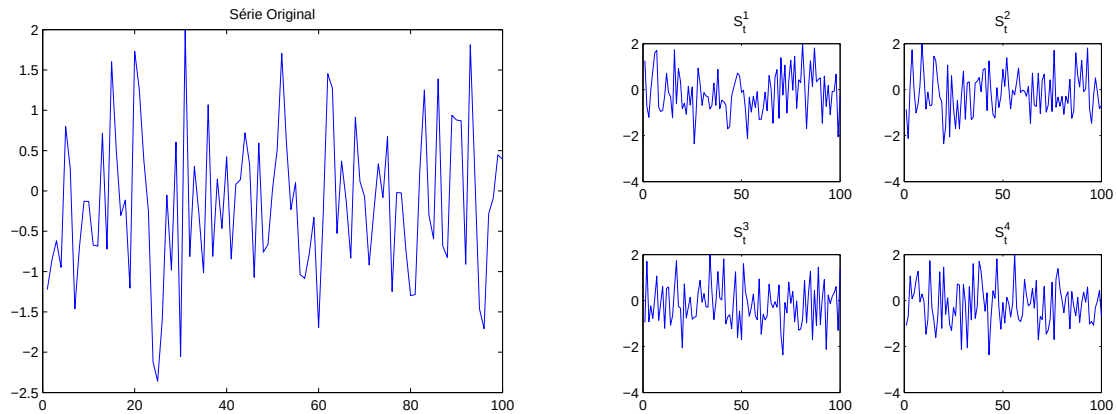


Figura 3.17: Série original aleatória e quatro séries de dados sub-rogados.

A Figura 3.18 apresenta os resultados obtidos. Observa-se que, para ambas as estatísticas discriminantes, a significância nunca atinge valores maiores do que 2σ ⁷.

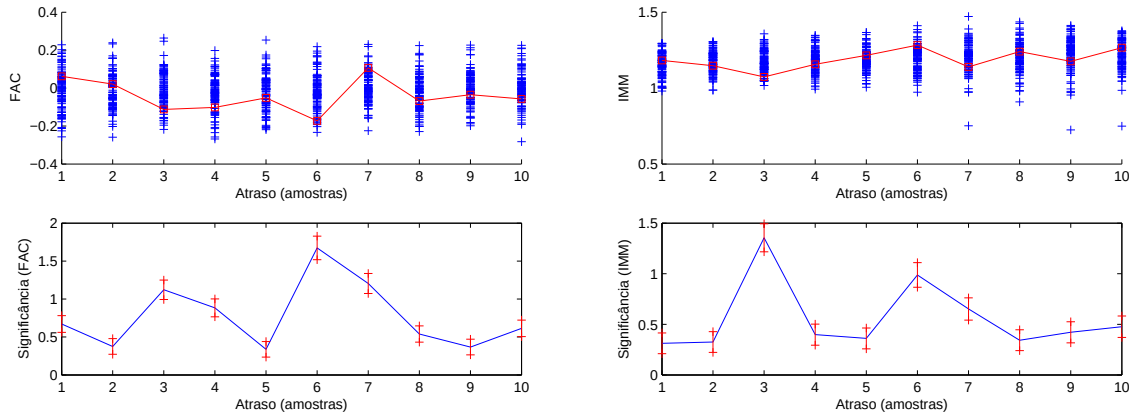


Figura 3.18: Função de autocorrelação e informação mútua média da série original aleatória e das séries artificiais

Agora, seja uma série Y_t em que haja dependência temporal entre as amostras. Seguindo-se os mesmos passos expostos acima com a série X_t , chega-se aos resultados representados nas Figuras 3.19 e 3.20.

Observa-se que a significância fica acima de 2σ para diversos atrasos de tempo. Logo, a hipótese nula é rejeitada, indicando alta probabilidade de que exista depen-

⁶Para esse exemplo, não foram calculados os p -valores.

⁷É válido ressaltar que dentro de uma faixa de mais ou menos 2σ encontram-se 95.0% das realizações de uma distribuição gaussiana. Qualquer valor fora dessa faixa tem grande probabilidade de não pertencer à amostra descrita por essa distribuição.

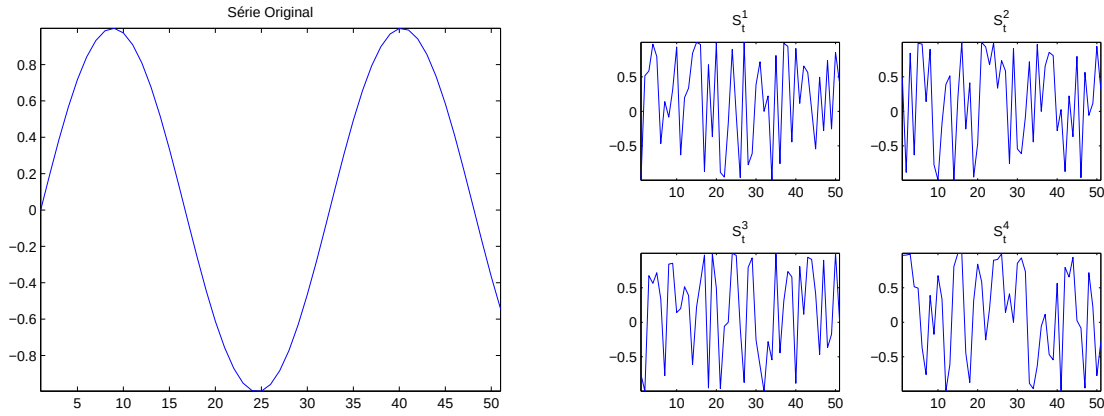


Figura 3.19: Série original com dependência temporal e quatro séries de dados sub-rogados.

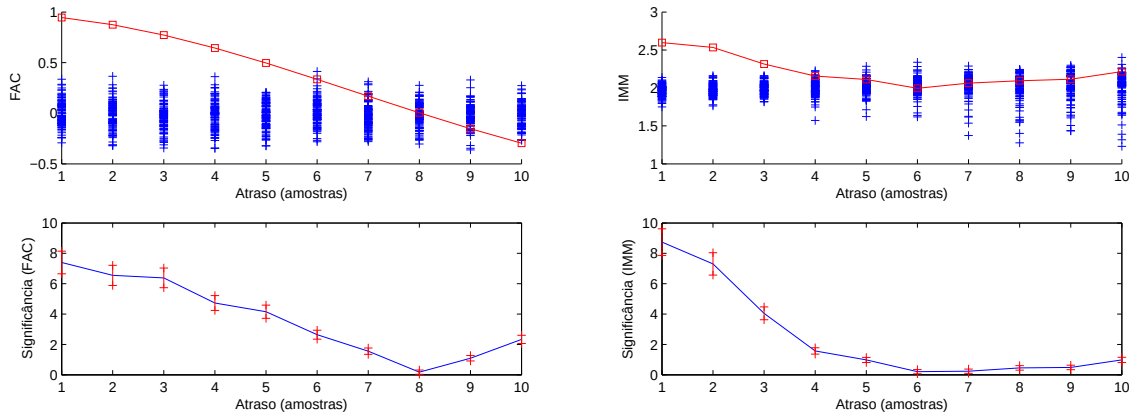


Figura 3.20: Função de autocorrelação e informação mútua média da série original com dependência temporal e das séries artificiais

dência temporal entre os dados da série Y_t . Portanto, conseguiu-se identificar corretamente se a série é ou não aleatória com base em limites de confiança de $\pm 2\sigma$.

□

No segundo algoritmo (Theiler et al., 1992), a hipótese nula diz que os dados originais são de natureza linear e estão imersos em ruído gaussiano. Portanto, o modelo que se pode usar para gerar os dados sub-rogados é auto-regressivo⁸ (AR) dado por

$$x_t = a_0 + \sum_{k=1}^q a_k x_{t-k} + \sigma e_t \quad (3.11)$$

em que os coeficientes a_k e o desvio padrão σ são calculados a partir da série original. Para geração das séries de dados sub-rogados, basta que se realize a simulação livre da Eq. (3.11). Uma maneira alternativa de se gerar esses dados sub-rogados é via

⁸Construído a partir de valores passados do próprio sinal para cujo sistema subjacente se deseja levantar um modelo.

embaralhamento de fase da transformada de Fourier do sinal original.

O terceiro teste proposto em (Theiler et al., 1992) tem como hipótese nula que a função de observação da série seja não-linear e que a série subjacente obedeça à hipótese do segundo algoritmo referenciada no parágrafo anterior. Portanto, dada um série X_t que foi medida de um processo qualquer, assume-se que existe uma série Y_t consistente com a hipótese de ser constituída de dados lineares imersos em ruído gaussiano, tal que

$$X_t = h(Y_t) \quad (3.12)$$

em que $h(\cdot)$ é uma função estática não-linear. O algoritmo de geração de dados sub-rogados consiste em embaralhar a ordem dos valores de X_t de tal forma que se preserve a correlação linear da série subjacente $Y_t = h^{-1}(X_t)$. Se houver rejeição da hipótese nula, então é alta a probabilidade de que a não-linearidade esteja presente na dinâmica do sistema.

É interessante perceber que há uma gradação nas hipóteses propostas para o teste de dados sub-rogados. Parte-se de uma estrutura mais simples, como a de dados temporalmente independentes, e caminha-se em direção a uma hipótese de maior generalidade, tal qual a da transformação estática não-linear de um conjunto de dados linear imerso em ruído gaussiano.

Depois desse artigo pioneiro, muitos outros trabalhos vieram complementar e desenvolver novas técnicas de dados sub-rogados para análise de séries temporais no âmbito da abordagem de sistemas dinâmicos. Em (Theiler and Prichard, 1996), é levantada uma discussão sobre a *pivotalidade da ED*. Para o entendimento desse conceito, suponha-se um espaço de processos $\mathcal{F}$ em que $\mathcal{F}_\phi \subset \mathcal{F}$ é o conjunto de processos que atende a uma dada hipótese nula ϕ . Se a distribuição da estatística discriminante T for pivotal, então isto significa que ela será a mesma para todos os elementos F da família de processos $\mathcal{F}_\phi$. Daí a importância de se escolher adequadamente a ED de maneira a se garantir um resultado coerente na análise de dados sub-rogados. Vê-se que essa necessidade de escolher um discriminante estatístico que atenda à condição de pivotalidade limita a flexibilidade em sua escolha. Por essa razão, ainda nesse mesmo artigo, os autores propõem uma maneira diferente de geração dos dados sub-rogados. A abordagem convencional é ajustar um modelo aos dados originais e a partir dele gerar os dados sub-rogados que seriam “realizações típicas” do modelo. Na nova abordagem proposta, tenta-se compelir os conjuntos de dados sub-rogados a concordarem exatamente com os dados originais. Dessa forma, para esta última abordagem, a condição de pivotalidade não precisa ser atendida pela ED, o que torna a sua escolha mais livre. Para fins práticos, no caso do algoritmo 0 proposto por Theiler, isso é feito gerando-se séries de dados sub-rogados cujas distribuições (descritas pela média e pela variância amostrais) sejam idênticas à da série original. Para tanto, a geração das séries artificiais pode ser realizada por meio da reamostragem da série original sem substituição de valores, por exemplo.

Um novo algoritmo de análise de dados sub-rogados pode ser encontrado em (Small et al., 2001), em que os autores realizam o estudo de séries que apresentam estrutura pseudo-periódica. O algoritmo é capaz de distinguir se um sinal é periódico imerso

em ruído ou se ele é pseudo-periódico com dinâmica entre ciclos. Portanto, para gerar dados sub-rogados que atendam à hipótese nula de que os dados originais constituem uma órbita periódica ruidosa, eles utilizam o algoritmo PPS (*pseudo-periodic surrogate*). Esses dados PPS gerados exibem a mesma estrutura periódica dos dados originais, e nenhum outro tipo de determinismo. Com essa ferramenta, é possível diferenciar, por exemplo, o sistema de Rössler caótico (Thomson and Stewart, 1986) de uma realização pseudo-periódica das equações de Rössler com ruído dinâmico. A ED utilizada foi a dimensão de correlação, por questões de pivotalidade.

Em (Nakamura and Small, 2005) é proposto um algoritmo para a identificação de dinâmica em séries temporais que apresentam flutuações irregulares (variabilidade local), chamado SSS (*small-shuffle surrogate*). A hipótese nula para esse caso é que não existe dinâmica local nos dados⁹. Dessa forma, o mecanismo de geração dos dados é extremamente simples: basta embaralhar os dados localmente, de modo que a dinâmica de longo prazo seja mantida, ao passo que a dinâmica local é destruída. De maneira mais formal, sejam $x(k)$ o sinal original, $i(k)$ os índices de $x(k)$ (portanto, $i(k) = k$), $g(k)$ uma amostra de uma distribuição gaussiana de média zero e variância unitária e $s(k)$ uma série de dados sub-rogados. Eis os passos do algoritmo para a geração de $s(k)$:

- Obtenha $i'(k) = i(k) + Ag(k)$, em que A é o desvio padrão do ruído gaussiano (os criadores do algoritmo argumentaram ser razoável que A valha 1).
- Reagrupe $i'(k)$ por ordem crescente de magnitude. Seja, então, $\hat{i}(k)$ a nova sequência dos índices de $i'(k)$ que propicie esse ordenamento.
- A série de dados sub-rogados será dada por $s(k) = x(\hat{i}(k))$.

Um exemplo numérico ajudará no melhor entendimento do algoritmo de geração dos dados sub-rogados. Seja $x(k) = (12,13,14,15,16)$, em que $i(k) = (1,2,3,4,5)$. Para $A=1$, obtem-se $i'(k) = (0.3,2.2,0.9,4.6,3.7)$. Reordenando-se $i'(k)$, chega-se a $(0.3,0.9,2.2,3.7,4.6)$. Logo, $\hat{i}(k) = (1,3,2,5,4)$, e portanto, $s(k) = (12,14,13,16,15)$.

Small e Nakamura utilizaram como EDs a função de autocorrelação (FAC) e a informação mútua média (IMM). Segundo os autores, a razão de tal escolha reside no fato de que o embaralhamento local promovido pelo SSS modifica o fluxo de informação nos dados, e o cálculo da FAC seria um bom indicativo da correlação linear que venha a ser destruída por esse procedimento; da forma análoga, a IMM se trata de uma versão não-linear geral da FAC, sendo capaz portanto de identificar possíveis aniquilamentos de dependência não-linear nos dados.

Com esse algoritmo (SSS), é possível identificar se as irregularidades presentes em uma série são unicamente aleatórias ou se há algum tipo de dinâmica nessas flutuações capaz de ser explicada via modelagem.

Algumas aplicações dos algoritmos expostos acima podem ser encontradas em (Nakamura and Small, 2006), (Soofi and Cao, 2002) e (Small, 2005).

Um outro teste para aplicação em séries que apresentam uma componente periódica de amplitude significativa foi proposto por Theiler (1995). A geração das séries de dados

⁹Em outras palavras, retirando-se a componente de tendência e de sazonalidade da série original, os resíduos resultantes são puramente aleatórios.

sub-rogados (*cycle-shuffle surrogate* - CSS)¹⁰ consiste basicamente em se embaralhar de maneira aleatória os ciclos da série original, sendo que cada ciclo desses corresponde a um período fundamental do sinal. A hipótese nula verificada em tal teste é a de que não existe nenhum tipo de dinâmica entre-ciclos, diferentemente do SSS proposto por (Nakamura and Small, 2005), o qual se aplica à investigação da existência de determinismo intra-ciclo. A ED utilizada nesse algoritmo é a FAC, considerando-se atrasos múltiplos do período da componente cíclica da série original. É importante frisar que o CSS só é aplicável a séries temporais estacionárias na média, diferentemente do que acontece no caso do SSS.

3.4.2 Aplicação do Teste de Dados Sub-rogados SSS

Para os primeiros testes realizados com os dados da CEMIG e de New England utilizou-se o algoritmo SSS (descrito logo acima), proposto em Nakamura and Small (2005), para verificar se as flutuações nas séries são aleatórias ou se realmente há dinâmica local. Importante é que se tenha em mente que a hipótese nula considerada por esse algoritmo é de que não existe dinâmica local nas séries, ou seja, as flutuações irregulares são aleatórias¹¹. Diferentemente da sequência adotada na Seção 3.2, foram tratados primeiramente os dados mensais de consumo total.

Séries de consumo total da CEMIG

Conforme já foi apresentado anteriormente, os dados de consumo total da CEMIG estão agrupados por hora, e a partir desses dados é possível portanto construir séries de consumo total por dia, por semana, por quinzena e por mês. Para tanto, basta a realização de uma soma cumulativa com os dados da primeira série.

Da maneira como foi proposto, o SSS aplica-se a séries temporais não-estacionárias na média, porém considerou-se sensata a decisão de subtrair a componente de tendência da série da CEMIG em virtude do número muito reduzido de 84 amostras desta série em relação às séries de 300 a 1000 amostras estudadas pelos criadores do método em (Nakamura and Small, 2005). Para tanto, no caso mensal, utilizou-se o SSA para decompor a série em 30 sub-séries e considerou-se a soma das quatro componentes mais significativas em amplitude como o termo de tendência, uma vez que a primeira componente não foi suficiente devido à variação brusca na série decorrente da crise do setor elétrico brasileiro em maio de 2001¹². A série original juntamente com a componente de tendência estimada e a série resultante da subtração das duas primeiras são apresentadas na Figura 3.21.

Assim sendo, os dados foram agrupados da forma descrita acima, extraiu-se a sua componente de tendência, e a eles foram aplicados o SSS considerando-se a geração de 1000 séries de dados sub-rogados em cada teste. A Figura 3.22 ilustra as estatísticas

¹⁰Neste trabalho esse algoritmo será sempre referenciado como CSS.

¹¹Justamente por isso foram considerados apenas atrasos de quatro amostras no cálculo das EDs.

¹²No caso das séries de consumo por semana e por quinzena, consideraram-se as duas primeiras componentes para aproximar a tendência, e para as séries de consumo horário e diário, só o a primeira componente foi utilizada.

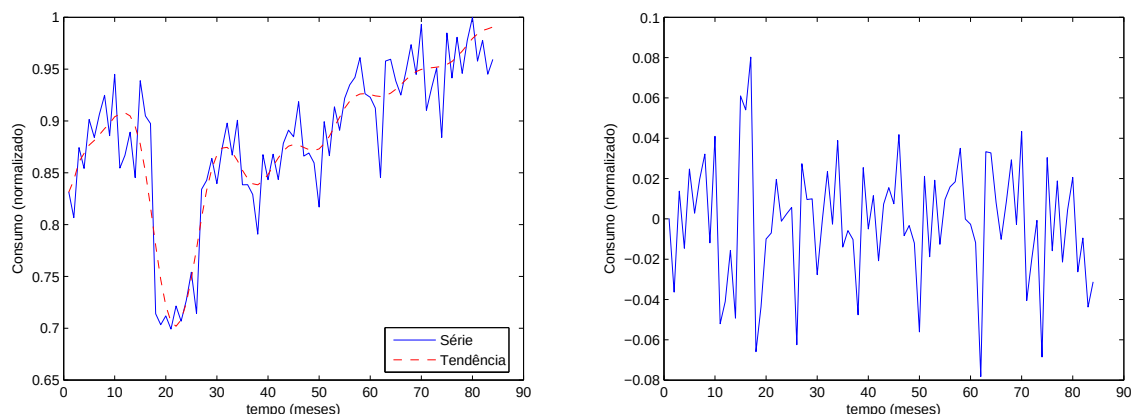


Figura 3.21: À esquerda, série original e componente de tendência estimada; à direita, série resultante após a extração da componente de tendência.

discriminantes associadas à série mensal juntamente com a significância que mede a proximidade entre a série original e as séries sub-rogadas. A aplicação do SSS no caso das séries de consumo quinzenal, diário e horário conduziu aos resultados apresentados na Tabela 3.3. Nela encontram-se os valores de S associados à FAC e à IMM em função do número de atrasos que variou entre 1 e 4. Em negrito estão evidenciados os valores de S que superaram o limiar do intervalo de confiança de 2σ , e que portanto definem a rejeição da hipótese em nível de confiança de 95%.

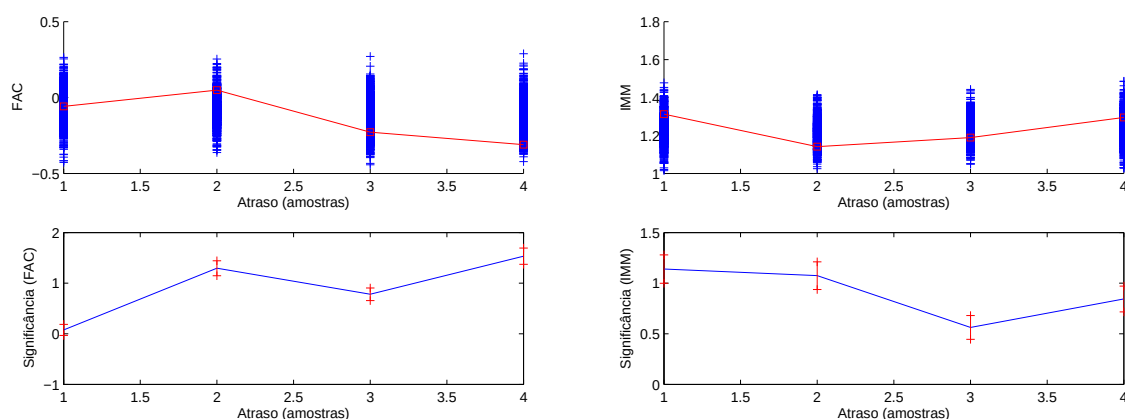


Figura 3.22: FAC e IMM das séries de consumo mensal de baixa tensão da CEMIG e das respectivas séries sub-rogadas. A significância define a distância entre a série original e as séries artificiais.

Para confirmar ou retificar a informação obtida com o cálculo de significância, a qual assume gaussianidade das EDs dos dados sub-rogados, realizou-se o cálculo dos p -valores para a mesma faixa de atrasos, os quais são apresentados na Tabela 3.4. Os p -valores menores do que 0,01, a definirem uma evidência muito forte de rejeição da hipótese nula, e entre 0,01 e 0,05, que estabelecem a faixa de forte evidência de rejeição (ver Tabela 3.2), estão destacados em negrito.

Tabela 3.3: Séries de consumo total da CEMIG - Significância associada à FAC e à IMM para atrasos de 1 a 4 amostras.

Série vs Atraso	S_{FAC}				S_{IMM}			
	1	2	3	4	1	2	3	4
Mensal	0,07	1,29	0,78	1,53	1,14	1,07	0,56	0,84
Quinzenal	3,03	1,40	0,16	1,10	2,18	0,59	0,14	0,83
Semanal	3,31	0,21	1,02	1,21	2,76	0,62	0,44	1,55
Diária	16,51	10,99	6,31	6,75	31,56	11,42	4,51	5,70
Horária	30,05	17,66	5,42	7,00	59,20	33,51	29,13	30,21

Tabela 3.4: Séries de consumo total da CEMIG - p -valores associados à FAC e à IMM para atrasos de 1 a 4 amostras.

Série vs Atraso	Pv_{FAC}				Pv_{IMM}			
	1	2	3	4	1	2	3	4
Mensal	0,90	0,18	0,50	0,09	0,24	0,29	0,60	0,40
Quinzenal	0,002	0,50	0,94	0,41	0,006	0,16	0,90	0,27
Semanal	0,002	0,81	0,31	0,22	0,004	0,52	0,67	0,11
Diária	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002
Horária	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0,002

Observa-se pelos valores em negrito das duas tabelas que a rejeição de hipótese com base em limites de confiança de 2σ foi confirmada pelos respectivos p -valores associados. É interessante ressaltar que só não houve rejeição de hipótese no caso da série mensal, sendo que no caso quinzenal a rejeição se deveu à detecção de alguma correlação não-linear significativa, já que a significância associada à IMM superou o limiar de 2σ . É curioso o fato de que existe uma tendência a haver maior correlação entre as flutuações locais conforme se consideram os dados agrupados em unidades de tempo menores. Talvez isso se deva ao fato de que os dados agrupados por hora trazem informações acerca do processo de consumo energético com maior nível de detalhe do que os dados agrupados por mês, por exemplo. No caso das séries diária e horária, houve rejeição para todos os quatro atrasos considerados no cálculo de ambas as estatísticas discriminantes, tanto em termos de análise de significância como de p -valores.

Séries de consumo mensal de baixa tensão da CEMIG

No caso das séries de consumo mensal de baixa tensão da CEMIG, foi aplicado o SSS a cada uma das oito séries associadas às diferentes classes de consumidores, além da própria série de consumo total de baixa tensão. Em cada teste foram geradas 1000 séries de dados sub-rogados. Na Figura 3.23, são apresentados os gráficos da série original de consumo de baixa tensão total juntamente com a componente de tendência estimada e a série resultante do procedimento de extração da tendência.

As Tabelas 3.5 e 3.6 apresentam as significâncias e os p -valores obtidos para o teste

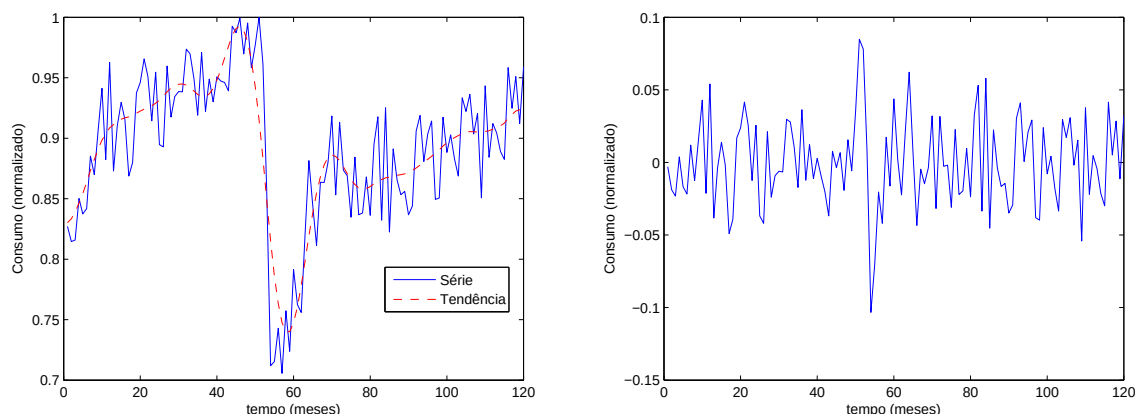


Figura 3.23: Dados de consumo mensal de baixa tensão da CEMIG - À esquerda, série original e componente de tendência estimada; à direita, série resultante após a extração da componente de tendência.

de dados sub-rogados realizado com as oito séries de consumo mensal por setor e a série de consumo mensal total de baixa tensão. Em todos os casos foram consideradas as primeiras quatro componentes obtidas via SSA para a aproximação da tendência, à exceção das séries dos setores ‘poder público’ e ‘rural’, para as quais se utilizaram a duas primeiras componentes e a primeira componente, respectivamente.

A observação das referidas tabelas revela que os setores para os quais se patenteou a rejeição da hipótese nula, que pressupõe aleatoriedade nas flutuações irregulares das séries, foram o ‘poder público’, o ‘rural’, o ‘industrial’ e o ‘serviço público’, em termos de limites confiança de 2σ . A análise de p -valores conduz à mesma conclusão acrescentando-se o setor ‘iluminação pública’, e com a ressalva de que para este setor e para o setor ‘industrial’ a rejeição se dá em nível forte ($0,01 < Pv < 0,05$), enquanto que nos outras ela se verifica em nível muito forte ($Pv < 0,01$), de acordo com a escala apresentada na Tabela 3.2. No caso do setor ‘iluminação pública’ a análise do p -valor indicou rejeição em nível forte, sendo que não houve rejeição analisando-se a significância com intervalos de confiança de 2σ . Isso parece ser um indicativo que a distribuição de FACs, pelo menos para o atraso de uma amostra, não é gaussiana.

As séries para as quais não se acusou a rejeição da hipótese respondem por algo em torno de 75% do total de consumo de baixa tensão, o que se reflete na não rejeição da hipótese para o caso da série de consumo total de baixa tensão.

Série de consumo total de New England

Do mesmo modo que a série de consumo total da CEMIG, a série de consumo de New England também está originalmente agrupada por hora, e a partir dela é possível construir as séries de consumo total por dia, por semana, por quinzena e por mês. Na Figura 3.24 são apresentados os gráficos da série original de consumo mensal juntamente com a componente de tendência estimada e a série obtida após extração da tendência. Para estimação da tendência, foi considerada a combinação aditiva das

Tabela 3.5: Séries de consumo mensal de baixa tensão da CEMIG - Significância associada à FAC e à IMM para atrasos de 1 a 4 amostras.

Série vs Atraso	S_{FAC}				S_{IMM}			
	1	2	3	4	1	2	3	4
Ilum. Pub.	1,64	0,48	1,36	1,05	1,72	0,52	0,91	0,13
Pod. Pub.	3,33	2,25	1,67	2,85	1,06	1,04	2,26	0,05
Rural	4,36	2,46	0,29	2,69	3,92	1,40	0,22	1,27
Resid.	0,62	1,20	0,28	0,19	0,11	0,49	0,05	0,66
Ind.	1,59	1,15	0,27	2,10	1,40	0,57	1,34	0,59
Serv. Pub.	0,82	0,02	0,97	0,03	2,40	0,61	0,51	0,69
Cons. Prop.	1,35	1,00	0,65	0,64	0,57	1,21	0,44	0,09
Comer.	1,21	0,42	1,19	0,66	1,25	0,09	0,25	0,94
Total BT	0,32	0,81	0,69	0,84	1,15	0,54	1,09	1,65

Tabela 3.6: Séries de consumo mensal de baixa tensão da CEMIG - p -valores associados à FAC e à IMM para atrasos de 1 a 4 amostras.

Série vs Atraso	Pv_{FAC}				Pv_{IMM}			
	1	2	3	4	1	2	3	4
Ilum. Pub.	0,02	0,61	0,12	0,25	0,09	0,58	0,38	0,92
Pod. Pub.	0,002	0,02	0,07	0,002	0,30	0,30	0,01	0,99
Rural	0,002	0,002	0,75	0,002	0,002	0,16	0,83	0,08
Resid.	0,53	0,21	0,77	0,87	0,86	0,66	0,99	0,52
Ind.	0,09	0,24	0,82	0,02	0,14	0,60	0,15	0,56
Serv. Pub.	0,50	0,96	0,49	0,79	0,006	0,54	0,52	0,57
Cons. Prop.	0,17	0,30	0,56	0,56	0,60	0,24	0,68	0,97
Comer.	0,22	0,65	0,24	0,54	0,22	0,88	0,75	0,35
Total BT	0,79	0,41	0,50	0,42	0,25	0,62	0,30	0,10

quatro primeiras componentes obtidas por meio da decomposição da série original em 30 componentes via SSA. A Figura 3.25 ilustra as estatísticas discriminates associadas à série mensal juntamente com as respectivas significâncias obtidas ao se realizar o teste de dados sub-rogados com a geração também de 1000 séries artificiais.

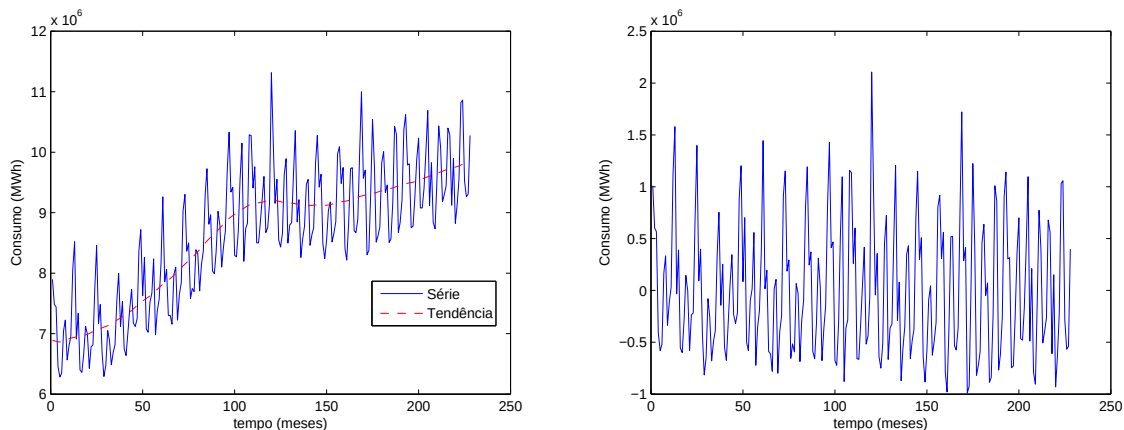


Figura 3.24: Dados de consumo mensal de New England - À esquerda, série original e componente de tendência estimada; à direita, série resultante após a extração da componente de tendência.

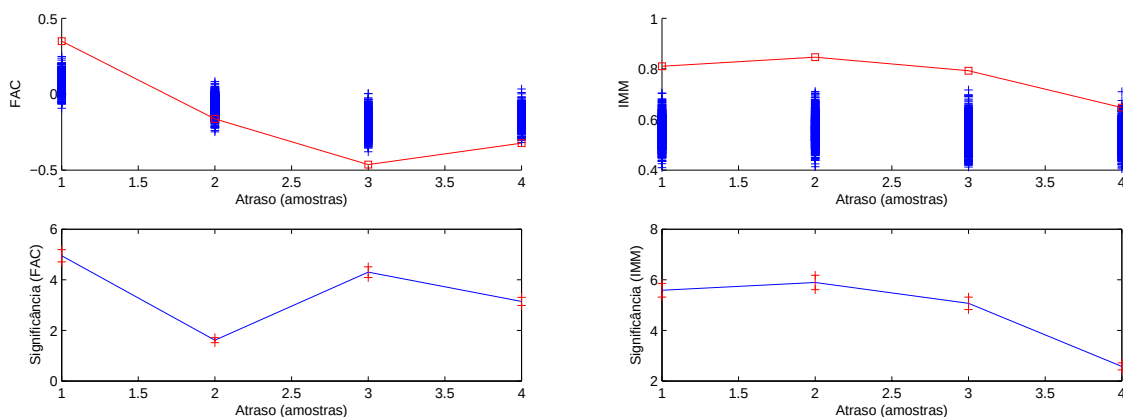


Figura 3.25: FAC e IMM das séries de consumo mensal de New England e das respectivas séries sub-rogadas. A significância define a distância entre a série original e as séries artificiais.

As Tabelas 3.7 e 3.8 apresentam as significâncias e os p -valores do teste de dados sub-rogados para o caso mensal. É inequívoca a rejeição da hipótese nula tanto para a FAC como para a IMM, o que nos permite concluir quanto à existência de determinismo nas flutuações irregulares dessa série. Com base nesta constatação e na observação feita anteriormente, no caso das séries de consumo de baixa tensão da CEMIG, acerca da maior tendência em se rejeitar a hipóteses em séries com intervalos menores entre amostras adjacentes, poupou-se do trabalho de realização do teste para as séries semanal, quinzenal, diária e horária. Importante lembrar mais uma vez que a série mensal

de New England apresenta um padrão muito mais regular e detectável a olho nu do que as séries da CEMIG, e talvez seja por esse fato que a rejeição de hipótese no caso da primeira seja mais patente no que no caso das últimas.

Tabela 3.7: Aplicação do SSS às séries de consumo mensal de New England - Significância associada à FAC e à IMM para atrasos de 1 a 4 amostras.

Série vs Atraso	S_{FAC}				S_{IMM}			
	1	2	3	4	1	2	3	4
Mensal	4,95	1,61	4,30	3,15	5,59	5,89	5,07	2,58

Tabela 3.8: Aplicação do SSS às séries de consumo mensal de New England - p -valores associados à FAC e à IMM para atrasos de 1 a 4 amostras.

Série vs Atraso	Pv_{FAC}				Pv_{IMM}			
	1	2	3	4	1	2	3	4
Mensal	0,002	0.09	0,002	0,002	0,002	0,002	0,002	0,01

3.4.3 Aplicação do Teste de Dados Sub-rogados CSS

O mesmo procedimento de estimação de tendência e de geração de 1000 séries de dados sub-rogados para cada teste foi repetido aqui na aplicação do algoritmo CSS, a fim de se caracterizar a dependência temporal entre-ciclos das séries de consumo de energia elétrica estudadas. Pelo objetivo desse teste, a diferença fundamental é que no cálculo das EDs foram considerados os quatro primeiros atrasos múltiplos do ciclo anual, portanto atrasos de 12, 24, 36 e 48 meses.

Série de consumo total da CEMIG

As Tabelas 3.9 e 3.10 apresentam os resultados obtidos com o CSS para a série de consumo mensal total da CEMIG. Observa-se que tanto em termos de significância S como de p -valores, não há evidência de rejeição da hipótese da nula, que assume a inexistência de determinismo entre ciclos na série.

Tabela 3.9: Aplicação do CSS às séries de consumo total da CEMIG - Significância associada à FAC e à IMM para atrasos de 12, 24, 36 e 48 amostras.

Série vs Atraso	S_{FAC}				S_{IMM}			
	12	24	36	48	12	24	36	48
Mensal	0,11	0,93	0,71	0,47	1,13	1,55	0,59	0,29

Tabela 3.10: Aplicação do CSS às séries de consumo total da CEMIG - p -valores associados à FAC e à IMM para atrasos de 12, 24, 36 e 48 amostras.

Série vs Atraso	Pv_{FAC}				Pv_{IMM}			
	12	24	36	48	12	24	36	48
Mensal	0,87	0,37	0,45	0,67	0,26	0,13	0,64	0,77

Séries de consumo mensal de baixa tensão da CEMIG

Aplicando-se o CSS às séries das oito categorias de consumidores de baixa tensão e à série de consumo de baixa tensão total da CEMIG, obtêm-se os resultados do teste de dados sub-rogados em termos de significância e de p -valores apresentados nas Tabelas 3.11 e 3.12.

Tabela 3.11: Séries de consumo mensal de baixa tensão da CEMIG - Significância associada à FAC e à IMM para atrasos de 12, 24, 36 e 48 amostras.

Série vs Atraso	S_{FAC}				S_{IMM}			
	12	24	36	48	12	24	36	48
Ilum. Pub.	2,31	0,16	0,95	1,07	0,63	0,57	0,69	1,06
Pod. Pub.	2,06	1,43	0,00	0,96	1,04	0,97	0,99	2,26
Rural	0,70	0,69	0,91	1,25	1,05	0,72	0,05	0,53
Resid.	1,62	1,42	0,49	0,29	0,24	1,19	0,97	1,49
Ind.	2,23	1,19	0,85	0,21	0,07	0,99	0,83	1,59
Serv. Pub.	2,31	0,89	1,47	0,95	2,32	0,39	0,53	0,17
Cons. Prop.	1,24	1,48	1,55	0,04	0,57	0,47	0,79	1,59
Comer.	2,27	0,36	2,06	0,77	0,08	0,08	0,53	0,00
Total BT	2,00	1,72	1,25	1,03	0,22	1,71	1,59	1,93

Tabela 3.12: Séries de consumo mensal de baixa tensão da CEMIG - p -valores associados à FAC e à IMM para atrasos de 12, 24, 36 e 48 amostras.

Série vs Atraso	Pv_{FAC}				Pv_{IMM}			
	12	24	36	48	12	24	36	48
Ilum. Pub.	0,04	0,89	0,25	0,13	0,50	0,57	0,48	0,27
Pod. Pub.	0,05	0,13	0,92	0,36	0,33	0,36	0,34	0,01
Rural	0,52	0,59	0,47	0,26	0,33	0,53	0,95	0,66
Resid.	0,11	0,15	0,71	0,84	0,91	0,10	0,44	0,16
Ind.	0,03	0,23	0,40	0,77	0,76	0,32	0,50	0,11
Serv. Pub.	0,02	0,40	0,11	0,38	0,01	0,73	0,36	0,61
Cons. Prop.	0,23	0,13	0,15	0,98	0,75	0,81	0,52	0,03
Comer.	0,03	0,71	0,02	0,47	0,81	0,96	0,44	0,80
Total BT	0,06	0,04	0,17	0,26	0,98	0,02	0,09	0,07

Observa-se que os resultados de ambas as tabelas coincidiram em termos de rejeição da hipótese. Para os setores 'iluminação pública', 'poder público', 'industrial', 'serviço

público' e 'comercial', os quais correspondem a 35% do consumo total de baixa tensão, houve rejeição da hipótese pelo menos em nível forte, de acordo com o cálculo dos p -valores. Os resultados revelam que a série total atingiu o limiar de significância de 2σ para atraso de uma amostra em termos de FAC e um p -valor de 0,06, indicativo de evidência fraca de rejeição da hipótese. Um outro aspecto a ser considerado e que pode ser enfatizado com esses resultados é que o fato de haver rejeição no SSS não implica rejeição no CSS. Foi o que o correu com o setor 'rural', em cuja série foi identificado determinismo nas flutuações por meio do SSS (dinâmica intra-ciclo) porém não foi identificada dependência entre um ciclo e outro (dinâmica entre ciclos), o que não significa que essa dependência não exista, uma vez que o teste de dados sub-rogados só é capaz de revelar se a hipótese nula pode ser rejeitada ou não. Só se pode dizer que o teste não foi capaz de identificá-la.

Série de consumo total de New England

Por fim, aplicou-se o CSS à série de consumo mensal de New England. Os resultados de significância e p -valores obtidos são apresentados nas Tabelas 3.13 e 3.14. Observa-se que houve rejeição da hipótese nula considerando-se ambas as estatísticas discriminantes (FAC e IMM) e ambas as medidas do teste (significância e p -valores). Como houve significativa rejeição da hipótese para o caso da série de New England tanto na aplicação do SSS como do CSS, pode-se dizer que essa série é caracterizada tanto pela presença de dinâmica em suas flutuações intra-ciclo como pela dependência temporal existente entre ciclos diferentes.

Tabela 3.13: Aplicação do CSS às séries de consumo mensal de New England - Significância associada à FAC e à IMM para atrasos de 1 a 4 amostras.

Série vs Atraso	S_{FAC}				S_{IMM}			
	12	24	36	48	12	24	36	48
Mensal	2,68	1,34	2,61	2,38	2,58	0,27	1,66	1,00

Tabela 3.14: Aplicação do CSS às séries de consumo mensal de New England - p -valores associados à FAC e à IMM para atrasos de 1 a 4 amostras.

Série vs Atraso	Pv_{FAC}				Pv_{IMM}			
	12	24	36	48	12	24	36	48
Mensal	0,006	0,17	0,01	0,02	0,01	0,81	0,06	0,31

3.5 Conclusão

No presente capítulo foram apresentadas as séries de consumo de energia elétrica da CEMIG e de New England, fez-se a descrição do procedimento de retirada de

valores espúrios por meio do Teste de Grubbs, foi abordada uma técnica decomposição das séries temporais por meio da ferramenta SSA e realizou-se o teste de dados sub-rogados para caracterização da dinâmica intra-ciclo e da dinâmica entre ciclos das referidas séries.

Por sua simplicidade de implementação e flexibilidade na escolha de em quantas componentes se deseja decompor as séries, o que possibilita um ajuste da suavidade da sua tendência, por exemplo, o SSA foi utilizado nesse trabalho como ferramenta para a decomposição de séries temporais. A utilização do SSA permitiu que se fizesse uma estimativa das componentes de tendência e até mesmo de comportamento cíclico ou sazonal, a depender do número de componentes consideradas na reconstrução da série original. Essa decomposição foi utilizada para extrair a tendência dos dados antes de se executar a análise de dados sub-rogados. Das várias técnicas de análise de dados sub-rogados apresentadas e descritas ao longo do capítulo, apenas o SSS e o CSS foram aplicados a essas séries na identificação de determinismo entre ciclos e intra-ciclo, respectivamente. Para a realização dos testes utilizaram-se como EDs a FAC e a IMM, ao passo que a validade do teste foi aferida por meio do cálculo da significância S e do p -valor (ou probabilidade de significância). Observou-se que nas séries da CEMIG agrupadas por mês, visualmente mais irregulares do que a série de New England, não houve rejeição significativa das hipóteses nulas, ao passo que para os dados mensais norte americanos a rejeição foi patente em ambos os algoritmos. No caso do CSS, para as séries da CEMIG divididas por setor, identificou-se que para algumas houve rejeição significativa da hipótese nula de inexistência de determinismo entre ciclos, o que implicou fraca evidência de rejeição no caso da série total de baixa tensão ($S_{FAC} = 2\sigma$ e $Pv_{FAC} = 0,06$ para defasamento de 12 amostras).

No Capítulo 5 são registrados os resultados de previsão alcançados com as ferramentas preditivas usadas neste trabalho para as três séries apresentadas neste Capítulo 3. Conforme se verificará, no caso das séries de New England, em que a rejeição das hipóteses nulas no teste de dados sub-rogados sempre foi verificada, a forma de onda das previsões de longo prazo se aproximaram mais do comportamento real de consumo do que no caso de ambas as séries da CEMIG, para as quais a detecção de determinismo muitas vezes não foi verificada, a não ser no caso das séries de alguns setores pouco significativos no cômputo global de consumo ou nas séries de consumo total por quinzena, semana, dia ou hora. Dentro desta tônica, uma pequena discussão acerca de predição dinâmica e mapeamento de padrões será levantada no já referido Capítulo 5.

Metodologias de Previsão

4.1 Introdução

Este capítulo consiste na exposição das metodologias utilizadas neste trabalho para a construção de modelos matemáticos dinâmicos com finalidades preditivas a partir das séries temporais de consumo de energia elétrica já apresentadas (ver capítulo 3). Três foram os tipos de modelos utilizados: (i) modelos auto-regressivos múltiplos (ARM); (ii) rede neural artificial do tipo MLP; (iii) rede neuro-fuzzy linear local (RNFL). Antes de apresentar as três metodologias preditivas, todavia, apresentaremos as medidas de erro para avaliar a qualidade preditiva dos modelos e definiremos um preditor trivial com o objetivo de termos um elemento de comparação simples que permita aferir a qualidade dos resultados obtidos com as três estratégias de previsão enunciadas.

4.2 Medidas de Erro

Com o objetivo de quantificar os resultados obtidos com os modelos de previsão, foram escolhidos três índices como medidas de erro (Gooijer and Hyndman, 2006):

- MAPE (Mean Absolute Percentual Error)

$$MAPE = \frac{100}{n} \sum_{k=1}^n \frac{|\hat{y}(k) - y(k)|}{y(k)} \quad (4.1)$$

em que $\hat{y}(k)$ é o consumo predito na amostra k , $y(k)$ é o consumo medido na amostra k , e n é o horizonte de previsão.

- MPE (Mean Percentual Error)

$$MPE = \frac{100}{n} \sum_{k=1}^n \frac{\hat{y}(k) - y(k)}{y(k)} \quad (4.2)$$

- TPE (Total Percentual Error)

$$TPE = 100 \frac{\sum_{k=1}^n \hat{y}(k) - \sum_{k=1}^n y(k)}{\sum_{k=1}^n y(k)} \quad (4.3)$$

O MAPE é um índice que mede a acurácia com que $\hat{y}(k)$ se ajusta a $y(k)$. Um ajuste perfeito implica em MAPE igual a zero, da mesma forma que um MAPE nulo implica em ajuste perfeito. Qualquer desvio ou diferença entre a curva $\hat{y}(k)$ e $y(k)$ é contabilizado positivamente no MAPE. Por essa razão, a partir do MAPE não se é possível distinguir um erro de previsão por subestimação de um erro de previsão por superestimação.

Já o MPE se trata de uma espécie de medida da média de erros percentuais locais, de maneira que erros percentuais ponto a ponto da série que sejam diferentes de zero podem se combinar e gerar um MPE nulo. Em outras palavras, erro pontuais positivos e negativos podem se cancelar e implicar em $MPE = 0$. Se o MPE for menor do que zero, significa que a previsão da série foi subestimada.

O TPE, por fim, parece-se com o MPE no sentido de poder se anular para um ajuste de curvas que não seja perfeito, porém ele se trata de uma medida do erro médio global, diferentemente do MPE que é uma média dos erros percentuais locais (ou ponto a ponto). Se o TPE for menor do que zero, assim como o MPE, então conclui-se que a previsão da série foi subestimada.

Essas três medidas foram utilizadas para acessar a qualidade dos modelos obtidos.

4.3 Definição de um preditor trivial

Para fins de verificação do valor dos resultados obtidos a partir de quaisquer modelos preditivos, é interessante que eles sejam comparados com os resultados de um preditor dito trivial. Costumeiramente, ele é expressado da seguinte maneira:

$$\hat{y}(k+1) = y(k) \quad (4.4)$$

Trata-se, portanto, de um modelo que diz que o valor previsto para a série no instante $(k+1)$ é simplesmente o valor real da série em um instante anterior (k) . Como o presente trabalho trata da previsão de consumo de energia em um horizonte de previsão da ordem de unidades de anos (previsão de longo prazo), seria interessante definir um preditor mais rigoroso que assumisse que só se conhece a série até o instante (N) e desejasse estimar o seu valor no instante $(N+n)$, em que n varia de 1 até o horizonte de previsão desejado h . Matematicamente:

$$\hat{y}(N+n) = y(\text{mod}(N+n, T)), \quad y(0) = y(T), \quad n \in \{1, 2, \dots, h\} \quad (4.5)$$

em que $\text{mod}(a, b)$ é o resto da divisão de a por b e T é o período fundamental da série em meses (nesse caso, $T = 12$). Portanto, assumindo-se que a estimativa de consumo de energia para os anos de 1993 a 1997 em New England seja uma repetição do consumo em 1992 ($n = 1, 2, 3, \dots, 60$, ou seja, $h = 60$ meses), chega-se aos resultados apresentados na Tabela 4.1. A Figura 4.1 ilustra o fato óbvio de que a curva predita é constituída de cinco ciclos idênticos. Além disso, observa-se que a previsão fica em média abaixo da curva real, uma vez que os dados reais apresentam uma leve tendência de crescimento, enquanto que a série prevista, *por construção*, se mantém num mesmo patamar médio. Tal fato se evidencia verificando-se na Tabela 4.1 que os valores de MPE e TPE são todos

negativos (previsão subestimada).

Tabela 4.1: Erros da previsão de consumo de 1993 a 1997 a partir do consumo de 1992 com o preditor ingênuo.

Ano	MAPE	MPE	TPE
1993	2,04%	-1,50%	-1,54%
1994	3,95%	-3,08%	-3,36%
1995	4,46%	-3,70%	-3,85%
1996	5,37%	-5,37%	-5,37%
1997	6,33%	-5,77%	-5,84%

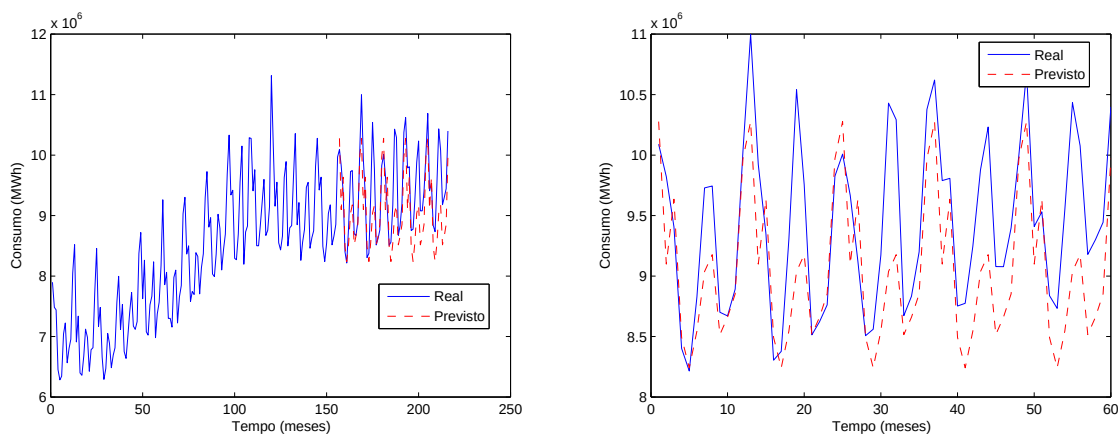


Figura 4.1: Resultados da previsão de 1993 a 1997 a partir do consumo de 1992 com o preditor ingênuo.

Justamente por esse fato decorrente da não-estacionariedade das séries em termo de média, uma maneira mais interessante de se realizar a previsão trivial seria combinar aditivamente uma extrapolação de um ajuste polinomial (uma reta) com a repetição do último ciclo sazonal anual da série, por exemplo. Matematicamente, podemos expressar essas idéia da seguinte forma:

$$\hat{y}(N + n) = s_1(\text{mod}(N + n, T)) + p(N + n), \quad s_1(0) = s_1(T), \quad n \in \{1, 2, \dots, h\} \quad (4.6)$$

em que $s_1(\cdot)$ é o último ciclo sazonal do conjunto de dados de identificação e $p(\cdot)$ é um polinômio utilizado para o ajuste de tendência. No próximo capítulo, apresentaremos os resultados para todas as séries obtidas com o preditor trivial dado pela Eq. (4.6). Convém ressaltar que o preditor trivial será considerado a pedra de toque para a verificação da eficácia dos modelos propostos para a previsão de consumo de energia.

4.4 Modelos Auto-Regressivos Múltiplos

Foi proposto por Al-Hamadi and Soliman (2005) um modelo para previsão de consumo de energia elétrica a médio e longo prazo o qual consiste na construção de

uma matriz $L(i,k)$ de 52 linhas e 24 colunas, em que cada linha corresponde a uma semana do ano e cada coluna a uma hora do dia. Portanto, essa matriz é da forma

$$L(i,k) = \begin{bmatrix} l_{1,1} & l_{1,2} & \dots & l_{1,24} \\ l_{2,1} & l_{2,2} & \dots & l_{2,24} \\ \vdots & \vdots & \ddots & \vdots \\ l_{52,1} & l_{52,2} & \dots & l_{52,24} \end{bmatrix} \quad (4.7)$$

em que $L \in \mathbb{R}^{52 \times 24}$ e o elemento $l_{i,k}$ representa o consumo médio de energia da k -ésima hora do dia ao longo da semana i . Por exemplo, o elemento $l_{36,9}$ representa o consumo médio da nona hora do dia ao longo da trigésima sexta semana do ano. Cada linha da matriz L , portanto, corresponde ao dia médio de uma determinada semana do ano.

A partir dos dados de consumo total por hora da CEMIG, foram montadas duas matrizes, uma correspondente ao ano de 2004 e outra ao ano de 2005, e em seguida gerados alguns gráficos. Eles estão apresentados nas Figuras 4.2 e 4.3. O primeiro gráfico da Figura 4.2 apresenta a primeira e a segunda linhas da matriz do ano de 2004, ou seja, o consumo médio ao longo das horas do dia para a primeira e segunda semana de 2004, ao passo que o segundo gráfico traz as curvas de consumo médio ao longo das horas do dia para a primeira semana dos anos de 2004 e 2005 (primeira linha de cada matriz). O grau de semelhança entre as duas curvas de consumo em cada gráfico indica que elas estão fortemente correlacionadas. Já o primeiro gráfico da Figura 4.3 ilustra o consumo médio para a primeira e a segunda hora do dia ao longo das semanas do ano de 2004, portanto correspondem às duas primeiras colunas da matriz de 2004, enquanto que o segundo gráfico apresenta o consumo médio ao longo das semanas para a primeira hora do dia dos anos de 2004 e 2005 (primeira coluna de cada matriz). É visivelmente notório, nesse último gráfico, que a correlação entre as duas curvas mostradas é bem menor do que no caso dos três gráficos anteriores.

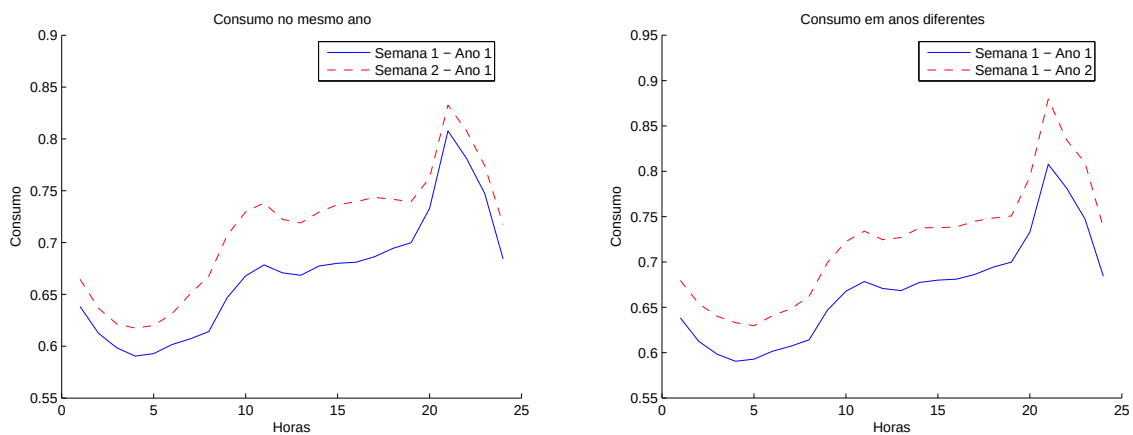


Figura 4.2: Séries de consumo evidenciando correlações entre semanas do ano existentes nos dados de consumo dos anos de 2004 e 2005.

Baseados nessas correlações, os autores do método propõem que se construam

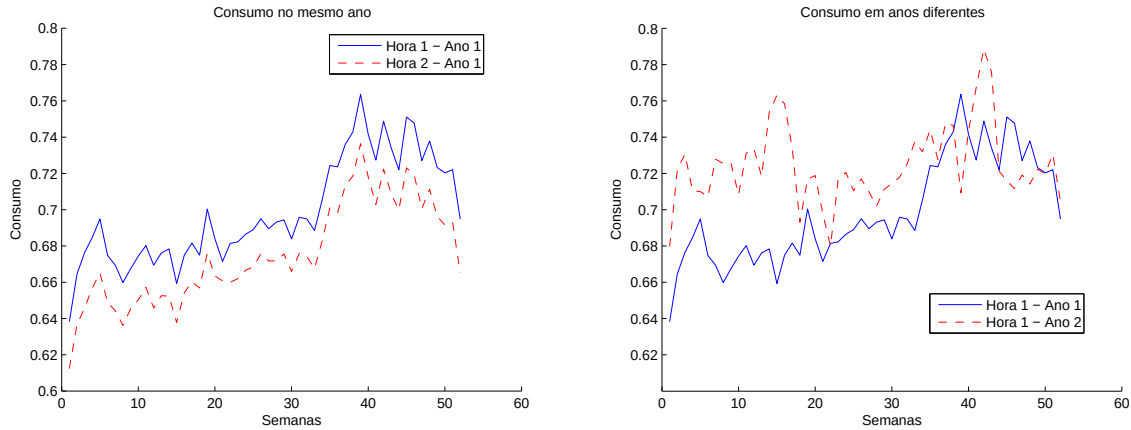


Figura 4.3: Séries de consumo evidenciando correlações entre horas do dia existentes nos dados de consumo dos anos de 2004 e 2005.

modelos “auto-regressivos”¹ da forma

$$L(i,k) = a(i)L(i-1,k) + b(i), \quad k = 1,2,\dots,24 \quad (4.8)$$

em que $L(i,k), k = 1,2,\dots,24$ representa a i -ésima linha da matriz L , e $L(i-1,k), k = 1,2,\dots,24$ representa a $(i-1)$ -ésima linha. De maneira análoga, são construídos modelos

$$L(i,k) = c(k)L(i,k-1) + d(k), \quad i = 1,2,\dots,52 \quad (4.9)$$

em que $L(i,k), i = 1,2,\dots,52$ representa a k -ésima coluna da matriz L , e $L(i,k-1), i = 1,2,\dots,52$ representa a $(k-1)$ -ésima coluna. Os termos não-lineares $b(i)$ e $d(k)$ são de fundamental importância para a capacidade de ajuste dos modelos ARM aos dados de treinamento, como será discutido no próximo capítulo.

Para ilustração do funcionamento do algoritmo, seja L_{2004} a já referida matriz dos dados de consumo da CEMIG relativos ao ano de 2004. Primeiramente, faz-se $i = 1$ na Equação 4.8 (nesse caso, $L_{2004}(i-1,k) = L_{2004}(0,k) = L_{2003}(52,k), k = 1,2,\dots,24$)², depois monta-se a matriz de regressores $\Psi_1 \in \mathbb{R}^{24 \times 2}$ e em seguida estima-se, via mínimos quadrados, os valores de $a(1)$ e $b(1)$, a partir da expressão

$$\begin{bmatrix} L_{2004}(1,1) \\ L_{2004}(1,2) \\ \vdots \\ L_{2004}(1,24) \end{bmatrix} = \Psi_1 \times \begin{bmatrix} a(1) \\ b(1) \end{bmatrix} = \begin{bmatrix} L_{2003}(52,1) & 1 \\ L_{2003}(52,2) & 1 \\ \vdots & \vdots \\ L_{2003}(52,24) & 1 \end{bmatrix} \begin{bmatrix} a(1) \\ b(1) \end{bmatrix}. \quad (4.10)$$

Em seguida, calcula-se a primeira linha da matriz de 2005, conforme abaixo:

¹O termo está entre aspas porque, em modelos AR, o mesmo conjunto de dados é usado para comporem-se os regressores (a menos dos respectivos atrasos, obviamente). No presente caso, todavia, isso não acontece. Para cada regressor do modelo, considera-se um conjunto diferente de dados, quais sejam as diferentes linhas ou colunas da matriz L .

²Um típico problema de condição inicial. Uma vez que não existe linha anterior à primeira da matriz L_{2004} , utiliza-se a última linha da matriz de 2003 (L_{2003}) como condição inicial.

$$\hat{L}_{2005}(1,k) = a(1)L_{2004}(52,k) + b(1), \quad k = 1, 2, \dots, 24 \quad (4.11)$$

Concluindo esse passo, faz-se $k = 1$ na Equação 4.9 (nesse caso, $L_{2004}(i,k-1) = L_{2004}(i,0) = L_{2003}(i,24)$, $i = 1, 2, \dots, 52$, depois monta-se a matriz de regressores $\Psi_2 \in \mathbb{R}^{51 \times 2}$ e em seguida estima-se, via mínimos quadrados, os valores de $c(1)$ e $d(1)$, a partir da expressão

$$\begin{bmatrix} L_{2004}(2,1) \\ L_{2004}(3,1) \\ \vdots \\ L_{2004}(52,1) \end{bmatrix} = \Psi_2 \times \begin{bmatrix} c(1) \\ d(1) \end{bmatrix} = \begin{bmatrix} L_{2003}(2,24) & 1 \\ L_{2003}(3,24) & 1 \\ \vdots & \vdots \\ L_{2003}(52,24) & 1 \end{bmatrix} \begin{bmatrix} c(1) \\ d(1) \end{bmatrix} \quad (4.12)$$

. O cálculo da primeira coluna da matriz de 2005 é direto:

$$\hat{L}_{2005}(i,1) = c(1)L_{2004}(i,24) + d(1), \quad i = 1, 2, \dots, 52 \quad (4.13)$$

Repetindo-se o procedimento para $i = 2$, calculam-se $a(2)$ e $b(2)$ via MQ, para uma nova matriz de regressores $\Psi_3 \in \mathbb{R}^{23 \times 2}$:

$$\begin{bmatrix} L_{2004}(2,2) \\ L_{2004}(2,3) \\ \vdots \\ L_{2004}(2,24) \end{bmatrix} = \Psi_3 \times \begin{bmatrix} a(2) \\ b(2) \end{bmatrix} = \begin{bmatrix} L_{2004}(1,2) & 1 \\ L_{2004}(1,3) & 1 \\ \vdots & \vdots \\ L_{2004}(1,24) & 1 \end{bmatrix} \begin{bmatrix} a(2) \\ b(2) \end{bmatrix} \quad (4.14)$$

e tem-se então a previsão

$$\hat{L}_{2005}(2,k) = a(2)\hat{L}_{2005}(1,k) + b(2), \quad k = 2, 3, \dots, 24 \quad (4.15)$$

Para $k = 2$, calculam-se $c(2)$ e $d(2)$ via MQ, para uma nova matriz de regressores $\Psi_4 \in \mathbb{R}^{50 \times 2}$

$$\begin{bmatrix} L_{2004}(3,2) \\ L_{2004}(4,2) \\ \vdots \\ L_{2004}(52,2) \end{bmatrix} = \Psi_4 \times \begin{bmatrix} c(2) \\ d(2) \end{bmatrix} = \begin{bmatrix} L_{2004}(3,1) & 1 \\ L_{2004}(4,1) & 1 \\ \vdots & \vdots \\ L_{2004}(52,1) & 1 \end{bmatrix} \begin{bmatrix} c(2) \\ d(2) \end{bmatrix} \quad (4.16)$$

e obtém-se a previsão a partir de

$$\hat{L}_{2005}(i,2) = c(2)\hat{L}_{2005}(i,1) + d(2), \quad i = 3, 4, \dots, 52 \quad (4.17)$$

e assim sucessivamente, até se atingir $i = k = 24$. Depois de levantados os 48 modelos “auto-regressivos de primeira ordem”, tem-se a matriz de 2005 prevista.

Para melhorar o resultado obtido, os autores ainda propõem um ajuste para crescimento anual. Seja o seguinte modelo “auto-regressivo”:

$$LG(i,k) = \beta_0(k) + \beta_1(k)L(i,k-1) + \beta_2(k)L^2(i,k-1) + \beta_3(k)L^3(i,k-1), \quad i = 1, 2, \dots, 52. \quad (4.18)$$

Fazendo-se k variar de 1 a 24, encontra-se uma matriz LG_{2004} a partir da matriz L_{2004} e uma matriz LG_{2005} a partir da matriz $\hat{L}_{2005}$ prevista. Para se obterem os resultados finais da previsão, computa-se a simples expressão matricial

$$L_{2005_{final}} = \hat{L}_{2005} + (LG_{2005} - LG_{2004}) \quad (4.19)$$

Ao término, computam-se as medidas de erro da previsão (ver Seção 4.2).

4.5 Rede Neural Artificial MLP

As redes neurais artificiais (RNAs) são ferramentas computacionais cujo mecanismo de funcionamento é inspirado do comportamento das redes neurais biológicas. Elas tratam-se basicamente de sistemas compostos por unidades de processamento simples (nodos, nós ou neurônios) as quais realizam o cálculo de determinadas funções matemáticas tipicamente não-lineares (Braga et al., 2000). De acordo com Nelles (2001), suas principais características são: (i) grande número de unidades simples; (ii) existência de forte conexão entre essas unidades; (iii) robustez a falha de alguma de suas unidades; (iv) capacidade de aprendizado a partir de dados. Segundo Braga et al. (2000), os principais atrativos da solução de problemas via RNAs consistem em sua capacidade de aprender por meio de exemplos e de generalizar a informação aprendida; em outras palavras, ao se apresentarem dados às RNAs, elas extrairão automaticamente deles as características necessárias para representar a informação fornecida (essa é a chamada fase de aprendizagem) e ao se lhes apresentarem dados não conhecidos previamente, elas exibirão respostas coerentes. Isso mostra que as RNAs não se restringem a realizar um mero mapeamento entrada-saída, pois são capazes de realizar interpolações e extrapolações a partir do seu aprendizado.

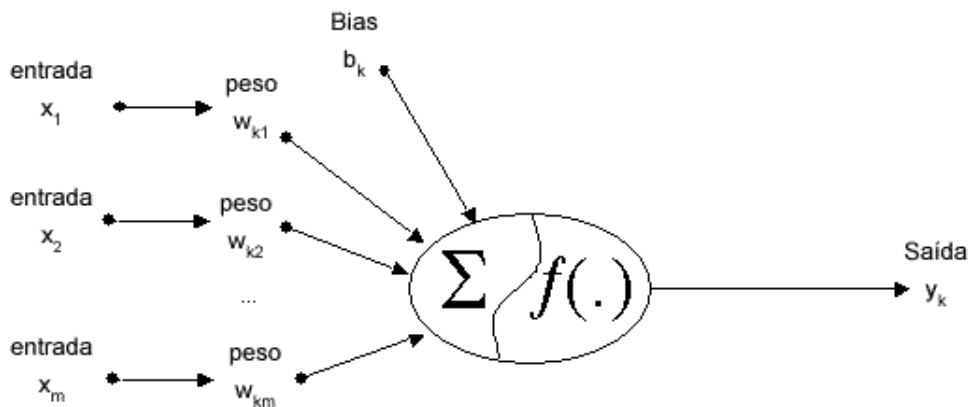


Figura 4.4: Diagrama esquemático de um neurônio artificial.

Na Figura 4.4 visualiza-se o diagrama esquemático de um neurônio artificial cuja

saída é dada pela expressão

$$y_k = f\left(\sum_{i=1}^m x_i w_{ki} + b_k\right) \quad (4.20)$$

em que f é uma função de ativação comumente não-linear, w_{ki} são os pesos e b_k é o termo de polarização.

Em termos de aplicações, existe uma gama muito vasta abarcada pelas RNAs dentro da qual se podem enumerar:

1. Modelagem e identificação de sistemas (Worden and Tomlinsom, 2001);
2. Controle (Carvalho and Hemerly, 2008);
3. Otimização (Curteanu and Leon, 2007);
4. Processamento de imagens e sinais (Masters, 1994);
5. Classificação (Zhang, 2000);
6. Previsão ou predição (Kermanshahi and Iwamiya, 2002).

No que diz respeito ao problema de previsão de séries temporais contemplado por este trabalho, foram utilizadas as chamadas redes perceptron multicamadas, ou MLPs (do inglês *multilayer perceptron*), ou seja, redes que possuem, além das camadas de entrada e de saída, camada(s) escondida(s) ou intermediária(s). A Figura 4.5 esboça uma típica topologia de uma rede MLP.

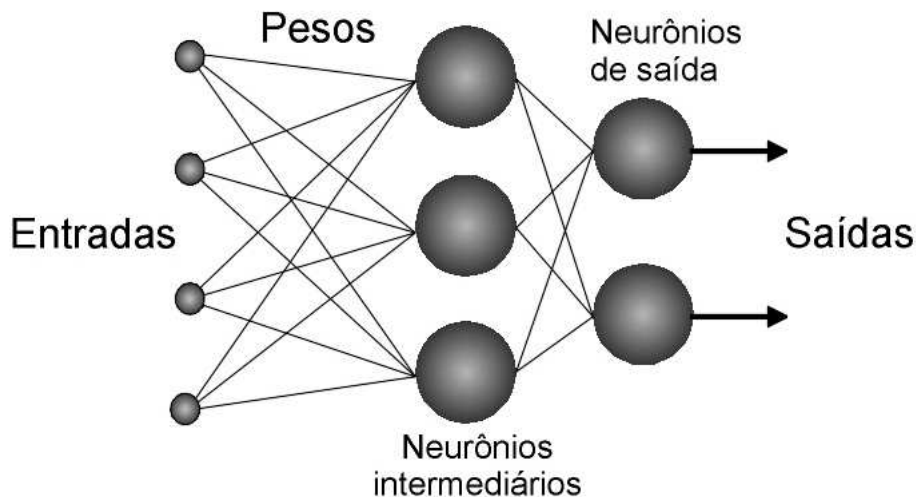


Figura 4.5: Diagrama esquemático de uma rede neural artificial MLP com uma camada escondida.

Uma das razões de tal escolha reside no fato de que uma rede MLP com pelo menos uma única camada escondida que contenha um número suficiente de neurônios é capaz de implementar qualquer função contínua (Braga et al., 2000). A saída de uma rede com tal configuração, em que o neurônio da camada de saída combina aditivamente o

seu termo de polarização b_0 e as saídas de cada neurônio da camada escondida, é dada por

$$y = \sum_{i=1}^M w_i f_i \left(\sum_{j=1}^m w_{ij} u_j + b_i \right) + b_0 \quad (4.21)$$

em que M é o número de neurônios na camada escondida, m é o número de entradas, w_i s são os pesos da camada de saída, w_{ij} s são os pesos da camada escondida e $f_i(\cdot)$ é a função de ativação do i -ésimo neurônio da camada escondida. Nesse trabalho, a tangente hiperbólica foi a função de ativação considerada em todos os neurônios da camada escondida:

$$f_i(x) = \frac{2}{1 + e^{-2x}} - 1, \quad (4.22)$$

enquanto que no único neurônio da camada de saída considerou-se a função de ativação linear

$$f(x) = x. \quad (4.23)$$

Uma rede MLP com essa topologia apresenta $M(m + 1) + M + 1$ parâmetros a serem estimados. O algoritmo mais conhecido para o treinamento de redes MLP (isto é, para o ajuste de seus parâmetros) é o *back-propagation*. Os outros métodos existentes para redes MLP consistem em variações desse algoritmo. O *back-propagation* é dividido em duas fases: *forward* e *backward*. Em síntese, a primeira fase consiste em calcular, para um dado padrão de entrada, a saída da rede. Na segunda fase, utiliza-se o erro entre a saída calculada e a saída desejada para que se atualizam os pesos das conexões entre os neurônios. Esse procedimento é repetido reiteradas vezes, para diferentes padrões de treinamento, até que se atinja um critério de parada, quer seja um número máximo de repetições ou um valor mínimo de erro. Uma descrição cuidadosa desse algoritmo pode ser encontrada em (Braga et al., 2000).

Um problema comumente enfrentado em treinamento de RNAs é a questão de *overfitting*, ou seja, o fato de a rede se ajustar tão bem aos dados de treinamento ao ponto de perder a sua capacidade de generalização. Esse é um problema quando se trata da modelagem ou identificação de um sistema dinâmico, em que a capacidade de generalização da rede é uma característica imprescindível, uma vez que um ajuste quase exato aos dados de treinamento não implica necessariamente boa qualidade do modelo preditivo.

Muitas modificações do algoritmo *back-propagation* tem sido propostas no sentido de aumentar a velocidade do treinamento e de melhorar o seu desempenho. No presente trabalho, foi utilizado para treinamento o algoritmo de Levenberg-Marquardt (LM), o qual se trata de um algoritmo de minimização de erro quadrático não-linear. Pode-se definir que ele se adequa à resolução do problema de ajuste de curvas, qual seja: dado um conjunto de dados de variáveis dependentes e independentes (x_i, y_i) , otimize os parâmetros θ do modelo não-linear $f(x_i, \theta)$ de modo que a soma dos quadrados dos desvios

$$J(\theta) = \sum_{i=1}^m [y_i - f(x_i, \theta)]^2 \quad (4.24)$$

seja mínima. Talvez a grande vantagem de se utilizar o algoritmo LM resida no fato de ele reunir em si a velocidade do algoritmo de Newton e a estabilidade do método do gradiente descendente (Madsen et al., 2004).

Uma outra questão a ser considerada no projeto de redes MLP é a determinação do número de camadas escondidas e do número de neurônios em cada uma dessas camadas. Como já foi citado anteriormente, uma rede com uma única camada é capaz de aproximar qualquer função contínua a depender da sua quantidade de neurônios. Resta, portanto, somente a definição de quantos neurônios irão compor a camada escondida, o que comumente se faz por meio de testes e simulações, como se verá mais adiante no capítulo de resultados deste trabalho.

4.6 Rede Neuro-Fuzzy Linear Local

Em dois trabalhos bastante semelhantes entre si (Ghoulipour et al., 2005, 2007), os autores aplicam uma rede neuro-fuzzy ao problema de modelagem preditiva de longo prazo da série de manchas solares. A ferramenta utilizada trata-se de um modelo neuro-fuzzy linear por partes cujo ajuste de parâmetros é realizado pelo algoritmo *Local Linear Model Tree* (LOLIMOT), o qual está detalhadamente descrito e analisado em (Nelles, 2001).

Uma rede neuro-fuzzy nada mais é do que um sistema baseado em regras nebulosas que é otimizado ou treinado diretamente a partir de dados (Nelles, 2001). No caso de um modelo neuro-fuzzy linear local ou linear por partes, tem-se uma estrutura como a apresentada na Figura 4.6 (Hoffman and Nelles, 2001). A cada neurônio ou nó da rede que implementa um Modelo Linear Local (MLL) está associado uma função de validade (FV), a qual define a região de validade do MLL. A saída de cada MLL, portanto, é dada por

$$\hat{y}_i = w_{i0} + w_{i1}u_1 + w_{i2}u_2 + \dots + w_{ip}u_p \quad (4.25)$$

em que w_{ij} são os parâmetros do MLL do i -ésimo neurônio. É de bom alvitre frisar que esses são os parâmetros lineares da rede, também chamados de parâmetros consequentes.

A saída global do modelo é dada pela expressão

$$\hat{y} = \sum_{i=1}^M \hat{y}_i \Phi_i(\vec{u}) \quad (4.26)$$

em que M é o número total de neurônios, $\vec{u} = \begin{bmatrix} u_1 & u_2 & \dots & u_p \end{bmatrix}^T$ é o vetor de entradas da rede e $\Phi_i(\cdot)$ é a i -ésima função de validade. A saída da rede, portanto, é calculada como uma soma ponderada da saída de cada um dos MLLs em que as funções $\Phi_i(\cdot)$ podem ser interpretadas como fatores de ponderação que dependem do ponto $\vec{u}$ em torno do qual a rede opera. Para garantir que a contribuição de todos os MLLs totalizem

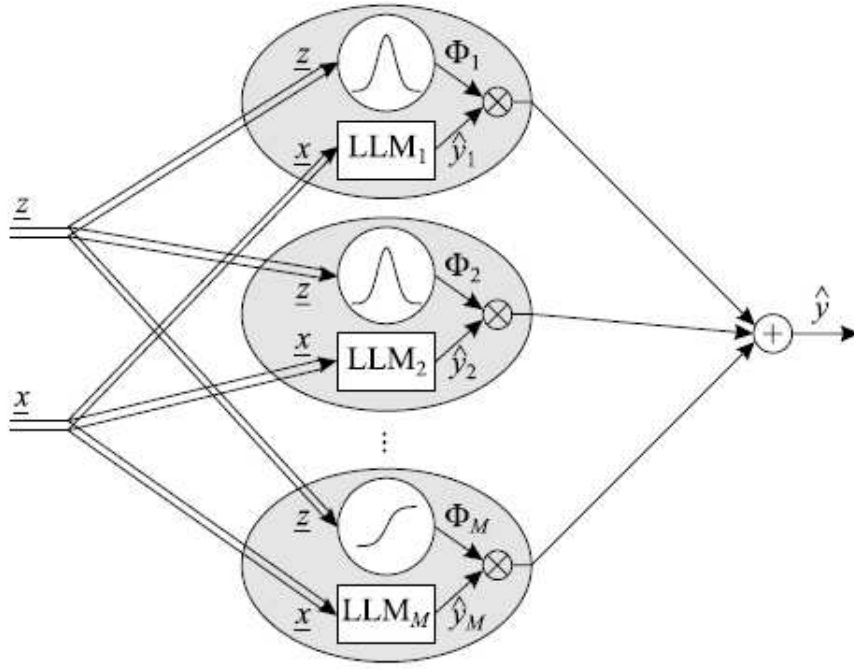


Figura 4.6: Diagrama esquemático de uma RNFL, em que $\vec{z}$ é o vetor de entradas das funções de validade e $\vec{x}$ é o vetor de entradas dos MLLs (Hoffman and Nelles, 2001). No caso geral considerado nessa dissertação, $\vec{u} = \vec{z} = \vec{x}$.

100%, é necessário que

$$\sum_{i=1}^M \Phi_i(\vec{u}) = 1 \quad (4.27)$$

ou seja, para $\vec{u}$ fixado em um determinado valor qualquer do espaço de entrada, as funções $\Phi_i(\vec{u})$ precisam ser tais que o seu somatório seja unitário.

Tipicamente, as funções de validade são Gaussianas normalizadas:

$$\Phi_i(\vec{u}) = \frac{\mu_i(\vec{u})}{\sum_{j=1}^M \mu_j(\vec{u})} \quad (4.28)$$

em que as funções de pertinência são dadas por

$$\begin{aligned} \mu_i(\vec{u}) &= \exp \left(-\frac{1}{2} \left(\frac{(u_1 - c_{i1})^2}{\sigma_{i1}^2} + \dots + \frac{(u_1 - c_{p1})^2}{\sigma_{p1}^2} \right) \right) \\ &= \exp \left(-\frac{1}{2} \frac{(u_1 - c_{i1})^2}{\sigma_{i1}^2} \right) \times \dots \times \exp \left(-\frac{1}{2} \frac{(u_1 - c_{p1})^2}{\sigma_{p1}^2} \right). \end{aligned} \quad (4.29)$$

As funções Gaussianas normalizadas dependem dos centros c_{ij} e dos desvios padrão σ_{ij} . Esses parâmetros são não-lineares e representam, por sua vez, os parâmetros da camada escondida de uma rede neural artificial. As funções de validade também são chamadas de funções de ativação, uma vez que elas definem ou controlam a atividade do MLL associado ao i -ésimo neurônio.

Para a estimação dos parâmetros consequentes (w_{ij}), duas alternativas se apresentam: estimação *global* e *local*.

Na estimação global, o que se faz é ajustar *todos* os parâmetros lineares de uma só vez via estimador de mínimos quadrados (MQ) (Aguirre, 2007). Dessa forma, o vetor de parâmetros é constituído de $M(p + 1)$ elementos, uma vez que a rede é composta por M neurônios e é alimentada com p entradas (ver Eqs. (4.25) e (4.26)):

$$\vec{w} = \begin{bmatrix} w_{10} & w_{11} & \dots & w_{1p} & w_{20} & w_{21} & \dots & w_{2p} & \dots & w_{M0} & w_{M1} & \dots & w_{Mp} \end{bmatrix}^T \quad (4.30)$$

A matriz de regressores global X para N medições é da forma

$$X = \begin{bmatrix} X_1^{(sub)} & X_2^{(sub)} & \dots & X_M^{(sub)} \end{bmatrix} \quad (4.31)$$

em que as submatrizes X_i são dadas por

$$X_i^{(sub)} = \begin{bmatrix} \Phi_i(\vec{u}(1)) & u_1(1)\Phi_i(\vec{u}(1)) & u_2(1)\Phi_i(\vec{u}(1)) & \dots & u_p(1)\Phi_i(\vec{u}(1)) \\ \Phi_i(\vec{u}(2)) & u_1(2)\Phi_i(\vec{u}(2)) & u_2(2)\Phi_i(\vec{u}(2)) & \dots & u_p(2)\Phi_i(\vec{u}(2)) \\ \vdots & \vdots & \vdots & & \vdots \\ \Phi_i(\vec{u}(N)) & u_1(N)\Phi_i(\vec{u}(N)) & u_2(N)\Phi_i(\vec{u}(N)) & \dots & u_p(N)\Phi_i(\vec{u}(N)) \end{bmatrix} \quad (4.32)$$

A saída do modelo $\hat{\vec{y}} = \begin{bmatrix} \hat{y}(1) & \hat{y}(2) & \dots & \hat{y}(N) \end{bmatrix}$, portanto, é calculada pela operação matricial

$$\hat{\vec{y}} = X\vec{w}. \quad (4.33)$$

Assim sendo, assumindo-se que $N \geq M(p + 1)$, os parâmetros lineares podem ser determinados pelo MQ:

$$\hat{\vec{w}} = \left[X^T X \right]^{-1} X^T \vec{y} \quad (4.34)$$

em que $\vec{y} = \begin{bmatrix} y(1) & y(2) & \dots & y(N) \end{bmatrix}^T$ é o vetor de medições da saída do processo. Essa é, portanto, a estimação global dos parâmetros consequentes.

Já na estimação local, os parâmetros de cada MLL são calculados *separadamente*. Se estivéssemos lidando com o caso em que σ é pequeno, de modo que gaussianas com médias diferentes praticamente não se superpõem, situação em que as funções de validade seriam pulsos retangulares que não se interpenetrariam, seria possível ignorar completamente as interações entre modelos e executar o MQ simples para uma única das submatrizes definidas na Eq. (4.32). Entretanto, no caso geral em que σ não é muito pequeno, ignorar as regiões de interseção entre as funções de validade pode comprometer o ajuste de parâmetros da rede como um todo. Para contornar esse problema, o que se pode fazer é aplicar o estimador de mínimos quadrados ponderado (MQP) (Aguirre, 2007), de tal forma que o ajuste de um MLL não comprometa o ajuste de um outro MLL adjacente. Como fatores de ponderação, claramente as funções de validade se tratam da escolha mais imediata.

Definindo-se então a matriz de pesos $N \times N$

$$Q_i = \begin{bmatrix} \Phi_i(\vec{u}(1)) & 0 & \dots & 0 \\ 0 & \Phi_i(\vec{u}(2)) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Phi_i(\vec{u}(N)) \end{bmatrix} \quad (4.35)$$

e a matriz de regressores

$$X_i = \begin{bmatrix} 1 & u_1(1) & u_2(1) & \dots & u_p(1) \\ 1 & u_1(2) & u_2(2) & \dots & u_p(2) \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & u_1(N) & u_2(N) & \dots & u_p(N) \end{bmatrix}, \quad (4.36)$$

a solução do MQP é dada por

$$\hat{w}_i = [X_i^T Q_i X_i]^{-1} X_i^T Q_i \vec{y}. \quad (4.37)$$

No que diz respeito ao ajuste dos parâmetros não-lineares c_{ij} e σ_{ij} , também chamados de parâmetros de premissa, deve-se levar em consideração o *particionamento do espaço de entrada*. Isso justamente porque o número de funções de validade está intimamente ligado à quantidade e aos valores associados aos parâmetros de premissa.

Para a realização de tal cometimento, Nelles (2001) propôs o algoritmo LOLIMOT, que é, em suma, um algoritmo incremental de particionamento do espaço de entrada em divisões ortogonais aos eixos do sistema de coordenadas euclidiano, o que define sub-regiões que claramente são hiper-retângulos. Em cada iteração, um novo MLL é adicionado à rede, as funções de validade correspondentes ao novo particionamento são calculadas e os correspondentes parâmetros consequentes são re-estimados pelo MQP (ver Eq. (4.37)). O único parâmetro que precisa ser ajustado *a priori* é o fator de proporcionalidade entre a extensão dos hiper-retângulos e os desvios padrão. De acordo com Nelles, um valor que costuma apresentar bons resultados é

$$k_\sigma = \frac{1}{3} \quad (4.38)$$

A partir dele, os desvios padrão são calculados por

$$\sigma_{ij} = k_\sigma \Delta_{ij} \quad (4.39)$$

em que Δ_{ij} representa a extensão do hiper-retângulo do i -ésimo modelo local na dimensão j .

O parâmetro c_{ij} será definido como a coordenada central do hiper-retângulo, na j -ésima dimensão, associado ao i -ésimo MLL.

Eis a seguir uma descrição mais detalhada do algoritmo LOLIMOT:

1. *Comece com um modelo inicial:* Construa as funções de validade para o particio-

namento inicial do espaço de entrada dado e estime os parâmetros dos MLLs via MQP. Ajuste M para o valor inicial de MLLs. Se não houver nenhum particionamento do espaço de entrada *a priori*, então faça $M = 1$ e comece com um MLL, que será de fato um modelo linear global, já que sua função de validade varrerá todo o espaço de entrada com $\Phi_1(\vec{u}) = 1$.

2. *Encontre o pior MLL*: Calcule uma função de custo para cada um dos $i = 1, \dots, M$ MLLs. Uma maneira de se fazer isso é calcular o somatório dos erros quadráticos do modelo ponderados pela função de validade do MLL correspondente:

$$I_i = \sum_{j=1}^N (y(j) - \hat{y}(j))^2 \Phi_i(\vec{u}(j)). \quad (4.40)$$

Encontre o MLL de pior performance ($\max(I_i)$), e o indexe como l -ésimo.

3. *Verifique todas as divisões*: O hiper-retângulo associado ao l -ésimo MLL é dividido em duas metades com um corte ortogonal a um dos eixos do sistema de coordenadas. Tente divisões em todas as dimensões. Para cada uma das divisões $dim = 1, \dots, p$, os seguintes passos devem ser executados:
 - construa as funções de pertinência multidimensionais para cada um dos dois novos hiper-retângulos;
 - construa as funções de validade de todos os MLLs;
 - estime localmente os parâmetros consequentes para ambos os novos MLLs;
 - calcule a função de custo para o novo modelo global (somatório das funções de custo de cada MLL).
4. *Encontre a melhor divisão*: A melhor das p alternativas verificadas no Passo 3 é selecionada. As funções de validade e os dois novos MLLs são integrados à rede. Incrementa-se o número de neurônios, portanto: $M \rightarrow M + 1$.
5. *Teste de convergência*: Se o critério de parada for atingido, então finalize o algoritmo, caso contrário, vá para o Passo 2.

A Figura 4.7 (Hoffman and Nelles, 2001) apresenta um exemplo de como seria o particionamento incremental do espaço de entrada da RNFL para o caso bidimensional ($p = 2$). Observa-se que a cada iteração, conforme apresentado no detalhamento do algoritmo acima, apenas uma única região é particionada em duas, o que é um fator determinante no baixo custo computacional do procedimento.

A seguir tem-se uma demonstração da utilização da RNFL para aproximação de uma curva estática por modelos afins. A expressão da curva que se deseja aproximar é

$$y = \frac{1}{0,1 + u} \quad (4.41)$$

Observa-se na Figura 4.8 que com o passar das iterações a saída da rede se aproxima cada vez mais da curva real. Cada um dos MLLs implementa uma reta e a cada

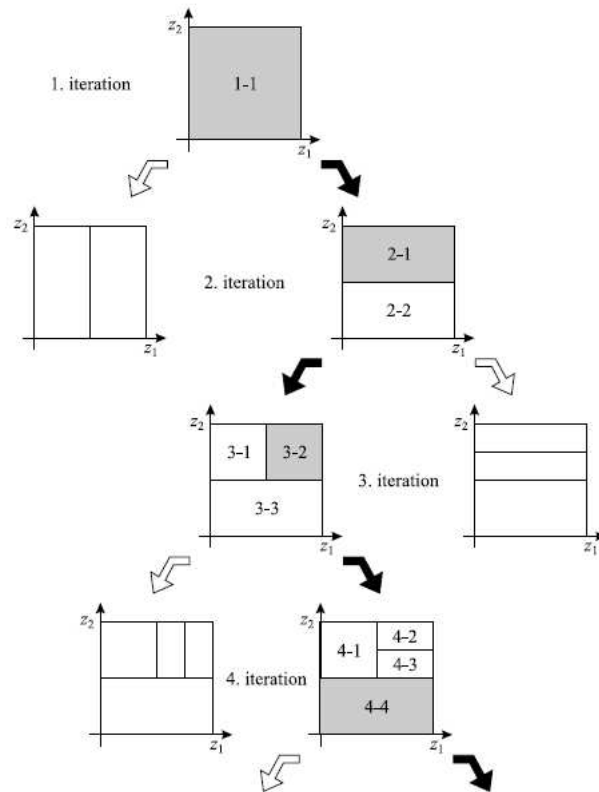


Figura 4.7: Diagrama esquemático do particionamento do espaço de entrada bidimensional ($p = 2$) de uma rede RNFL realizado pelo algoritmo LOLIMOT (Hoffman and Nelles, 2001).

neurônio, como se observa na referida figura, está associada uma função de validade diferente. Para esse ajuste, o critério de parada consiste em observar se, com uma nova divisão do pior neurônio, o erro global da aproximação aumenta. Em caso afirmativo, interrompe-se o algoritmo.

Conforme se verá no próximo capítulo, o critério de parada para o caso dinâmico da modelagem preditiva das séries temporais de consumo de energia será muito semelhante ao do caso estático, com a diferença de que a função de custo será computada sobre os primeiros valores do conjunto de validação. Em caso de aumento do erro global do modelo sobre essa parcela do conjunto de dados de validação após uma divisão do pior neurônio, o algoritmo chega ao seu fim.

4.7 Conclusão

Neste capítulo foram apresentadas as metodologias que serão aplicadas neste trabalho ao problema de previsão de longo prazo de séries temporais e as medidas de erro de previsão escolhidas. As ferramentas matemáticas utilizadas serão: modelos auto-regressivos múltiplos (ARM), rede neural artificial (RNA) do tipo MLP e rede neuro-fuzzy linear local (RNFL). Para sua validação, definiu-se um preditor trivial que consiste na combinação aditiva de uma extrapolação polinomial de primeira ordem

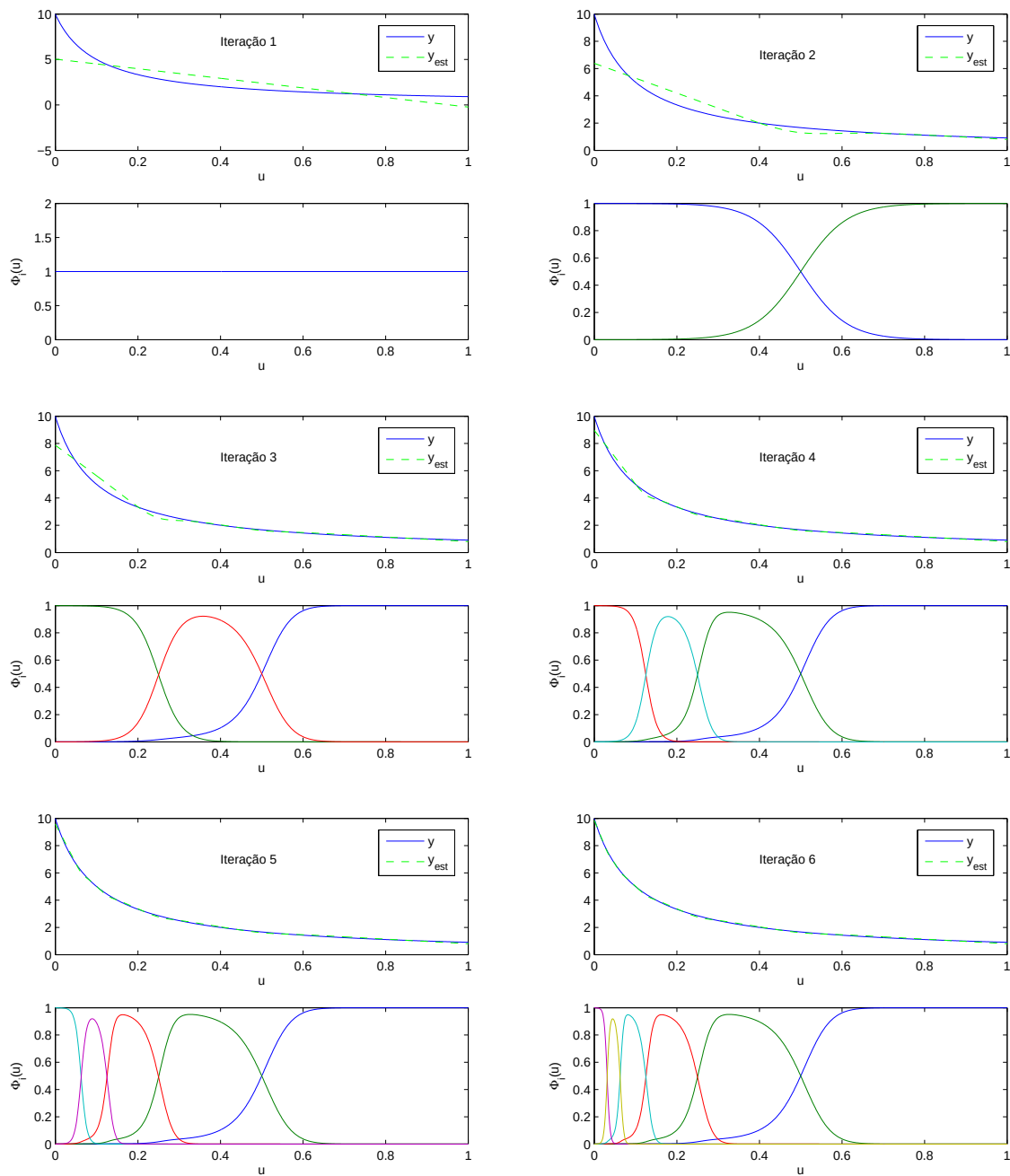


Figura 4.8: Funcionamento do algoritmo LOLIMOT para ajuste dos parâmetros da rede neuro-fuzzy linear local. Da esquerda para a direita, de cima para baixo, são exibidas as seis primeiras iterações do algoritmo para aproximar uma função não-linear estática.

da componente de tendência da série com o último ciclo sazonal do conjunto de dados de treinamento. As medidas definidas para aferir a qualidade dos modelos foram o MAPE, o MPE e o TPE.

Resultados de Previsão

“Os sábios mais supõem do que conhecem.”

Michel de Montaigne (1533-1592)

5.1 Introdução

Neste capítulo são apresentados os resultados obtidos com todos os modelos de previsão descritos no capítulo 4. Os modelos foram aplicados às séries analisadas no capítulo 3: (i) série de consumo de New England (horizonte de 5 anos); (ii) série de consumo de baixa tensão da CEMIG (horizonte de 2 anos); e (iii) séries de consumo total da CEMIG (horizonte de 2 anos). Para a série de consumo de New England, aplicaram-se os modelos ARM, a RNA e a RNFL; a série de consumo de baixa tensão da CEMIG foi modelada pela RNA e pela RNFL; e, por fim, a série de consumo total da CEMIG foi modelada somente pelos modelos ARM.

5.2 Preditor trivial

É importante ressaltar, mais uma vez, que o preditor trivial será usado como referência para todas as três séries, a fim de se quantificar a eficiência dos outros modelos mais sofisticados que foram aplicados ao problema abordado nesta dissertação.

5.2.1 Série de Consumo Mensal de New England

O preditor trivial foi utilizado para prever o consumo de energia nos anos de 1994 a 1998 a partir da série histórica de consumo de New England. Para tanto, ajustou-se uma reta aos dados mensais dos anos de 1991 a 1993 via interpolação polinomial e a essa reta adicionou-se a oscilação sazonal do ano de 1993. A Figura 5.1 e a Tabela 5.1 revelam os resultados encontrados.

Entre 1991 e 1993 praticamente não há mudança de patamar na série, o que implica uma reta quase horizontal para aproximar o seu comportamento em termos de tendência. Observa-se, todavia, uma pequena elevação na tendência nos anos subsequentes, o que explica o fato de os valores de MPE e TPE apresentados na Tabela 5.1 serem todos negativos e crescentes em módulo com o passar dos anos. Em outras palavras, a

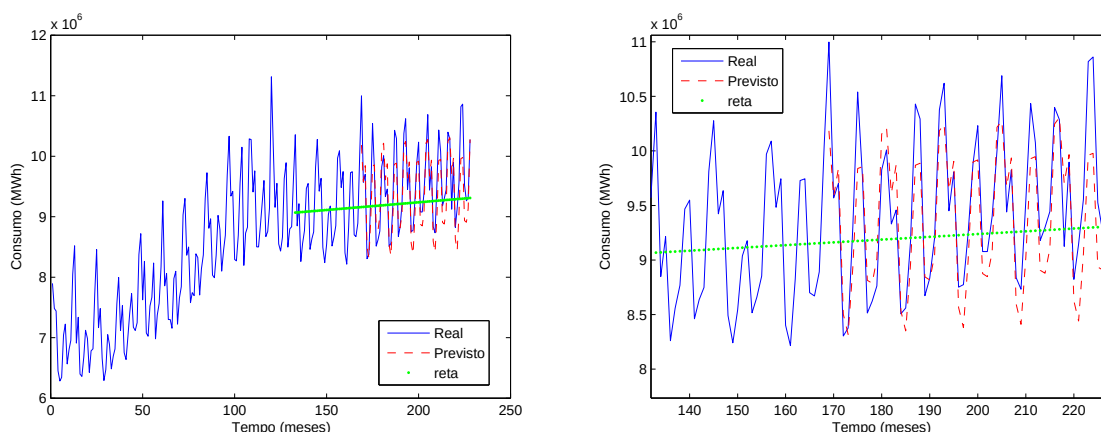


Figura 5.1: Resultados da previsão de consumo em New England de 1994 a 1998 utilizando-se o preditor trivial.

Tabela 5.1: Erros da previsão de consumo em New England de 1994 a 1998 utilizando-se o preditor trivial.

Ano	MAPE	MPE	TPE
1994	2,95%	-0,18%	-0,40%
1995	2,47%	-0,50%	-0,58%
1996	2,46%	-1,57%	-1,54%
1997	3,17%	-2,58%	-2,57%
1998	4,20%	-2,99%	-3,08%

não-estacionariedade de média da série, aliada ao fato de que a reta estimada refere-se a uma janela de crescimento menor, faz com que os valores de consumo previstos no horizonte desejado sejam subestimados.

5.2.2 Série de Consumo Mensal de Baixa Tensão da CEMIG

Os dados de consumo de baixa tensão da Cemig se estendem de Dezembro de 1996 a Setembro de 2007. Unicamente para tratar dados que fechem o ciclo anual, truncou-se a série a partir de Dezembro de 2006. Dessa forma, ajustou-se uma reta aos dados de 2003 e 2004 e adicionou-se a ela o ciclo sazonal referente ao ano de 2004. Essa composição é justamente a previsão trivial, a qual pode ser observada na Figura 5.2. A Tabela 5.2 sintetiza os resultados obtidos com as medidas de erro de previsão.

Tabela 5.2: Erros da previsão de consumo de baixa tensão da CEMIG de 2005 e 2006 utilizando-se o preditor trivial.

Ano	MAPE	MPE	TPE
2005	2,26%	0,07%	0,03%
2006	1,72%	0,33%	0,30%

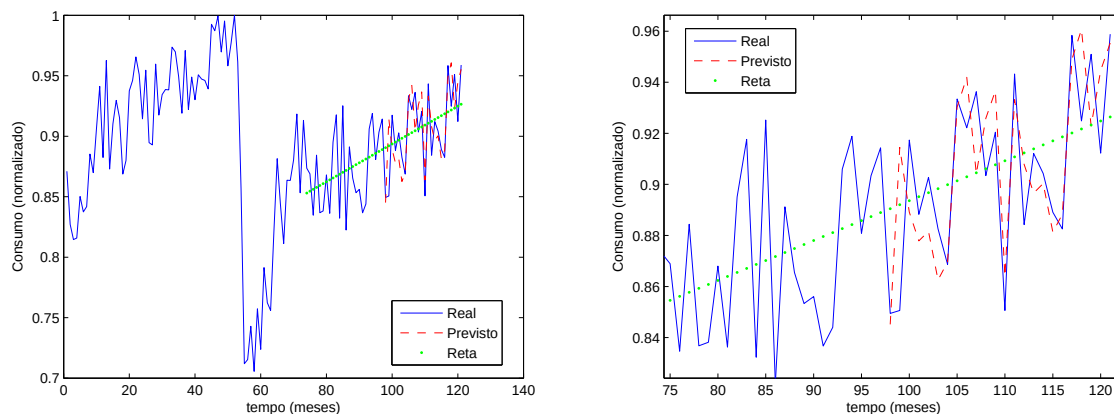


Figura 5.2: Resultados da previsão de consumo de baixa tensão da CEMIG de 2005 e 2006 utilizando-se o preditor trivial.

Não é difícil perceber que o ciclo dessa série é irregular, mas ainda assim o ajuste se mostrou de grande acurácia devido ao fato de o crescimento da série ser praticamente linear a partir de 2003.

5.2.3 Série de Consumo Mensal Total da CEMIG

A série de consumo total da CEMIG corresponde a um período que varia de janeiro de 2000 a dezembro de 2006. Exatamente o mesmo procedimento realizado para os dados de baixa tensão foi repetido com esses dados, e obtiveram-se os resultados que podem ser visualizados na Figura 5.3 e na Tabela 5.3

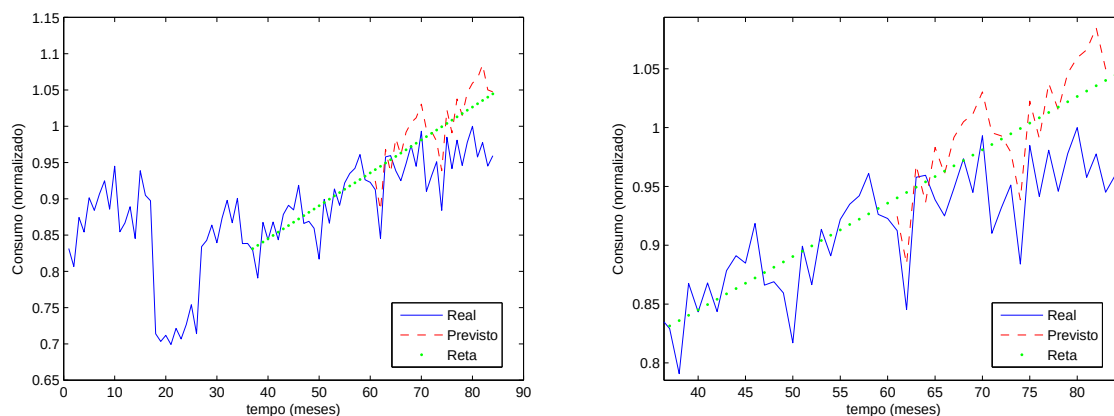


Figura 5.3: Resultados da previsão de consumo total da CEMIG de 2005 e 2006 utilizando-se o preditor trivial.

É válido notar que existe uma pequena redução na taxa de crescimento médio a partir de 2004 (ver Figura 3.16), o que explica o fato de a previsão ter divergido significativamente dos valores reais de consumo nos anos de 2005 e 2006.

Tabela 5.3: Erros da previsão de consumo total da CEMIG de 2005 e 2006 utilizando-se o preditor trivial.

Ano	MAPE	MPE	TPE
2005	4,38%	3,98%	3,95%
2006	7,23%	7,23%	7,23%

5.3 Modelos Auto-Regressivos Múltiplos

5.3.1 Série de Consumo Horário de New England

Para se verificar a capacidade preditiva do modelo obtido pelo algoritmo descrito na Seção 4.4, foram realizados alguns testes preliminares com a base de dados da companhia elétrica de New England, uma vez que o histórico de dados da CEMIG é muito pequeno (em especial por causa do racionamento, que impossibilita a utilização de dados anteriores ao ano de 2002). Por outro lado, o histórico de dados de New England se estende de 1980 a 1998.

Primeiramente, foi realizado todo o procedimento descrito para montar-se a matriz do ano de 1993 e a partir dela foi feita uma previsão para o ano de 1994. A Figura 5.4 ilustra a concatenação de todas as linhas da matriz de 1994 prevista, bem como a concatenação da matriz de 1994 real. Portanto, cada um dos 52 ciclos representa o perfil médio de consumo semanal ao longo das 24 horas do dia. O gráfico à direita, na referida figura, permite uma melhor visualização do gráfico à esquerda, apresentado o perfil de consumo médio nas primeiras semanas de 1994. A Tabela 5.4 apresenta as medidas de erro obtidas com a previsão.

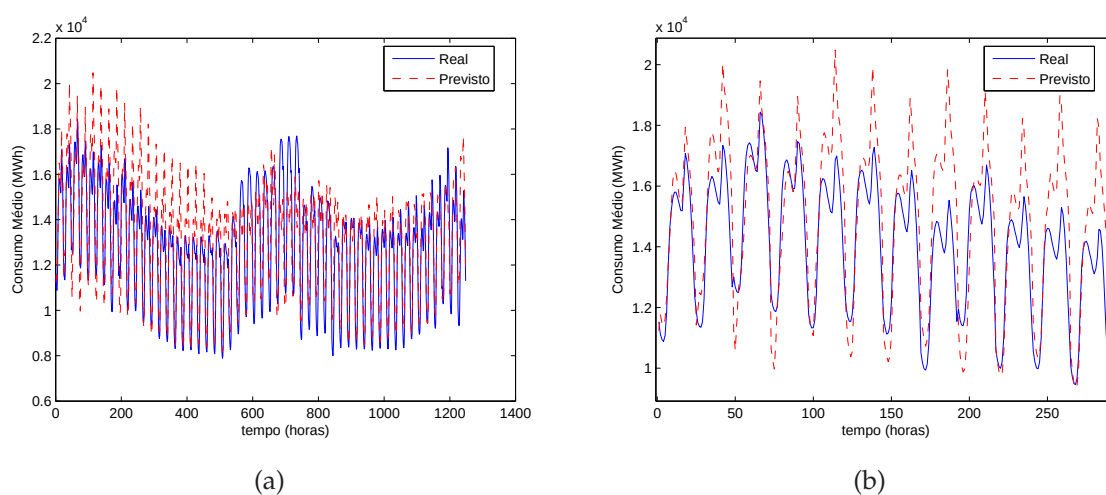


Figura 5.4: Resultados da previsão de 1994 a partir da matriz de 1993. (a) Todo o ano de 1994 e (b) as primeiras semanas de 1994.

Realizando-se o mesmo procedimento, porém sem a execução do ajuste de correção anual, obtêm-se os resultados apresentadas na Figura 5.5 e na Tabela 5.5, que se mostram melhores que os primeiros resultados atingidos considerando-se a correção anual.

Tabela 5.4: Erros da previsão de consumo de 1994 a partir do consumo de 1993

Ano	MAPE	MPE	TPE
1994	6,48%	4,22%	4,69%

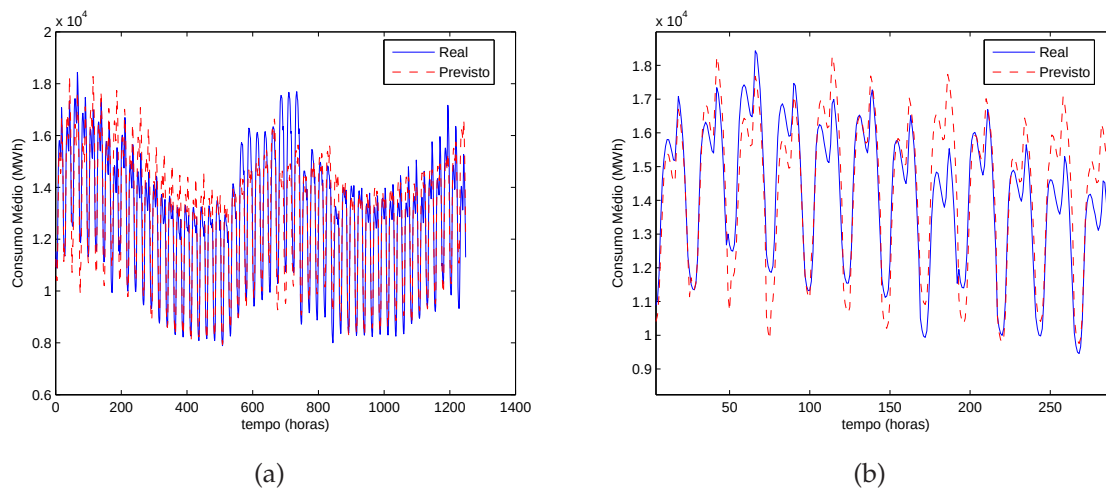


Figura 5.5: Resultados da previsão de 1994 a partir da matriz de 1993 sem o ajuste de correção anual. (a) Todo o ano de 1994 e (b) as primeiras semanas de 1994.

Uma característica evidente desse algoritmo é a sua simplicidade, todavia o resultado do preditor trivial, que é ainda mais simples, se mostrou superior para um horizonte de um ano em termos de redução do erro de previsão (comparar Tabelas 5.1 e 5.5). Não seria possível, portanto, melhorar o desempenho dos modelos ARM, ou seja, diminuir as medidas de erro atingidas? Por exemplo, da maneira como o algoritmo está proposto, são realizadas previsões de linhas e colunas, alternadamente (devido à correlação existente entre as linhas e colunas consecutivas de uma mesma matriz), até que seja preenchida por completo a matriz do ano para o qual se deseja estimar o consumo de energia. Observando-se a Figura 4.3, mais especificamente o segundo gráfico, nota-se que o comportamento da primeira hora ao longo de um ano não guarda correlação tão forte com a mesma hora do ano seguinte. Portanto, uma maneira diferente de se executar o algoritmo seria varrer a matriz unicamente na direção vertical, ou seja, realizar previsão apenas de linhas, já que elas guardam mais forte correlação entre si do que o fazem as colunas entre si. Nesse caso, seriam construídos então, para um mesmo ano, 52 modelos da forma representada na Equação 4.8. Surge aqui, entretanto, uma outra questão: pode ocorrer uma maior propagação de erros entre modelos nesse caso, já que em sua forma original o algoritmo gera apenas 24 modelos de linhas. Para contornar isso, o que se pode fazer é, ao invés de se montar uma matriz como a da Equação 4.7, propor-se uma nova matriz M , no seguinte formato,

Tabela 5.5: Erros da previsão de consumo de 1994 a partir do consumo de 1993 sem o ajuste de correção anual

Ano	MAPE	MPE	EP
1994	4,57%	1,68%	1,36%

$$M(i,k) = \begin{bmatrix} m_{1,1} & m_{1,2} & \dots & m_{1,24} \\ m_{2,1} & m_{2,2} & \dots & m_{2,24} \\ \vdots & \vdots & \ddots & \vdots \\ m_{12,1} & m_{12,2} & \dots & m_{12,24} \end{bmatrix} \quad (5.1)$$

em que o elemento $m_{i,k}$ é o consumo, em média, da k -ésima hora do dia ao longo do mês i .

Portanto, trabalha-se agora com 12 modelos do tipo

$$M(i,k) = a(i)L(i-1,k) + b(i), \quad k = 1,2,\dots,24 \quad (5.2)$$

em que $M(i,k), k = 1,2,\dots,24$ representa a i -ésima linha da matriz M , e, logicamente, $M(i-1,k), k = 1,2,\dots,24$ representa a $(i-1)$ -ésima linha. Fazendo-se $i = 1,2,\dots,12$ na Eq. (5.2), tem-se os doze modelos ARM. Para a estimação dos parâmetros $a(i)$ e $b(i)$ de cada um dos modelos, constrói-se a matriz de regressores¹ $\Psi \in \mathbb{R}^{24 \times 2}$ e aplica-se o MQ a partir da expressão:

$$\begin{bmatrix} M_{1993}(i,1) \\ M_{1993}(i,2) \\ \vdots \\ M_{1993}(i,24) \end{bmatrix} = \Psi \times \begin{bmatrix} a(i) \\ b(i) \end{bmatrix} = \begin{bmatrix} M_{1993}(i-1,1) & 1 \\ M_{1993}(i-1,2) & 1 \\ \vdots & \vdots \\ M_{1993}(i-1,24) & 1 \end{bmatrix} \begin{bmatrix} a(i) \\ b(i) \end{bmatrix}, \quad i = 1,2,\dots,12. \quad (5.3)$$

Com essas modificações do algoritmo proposto por Al-Hamadi and Soliman (2005), foram utilizados os mesmos dados de 1993 para se construir uma matriz M_{1993} e, em seguida, relizar-se a previsão da matriz M_{1994} . Os resultados encontrados estão apresentados na Figura 5.6 e na Tabela 5.6. O gráfico à esquerda ilustra o consumo médio em cada mês de 1994 (são doze ciclos) ao longo das 24 horas de um dia, enquanto que o gráfico à direita apresenta os valores de consumo real e previsto em cada mês do ano de 1994.

Tabela 5.6: Erros da previsão de consumo de 1994 a partir do consumo de 1993 com o algoritmo modificado

Ano	MAPE	MPE	EP
1994	2,76%	-0,72%	-0,93%

¹Diferentemente da proposta original do algoritmo, as dimensões da matriz de regressores não se alteram conforme se varre a matriz $M(i,k)$.

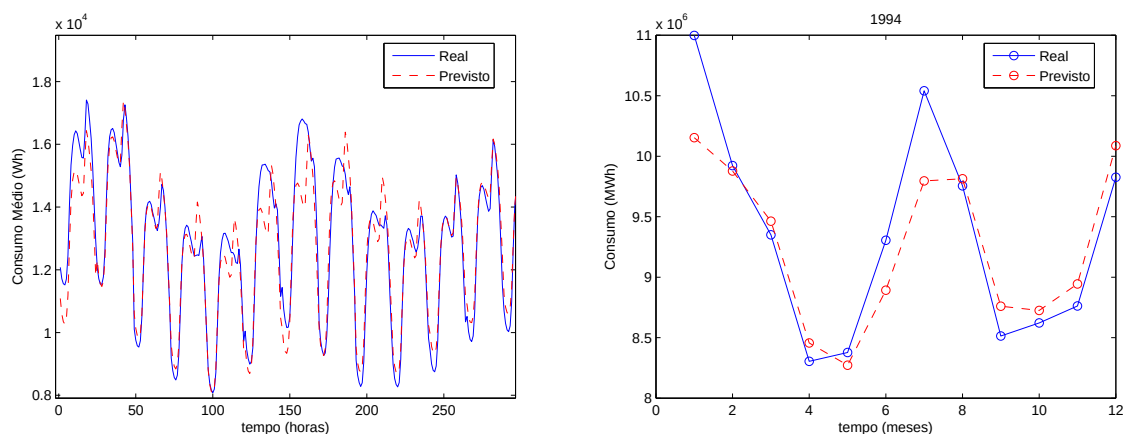


Figura 5.6: Resultados da previsão de 1994 a partir da matriz de 1993 com o algoritmo modificado. No gráfico à direita estão as séries real e prevista de consumo mensal.

Realmente, para esse conjunto de dados, os resultados do algoritmo modificado são melhores que os obtidos com o algoritmo original. Um ponto interessante a ser observado no gráfico à esquerda, na Figura 5.6, é que o formato de todos os ciclos previstos é o mesmo, com maior ou menor amplitude ou deslocados de um certo valor de *offset*. Isso se revela de fácil compreensão quando se observa a Equação 4.8, a qual diz que o valor da próxima linha prevista é o valor da anterior multiplicado por uma constante e deslocado de um termo também constante. Portanto, todos os ciclos terão a mesma forma que o primeiro previsto. Não obstante, isso não implica resultados de previsão ruins, uma vez que a sucessão dos ciclos previstos consegue acompanhar bem os ciclos reais. Para modificar-se o formato dos ciclos, uma alternativa seria o acréscimo de “regressores” de mais alta ordem ao modelo da Equação 4.8.

Verificado que o algoritmo consegue bons resultados de previsão com um horizonte de um ano, é válido que se verifique a possibilidade de expansão desse horizonte. De que forma isso seria feito? Basta utilizar a matriz M_{1994} estimada para se prever a matriz M_{1995} ; utilizar a matriz M_{1995} estimada para se prever a matriz M_{1996} , e assim por diante, até que se atinja o horizonte de previsão que se queira alcançar. Realizando-se esse procedimento recursivo, adicionando-se às previsões a mesma reta estimada no caso do preditor trivial com o objetivo de tentar rastrear a tendência de crescimento, obtêm-se os resultados mostrados na Figura 5.7 e na Tabela 5.7. A partir dos dados de 1993, portanto, foi feita a previsão para os anos de 1994 a 1998 (horizonte de previsão de 1 a 5 anos).

Em comparação com os resultados do preditor trivial (ver Tabela 5.1), vê-se que em termos de MAPE só o valor relativo ao horizonte de previsão de um ano foi maior para os modelos ARM, o que implica uma maior proximidade entre as curvas real e estimada, enquanto que o MPE e o TPE sempre foram inferiores a 1 em módulo, ou seja, o comportamento da média da série foi aproximado com erros positivos para 1994, 1995 e 1996 e erros negativos para 1997 e 1998, porém muito próximos de zero.

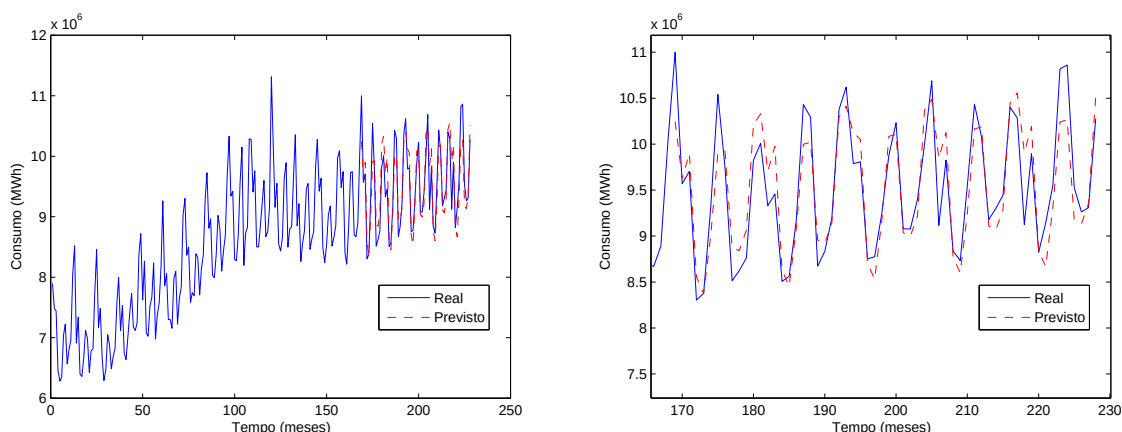


Figura 5.7: Resultados da previsão de 1994 a 1998 a partir da matriz de 1993 com o algoritmo modificado.

Tabela 5.7: Erros da previsão de consumo de 1994 a 1998 a partir dos dados de consumo de 1993 com o algoritmo modificado.

Ano	MAPE	MPE	EP
1994	3,15%	0,45%	0,22%
1995	2,41%	0,67%	0,60%
1996	1,87%	0,09%	0,13%
1997	2,29%	-0,18%	-0,21%
1998	3,39%	-0,54%	-0,62%

5.3.2 Série de Consumo Mensal Total da CEMIG

Montando-se uma matriz M_{2004} e executando-se o algoritmo para o horizonte de previsão de dois anos sem considerar a reta de ajuste de tendência, são encontrados para 2005 e 2006 os resultados apresentados na Figura 5.8 e na Tabela 5.8.

Tabela 5.8: Erros da previsão de consumo de 2005 e 2006 a partir dos dados de consumo de 2004 com o modelos AR múltiplos.

	MAPE	MPE	TPE
2005	3,38%	2,71%	2,72%
2006	6,08%	6,07%	6,12%

Em comparação com os resultados obtidos pelo preditor trivial (ver Tabela 5.3), os modelos ARM apresentaram melhor desempenho, a despeito da evidente discrepância entre as curvas apresentadas na Figura 5.8. Uma razão para tal fato pode ser proposta ao se observar os gráficos de tendência de consumo mensal total da CEMIG (construída utilizando-se as primeiras três componentes da série original obtidas via SSA) e de sua respectiva primeira diferença na Figura 5.9. Vê-se que no trecho em destaque de ambos os gráficos, correspondente ao consumo no ano de 2004 (dados de treinamento dos

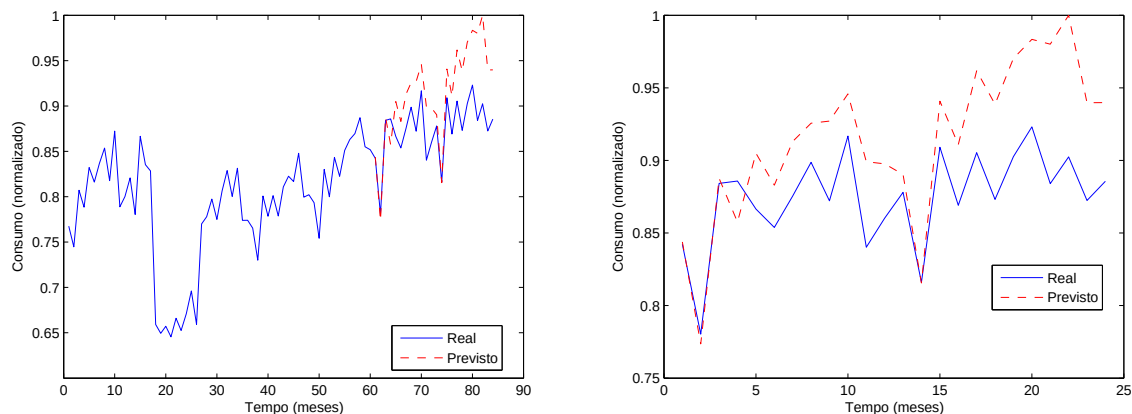


Figura 5.8: Resultados da previsão de 2005 e 2006 a partir da matriz de 2004 com o algoritmo modificado.

modelos ARM), está evidenciada a redução da taxa de crescimento da tendência. O que pode ter acontecido é que o modelo tentou extrapolar a tendência de crescimento da série de forma linear, todavia esse decrescimento na tendência fez com que houvesse uma superestimativa do consumo, implicando valores de MPE e TPE positivos e consequentemente um MAPE elevado também.

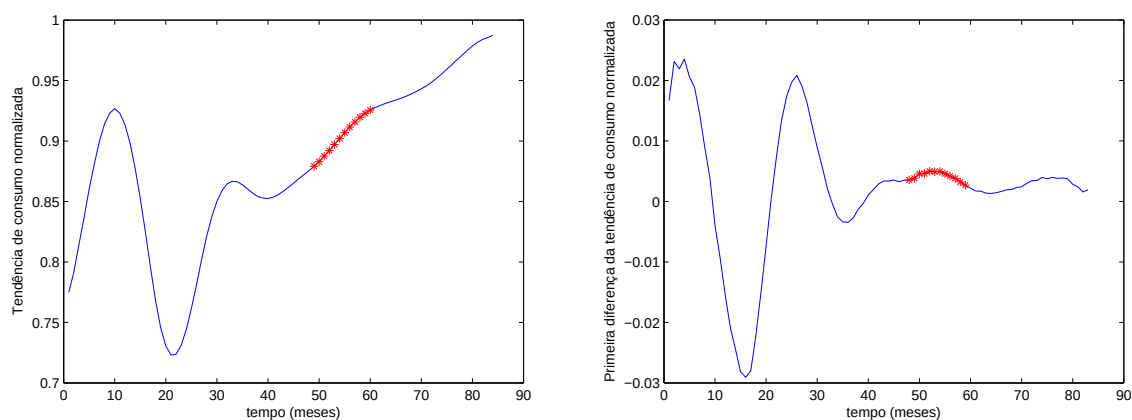


Figura 5.9: À esquerda, série de tendência de consumo mensal total da CEMIG; à direita, a primeira diferença da série de tendência de consumo mensal total da CEMIG. O trecho em destaque corresponde ao ano de 2004.

5.4 Rede Neural Artificial MLP

5.4.1 Série de Consumo Mensal de New England

O primeiro procedimento que se executou nesta etapa foi a divisão do conjunto de dados de consumo mensal de New England, constituído de 228 amostras, em três sub-conjuntos:

1. Dados de treinamento: $y_t = y(k)$, $k = 1, \dots, 156$ (Janeiro de 1980 a Dezembro de 1992);
2. Dados de seleção: $y_s = y(k)$, $k = 157, \dots, 168$ (Janeiro de 1993 a Dezembro de 1993);
3. Dados de validação: $y_v = y(k)$, $k = 169, \dots, 228$ (Janeiro de 1994 a Dezembro de 1998);

Para a etapa de ajuste dos pesos da rede, foram utilizados os dados de treinamento. Na etapa de escolha da melhor rede, variou-se o número de neurônios na camada escondida de 2 a 5 e para cada uma dessas quatro diferentes topologias foram realizados vinte treinamentos, com o objetivo de se contornar o problema da existência de mínimos locais no processo de ajuste dos pesos da rede, e utilizou-se cada uma das oitenta redes assim obtidas para prever o consumo no intervalo de tempo correspondente ao do conjunto de dados de seleção, isto é, o consumo no ano de 1993.

Uma das questões mais difíceis de se resolver em identificação de sistemas dinâmicos é a seleção de estrutura do modelo, o que no presente caso se restringe à escolha das variáveis regressoras. Em (Mauro, 2007) e (Campos, 2008), essa questão é discutida e os autores optam por considerar como regressores no modelo a série original com atrasos de 1, 2, 3, 4 e 12. Ao aplicar o algoritmo de taxa de redução de erro (ERR) (Billings and Chen, 1989; Aguirre, 2007), observamos que as cinco variáveis regressoras escolhidas foram 12, 1, 5, 3, 11 e 7, nesta ordem. Uma outra forma de selecionar estrutura do modelo pode ser realizada analisando-se o módulo da transformada de Fourier da série original subtraída da sua tendência estimada via SSA, o qual pode ser visualizado na Figura 5.10. Observa-se que os dois maiores picos presentes no sinal são correspondentes aos períodos de 6 e 12 meses.

Cabe aqui uma discussão acerca da problemática que envolve a questão da previsão dinâmica e do mapeamento de padrões (Aguirre et al., 2008). A observação do módulo da transformada de Fourier da série de New England (Figura 5.10) revela que a existência de picos significativos em frequências bem definidas pode ser um indicativo de que um mero mapeamento de padrões, isto, a projeção da “forma de onda” para o futuro, seja suficiente para se ter uma boa previsão do comportamento do sistema. Em outras palavras, considerando-se que a série sem tendência apresenta um comportamento predominantemente cíclico, a repetição deste(s) ciclo(s) combinada com a tendência retirada já se constituiria em uma boa estimativa do comportamento futuro do sistema. Isso será verificado ao se comparar os resultados obtidos pelo preditor trivial e pelas outros modelos, inclusive pela rede MLP.

Para validar a melhor rede selecionada, o que se fez foi utilizá-la para prever o consumo entre os meses 169 e 228 (definindo-se assim um horizonte de previsão de 60 meses ou 5 anos) e então comparar os resultados obtidos com o conjunto de dados de validação. De todas as configurações de estrutura de modelos dinâmicos apresentadas no parágrafo anterior, e de algumas variações delas, a que apresentou menores medidas

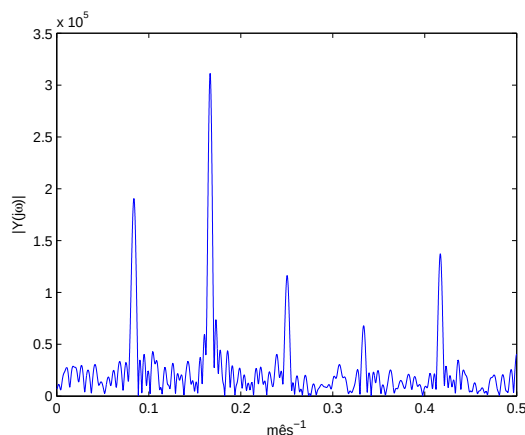


Figura 5.10: Módulo da transformada de Fourier da série de consumo mensal de New England.

de erro de previsão foi aquela formada apenas pelas séries defasadas $y(k-6)$ e $y(k-12)^2$. Os resultados apresentados na Figura 5.11 e na Tabela 5.9 evidenciam o desempenho obtido com a melhor rede, a qual teve a sua camada escondida composta por apenas dois neurônios.

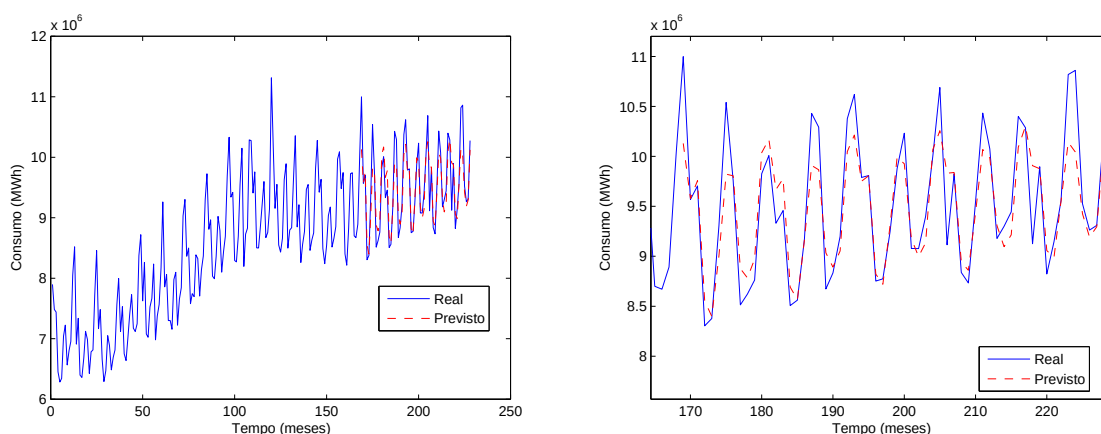


Figura 5.11: Resultados da previsão de 1994 a 1998 utilizando rede neural artificial MLP.

A comparação desses resultados com os resultados do preditor trivial apresentados na Tabela 5.1 demonstram que a rede atingiu um melhor desempenho. É imperioso notar que os valores de MAPE ficaram todos abaixo de 3%.

5.4.2 Série de Consumo Mensal de Baixa Tensão da CEMIG

Um procedimento equivalente ao da sub-seção anterior foi aqui repetido para se realizar as previsões da série de consumo mensal de baixa tensão da CEMIG. No

²É importante salientar que a consideração de poucas variáveis regressoras no modelo costuma evitar o tão conhecido problema de sobreparametrização.

Tabela 5.9: Erros da previsão de consumo de 1994 a 1998 a utilizando rede MLP (regressores 6 e 12).

Ano	MAPE	MPE	TPE
1994	2,75%	-0,24%	-0,50%
1995	2,48%	0,14%	0,00%
1996	1,33%	-0,60%	-0,65%
1997	2,43%	-0,43%	-0,56%
1998	2,55%	-0,65%	-0,84%

presente caso, temos os seguintes sub-conjuntos de dados:

1. Dados de treinamento: $y_t = y(k)$, $k = 1, \dots, 85$ (Dezembro de 1996 a Dezembro de 2003);
2. Dados de seleção: $y_s = y(k)$, $k = 86, \dots, 97$ (Janeiro de 2004 a Dezembro de 2004);
3. Dados de validação: $y_v = y(k)$, $k = 98, \dots, 121$ (Janeiro de 2005 a Dezembro de 2006);

Observa-se pela comparação entre as Figuras 5.10 e 5.12 (obtido com a série subtraída de sua tendência estimada pelo SSA) que o espectro da série da CEMIG é mais rico que o da série de New England, fato esse já evidenciado anteriormente no capítulo 3 quando se identificou que a primeira série era visualmente muito mais irregular que a segunda. Não obstante esse fato, o espectro da série CEMIG revela a presença de alguns picos de frequência, dentre eles os correspondentes aos períodos de 4, 6, 12 meses. Também foi identificado um pico correspondente a um período de 18,5 meses, o qual foi desconsiderado na determinação da estrutura do modelo. Portanto, definiram-se $y(k-4)$, $y(k-6)$ e $y(k-12)$ como variáveis regressoras³.

Seguindo-se os mesmos passos já expostos na seção anterior, seleciona-se uma rede de 3 neurônios na camada escondida que apresenta o desempenho representado graficamente na Figura 5.13 e numericamente na Tabela 5.10.

Tabela 5.10: Erros da previsão de consumo de 2005 a 2006 a utilizando rede MLP (regressores 4, 6 e 12).

Ano	MAPE	MPE	TPE
2005	2,09%	-0,13%	-0,21%
2006	2,40%	-0,33%	-0,42%

Vê-se que os resultados obtidos com o preditor trivial (ver Tabela 5.2) são superiores aos da rede neural. Para tentar explicar o fato de se atingir um ajuste tão bom com o

³A mesma comparação feita com a estrutura dada pelo *ERR* e a estrutura proposta por (Mauro, 2007) e (Campos, 2008) no caso dos dados de New England foi realizada aqui, e elegeram-se em definitivo a obtida com a análise do espectro de frequência da série da CEMIG uma vez que ela conduziu aos melhores resultados em termos de minimização das medidas de erro.

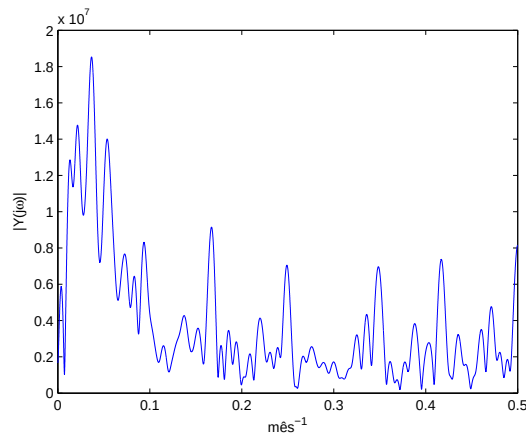


Figura 5.12: Módulo da transformada de Fourier da série de consumo mensal de baixa tensão da CEMIG.

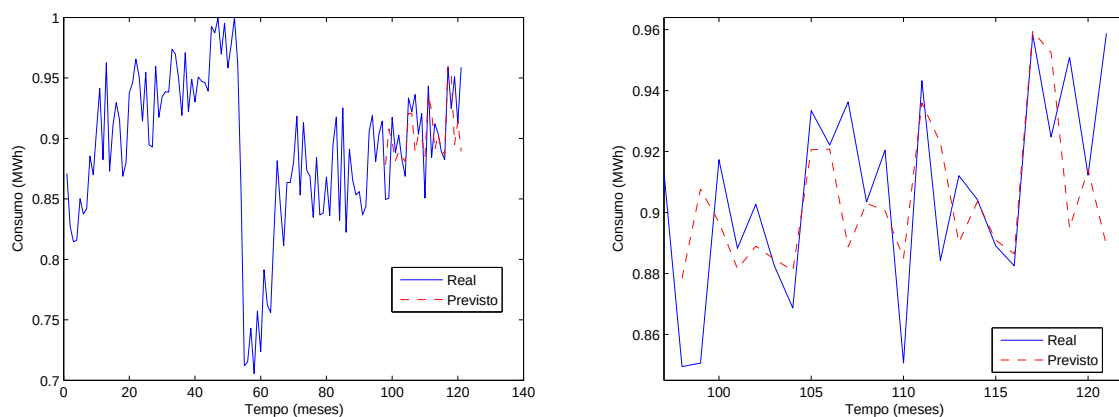


Figura 5.13: Resultados da previsão de 2005 e 2006 utilizando rede neural artificial MLP.

preditor trivial, convém observar a Figura 5.14, a qual apresenta a série de tendência dos dados de consumo mensal de baixa tensão da CEMIG e sua respectiva série de primeiras diferenças. O trecho em destaque corresponde aos anos 2004, 2005 e 2006. Como a tendência é linear, o crescimento da tendência nesse trecho é praticamente constante (da ordem de milésimos de unidades de consumo normalizado por mês), e como o ajuste do preditor trivial consiste na combinação aditiva de uma reta e um ciclo, atingiu-se com ele um resultado muito bom em termos de redução das medidas de erro, o que no presente caso o fez superar uma rede neural MLP.

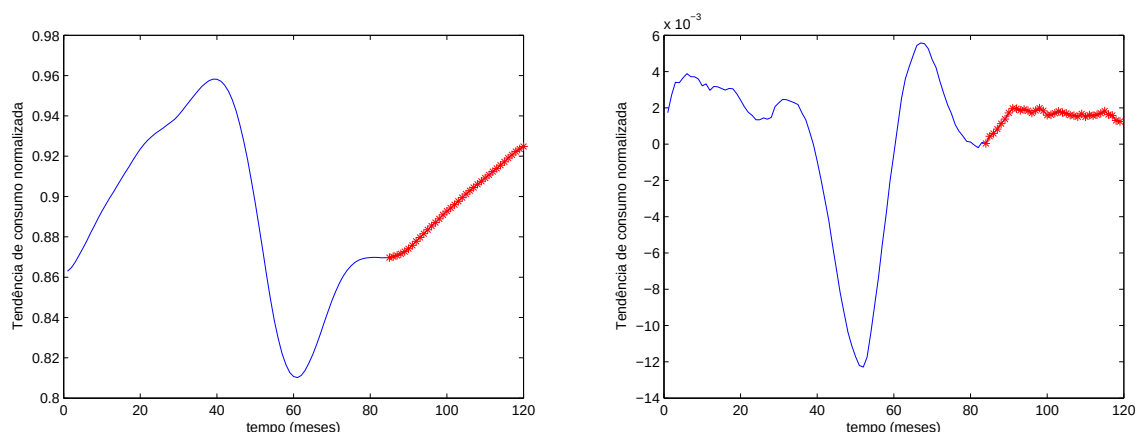


Figura 5.14: À esquerda, série de tendência de consumo mensal total de baixa tensão da CEMIG; à direita, a primeira diferença da série de tendência de consumo mensal total de baixa tensão da CEMIG. O trecho em destaque corresponde aos anos de 2004, 2005 e 2006.

5.5 Rede Neuro-Fuzzy Linear Local

5.5.1 Série de Consumo Mensal de New England

Para a modelagem da série de New England utilizando a rede neuro-fuzzy linear local, os dados foram divididos em dois sub-conjuntos:

1. Dados de treinamento: $y_t = y(k)$, $k = 1, \dots, 168$ (Janeiro de 1980 a Dezembro de 1993);
2. Dados de validação: $y_v = y(k)$, $k = 169, \dots, 228$ (Janeiro de 1994 a Dezembro de 1998);

Para o critério de parada no ajuste automático do número de neurônios da rede, utilizou-se o somatório dos erros quadráticos sobre os primeiros 12 valores do conjunto de validação. Para definir os regressores a serem usados na modelagem, foram feitas algumas simulações considerando-se as mesmas estruturas apresentadas no caso das redes MLP, inclusive a que foi identificada na análise do espectro de frequência da série original (ver Figura 5.10), porém constatou-se que o melhor resultado foi atingido com apenas um único regressor: $y(k - 12)$. Após o treinamento, chegou-se a uma topologia de rede com seis neurônios.

Conforme se verifica na Figura 5.15 e na Tabela 5.11, o desempenho da rede foi bastante satisfatório, tendo em vista que o MAPE extrapolou a faixa de 3% por muito pouco, e todas as medidas de MPE e TPE foram menores que 1% em módulo. É interessante observar, em especial no segundo gráfico da referida figura, que a tendência de crescimento do consumo foi absorvida pela rede, como se observa pelos vales da curva pontilhada, porém houve uma saturação nos picos, o que os situou todos em um mesmo patamar. Essa é uma limitação da rede, uma vez que os limites do espaço de entrada são definidos a partir da excursão do conjunto de dados de treinamento. Em

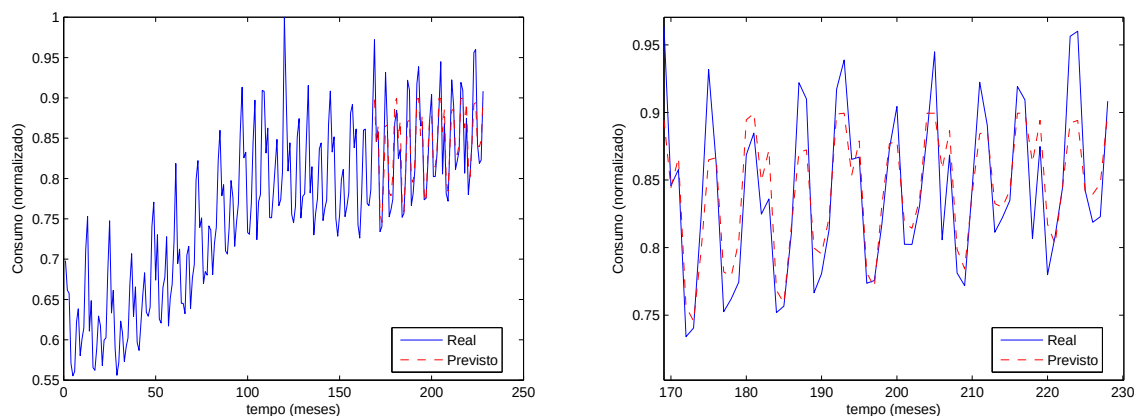


Figura 5.15: Resultados da previsão de 1994 a 1998 utilizando a rede neuro-fuzzy.

Tabela 5.11: Erros da previsão de consumo de 1994 a 1998 a utilizando a rede neuro-fuzzy (regressor 12).

Ano	MAPE	MPE	TPE
1994	3,02%	0,01%	-0,25%
1995	2,59%	0,63%	0,48%
1996	1,60%	0,10%	0,03%
1997	2,41%	0,41%	0,25%
1998	3,01%	0,26%	0,03%

outras palavras, séries não-estacionárias na média levam a rede a operar durante a etapa de validação em regiões nas quais ela não foi treinada. Não obstante isso, os resultados atingidos foram muito bons em termos numéricos, como já exposto acima. Um teste interessante a se fazer seria inserir não-estacionariedade de média na série e observar se a saída da rede na etapa de validação consegue varrer essas “regiões artificiais” geradas arbitrariamente. Somando-se, por exemplo, um *offset* de 0,4 às primeiras 84 amostras da série de New England, e acrescentando-se ao modelo o regressor $y(k-6)$, consegue-se uma rede de 4 neurônios que apresenta o desempenho expresso nas Figuras 5.16 e na Tabela 5.12. É interessante observar que a saturação que se via em alguns picos na Figura 5.15 não se repete nesse caso.

Tabela 5.12: Erros da previsão de consumo de 1994 a 1998 a utilizando a rede neuro-fuzzy com inserção arbitrária de não-estacionariedade na série (regressores 6 e 12).

Ano	MAPE	MPE	TPE
1994	2,72%	-0,17%	-0,41%
1995	2,48%	0,35%	0,23%
1996	1,32%	-0,19%	-0,20%
1997	2,15%	0,24%	0,15%
1998	2,55%	0,33%	0,20%

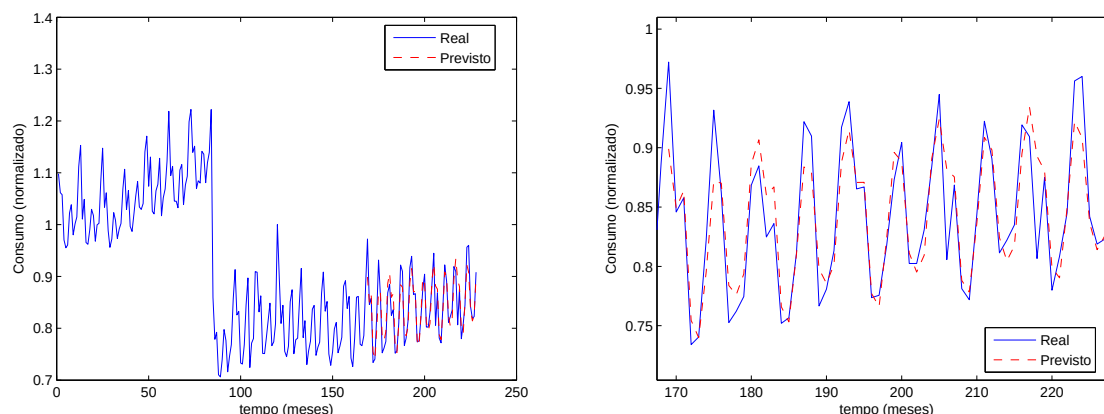


Figura 5.16: Resultados da previsão de 1994 a 1998 utilizando a rede neuro-fuzzy com inserção arbitrária de não-estacionariedade na série.

Ainda que os índices de desempenho atingidos tenham sido muito bons, é de bom alvitre frisar que o objetivo de se realizar a mudança arbitrária de não-estacionariedade da média em um trecho da série era unicamente o de se mostrar que é possível contornar o problema de saturação da saída da RNF, o que não garantiria em princípio uma melhoria dos índices de desempenho.

5.5.2 Série de Consumo Mensal de Baixa Tensão da CEMIG

Procedimento semelhante ao realizado na sub-seção anterior é aqui repetido para o caso da série de consumo da CEMIG. Em primeiro lugar, a subdivisão da série:

1. Dados de treinamento: $y_t = y(k)$, $k = 1, \dots, 97$ (Dezembro de 1996 a Dezembro de 2004);
2. Dados de validação: $y_v = y(k)$, $k = 98, \dots, 121$ (Janeiro de 2005 a Dezembro de 2006);

Em seguida, definindo-se novamente apenas o regressor $y(k-12)$ para a modelagem, tem-se como resultado final uma rede de 5 neurônios. Os resultados atingidos são apresentados na Figura 5.17 e na Tabela 5.13

Tabela 5.13: Erros da previsão de consumo de 2005 a 2006 utilizando a rede neuro-fuzzy (regressor 12).

Ano	MAPE	MPE	TPE
2005	2,16%	0,14%	0,07%
2006	2,26%	0,02%	-0,07%

Observa-se que, em comparação com os resultados da RNA e do preditor trivial (ver Tabelas 5.10 e 5.8), a rede neuro-fuzzy apresentou um desempenho levemente superior à primeira e um pouco inferior ao segundo.

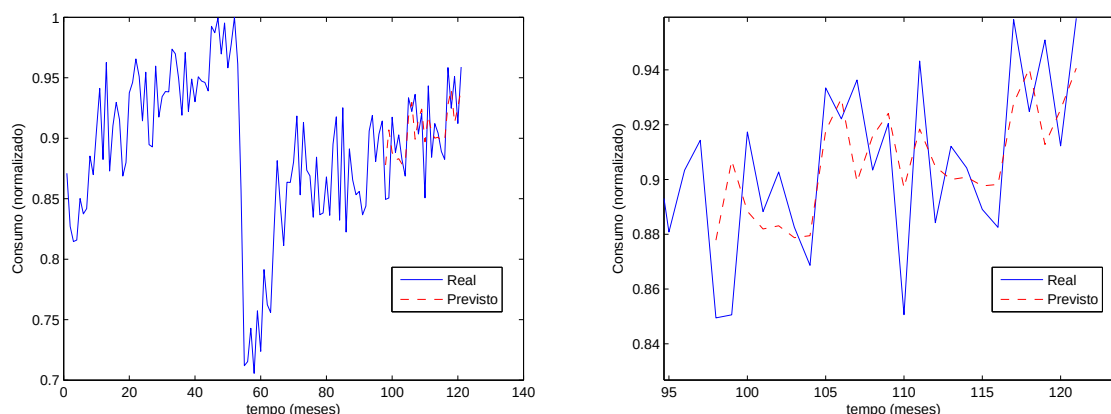


Figura 5.17: Resultados da previsão de 2005 e 2006 utilizando a rede neuro-fuzzy.

5.6 Síntese comparativa dos resultados de previsão

De posse de todos os resultados expostos ao longo deste capítulo, é possível realizar uma síntese calculando-se as médias de todos os três índices de erro ao longo de todo o período correspondente ao horizonte de previsão, isto é, de Janeiro de 1994 a Dezembro de 1998 no caso da série de New England e de janeiro de 2005 a dezembro de 2006 no caso das séries da CEMIG. Tal síntese pode ser visualizada na Tabela 5.14.

Tabela 5.14: Medidas de erro percentuais ao longo dos respectivos horizontes de previsão para cada um dos modelos e cada uma das séries.

	CEMIG Total			CEMIG BT			NE		
	MAPE	MPE	TPE	MAPE	MPE	TPE	MAPE	MPE	TPE
Trivial	5,80	5,60	5,60	1,99	0,20	0,16	3,05	-1,56	-1,63
ARM	4,73	4,39	4,42	-	-	-	2,62	0,10	0,02
RNA	-	-	-	2,24	-0,23	-0,31	2,31	-0,36	-0,51
RNFL	-	-	-	2,21	0,08	0,00	2,53	0,28	0,11
RNFLLE	-	-	-	-	-	-	2,24	0,11	-0,01

Os campos com um traço na Tabela 5.14 dizem respeito aos casos em que não foi realizada a previsão com aquele tipo de modelo para aquela determinada série. A sigla RNFLLE refere-se à RNFL estendida ao caso em que se inseriu mudança arbitrária de não-estacionariedade de média na série. Frisamos mais uma vez que o fato de se realizar tal procedimento não é garantia de que ocorra redução dos índices de erro, mas registramos assim mesmo os resultados encontrados, uma vez que, para o caso da série de New England (MAPE de 2,24%), eles foram equiparáveis e até superiores aos resultados apresentados pela RNA (MAPE de 2,31%) e pela RNFL (MAPE de 2,53%).

É interessante verificar que não existe o melhor modelo. Para cada situação, vê-se que um modelo se adequou mais do que outro em termos de redução dos índices de erro. No caso da série da Cemig de baixa tensão, por exemplo, o preditor trivial foi o que apresentou melhor desempenho em termos de MAPE, porém o MPE e o TPE da

RNFL é que foram os menores. Na modelagem da série de New England, a RNA apresentou o menor MAPE dentre todos os quatro modelos (exceção feita à RNFLLE)), porém os modelos ARM se revelaram os mais eficientes no que diz respeito ao MPE e o TPE, e nem por isso o resultado da RNFL para esse caso deixou de ser satisfatório, pois superou a RNA em MPE e TPE (em módulo) e os ARM em MAPE.

Cabe aqui uma comparação com os resultados obtidos por Mauro (2007) e Campos (2008), que desenvolveram outros dois trabalhos dentro do mesmo projeto de pesquisa no qual se insere o escopo do presente trabalho. Em (Mauro, 2007), as séries de consumo de baixa tensão da CEMIG foram modeladas por meio de redes neurais MLP e duas configurações conduziram o autor a resultados bastante interessantes: i) considerando-se a série de consumo total de baixa tensão, ele chegou a um MAPE de 2,02% e um MPE de -0,28% para uma rede com 10 neurônios na camada escondida e regressores $y(k-1)$, $y(k-2)$, $y(k-3)$, $y(k-4)$ e $y(k-12)$; ii) construindo um modelo neural para cada um dos oito setores diferentes (ou um modelo que é simplesmente uma reta no caso de o modelo neural não se ajustar bem), ele atingiu um MAPE de 1,78% e um MPE de -0,15%. O primeiro resultado de Mauro (2007) superou a rede neural menos complexa (dois neurônios na camada escondida e regressores $y(k-6)$ e $y(k-12)$) cujo resultado está apresentado na Tabela 5.14 em termos de MAPE, porém lhe foi inferior em termos de MPE. O preditor trivial aqui definido superou ambas as redes MLP, com um MAPE de 1,99% e um MPE de 0,20%. O segundo resultado de Mauro (2007) superou as duas RNAs e o preditor trivial, ao preço de ajustar um modelo diferente a cada uma das oito componentes do consumo total de baixa tensão da CEMIG. Isso ratifica o resultado apresentado nesse trabalho com a análise de dados sub-rogados, por meio da qual se identificou determinismo intra-ciclo e entre ciclos nas séries por setor, porém não se foi capaz de fazer o mesmo no caso da série de consumo de baixa tensão total da CEMIG.

Já em (Campos, 2008), foi modelada a série de consumo de New England por meio de quatro ferramentas diferentes: i) modelos ARIMA; ii) modelos NARIMA; iii) redes neurais MLP; iv) redes neuro-fuzzy. Para um horizonte de 60 meses (janeiro de 1993 a dezembro de 1998), o melhor resultado em termos de MAPE foi obtido com a rede neuro-fuzzy (2,67%) e o melhor resultado em termos de MPE foi alcançado com o modelo ARIMA (0,13%). Em termos de MAPE, os modelos ARM (2,62%), RNA (2,31%), RNFL (2,53%) e RNFLLE (2,24%) desenvolvidos neste trabalho foram superiores à rede neuro-fuzzy apresentada em (Campos, 2008) e o preditor trivial foi-lhe inferior (MAPE de 3,05%). Em termos de MPE, o preditor trivial (-1,56%), a RNA (-0,36%) e a RNFL (0,28%) ficaram abaixo do modelo ARIMA, o qual foi levemente superado pelos ARM (0,10%) e pela RNFLLE (0,11%).

No caso dos modelos ditos mapeadores universais de funções (a exemplo a RNA e a RNFL), uma comparação estatística entre os seus desempenhos poderia ser realizada para se verificar até que ponto os resultados obtidos com elas são realmente diferentes entre si. A Tabela 5.15 mostra em especial que, em termos de MAPE, tanto a RNA ($3,25\% \pm 0,61\%$) como a RNFL ($3,81\% \pm 0,39\%$) e a RNFLLE ($3,36\% \pm 0,40\%$) apresentaram desempenhos estatisticamente equivalentes para a série de New England, porém a RNA revelou-se superior no caso da série de consumo de baixa tensão da CEMIG ($2,34\% \pm 0,09\%$ contra $3,01\% \pm 0,33\%$ da RNFL). Para se definirem essas medidas de

Tabela 5.15: Média e desvio padrão das medidas de erro percentuais ao longo dos respectivos horizontes de previsão para os modelos RNA, RNFL e RNFLLE e para as séries de New england e de consumo de baixa tensão da CEMIG.

	CEMIG BT			NE		
	MAPE	MPE	TPE	MAPE	MPE	TPE
RNA	2,34±0,09	0,52±0,57	0,43±0,57	3,25±0,61	-0,70±0,61	-0,92±0,61
RNFL	3,01±0,33	-0,13±0,24	-0,24±0,24	3,81±0,39	0,22±0,22	-0,06±0,21
RNFLLE	-	-	-	3,36±0,40	-0,06±0,20	-0,29±0,20

erro médias juntamente com suas respectivas margens de erro de $\pm\sigma$, no caso da RNA o que se fez foi calcular a média e o desvio padrão amostrais das medidas de erro de previsão obtidas com as 20 melhores dentre as 80 redes treinadas (ver Capítulo 5); e no caso da RNFL e da RNFLLE o que se fez foi subtrair dos últimos 12 valores do conjunto de dados de treinamento os seus respectivos resíduos e, após embaralhá-los, readicioná-los ao conjunto de dados que inicializam o modelo na etapa de validação, de modo a se obter uma coleção de 100 previsões para cada série e a partir de tal coleção se realizar o cálculo da média e do desvio padrão amostrais das medidas de erro de previsão atingidas.

Conclusões

Técnicas de análise de dados sub-rogados recentemente propostas foram aplicadas às séries temporais de consumo de energia elétrica da CEMIG e de New England com o fim de se caracterizar a natureza da dinâmica das flutuações irregulares (variabilidade local ou intra-ciclo) inerentes a essas séries (SSS) e da dinâmica de ciclo a ciclo devido ao seu padrão sazonal (CSS). Verificou-se que nas séries mensais da CEMIG divididas por setor, a probabilidade de haver dinâmica intra-ciclo foi superior a 99% (maior do que 3σ) apenas para os setores ‘poder público’ e ‘rural’. No caso do consumo dos setores ‘industrial’ e ‘serviço público’, superou-se o limite de confiança de 95%. Em todos os outros setores, o cálculo de significância foi inferior a 2σ para os atrasos de até quatro amostras considerados no teste. Um outro aspecto digno de nota é que os resultados do teste de hipóteses utilizando os p -valores e a significância estão em concordância entre si para o caso dessa séries. Já nas séries da CEMIG agrupadas por hora, a significância chega a atingir 30 em termos de FAC e mais de 50 em termos de IMM, indicando claramente a existência de dinâmica determinística intra-ciclo. Para os dados de New England, o teste foi realizado apenas para a série mensal e a hipótese nula foi rejeitada porque a significância chegou a atingir 5σ e o menor p -valor esteve dentro da faixa indicativa de evidência muito forte contra H_0 para o caso de ambas as EDs consideradas. Na análise entre ciclos por meio da aplicação do CSS, não houve nenhum indicativo de rejeição da hipótese para a série de consumo mensal total da CEMIG. Já nas séries da CEMIG por setor, houve evidência forte contra H_0 para as categorias ‘iluminação pública’, ‘poder público’, ‘industrial’, ‘serviço público’ e ‘comercial’, definindo-se assim um indicativo de dependência temporal entre ciclos. No caso do setor ‘rural’ observou-se um acontecimento interessante: houve indicativo de dinâmica intra-ciclo porém não se detectou indícios de dinâmica entre ciclos. Já em se tratando da série de New England mensal, o algoritmo CSS acusou rejeição da hipótese nula tanto em termos de significância e p -valores para ambas as EDs.

Um aspecto interessante que se observou na análise de dados sub-rogados é que conforme a série se apresentasse mais irregular, mais difícil era que acontecesse a rejeição da hipótese, em especial na aplicação do SSS. Já no caso das série com padrão de periodicidade mais bem definido, como as séries agrupadas por hora da CEMIG e a série mensal de New England, a rejeição da hipótese acusando determinismo se tornava muito mais patente.

Ainda na etapa de análise de dados, foi mostrado que o SSA se revelou uma ferramenta eficiente para a decomposição de séries temporais, e em especial para caracterizar

e extrair a componente de tendência das séries. Por essa razão, a sua utilização foi fundamental na etapa de realização das previsões utilizando o chamado preditor trivial (o ciclo $s_1(\cdot)$ era calculado subtraindo-se do conjunto de dados de treinamento a componente de tendência estimada), no pré-processamento das séries para a aplicação do teste de dados sub-rogados e na escolha dos regressores para a RNA e a RNFL.

No que diz respeito aos modelos construídos, o preditor trivial só foi superior aos modelos ARM, à RNA e à RNFL na previsão da série de consumo de baixa tensão da CEMIG, e ainda assim a RNFL o superou em termos de MPE (0,00% contra 0,16%) e TPE (0,08% contra 0,20%), perdendo apenas no MAPE (2,21% contra 1,99%). Em todos os outros casos, as três estratégias de previsão estudadas nesse trabalho apresentaram melhor desempenho em termos de redução do MAPE, do MPE e do TPE. Foi demonstrado por comparação que os resultados obtidos neste trabalho foram equiparáveis aos apresentados em (Mauro, 2007) e (Campos, 2008), algumas vezes superando-os e outras sendo-lhe inferiores. Em especial pôde-se constatar que o resultado da análise de dados sub-rogados que indicou determinismo intra-ciclo e entre ciclos nas séries de consumo dos setores de baixa tensão da CEMIG foi referendado com o resultado de previsão atingido por Mauro (2007) ao propor um modelo diferente para cada um dos oito setores (MAPE = 1,78% e MPE = -0,15%). Uma última análise comparativa ainda revelou que os resultados obtidos com a RNA e a RNFL utilizadas neste trabalho foram estatisticamente equivalentes em termos de MAPE para a série de New England, enquanto que houve um indicativo de superioridade da RNA no caso da série de consumo de baixa tensão da CEMIG ($2,34\% \pm 0,09\%$ contra $3,01\% \pm 0,33\%$ da RNFL).

Pode-se dizer que duas questões ficaram claras nesta pesquisa: (i) a natureza das séries e a quantidade de amostras que lhes constituem são fatores que influenciam fortemente a qualidade das previsões obtidas e as possibilidades de estudo e pesquisa. O algoritmo PPS, por exemplo, capaz de detectar a existência de determinismo entre os ciclos de uma série temporal, deixou de ser aplicado nesse trabalho em virtude do reduzido número de amostras das séries que impossibilitou o cálculo da dimensão de correlação, que é a ED utilizada nesse algoritmo. Por essa restrição, portanto, aplicou-se o CSS no lugar do PPS; (ii) o fato de os modelos serem complexos não garante que eles sejam bons previsores. Veja-se o exemplo deste trabalho ao se estudar a série de New England, em que a rede MLP que melhor se adequou ao problema de previsão apresentava apenas dois neurônios na camada escondida e foi alimentada somente com os regressores $y(k-6)$ e $y(k-12)$, sem mencionar o uso do preditor trivial que no caso da série de baixa tensão da CEMIG se mostrou o modelo de melhor desempenho preditivo dentre todos os outros em termos de MAPE, justamente uma das séries para as quais não houve detecção de determinismo intra-ciclo e entre ciclos por meio da análise de dados sub-rogados.

Como perspectiva futura, existe a possibilidade de se realizar um estudo de forma a se tentar quantificar a incerteza preditiva inerente aos dados e aos modelos por meio de algum tipo de simulação de Monte Carlo, por exemplo, de modo a se ter um argumento técnico peremptório que diga se é possível satisfazer, de forma confiável e não acidental, a margem de erro de +3%, estabelecida pelo governo brasileiro, para um horizonte de previsão de cinco anos. Talvez a aplicação de ferramentas de *bootstrap*

(Efron and Tibshirani, 1986) conduza a resultados interessantes, como a determinação de limites de confiança para as previsões de consumo obtidas.

Referências Bibliográficas

- Aguirre, L. A. (2007). *Introdução à Identificação de Sistemas - Técnicas Lineares e Não-Lineares Aplicadas a Sitemas Reais*. Editora UFMG, Av. Antônio Carlos, 6627 - Ala direita da Biblioteca Central, térreo - Campus Pampulha - 31270-901 - Belo Horizonte/MG.
- Aguirre, L. A., Aguirre, A., Martinez, C. B., and Mendes, E. M. A. M. (2006). *Pesquisa de modelo de previsão de energia, para curto e médio prazo, utilizando inteligência computacional - Relatório inicial*.
- Aguirre, L. A., Rodrigues, D. D., Lima, S. T., and Martinez, C. B. (2008). Dynamical prediction and pattern mapping in short-term load forecasting. *Electrical Power Energy Systems*, 30:73–82.
- Al-Hamadi, H. M. and Soliman, S. (2005). Long-term/mid-term load forecasting based on short-term correlation and annual growth. *Electric Power Systems Research*, 74:353–361.
- Al-Hamadi, H. M. and Soliman, S. (2006). One year long-term electric load forecasting based on multiple regression models and Kalman filtering algorithm. *Engineering Intelligent Systems*, 2:79–88.
- Billings, S. A. and Chen, S. (1989). Identification of nonlinear rational systems using a prediction-error estimation algorithm. *International Journal of Systems Science*, 20:467–494.
- Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis - Forecasting and Control*. Holden Day, San Francisco.
- Braga, A. d. P., Carvalho, A. C. P. d. L. F., and Ludermir, T. B. (2000). *Redes Neurais Artificiais: Teoria e Aplicações*. Livros Técnicos e Científicos Editora S.A.
- Campos, R. J. (2008). *Previsão de séries temporais com aplicações a séries de consumo de energia elétrica*. Dissertação de Mestrado em Engenharia Elétrica, UFMG, Brasil.
- Carvalho, F. G. M. d. and Hemerly, E. M. (2008). Controle de sistemas dinâmicos não-lineares utilizando realimentação de saída e redes neurais não-linearmente parametrizadas. *Anais do XVII Congresso Brasileiro de Automática - Juiz de Fora (MG)*.

- Coelho, M. C. d. S. (2006). *Análise de Determinismo em Séries Temporais Quase-Periódicas*. Tese de Doutorado em Engenharia Elétrica, UFMG, Brasil.
- Costa, C. P. (2008). *Métodos de Dados Sub-rogados Aplicados a Séries Temporais*. Dissertação de Mestrado em Engenharia Elétrica, UFMG, Brasil.
- Cullen, K. A. (1999). *Forecasting Electricity Demand Using Regression and Monte Carlo Simulation Under Conditions of Insufficient Data*. Master's Thesis, West Virginia University, USA.
- Curteanu, S. and Leon, F. (2007). Optimization strategy based on genetic algorithms and neural networks applied to a polymerization process. *International Journal of Quantum Chemistry*, 108:617–630.
- Decreto (2004). Decreto 5.163. Diário Oficial da União, República Federativa do Brasil, 30 de julho de 2004.
- Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1:54–77.
- Frizlewics, P., Van Belleghem, S. V., and von Sachs, R. (2003). Forecasting non-stationary time series by wavelet process modelling. *Ann Inst Stat Math*, 55:737–764.
- Ghiassi, Z. and Saidane (2006). Medium term system load forecasting with dynamic artificial neural network model. *Electric Power Systems Research*, 76:302–316.
- Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E., Robertson, A. W., Saunders, A., Tian, Y., Varadi, F., and Yiou, P. (2002). Advanced spectral methods for climatic time series. *Reviews of Geophysics*, 40.
- Ghoulipour, A., Lucas, C., Araabi, B. N., Mirmomeni, M., and Shafiee, M. (2007). Extracting the main patterns of natural time series for long-term neuro-fuzzy prediction. *Neural Comput & Applic*, 16:383–393.
- Ghoulipour, A., Lucas, C., Araabi, B. N., and Shafiee, M. (2005). Solar activity forecast: Spectral analysis and neurofuzzy prediction. *Journal of Atmospheric and Solar-Terrestrial Physics*, 67:595–603.
- Gomes, M. E. D. (2001). *Técnicas de Sistemas Dinâmicos Não-Lineares na Análise da Modulação Autônômica da Variabilidade da Frequência Cardíaca*. Tese de Doutorado em Engenharia Elétrica, UFMG, Brasil.
- González-Romera, E., Jaramillo-Moran, M. A., and Carmona-Fernández, D. (2007). Forecasting of the electric energy demand trend and monthly fluctuation with neural networks. *Computers Industrial Engineering*, 52:336–343.
- Gooijer, J. G. D. and Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, 22:443–473.

- Hinich, M. J., Mendes, E. M. A. M., and Stone, L. (2005). Detecting nonlinearity in time series: Surrogate and bootstrap approaches. *Berkeley Electronic Press*, 9:1–14.
- Hoffman, F. and Nelles, O. (2001). Genetic programming for model selection of ts-fuzzy systems. *Information Sciences - An International Journal*, 136:7–28.
- Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20:5–10.
- Jesus, T. A. (2006). *Detecção de Determinismo em Séries Temporais via Análise de Dados Sub-Rogados e Modelagem Preditiva Baseada em Modelos Auto-Regressivos Híbridos*. Monografia de Graduação em Engenharia de Controle e Automação, UFMG, Brasil.
- Jesus, T. A. and Mendes, E. M. A. M. (2008). Long-term load forecasting using multiple auto-regressive-like models. *8th Portuguese Conference on Automatic Control - Vila Real - Portugal*.
- Jesus, T. A., Mendes, E. M. A. M., and Campos, R. J. (2007). Análise de previsibilidade em séries temporais via testes de dados sub-rogados. *XXX CNAMC - Congresso Nacional de Matemática Aplicada e Computacional - Florianópolis - SC*.
- Jia, N. X., Yokoyama, R., Zhou, Y. C., and Gao, Z. Y. (2001). A flexible long-term load forecasting approach based on a new dynamic simulation theory - GSIM. *Electrical Power Energy Systems*, 23:549–556.
- Johnston, S., Kramer, M., Lorimer, D. R., Lyne, A. G., McLaughlin, M., Klein, B., and Manchester, R. N. (2006). Discovery of two pulsars towards the galactic centre. *Monthly Notices of the Royal Astronomical Society: Letters*, 373:L6–L10.
- Kandil, M. S., El-Debeiky, S. M., and Hassanien, N. E. (2002). Long-term load forecasting for fast developing utility using a knowledge-based expert system. *IEEE Trans on Power Systems*, 17:491–496.
- Kermanshahi, B. and Iwamiya, H. (2002). Up to year 2020 load forecasting using neural nets. *Electrical Power Energy Systems*, 24:789–797.
- Lei (1995). Lei 9.074. Diário Oficial da União, República Federativa do Brasil, 7 de julho de 1995.
- Lei (2004). Lei 10.848. Diário Oficial da União, República Federativa do Brasil, 15 de março de 2004.
- Liao, Z. (2001). *Methodology for Forecasting Residential Energy Demand in China: An Approach to Technological Impacts*. Ph.D. Thesis, Hong Kong Polytechnic University, China.
- Madsen, K., Nielsen, H. B., and Tingleff, O. (2004). *Methods for Non-Linear Least Squares Problems*. Informatics and Mathematical Modelling - Technical University of Denmark.

- Masters, T. (1994). *Signal and Image Processing With Neural Networks*. John Wiley.
- Mauro, D. F. (2007). *Desenvolvimento de um módulo baseado em redes MLP de um sistema de previsão de consumo de energia*. Monografia de Graduação em Engenharia de Controle e Automação, UFMG, Brasil.
- Nakamura, T. and Small, M. (2005). Small-shuffle surrogate data: Testing for dynamics in fluctuating data with trends. *Phys Rev E*, 72:056216.
- Nakamura, T. and Small, M. (2006). Testing for dynamics in the irregular fluctuations of financial data. *Physica A*, 366:377–386.
- Nakashima, T. (2000). *Coordination and Decision Making of Regulation, Operation, and Market Activities in Power Systems*. Ph.D. Thesis, The University of British Columbia, Canada.
- Narasimha, P. L. (2003). *Temperature Forecasting Using Neural Networks*. Master's Thesis, University of Texas at Arlington, USA.
- Nelles, O. (2001). *Nonlinear System Identification - From Classical Approaches to Neural Networks and Fuzzy Models*. Springer-Verlag.
- Nunn, I. and White, T. (2005). The application of antigenic search techniques to time series forecasting. GECCO.
- Ozkurt, N. and Savaci, F. A. (2006). The implementation of nonlinear dynamical systems with wavelet network. *Int. J. Electron. Commun.*, 60:338–344.
- Papoulis, A. and Pillai, S. U. (2002). *Probability, Random Variables and Stochastic Processes*. Tata McGraw-Hill Publishing Company Limited, New Delhi.
- Perepelitsa, V. A., Maksishko, N. K., and Kozin, I. V. (2006). Using a model of cellular automata and classification methods for prediction of time series with memory. *Cybernetics and Systems Analysis*, 42:807–816.
- Postalcioglu, S. and Becerikli, Y. (2007). Wavelet networks for nonlinear system modeling. *Neural Comput & Applic*, 16:433–441.
- Punjabi, M. (2005). *Integrating Econometric Analysis with Scenario Analysis for Forecasting in Rapidly Changing Environment; Case Study of the U.S. Dry Bean Industry*. Ph.D. Thesis, Michigan State University, USA.
- Ramsey, J. B. (1999). The contribution of wavelets to the analysis of economic and financial data. *Phil. Trans. R. Soc. Lond. A*, 357:2593–2606.
- Small, M. (2005). *Applied Nonlinear Time Series Analysis: Applications in Physics, Physiology and Finance*. World Scientific.
- Small, M., Harrison, R. G., and Ts, C. K. (2001). A surrogate test for pseudo-periodic time series data. *Phys Rev Lett*, 87.

- Small, M., Judd, K., and Lowe, M. and Stick, S. (1999). Is breathing in infants chaotic? dimension estimates for respiratory patterns during quiet sleep. *J Appl Physiol*, 86:359–376.
- Soofi, A. and Cao, L. (2002). *Modelling and Forecasting in Financial Data - Techniques of Nonlinear Dynamics*. Kluwer Academic Publishers, Assinippi Park, Norwell, Massachusetts - USA.
- Takagi, T. and Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, 15:116–132.
- Teixeira, B. O. S. (2004). *Modelagem e Sintonia de Controladores para um Forno de Reaquecimento*. Monografia de Graduação em Engenharia de Controle e Automação, UFMG, Brasil.
- Theiler, J. (1995). On the evidence for how-dimensional chaos in an epileptic electroencephalogram. *Phys Lett A*, 196:335–341.
- Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., and Farmer, D. (1992). Testing for nonlinearity in time series: the method of surrogate data. *Physica D*, 58:77–94.
- Theiler, J. and Prichard, D. (1996). Constrained-realization monte-carlo method for hypothesis testing. *Physica D*, 94:221–235.
- Theiler, J. and Rapp, P. E. (1996). Re-examination of the evidence for low-dimensional, nonlinear structure in the human electroencephalogram. *Electroencephalography and Clinical Neurophysiology*, 98:213–222.
- Thomson, J. M. T. and Stewart, H. (1986). *Nonlinear Dynamics and Chaos: Geometrical Methods for Engineers and Scientists*. Chichester.
- Tong, H. (1996). *Non-linear Time Series - A Dynamical System Approach*. Oxford Science Publications.
- Tribble, A. N. (2003). *The Relationship between Weather Variables and Electricity Demand to Improve Short-Term Load Forecasting*. Ph.D. Thesis, University of Oklahoma, Norman, Oklahoma, USA.
- Tôrres, L. A. B. (2001). *Controle e Sincronismo de Osciladores Caóticos*. Tese de Doutorado em Engenharia Elétrica, UFMG, Brasil.
- Vautard, R., Yiou, P., and Ghil, M. (1992). Singular spectrum analysis: a toolkit for short noisy chaotic signals. *Physica D*, 58:95–126.
- Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics.
- Weigend, A. S. and Gershenfeld, N. A. (1996). *Time Series Prediction: Forecasting the Future and Understanding the Past*. Addison-Wesley Publishing Company, Inc, USA.

- Worden, K. and Tomlinsom, G. R. (2001). *Nonlinearity in Structral Dynamics - Detection, Identification and Modelling*. IOP - Institute of Physics Publishing.
- Yousefi, S., Weinrich, I., and Reinarz, D. (2005). Wavelet-based prediction of oil prices. *Chaos Soliton Fractal*, 25:265–275.
- Yule, G. U. (1927). On the method of investigating periodicities in disturbed series, with special reference to Wölfer's sunspot numbers. *Philosophical Transactions of the Royal Society London, Series A*, 226:267–298.
- Zhang, G. P. (2000). Neural networks for classification: a survey. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, 30:451–462.

Teste de Grubbs

O teste de Grubbs é usado para detecção de *outliers* em conjuntos de dados mono-variados. Em última análise, ele é um teste de hipóteses convencional que se baseia no fato de que o conjunto de dados a ser analisado seja razoavelmente aproximado por uma distribuição normal.

O teste de Grubbs é capaz de detectar um *outlier* por vez e o retirar. O teste é realizado iteradas vezes até que não se encontrem mais *outliers* no conjunto de dados em análise.

Seja a seguinte hipótese nula H_0 : ‘Não há outliers no conjunto de dados’.

Como em todo teste de hipótese, é necessário escolher uma estatística. No caso, define-se a estatística do teste de Grubbs como:

$$G = \frac{\max |Y_i| - \bar{Y}}{s} \quad (\text{A.1})$$

em que $\bar{Y}$ é a média amostral e s o desvio padrão amostral. G é o máximo desvio absoluto em relação à média, e é dado em unidades de desvio padrão amostral.

Estabelecem-se, então, limites de confiança de tal forma que, se o valor de G estiver fora desses limites, o teste pode ser rejeitado e o ponto Y_i é considerado um *outlier*. Esses limites podem ser de $\bar{Y} \pm 3s$, por exemplo, uma vez que a probabilidade de uma realização da distribuição normal estar compreendida nessa faixa é de 99,7%.