SISTEMA DE MONITORAÇÃO DE QUALIDADE DE DADOS

Augusto Cesar Heluy Dantas

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Aprovada por:

Prof. José Manoel de Seixas, D.Sc.

Prof. Fernando José Von Zuben, Dr.E.E.

Prof. Marcelo Cunha Medeiros, D.Sc.

Prof. Luiz Pereira Calôba, D.Ing.

Prof. Basílio de Bragança Pereira, Ph.D.

Prof. Antonio Petraglia, Ph.D.

RIO DE JANEIRO, RJ - BRASIL

MAIO DE 2007

DANTAS, AUGUSTO CESAR HELUY

Sistema de Monitoração de Qualidade de Dados [Rio de Janeiro] 2007

XV, 149 p 29,7 cm (COPPE/UFRJ, D.Sc., Engenharia Elétrica, 2007)

Tese - Universidade Federal do Rio de Janeiro, COPPE

1.Qualidade de Dados 2.Sistema de Monitoração 3.Análise Estatística 4.Redes Neurais Artificiais 5.Séries Temporais e Registros Multivariados

I.COPPE/UFRJ II.Título (série)

Aos meus queridos pais, Antonio e Odila.

À memória do meu querido avô Aziz.

Agradecimentos

Agradeço em primeiro lugar a Deus, nosso Pai. Por Ti, Senhor, qualquer esforço vale a pena!

Agradeço aos meus pais, Antonio e Odila, pelo imenso amor com que me têm tratado. Sem seu carinho e incentivo, esse trabalho não teria sido possível.

Agradeço aos meus irmãos Frederico, Cecilia e Ana, pelo apoio e amizade. Agradeço também à minha avó Eunice e à minha tia Angelina pelo carinho e presença constante.

Agradeço ao Prof. José Manoel de Seixas, meu orientador, pelo desvelo na orientação desta tese e, sobretudo, pela amizade cultivada ao longo desses anos.

Agradeço aos demais membros da banca, por sua disponibilidade em acompanhar esta tese e pelos comentários feitos sobre o trabalho, assim como pelas correções realizadas.

Quero também agradecer a tantos amigos... Mesmo correndo o risco de aqui me esquecer de alguém (no meu coração não os esqueço!), gostaria de mencioná-los: Luiz Fernando B. Brandão, José Gabriel R. C. Gomes, Charles Bezerra do Prado, Rafael Linhares Marinho, José Márcio Faier, Augusto Silberstein, Eugenio Suarez Caner, Pepe Almagro, Paulo Oriente, João Malheiro, Henrique Elfes, César Perez, Christian Paz-Trillo, Victor Hugo Lachos, Filipe Barbosa, Filipe Diniz, Tadeu Ferreira, Fábio Freeland e a todos os amigos do Laboratório de Processamento de Sinais e da Bayes Forecast. Todos, de alguma maneira, tem uma parcela de contribuição neste trabalho e, sobretudo, na minha vida.

Muito obrigado!

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

SISTEMA DE MONITORAÇÃO DE QUALIDADE DE DADOS

Augusto Cesar Heluy Dantas

Maio/2007

Orientador: José Manoel de Seixas

Programa de Engenharia Elétrica

O tema da Qualidade de Dados tem crescido rapidamente em importância nos últimos anos, devido à enorme quantidade de dados disponíveis atualmente e ao seu papel na economia da informação. Algumas técnicas isoladas vêm sendo desenvolvidas com o intuito de atuar na melhoria da Qualidade de Dados de bases específicas, sobretudo no meio empresarial. Porém, a área ainda se ressente da ausência de sistemas integrados de monitoração que sejam capazes de prover uma melhoria contínua da qualidade dos dados de grandes bases, cujo conteúdo é formado por dados de origens bastante diversas. Nesta tese, desenvolvemos um sistema de monitoração de qualidade de dados baseado em ferramentas estatísticas e de processamento de sinais, com destaque especial para o uso das redes neurais artificiais. O sistema em questão foi desenvolvido segundo as diretrizes que vêm norteando a pesquisa acadêmica em Qualidade de Dados, e aplicado a alguns grupos de dados (entre séries temporais e registros multivariados) como forma de avaliar o seu desempenho.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

DATA QUALITY MONITORING SYSTEM

Augusto Cesar Heluy Dantas

May/2007

Advisor: José Manoel de Seixas

Department: Electrical Engineering

In the past years, the Data Quality theme has grown rapidly in importance due to the enormous quantity of data available and its role in the information economy. Several isolated techniques have been developed in order to improve the Data Quality of specific databases, mainly within the business environment. However, this field still suffers with the lack of integrated monitoring systems capable of offering continuous improvement of data quality in large databases, whose content is formed by data of very different sources. In this thesis, a data quality monitoring system has been developed, based on statistical and signal processing tools, with special emphasis on the use of artificial neural networks. The system in question was developed according to the directions that orient academic research in Data Quality, and was applied to some data groups (time series and multivariate data) so as to assess its performance.

Sumário

1 Introdução **1**

1.1 Contexto 1

1.2 Motivação e Objetivos 2

1.3 Contribuições 4

1.4 Organização do Texto 5

2 Qualidade de Dados **7**

2.1 Visão Geral 7

2.1.1 Breve Histórico 7

2.1.2 Importância 8

2.1.3 Impacto Negativo da Má Qualidade de Dados 9

2.1.4 Aspecto Multidisciplinar 10

2.1.5 Um Programa de Referência 11

2.1.6 Dados ou Informação? 12

2.2 Conceitos e Definições 12

2.2.1 O Que é Qualidade de Dados? 12

2.2.1.1 Dependência do Contexto 13

2.2.1.2 Multidimensionalidade 13

2.2.1.3 Objetividade e Subjetividade 14

2.2.1.4 Métrica para Qualidade de Dados 15

2.2.2 Como um Dado Perde Qualidade? 18

2.3 Áreas de Aplicação 19

2.4 Gerenciamento de Qualidade de Dados 20

2.5 Qualidade de Dados, Estatística e Processamento de Sinais 22

2.5.1 Qualidade de Dados e Estatística 22

2.5.2 Qualidade de Dados e Processamento de Sinais 22

3 Análise Descritiva dos Dados 25

3.1 Tópicos em Séries Temporais 25

3.1.1 Estacionariedade em Séries Temporais 25

3.1.1.1 Conceito de Estacionariedade 26

3.1.1.2 Detecção da Não-Estacionariedade 27

3.1.2 Pré-processamento de Séries Temporais 38

3.1.3 Exemplos de Séries Temporais Utilizadas 41

3.1.3.1 Comentários Iniciais 41

3.1.3.2 Séries Temporais Financeiras 42

3.1.3.3 Outras Séries Temporais 44

3.2 Tópicos em Registros Multivariados 45

3.2.1 Exemplos de Registros Multivariados Utilizados 46

3.2.1.1 Os dados do ICPSR 46

3.2.1.2 A Base de dados FISCT 47

4 O Sistema de Monitoração de Qualidade de Dados (SMQD) 50

4.1 Modelos Propostos na Literatura 50

4.1.1 Modelo Baseado na Teoria de Controle 51

4.2 O SMQD 54

4.2.1 Metodologia 54

4.2.1.1 Interface do Sistema 55

4.2.1.2 Dimensões Monitoradas 56

4.2.1.3 Anomalias Buscadas 58

4.2.1.4 Tipos de Dados 58

4.2.1.5 Método de Correção 59

4.2.1.6 Ferramentas Utilizadas 59

4.2.2 Dinâmica de Atuação do SQMD 59

4.2.2.1 Validação de Novos Dados 60

4.3 Monitoração das Dimensões Objetivas 63

4.3.1 Monitoração da Acurácia 63

4.3.1.1 Detecção de *Outliers* 63

4.3.2 Monitoração da Completude 85
4.3.2.1 Completando Dados Faltantes 85
4.3.3 Monitoração da Pontualidade 88

5 Testes com o SMQD **91**
5.1 Monitoração de Séries Temporais 91
5.1.1 Séries Temporais Financeiras 92
5.1.1.1 Fenômeno do "Atraso" na Predição de Séries Financeiras 92
5.1.2 Séries Temporais Caóticas 106
5.1.3 Outras Séries Temporais 108
5.2 Monitoração de Registros Multivariados 121
5.2.1 Dados do ICPSR 121
5.2.2 Dados da Área Médica 127
5.2.2.1 Imputação Usando Algoritmo EM Regularizado . . . 130
5.2.2.2 Usando Imputação Múltipla 132

6 Conclusões e Sugestões para Trabalhos Futuros **134**
6.1 Estudo dos Tópicos de Qualidade de Dados 134
6.2 A Metodologia do Sistema de Monitoração 135
6.3 Testes Implementados 136
6.4 Continuações para esta Pesquisa 137

Referências Bibliográficas **139**

Lista de Figuras

2.1 Formação dos atributos de QD. 15

3.1 Passeio aleatório sem "tração" (acima) e FAC (abaixo). 29
3.2 Primeira diferença do passeio aleatório sem "tração" (acima) e FAC (abaixo). 29
3.3 Passeio aleatório com "tração" (acima) e FAC (abaixo). 30
3.4 Primeira diferença do passeio aleatório com "tração" (acima) e FAC (abaixo). 30
3.5 Série de demanda de energia elétrica nos EUA. Série original (acima) e após a remoção da tendência determinística (abaixo). 33
3.6 Série de demanda de energia elétrica nos EUA (acima) e após a transformação logarítmica (abaixo). 34
3.7 Série de demanda de energia elétrica (a); FAC da série diferenciada (b); e FAC da série dessazonalizada (c). 35
3.8 Número de nascimentos na cidade de Quebec entre 01/01/1977 e 31/12/1990. 37
3.9 Função de autocorrelação para a série de Quebec. 37
3.10 Espectro de Fourier para a série de Quebec, com a freqüência normalizada. 38
3.11 Espectro de Fourier para a série de Quebec, com a freqüência normalizada. Detalhe da região de baixa freqüência. 39
3.12 Ciclo de 15 anos e variação anual para a série de Quebec. 39
3.13 Procedimento para a estacionarização de séries temporais. 40
3.14 Valores (certificados) de fechamento para a série da IBM de 01/1998 a 06/2006. 43

3.15 Comparação entre duas séries para a AMD: diferença percentual entre as séries certificada e não-certificada. 44

4.1 Ciclo para a monitoração da Qualidade de Dados. 51
4.2 Fluxo de dados fonte-usuário. 52
4.3 Sistema de controle de qualidade de dados. 53
4.4 Controle clássico de uma planta. 53
4.5 Estrutura hierárquica em camadas para o SMQD. 55
4.6 Fluxo de dados dentro do SMQD. 61
4.7 Validação de um dado novo. 62
4.8 Corredor de validação para novas amostras. 65
4.9 Descontinuidade ("quebra estrutural") no valor das ações da Hawaiian Airlines em torno ao 11 de setembro de 2001 (o mercado de ações ficou fechado por uma semana). 66
4.10 Procedimento para a estacionarização de séries temporais. 68
4.11 Modelo básico para o estimador neural de séries temporais. 71
4.12 Esquema entrada-saída para o estimador de base $N = 3$. 72
4.13 Esquema geral para a arquitetura de Elman. 73
4.14 Estrutura das redes ESN em comparação com o aprendizado em Redes Neurais Recorrentes (RNRs) tradicionais. 74
4.15 Razão entre observações consecutivas para a série da IBM. Patamar em 6%. 77
4.16 Esquema genérico para o procedimento de Janela-Móvel. 78
4.17 Ilustrando quebras de tendência. 79
4.18 Exemplo de operação do SMQD durante a fase de pré-processamento. Acima, o trecho da série analisado juntamente com a retirada da tendência global. Abaixo, o espectro da série de média zero (com freqüência não normalizada). 80
4.19 Trecho da série já normalizada e após a quebra de tendência, com o espectro da série de média zero. 80
4.20 Detectando *outliers* através da distância de Mahalanobis. 82
4.21 Remoção de *outliers* através do algoritmo proposto. Os pontos removidos estão representados por "o". 83

4.22 Exemplo de substituição de valores faltantes pela média. 87
4.23 Exemplo de substituição de valores faltantes por imputação múltipla. 89

5.1 Arquitetura para a rede neural treinada com o algoritmo FFBP. . . . 93
5.2 Série da SUN predita pela rede neural vs. alvo (acima) e com deslocamento da saída da rede (abaixo). 94
5.3 Histograma da diferença entre a série alvo e a série deslocada (acima) e FAC da diferença (abaixo). 95
5.4 Erro médio quadrático durante o treinamento, para os conjunto de treino e validação. 95
5.5 Detalhes do "fenômeno do atraso" para algumas séries financeiras (a saída da rede está representada pela linha mais clara). 96
5.6 Tentativa de predição de ruído branco através de rede neural minimizando o EMQ. 97
5.7 Espectro de freqüência (acima) e função de autocorrelação para a amostra de ruído branco colocado na entrada da rede neural (abaixo). 97
5.8 Série do passeio aleatório (acima) e sua FAC (abaixo). 98
5.9 Série do passeio aleatório estacionarizado (acima) e sua FAC (abaixo). 99
5.10 Série do passeio aleatório predito pela rede neural vs. alvo (acima) e com deslocamento da saída da rede (abaixo). 99
5.11 Função de autocorrelação para 3 séries de mercado financeiro (séries estacionarizadas pela operador ∇). 100
5.12 Espectro em freqüência para as 3 séries analisadas. 101
5.13 Comparação da FAC da série estacionarizada da SUN com a de uma amostra de ruído branco do mesmo tamanho. 101
5.14 Qual destas 4 séries é o passeio aleatório? 102
5.15 Correlação entre a série real e a série predita (Microsoft), usando-se uma rede FFBP. 104
5.16 Correlação entre a série real e a série predita (Microsoft), usando-se uma rede Elman. 104
5.17 Predição da série da IBM através de modelo ARIMA(1,1,1). 106
5.18 Processo de validação da série da AMD. 107
5.19 Modelagem de uma série de Lorenz através de ESN. 109

5.20 Função de autocorrelação para a série de Lorenz (acima) e seu espectro de freqüência (abaixo). 109
5.21 Modelagem da série da Microsoft por uma rede ESN. 110
5.22 Consumo mensal de energia elétrica nos EUA (1973 a 2006) com as variâncias a cada trecho. 111
5.23 Homogeneização da variância da série de energia elétrica. 111
5.24 Função de autocorrelação para a série de energia elétrica já homoscedástica e sem tendência, mas ainda com sazonalidade. 112
5.25 Série do consumo elétrico estacionarizada (acima) e espectro em freqüência (abaixo). 112
5.26 Função de autocorrelação para a série de energia elétrica estacionarizada. 113
5.27 Predição da série de eletricidade pela rede FFBP. 114
5.28 Modelagem da série de eletricidade pela rede de Elman. 115
5.29 Determinação dos limites para *outliers* na série de eletricidade. 116
5.30 Correta remoção e substituição de 5 *outliers* através de corredor com k=1,5 (série da eletricidade). 117
5.31 Reposição de dados faltantes na série de eletricidade. 118
5.32 Série para a Produção Física Industrial - Produtos Alimentares (índice IPI, base Brasil) entre janeiro de 1985 e julho de 2000. 119
5.33 Correlograma para a série IPI antes e depois da retirada da tendência linear. 119
5.34 Série IPI desazonalizada (acima) e respectivo correlograma (abaixo). . 120
5.35 Modelagem da série do índice IPI pela rede de Elman, com k=3. O fator de sincronismo é R=1,22, e o erro de predição (conjunto de teste) vale 4,2%. 121
5.36 Detecção e correção de 2 *outliers* e preenchimento de 1 dado faltante, com k=3 (série IPI). 122
5.37 Histograma para a variável “tempo de sono” sem a atuação do SMQD. 125
5.38 Histograma para a variável “tempo de sono” após a atuação do SMQD. 126
5.39 *Outliers* removidos pelo SMQD. 127
5.40 Ajuste gaussiano para o conjunto sem monitoração. 128

5.41 Ajuste gaussiano para o conjunto monitorado. 128

Lista de Tabelas

3.1 Listagem de algumas das séries temporais não-financeiras utilizadas neste trabalho. 45
3.2 Descrição do resultados das ligações telefônicas para a base FISCT. . 47
3.3 Exemplo de resultado de entrevista para a base FISCT. 48
3.4 Descrição das colunas da Tabela 3.3. 49

4.1 Amostra com dados incompletos. 86

5.1 Especificações para a rede FFBP usada para gerar a Figura 5.27. . . . 114
5.2 Detecção e substituição de *outliers* na série de eletricidade. Os valores estão em 10^5 KWh. 117
5.3 Especificações para a rede Elman usada para modelar a série IPI. . . 120
5.4 Exemplo de resultado de entrevista para a base FISCT. 123
5.5 Descrição das colunas da Tabela 5.4. 123
5.6 Alguns eventos para o registro de casos de hepatite A. 129
5.7 Resultado da imputação via algoritmo EM para dados faltantes do conjunto da hepatite. 131

Capítulo 1

Introdução

Neste primeiro capítulo fazemos a introdução ao texto, situando a pesquisa feita em seu contexto original e descrevendo as principais motivações que guiaram o presente trabalho. São também comentadas as contribuições alcançadas por esta tese. Finalmente, apresentamos a organização deste texto, através de uma breve descrição de cada capítulo que o compõe.

1.1 Contexto

Desenvolver uma tese sobre Qualidade de Dados (QD) talvez não fosse possível há poucos anos atrás. Com efeito, trata-se de uma área de pesquisa bastante recente. Porém, uma área de pesquisa em estágio incipiente é sempre bastante convidativa, já que em geral está repleta de desafios e oportunidades. Em concreto, a área de QD tem propiciado a busca de novos rumos para o desenvolvimento tecnológico futuro, em que cada vez mais as bases de dados assumem o protagonismo na tomada de decisões. É hoje uma necessidade a criação de um novo paradigma para a modelagem de sistemas que utilizam grandes bases de dados.

Para entendermos melhor o contexto em que se encontra este trabalho, convém olharmos para o caminho trilhado pela produção de tecnologia nas últimas décadas. O avanço tecnológico da segunda metade do século XX teve como foco principal a geração de novos produtos, soluções ou facilidades para a indústria, as empresas e cada indivíduo. O gerenciamento dos dados envolvidos, porém, não acompanhou essa evolução. Quando realizado, dava-se de forma não estruturada e não conti-

nuada. Dessa forma, as informações que poderiam ser extraídas de tais dados permaneciam escondidas, fazendo com que fossem desperdiçadas inúmeras descobertas.

Paralelamente, dava-se o crescimento exponencial da capacidade de armazenamento digital de grandes bases de dados. A partir da década de 90, devido em grande parte à globalização da informação com a disseminação da *Internet*, deu-se início a uma nova era tecnológica, a chamada "era da informação" ou "era dos dados". Nesse momento, muitas instituições (acadêmicas ou comerciais) perceberam que, ao mesmo tempo em que aproveitavam muito pouco os recursos contidos nas informações escondidas nas suas próprias bases de dados, também sofriam muitos prejuízos em decorrência do seu mau gerenciamento.

Chegamos então a uma situação em que, de maneira cada vez mais intensa, a pesquisa acadêmica e a produção industrial são direcionadas pelos dados (enfoque *data driven*). Assim sendo, a preocupação com a *qualidade dos dados* utilizados tem também crescido, fazendo com que cada vez mais trabalhos sejam publicados nesta área.

1.2 Motivação e Objetivos

Verificado esse redirecionamento progressivo no desenvolvimento de novas tecnologias (*technology driven* $\longrightarrow$ *data driven*), abre-se um grande panorama de tópicos de pesquisa para o tratamento de dados. Em particular, a área de Qualidade de Dados surge como um dos principais ramos a serem investigados e, por se tratar de tema intrinsecamente multidisciplinar, diversos enfoques têm sido abordados.

O principal objetivo desta tese é desenvolver um sistema integrado em que se possa realizar, de forma semi-automática e dinâmica, a monitoração da Qualidade de Dados das bases que integram sistemas de informação diversos. Esta monitoração deve compreender não só a análise mas também a implementação de melhorias, visando sempre à obtenção de bases de dados mais ricas em informação segura. O sistema baseia-se na certeza de que, em se tratando de monitorar qualidade de dados, é sempre mais vantajoso realizar essa monitoração de forma contínua, detectando e corrigindo pequenas falhas, do que fazer, de tempos em tempos, uma grande análise da qualidade da base. Com isso, inibem-se propagações desnecessárias de erros, além

de se poder oferecer ao usuário do sistema (por exemplo, a própria empresa detentora dos dados) uma maior probabilidade de estar trabalhando com dados seguros.

Esta monitoração deve dar-se de forma semi-automática porque é indispensável, em alguns momentos, a ação direta de um especialista, seja para a introdução de conhecimento novo, seja para a resolução de problemas não previstos pelo sistema. No entanto, é desejável que o sistema possa ir ganhando, com o tempo, maior autonomia operacional, passando a depender cada vez menos de auxílio externo que, no entanto, deverá estar sempre presente.

Paralelamente ao desenvolvimento da estrutura de um Sistema de Monitoração de Qualidade de Dados (SMQD), temos também como objetivo estudar a aplicabilidade de um grande conjunto de técnicas na monitoração da QD. Numa área que é multidisciplinar por definição, o nosso desafio é testar a utilização de técnicas comumente usadas na área de Processamento de Sinais; entretanto, procuramos também aproveitar as vantagens de diversas outras técnicas tradicionalmente usadas no tratamento de bases de dados, sobretudo técnicas estatísticas. Fazer uma seleção das técnicas mais apropriadas para cada caso constitui parte dos nossos objetivos.

É interessante aqui tecer alguns comentários com relação à natureza das bases de dados a que se destina o sistema de monitoração proposto. De fato, é própria do sistema desenvolvido a habilidade de lidar com os mais diferentes tipos de dados, sejam eles relativos a fenômenos físicos (sinais de energia, sob a forma de tensão, corrente elétrica ou luz, ou ainda sinais de vídeo, áudio etc), sejam relativos a fenômenos sócio-econômicos, tais como séries temporais em geral, bases cadastrais, indicadores de mercado de ações etc. Neste trabalho, daremos grande ênfase ao tratamento deste último grupo de sinais (dados), por duas razões principais: é o tipo de dados que vem guiando as pesquisas na área de Qualidade de Dados (em virtude do grande impacto financeiro causado pela má gestão de dados nessa área), além de haver grande quantidade de dados reais desse tipo disponível para pesquisa. Esses dados serão sempre analisados em dois grandes grupos: o das séries temporais e o dos registros multivariados.

1.3 Contribuições

Acreditamos que um dos principais méritos desta tese tenha sido o de delinear uma metodologia própria para um sistema geral de monitoração de Qualidade de Dados. Isto adquire particular importância quando se percebe que, dado o estágio inicial em que se encontram as pesquisas nesta área, uma das lacunas mais significativas é justamente a ausência de sistemas que atuem no gerenciamento da qualidade dos dados de grandes bases de dados. Desta forma, o sistema proposto nesta tese surge como uma possibilidade real, podendo ser utilizado como ponto de partida de trabalhos futuros, em que ele seja aperfeiçoado e ampliado.

Isto confere a esta tese um papel diretivo, qual seja, o de propor uma linha de atuação para futuras pesquisas na área. Com efeito, o nosso foco não esteve tanto na execução de testes com o sistema proposto sobre extensas bases de dados, ainda que sejam relatados, no Capítulo 5, os resultados de simulações feitas com alguns grupos de dados específicos, porém representativos da maior parte dos dados presentes nos sistemas de informação atuais. Mais importante pareceu-nos conceber a dinâmica de atuação do sistema, desde a recepção de dados novos até sua validação e incorporação na base de dados, passando por todas as etapas de detecção e correção de falhas.

Outra característica desta tese é o seu aspecto ao mesmo tempo conceitual e prático. Primeiramente, assentam-se os fundamentos teóricos da área de qualidade de dados; posteriormente, são selecionadas as técnicas que devem compor os mecanismos de atuação do sistema de monitoração. Por exemplo, dentre as técnicas de Processamento de Sinais que são utilizadas neste trabalho, destacamos aqui as Redes Neurais Artificiais, implementadas com variadas arquiteturas. Com relação às técnicas estatísticas, utilizamos modelos ARIMA (para modelagem de séries temporais), e algumas ferramentas com aplicação a distribuições multidimensionais, como as distâncias de Mahalanobis e Kullback-Leibler, algoritmos EM (*Expectation-Maximization*) e imputação múltipla.

1.4 Organização do Texto

Além desta introdução, o presente texto contém mais 5 capítulos, organizados da forma descrita a seguir.

O Capítulo 2 discorre sobre o tema da Qualidade de Dados, assentando as bases teóricas para o desenvolvimento de um sistema de monitoração. Dá-se inicialmente uma visão geral do assunto, em que um pouco da história do tema é resgatada, a sua importância é constatada e o seu aspecto multidisciplinar é vislumbrado.

Continuando nesse capítulo, os conceitos e definições relativos à área de Qualidade de Dados são fornecidos. Aspectos como a multidimensionalidade[1] da QD, a dicotomia entre a objetividade ou subjetividade dessas dimensões (que são aqui definidas) e o processo de perda de qualidade em bases de dados são destacados. As áreas de aplicação de QD, as motivações para o seu gerenciamento em bases de dados e as relações entre Qualidade de Dados, Estatística e Processamento de Sinais completam o Capítulo 2.

O capítulo seguinte traz uma análise detalhada dos dados utilizados nesta tese, sobretudo das *séries temporais*, cuja teoria é analisada e cujo procedimento de estacionarização é descrito[2]. Também são dados alguns exemplos de séries utilizadas nesse trabalho, além de exemplos de registros multivariados, que constituem o segundo grande grupo de dados testado no sistema de monitoração desenvolvido.

O Capítulo 4 é o núcleo desta tese. Nele descrevemos os princípios de funcionamento do Sistema de Monitoração de Qualidade de Dados (SMQD), principal contribuição deste trabalho. Partindo de algumas indicações propostas na literatura com respeito aos atributos que seriam desejáveis dentro de um sistema de monitoração, construímos passo a passo a estrutura do nosso SMQD: desde a metodologia que o fundamenta, em que definimos quais as dimensões da QD são monitoradas (juntamente com a descrição das ferramentas utilizadas para realizar essa

[1]Neste texto, o termo "dimensão" é normalmente utilizado para significar cada uma das distintas características ou aspectos que compõem a Qualidade de Dados; desta forma, "multidimensional" quer dizer "multifacetado".

[2]Convém dizer aqui que o foco do nosso trabalho não está em desenvolver mecanismos preditores para séries temporais, mas em gerar um sistema de monitoração que faça a análise da qualidade de dados para bases grandes e dinâmicas.

monitoração), até a sua dinâmica de atuação, em que detalhamos como o SMQD atua na detecção e correção de anomalias em dados específicos.

Na seqüência, dedicamos um capítulo para ilustrar a atuação do SMQD sobre bases de dados reais, de maneira a poder quantificar a sua eficácia quando se depara com os problemas específicos de cada grupo de dados. Não se trata de fazer nesse capítulo uma validação do SMQD a partir de testes exaustivos[3], mas sobretudo de exemplificar a sua atuação em alguns casos importantes. Os testes com dados reais incluem a análise de dados faltantes e de *outliers*, além da modelagem de distribuições para as variáveis envolvidas nos registros multivariados. Aproveitamos este capítulo para discutir também alguns fenômenos observados em grupos de dados específicos, como as séries temporais financeiras.

Por fim, o Capítulo 6 é reservado para as conclusões do trabalho. Nele, comentamos as contribuições desta tese para a pesquisa na área de Qualidade de Dados. As principais dificuldades encontradas são analisadas e caminhos alternativos, além de extensões possíveis, são apontados. Como esta área de pesquisa tem um caráter ainda embrionário, muitas portas foram abertas durante o desenvolvimento desta tese, de maneira que recolhemos as principais idéias não implementadas e as deixamos, no fim do texto, à guisa de sugestões para novos trabalhos.

[3]Tais testes seriam necessários quando se quisesse, posteriormente, aplicar o SMQD a sistemas específicos, de maneira a verificar a sua eficiência nesses ambientes.

Capítulo 2

Qualidade de Dados

Qualidade de Dados (QD) é o tema central em estudo nesta tese. É uma área recente, do ponto de vista da pesquisa acadêmica, e tem sido alvo de um enfoque multidisciplinar. Neste capítulo, pretende-se reunir e discutir os principais aspectos que giram em torno do tema da Qualidade de Dados, de maneira a assentar as bases para o desenvolvimento do sistema de monitoração, que será proposto no Capítulo 4.

2.1 Visão Geral

Nesta seção, pretendemos fornecer uma primeira aproximação ao tema da Qualidade de Dados, através da exposição de alguns aspectos introdutórios e de algumas informações gerais.

2.1.1 Breve Histórico

De certa maneira, a preocupação com a melhoria dos processos em geral sempre existiu dentro do meio produtivo, seja empresarial ou acadêmico. A relação direta entre a qualidade dos meios (dados, sistemas) e a produtividade final sempre foi algo patente, de maneira que um desleixo formal nesta área seria difícil de se compreender.

Porém, esta preocupação somente veio a redundar em ações mais concretas nos últimos anos, em particular a partir da segunda metade da década de 90. Com efeito, é difícil encontrar alguma referência bibliográfica sobre o assunto com data anterior ao ano 2000. Antes desse surgimento, não havia programas estruturados

nem técnicas estabelecidas para atuar diretamente na melhoria da qualidade de determinados produtos e serviços. Existiam, quando muito, procedimentos *ad hoc* que atuavam de forma isolada e descontinuada [1].

Os primórdios do investimento estruturado em Qualidade de Dados se deram nos anos 80, quando o rápido crescimento na disponibilidade de ferramentas computacionais, tanto em *hardware* como em *software*, possibilitou que algumas grandes bases de dados norte-americanas, então digitalizadas, pudessem sofrer algum tipo de intervenção automática. Assim, uma das primeiras ações estruturadas de que se tem notícia, no campo da Qualidade de Dados, diz respeito à correção dos dados postais de centenas de milhares de cidadãos norte-americanos. De fato, tais dados estão sempre sujeitos a alterações (por diversos motivos, tais como mudança de endereço, casamento, falecimento etc), falhas no registro, incompletude etc. Estima-se que 2% do total de registros desse tipo tornam-se obsoletos a cada mês [2]. Por outro lado, empreender a correção e atualização manuais dessas bases de dados consumiria enormes quantidades de tempo e dinheiro. A partir da disponibilização, por parte de agências governamentais norte-americanas, de registros nacionais de mudanças de endereço, muitas companhias puderam beneficiar-se do cruzamento destes registros com os seus, atualizando e corrigindo suas bases com o auxílio de *mainframes* que executavam algumas regras pré-estabelecidas. O resultado imediato deste procedimento foi uma enorme economia em postagem de correspondências, como contas e publicidade direta, devido simplesmente a uma melhor acurácia nos dados dos destinatários.

Não obstante o crescimento observado nos últimos anos, um problema ainda hoje enfrentado pelas áreas de TI (Tecnologia da Informação) nas grandes empresas é o pouco uso de metodologias e técnicas consistentes no desenvolvimento de programas de Qualidade de Dados.

2.1.2 Importância

Uma das principais características do mundo atual é o incrível volume de dados gerados constantemente. Segundo estimativa feita pela Universidade da Califórnia, Berkeley [3], o volume de dados disponível na rede *Internet* ao final do ano de 2005 equivalia a 5 *Exabytes* (EB, 10^{18} *bytes* ou um bilhão de *Terabytes*). Para se

ter uma idéia do que essa quantidade representa, seriam necessários 40,7 mil anos de transmissão ininterrupta de televisão para se transmitir a mesma quantidade de dados.

Há que se admitir, então, que o século atual configura-se como o "século dos dados" [4]. Além da *Internet* e dos demais veículos de mídia, diversas outras fontes de dados contribuem para que a quantidade de informação armazenada nos dias de hoje tenha alcançado um nível inimaginável até poucos anos atrás. Imagens hiper-espectrais, dados financeiros de alta freqüência, mapeamentos genéticos e bases cadastrais são apenas alguns exemplos de como o mundo científico e o dos negócios funcionam como autênticas indústrias de dados.

A quantidade de material impresso no mundo de hoje está estimada em 5 EB. Até o ano 2000, tudo o que o ser humano havia produzido em mídia (incluindo áudio, vídeo e texto) somava cerca de 12 EB. Os experimentos em física de altas energias levados a cabo no CERN vão gerar, durante os próximos 10 anos, 60 TB/s [5]. No ano de 2002, foram criados aproximadamente 5 EB de espaço para armazenamento de dados, 92% em forma magnética. Finalmente, já houve quem estimasse em 5 EB o total das palavras pronunciadas por todos os homens de todos os tempos [6].

Mesmo com prováveis imprecisões nessas estimativas, é inegável que a enorme e crescente quantidade de dados existentes no mundo de hoje requer uma atenção proporcional à sua qualidade, sob pena de se converterem em problema, em vez de solução.

2.1.3 Impacto Negativo da Má Qualidade de Dados

O impacto econômico-social que pode causar um mau gerenciamento da qualidade de dados é muito significativo. De fato, cada dado errado, incompleto, inconsistente, enfim, com má qualidade, é um fator potencial de desperdício, decisões erradas, perdas de oportunidade e, em parte das vezes, de danos diretos à pessoa humana.

Do ponto de vista econômico, há pesquisas que tentam quantificar o custo-benefício das ações na área de QD, ou seja, procuram verificar se os benefícios financeiros provenientes de se ter dados com melhor qualidade compensam os investimentos feitos na melhoria da qualidade [7][8]. Esta análise foge ao escopo desta

tese. Claro está, também, que na maioria das vezes, os benefícios da QD transcendem o plano econômico, não sendo alvo portanto de análises deste tipo.

Nos EUA, a má qualidade dos dados de clientes já custava ao setor econômico, na virada do milênio, mais de 600 bilhões de dólares (cerca de 13% do faturamento total) por ano [2] (somente em gastos excedentes com postagem, impressão etc); entre 50% e 80% dos registros computadorizados de crimes cometidos nos EUA estavam inacurados, incompletos ou ambíguos [9]. Ao mesmo tempo, cerca de 50% das empresas cujo porte justificaria um maior controle da qualidade de seus dados não dispunham, àquela altura, de planos de gerenciamento ou melhoria de QD.

Na área de gestão da saúde, o mau gerenciamento das bases de dados envolvidas pode trazer algumas conseqüências graves, chegando até mesmo à morte de pacientes através de, por exemplo, erros na administração de medicamentos, no registro e/ou leitura dos resultados de exames e prontuários etc. Foi também amplamente divulgado o caso do Dr. Harold Shipman (conhecido depois como o "Dr. Morte"), responsável pela morte de 215 dos seus pacientes entre os anos de 1977 e 1998. Estudos publicados em [10][11] mostram, através de técnicas de análise de risco, que a monitoração da série do número de mortos dentre os pacientes do Dr. Shipman poderia ter levado à conclusão de que havia mortes em excesso, com o que, possivelmente, se poderia deter sua ação antes de 1998.

Estes exemplos nos fazem entender porque os dados têm hoje um valor crítico em diversos campos da sociedade, e que a qualidade dos dados de uma determinada instituição (seja acadêmica, comercial ou governamental) pode ser um bom indicador do seu desempenho atual e do seu sucesso futuro. De fato, estamos entrando na "era dos dados" e nossas bases de dados devem ser tratadas como patrimônios reais.

2.1.4 Aspecto Multidisciplinar

Uma das características que mais chama a atenção quando se estuda a Qualidade de Dados (QD) é a sua intrínseca multidisciplinariedade. Esta, por sua vez, está manifestada de três formas principais:

- *A origem diversa dos dados.* Com efeito, apesar de ter, atualmente, mais aplicações nas áreas econômica e empresarial, a teoria de QD não se restringe a um determinado tipo de dados, podendo ser aplicada a bases de dados

que guardem informações de fenômenos bastante distintos, sejam eles físicos (climáticos, geológicos etc), econômicos (macro/micro), médicos (informações do paciente, dados epidemiológicos, administração hospitalar etc), empresariais (dados da indústria, dos clientes, do comércio) ou de qualquer outra origem;

- *Inter-relação Academia-Indústria.* As motivações iniciais para o desenvolvimento de ferramentas de QD vinham das necessidades do mundo empresarial-industrial; entretanto, o ambiente acadêmico logo assumiu um papel fundamental na pesquisa de novos métodos para análise de QD, e os próprios dados de instituições acadêmicas passaram a ser objeto dessa análise;

- *Técnicas provenientes de diversas áreas.* Quando se desenvolvem os métodos para a análise de QD, as técnicas utilizadas costumam vir de diversas áreas do conhecimento, notadamente a Estatística, a Engenharia (em seus diversos ramos), a Informática e a Economia.

2.1.5 Um Programa de Referência

Criado dentro do MIT (*Massachusetts Institute of Technology*) no início da década de 90, o TDQM (*Total Data Quality Management Program*) reúne alguns dos principais pesquisadores da área de Qualidade de Dados. Atualmente, é dirigido pelo Prof. Richard Wang. Tem como principal objetivo "estabelecer uma sólida base teórica neste campo ainda embrionário e, a partir deste trabalho, planejar métodos práticos para que o mundo da indústria e dos negócios melhore a sua Qualidade de Dados" [12]. Deseja também "desenvolver ferramentas e outras capacidades necessárias para o gerenciamento da Qualidade de Dados nas fases técnica, econômica e organizacional das operações de negócios".

No seu sítio na *Internet* (`http://web.mit.edu/tdqm/`) estão reunidos os principais tópicos de interesse para os pesquisadores da área, como, por exemplo, coleções de artigos e livros já publicados, divulgação de eventos, publicações do próprio TDQM, *links* para outros fornecedores de material sobre QD e o sítio da ICIQ (*International Conference on Information Quality*), uma das principais conferências internacionais da área, organizada anualmente, desde 1996, pelo próprio

MIT.

Um serviço bastante útil recentemente disponibilizado pelo TDQM é o agrupamento das teses de doutorado em Qualidade de Dados, produzidas em todo o mundo, o que o torna um importante meio de atualização e intercâmbio para os estudantes e pesquisadores deste assunto.

2.1.6 Dados ou Informação?

Para encerrar esta seção, convém esclarecer uma questão bastante recorrente quando se fala de Qualidade de Dados: trata-se da diferença entre "dados" e "informação". Apesar de haver uma diferença conceitual intrínseca (pode-se dizer que "os dados são uma representação formal de um conjunto de informações, que podem estar ou não aparentes na base de dados", "informações são dados dotados de significado dentro de um contexto" [13], ou ainda "informações são dados processados" [14]), os termos Qualidade de Dados e Qualidade da Informação são normalmente utilizados para significar a mesma realidade. Quanto a nós, privilegiaremos o uso do termo Qualidade de Dados (QD). Nas referências pesquisadas, apenas em [15][16][17] omite-se o termo Qualidade de Dados. Com o decorrer dos anos, o uso do termo "dados" se estabeleceu entre os pesquisadores da área.

2.2 Conceitos e Definições

Tendo já exposto uma visão geral do tema "Qualidade de Dados", convém agora formalizar alguns conceitos e definições.

2.2.1 O Que é Qualidade de Dados?

Responder a essa pergunta é algo bastante semelhante a responder: "Quão boa é a qualidade dos dados de uma determinada instituição?". E, naturalmente, esta pergunta pode ser formulada sob diversos aspectos. Com efeito, um dado com qualidade não significa apenas um dado livre de erros, pois dados incorretos (inacurados) respondem apenas por uma parcela da equação da qualidade de dados [2][18]. Afinal, para que serviriam dados livres de erro se, por exemplo, não se lhes pudesse compreender e acessar de maneira adequada?

Uma antiga definição, derivada do senso comum e amplamente aceita, diz que a qualidade dos dados refere-se à sua "conformidade com as especificações" [19]. No entanto, a definição de Qualidade de Dados não costuma ser dada em uma fórmula única. As referências mais comumente citadas para definir a QD expõem uma série de características que dados de boa qualidade devem conter. Na seqüência, analisaremos essas características.

2.2.1.1 Dependência do Contexto

Uma característica que se sobressai, quando analisamos a qualidade de dados de uma determinada base ou conjunto, é o fato de que esta tarefa é bastante dependente do contexto em que se está inserido. Com efeito, "conformidade com as especificações" deixa margem para que um mesmo dado possa ser considerado de boa qualidade em um determinado contexto, e de má qualidade em outro.

Exemplos bastante elucidativos estão expressos em [20], para diversas dimensões da QD. Detenhamo-nos na análise da acurácia de uma base de dados. A pergunta que se pode fazer inicialmente é: "Que nível de acurácia deve haver para considerarmos de boa qualidade uma base de dados?". A resposta é: "depende da função que se quer dar a essa base de dados". Por exemplo, para o caso de uma base de dados contendo cadastros (nomes, endereços) dos médicos de uma determinada região, 85% de acurácia nos dados pode ser considerado bom quando se quer utilizá-la para fazer uma "mala direta" a respeito de um novo produto da área médica; entretanto, esse mesmo nível de 85%, para essa mesma base de dados, deve ser considerado baixo se o uso que se quer dar a essa base é avisar os médicos de mudanças na legislação para os profissionais da medicina.

Portanto, não basta nem mesmo definir a base de dados; para que se possa fazer um julgamento da sua qualidade, é necessário, também, ter claros qual é a sua função e o seu uso.

2.2.1.2 Multidimensionalidade

Imaginemos que a base de dados de que tratamos no exemplo anterior tivesse uma boa acurácia, mas, ao mesmo tempo, os seus dados estivessem guardados em um computador isolado, ou seja, sem acesso por rede. Ora, para alguém que não

pudesse deslocar-se até esse computador, essa base se tornaria de má qualidade, devido à sua baixa acessibilidade, que é uma outra dimensão da QD em geral. De uma maneira geral, é importante verificar diversos aspectos da base de dados antes de julgar a sua qualidade. Isso reflete a característica da multidimensionalidade da QD, ou seja, ela não pode ser medida por apenas um parâmetro [21]. O número de dimensões necessárias para expressar satisfatoriamente a QD de uma base varia de acordo com cada caso. Muitas vezes, falhas em algum desses aspectos podem comprometer toda a qualidade da base, por inviabilizar o seu uso. Mais adiante voltaremos a este assunto, quando tratarmos das dimensões da QD.

2.2.1.3 Objetividade e Subjetividade

Quando se trata de aferir a qualidade dos dados de uma instituição, logo se percebe que nem todas as dimensões podem ser medidas de forma objetiva. Muitas dizem respeito não tanto às medidas que se realizam diretamente sobre os dados, mas sim a percepções subjetivas das pessoas que lidam com eles. Também essas dimensões subjetivas devem ser levadas em conta à hora de analisar a QD de uma base [1].

Enquanto as dimensões objetivas costumam ser aferidas a partir dos próprios dados, as dimensões subjetivas, em geral, refletem as necessidades e experiências dos que, de alguma forma, estão envolvidos com os dados (coletores, mantenedores e usuários dos dados)[1] [22]. Uma das maneiras de quantificá-las é através de questionários aplicados a essas pessoas. É importante o aferimento dessas dimensões subjetivas, porque, caso se descubra que a percepção que eles têm, com relação à base de dados, é negativa, é muito provável que também o seu comportamento esteja sendo influenciado negativamente[2]. De fato, o conceito de *qualidade* deve refletir todos os aspectos do produto (no caso, o dado) que afete as opiniões dos usuários em quão bem o produto (dado) corresponde às suas expectativas [23]. Assim sendo, a qualidade de dados deve medir cada grandeza envolvida na relação dado-usuário.

[1]No termo original em inglês, os *stakeholders*.

[2]Por exemplo, se a percepção que os usuários de um determinado sítio da *Internet* têm com relação aos seus dados é negativa, é bastante provável que eles o estejam acessando cada vez com menor freqüência.

Uma proposta feita em [1] para combinar as dimensões subjetivas e objetivas da QD consiste em definir *parâmetros* e *indicadores*. Os *parâmetros* de QD correspondem às dimensões *subjetivas* ou *qualitativas* que acabamos de analisar. Exemplos de parâmetros de QD são a *credibilidade da fonte* e a *significância* (relevância) do conjunto de dados. Os *indicadores* de QD são dimensões extraídas dos próprios dados que fornecem informações *objetivas* sobre o conjunto de dados analisado. A *completude* e o *método de aquisição* são exemplos de indicadores.

Parâmetros e *indicadores* reunidos formam os *atributos*, conforme ilustrado na Figura 2.1. Com eles, pode-se construir uma métrica (conforme veremos a seguir) capaz de nos informar sobre o nível da QD de uma determinada base [14].

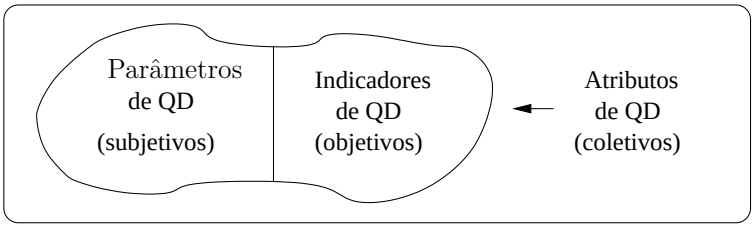

Figura 2.1: Formação dos atributos de QD.

2.2.1.4 Métrica para Qualidade de Dados

Há pouco falávamos sobre a multidimensionalidade inerente ao conceito de Qualidade de Dados, para depois classificar essas dimensões em subjetivas (parâmetros) e objetivas (indicadores). O passo seguinte é definir uma *métrica* para a avaliação da qualidade dos dados de uma base específica. Essa métrica, naturalmente, é construída a partir das dimensões de QD. Até hoje, já são cerca de 200 as dimensões mencionadas na literatura da área. Algumas, porém, são usadas com mais freqüência e constituem em núcleo a partir do qual se podem definir métricas para quaisquer bases de dados reais. É compreensível que surjam, com o tempo, novas dimensões de QD que podem ser incorporadas às métricas existentes. De fato, costuma-se dizer que há pelo menos tantas dimensões de QD quantas são as *ilities*[3]. Logicamente, para cada caso específico escolhem-se as dimensões pertinentes para o aferimento da QD.

[3]Alusão às palavras que, em língua inglesa, são terminadas em *-ility*. Em português, corresponderiam a "ilidades", tais como acessibilidade, confiabilidade, interpretabilidade etc.

A seguir, listamos as principais dimensões da QD, tal como são definidas em [1]–[25]. Indicamos também a sua objetividade ou subjetividade.

- *Acessibilidade.* O quão disponível está o dado, e o quão rapidamente ele pode ser acessado. Dimensão objetiva;
- *Quantidade apropriada de dados.* O volume de dados é suficiente para o que se pretende fazer? Dimensão subjetiva;
- *Credibilidade da fonte.* Até que ponto a origem dos dados é confiável? Dimensão subjetiva;
- *Completude.* Mede a quantidade de dados faltantes (*missing data*) e também se as dimensões da base de dados estão devidamente representadas. Dimensão objetiva;
- *Concisão.* A representação dos dados é compacta ou muito redundante? Dimensão objetiva;
- *Consistência.* Verifica se os dados são apresentados no mesmo formato. Dimensão objetiva;
- *Facilidade de manipulação.* Os dados são facilmente aplicáveis a tarefas distintas da que se tem em mãos? Dimensão subjetiva;
- *Acurácia.* Mede a correção dos dados, a sua conformidade às especificações de momento. Dimensão objetiva;
- *Interpretabilidade.* Considera até que ponto a linguagem, símbolos e unidades são apropriados, além de medir a clareza das definições utilizadas. Dimensão subjetiva;
- *Objetividade.* Os dados possuem algum viés? São totalmente imparciais? Dimensão subjetiva;
- *Relevância.* Os dados possuem aplicabilidade? São úteis para a aplicação em questão? Dimensão subjetiva;

- *Reputação.* Semelhante à credibilidade da fonte, mede a opinião geral sobre a veracidade do conteúdo da base. Dimensão subjetiva;

- *Segurança.* O acesso a base é restrito aos usuários com permissão? Que mecanismos controlam a sua segurança? Dimensão objetiva;

- *Inteligibilidade.* Mede a facilidade do usuário em compreender o conteúdo da base; Dimensão subjetiva;

- *Valor agregado.* Que benefícios pode trazer o uso de tais dados? Dimensão subjetiva;

- *Pontualidade.* Mede o quão recente é a base, com relação à última versão possível para tais dados. É o equivalente à completude, para o caso de dados temporais, como a maior parte das séries. Dimensão objetiva;

- *Método de aquisição.* Como são adquiridos os dados? Qual é o caminho da fonte à base de dados? Dimensão subjetiva;

A escolha das dimensões que farão parte de uma determinada métrica deve ser feita segundo o enfoque *data driven*, ou seja, em função dos dados cuja qualidade se deseja medir. É difícil definir nesta etapa um procedimento automático ou uma regra fixa. Uma vez definidos os parâmetros e indicadores que vão formar a métrica de QD para uma determinada base, é conveniente agrupá-los em alguma estrutura que expresse de forma condensada em que nível está a QD da base. A maneira de agrupá-los é também uma questão deixada aos analistas de QD, dentro do contexto da estratégia de QD [2][26].

Uma maneira que nos é conveniente é a formação de um vetor cujos elementos sejam os valores das métricas utilizadas, ou seja, os valores dos atributos. Os elementos do vetor devem admitir, entretanto, natureza mista, dado que os atributos podem ser qualitativos (os parâmetros) ou quantitativos (os indicadores). Assim, por exemplo, para um parâmetro como a credibilidade da fonte, podemos ter os valores "baixo", "médio" e "alto" (a menos que sejam depois quantificados, conforme descrito em 2.2.1.3), enquanto que para um indicador como a completude teremos um valor numérico que esteja dentro de um intervalo pré-determinado. Em geral, este intervalo é normalizado para [0;1]. Um vetor de métrica típico, portanto, tem

a seguinte forma: [0,8 'médio' 'alto' 1 0,4], correspondendo aos atributos acurácia, credibilidade da fonte, relevância, completude e precisão.

Outra consideração a ser feita com relação à métrica de QD de uma base diz respeito à independência do método de análise de cada atributo. A maneira de medir a acurácia é claramente distinta da maneira de medir a completude. Para quantificar esta última, poder-se-ia, numa primeira aproximação, tomar a razão entre campos preenchidos e o número total de campos de uma tabela, ao passo que para a acurácia seria necessário traçar uma distribuição dos erros das variáveis envolvidas e daí extrair o valor da métrica. Voltaremos a tratar da medição dos atributos no Capítulo 4.

2.2.2 Como um Dado Perde Qualidade?

Quando se trata de encontrar causas para possíveis inconsistências nos dados, há que se considerar que existem tantas maneiras de um dado estar "inacurado" quantas são as pessoas que o manipulam ou fases por que ele passa. Há aqui uma analogia com uma linha de produção: pode-se introduzir uma falha num produto a cada vez que ele é manuseado ou passa por algum processamento automático. Assim, para o caso dos dados, desde a sua aquisição (incluindo uma possivel medição), passando pelas diversas fases de transmissão (que podem requerer intervenção humana), registro, armazenamento etc, podem surgir falhas nos dados. Logicamente, sempre que possível deve-se evitar que tais falhas se propaguem e, por isso, é interessante que a proposta de um sistema de monitoração de qualidade de dados inclua uma inspeção contínua que analise os dados nas diversas fases da construção da respectiva base. Tal sistema deve levar em consideração quais são os possíveis problemas da base de dados da instituição em questão (empresa, universidade etc), a sua freqüência e os impactos que podem causar. Isto requer uma metodologia que utilize um enfoque direcionado pelos próprios dados (enfoque *data driven*) na avaliação da QD.

2.3 Áreas de Aplicação

Teoricamente, a aplicação de princípios de QD é indicada para qualquer instituição que possua uma base de dados e a utilize de alguma forma. Hoje em dia, é difícil imaginar uma organização que não dependa de seus dados. Cabe a cada uma julgar o melhor momento para começar um programa de QD, e em que nível o fará.

Todas as áreas do conhecimento têm se convertido, cada vez de forma mais acentuada, em geradoras de grandes "massas de dados", termo corrente que indica o armazenamento da informação contida nas atividades de empresas, centros de pesquisa, indústrias, universidades etc. É natural, pois, que tais dados apresentem natureza bastante heterogênea, e que a monitoração da sua qualidade tenha impactos em diversos setores, entre eles o tecnológico, o econômico e o social.

Do ponto de vista tecnológico, sabe-se que a geração de novos conhecimentos, a partir de relações existentes entre os dados analisados, estimula o avanço da indústria, que também se beneficia da maior "pureza" dos dados [2].

O setor econômico é hoje em dia, sem dúvida, bastante influenciado pela análise qualitativa de dados. Uma parcela dessa influência se dá no mercado financeiro, com a chamada "análise técnica" [27] do mercado de ações, em que se busca traçar horizontes de ação futura com base na predição de tendências para as séries das empresas envolvidas. Outro exemplo é a aplicação de ferramentas de QD em bases cadastrais de empresas, o que, em geral, resulta em grande diminuição das perdas decorrentes da aplicação de dados "sujos" [28]. Também tem se desenvolvido bastante o estudo da valoração de intangíveis, patrimônios que não podem ser mensurados diretamente através de uma quantidade material, mas que muitas vezes são decisivos na avaliação de uma empresa, como, por exemplo, o valor da marca, a imagem de um produto etc. Tais intangíveis devem ser avaliados a partir de alguns dados, cuja qualidade é então fundamental para a sua correta valoração.

A tendência atual é integrar todos os procedimentos relacionados à Inteligência de Negócios (*Business Intelligence* ou BI), como o gerenciamento das relações com o cliente (*Customer Relationship Management* ou CRM) e o gerenciamento das cadeias de provisão (*Supply Chain Management* ou SCM). Esta integração se dá através dos grandes repositórios de dados conhecidos como *Data Warehouses* (DW). Muitas vezes, as análises de QD se dão dentro do próprio DW. Segundo a nossa con-

cepção, essa em geral não é a melhor alternativa, conforme explicaremos quando da descrição do sistema de monitoração proposto neste trabalho.

O impacto social da análise de QD, apesar de ser menos aparente, é também extremamente importante. A razão primeira disto é que os dados estão sempre relacionados a pessoas, e muitas vezes a correção de falhas em dados ou simplesmente uma melhor compreensão do seu conteúdo possibilitam uma melhora qualitativa no dia-a-dia de muitos. A análise de QD contribui também para a descrição de padrões de comportamento (através da descoberta de algumas distribuições estatísticas), ajudando a descobrir problemas ou carências estruturais. Além disso, não se deve esquecer que o homem – o maior "gerador de dados" – possui um agir livre e é o destinatário último de todo o processo de QD. Portanto, não se deve fazer uma análise "fria" (meramente técnica) dos dados em geral. Nos sistemas de monitoração de QD, isso se dá com a figura do "supervisor", que pode intervir e tornar não-automática as decisões do sistema de monitoração da qualidade de dados.

Poderíamos citar ainda o auxílio da análise de QD nas ciências biomédicas. O Projeto Genoma [29], por exemplo, só foi possível devido à contínua monitoração da qualidade da informação gerada. Também no ramo da administração hospitalar, o gerenciamento inteligente de dados ajuda no cruzamento de informações, agilização de processos, além de minimizar a possibilidade de ocorrência de erros médicos ao garantir aos profissionais dados com maior qualidade. Nas pesquisas atuais com biodiversidade, a análise de QD presta grande apoio às áreas de taxonomia e museologia [30].

2.4 Gerenciamento de Qualidade de Dados

Nesta tese, ao estudarmos o tema da Qualidade de Dados, temos por objetivo propor um mecanismo que seja capaz de atuar no gerenciamento da QD de distintas bases, um sistema de monitoração responsável por manter a QD no patamar desejado.

São diversas as razões pelas quais vale a pena investir esforços para o desenvolvimento de ferramentas de QD. Analisemos aqui algumas delas:

- *As bases de dados poucas vezes são tratadas de forma sistemática e inteligente.*

De fato, o crescimento da indústria, sobretudo, mas também o da pesquisa científica, sempre foi muito dirigido pelo fator tecnológico, com as bases de dados ocupando invariavelmente uma posição secundária. Isto na prática significava que as bases de dados eram tratadas como um bloco estático, do qual se podiam extrair informações, mas no qual não se podia mexer. Desta forma, os possíveis erros existentes nas bases nunca eram detectados, e muitas vezes se propagavam ao longo da cadeia produtiva;

- *Praticamente não há métodos prontos para uma análise de dados consistente.* Ainda hoje, é muito difícil encontrar sistemas desenvolvidos para uma análise mais aprofundada da QD de bases de dados. Também, pelo fato de que cada base apresenta suas peculiaridades, o que se encontra atualmente são ferramentas específicas para determinados tipos de problemas. Mesmo essas ferramentas, como por exemplo o SAS [31], ainda têm um custo que as torna inacessíveis a grande parte das instituições que desejam monitorar a sua QD;

- *Novas tecnologias podem vir dos procedimentos desenvolvidos.* Longe de ser deixado de lado, devido a uma maior atenção dada à qualidade dos dados, o crescimento tecnológico se beneficia dos novos conhecimentos provenientes das análises descritiva e exploratória de grandes bases de dados;

- *O tratamento dos dados resulta geralmente em grandes aumentos para o retorno nos investimentos.* Dados tratados evitam perdas com o erro no foco e o mau dimensionamento dos investimentos, além de ampliar os horizontes de ação de uma determinada instituição. Além disso, assim como na linha de produção de uma indústria, é importante que o fator qualidade seja considerado o quanto antes para que sejam minimizadas as perdas. Na construção de uma base de dados, quando se prioriza desde o início a qualidade, podem-se esperar maiores retornos.

2.5 Qualidade de Dados, Estatística e Processamento de Sinais

Conforme visto na Subseção 2.1.4, a multidisciplinariedade é algo inerente à pesquisa em Qualidade de Dados. Um dos objetivos desta tese é selecionar algumas técnicas comuns à área de processamento de sinais na análise de QD, em auxílio a técnicas estatísticas tradicionalmente empregadas.

Concretamente, técnicas provenientes das Ciências Estatísticas e das áreas de Processamento de Sinais são combinadas no sistema de monitoração que propomos, de modo a contribuir para a análise e melhoria da qualidade de dados de bases diversas. Nas subseções seguintes, veremos alguns aspectos de como essas três áreas – Qualidade de Dados, Estatística e Processamento de Sinais – podem se relacionar.

2.5.1 Qualidade de Dados e Estatística

Toda a *análise de dados*, tradicionalmente, tem sido área de estudo da estatística, seja na parte descritiva, seja na analítica. Com o advento da pesquisa em Qualidade de Dados, era natural que fosse a estatística a primeira área procurada na busca de métodos eficientes para se fazer uma análise consistente da Qualidade de Dados. De fato, em não raras vezes o termo Qualidade de Dados pode ser trocado por Qualidade *das Estatísticas* [23]. Com isso, as primeiras técnicas utilizadas na medição[4] de QD baseavam-se em métodos estatísticos tradicionais [32]. Ainda hoje, o peso da análise estatística nesta área é bastante forte, e não podemos prescindir totalmente da sua colaboração, se queremos desenvolver um sistema robusto que atue sobre bases grandes e diversificadas. Assim sendo, áreas como análise econométrica e análise multidimensional de dados, entre outras, têm peso bastante grande no desenvolvimento de tais sistemas.

2.5.2 Qualidade de Dados e Processamento de Sinais

A área de Processamento de Sinais tem enorme aplicabilidade, visto que os sinais que manipula representam uma enorme gama de fenômenos físicos e econômicos.

[4]Feita de maneira *ad hoc*, não sistemática.

Tal como a área de Qualidade de Dados, lida com informações provenientes dos mais diversos fenômenos. Assim, a ligação entre as duas áreas pareceu-nos uma extensão natural de ambas, em que a QD se beneficia da eficiência das técnicas de Processamento de Sinais.

De fato, percebemos que os sinais a que nos referimos sempre podem ser entendidos como dados e, da mesma forma, uma determinada base de dados pode ser compreendida como sinais a serem analisados, ainda que não tenham domínio temporal ou tenham formato complexo (o que implicaria na formação de vários sinais a partir de uma mesma base). De fato, a teoria clássica diz que sinal, de uma forma bem ampla, é "aquilo que carrega informação" [33] e, conforme visto em 2.1.6, um dado é a representação formal de uma informação.

Uma das principais tarefas de processamento de sinais, a filtragem, é essencialmente um procedimento relacionado à Qualidade de Dados. Sob a ótica adotada neste trabalho, enxergamos um filtro como uma ferramenta capaz de detectar e retirar irregularidades (por exemplo, ruído) de um sinal (dados), ou mesmo destacar informações contidas nos dados, o que naturalmente faz com que a sua qualidade seja melhorada. Conforme destacado em [34], Processamento de Sinais aproxima-se também bastante de um campo importante para a Qualidade de Dados: a análise de séries temporais. De fato, tarefas como o reconhecimento de voz ou a detecção de padrões em eletrocardiogramas, por exemplo, lidam com sinais temporais, processados em geral por alguma técnica de filtragem e em que o fator qualidade está intimamente inserido. Como exemplo, a filtragem de Kalman é utilizada neste trabalho como meio de pré-processamento na modelagem de alguns tipos de séries temporais. Desta forma, torna-se imediato vislumbrar o uso de Processamento de Sinais na análise de séries temporais que representem fenômenos econômicos ou sociais, tipos de dados bastante típicos em grandes bases.

As Redes Neurais Artificiais (RNAs) são outro exemplo de ligação entre as áreas de Processamento de Sinais e Qualidade de Dados, pois constituem um conjunto de técnicas que podem ser aplicadas em problemas bastante diversos, abrangendo bases de dados de naturezas distintas. As RNAs podem ser classificadas, em sentido amplo, como filtros adaptativos não-lineares, o que as torna aptas para serem aplicadas na resolução de uma ampla gama de problemas, tais como reconhecimento

de padrões (classificação), predição e agrupamento [35]. Com efeito, na bibliografia recente podem ser encontrados alguns artigos que relacionam áreas relativas a QD e a Processamento de Sinais, como por exemplo [34].

Capítulo 3

Análise Descritiva dos Dados

Este capítulo será dedicado à análise dos dados que foram estudados nesta tese. Para efeito da descrição, dividiremos os dados em dois grandes blocos, a saber: séries temporais e registros multivariados. Para cada um deles, serão vistos os principais aspectos teóricos relacionados, além de exemplos de dados reais utilizados nesta tese.

3.1 Tópicos em Séries Temporais

Grande parte dos dados estudados nesta tese estão agrupados sob a forma de séries temporais. Nesta seção pretendemos fornecer alguns subsídios fundamentais para a modelagem desse tipo de série. Sendo verdade que o modelo escolhido para a representação de uma série temporal é influenciado pela natureza do fenômeno que a gerou, também é válido dizer que há determinadas características que devem ser verificadas em todos os tipos de séries temporais. Dedicaremos esta seção à análise destes aspectos comuns, com o objetivo de poder descrever, no final, um procedimento para a análise estrutural de séries temporais.

3.1.1 Estacionariedade em Séries Temporais

Boa parte dos procedimentos de análise estatística de séries temporais supõe que estas sejam estacionárias[1]. Desta forma, é fundamental verificar a estacionarie-

[1]Modelos estruturais (*Harvey*) e modelos de cointegração não fazem essa suposição. Porém, nos deteremos nos métodos que pressupõem a estacionariedade.

dade das séries que se deseja modelar, assim como definir um procedimento para tornar estacionárias as que não o sejam previamente.

A seguir, será definido o conceito de série temporal estacionária. Em seguida, alguns testes para detectar não-estacionariedade em séries temporais serão descritos. Finalmente, serão apresentados os procedimentos para transformar séries temporais não-estacionárias em estacionárias.

3.1.1.1 Conceito de Estacionariedade

Do ponto de vista matemático, o conceito de estacionariedade é definido inicialmente para *processos estocásticos*. Um processo estocástico pode ser definido como uma família Z de variáveis aleatórias, ou seja, $Z = \{z(t), t \in T\}$, tal que, para cada $t \in T$, $z(t)$ é uma variável aleatória, sendo T um conjunto arbitrário [36]. As variáveis aleatórias de um processo estocástico não são necessariamente função do tempo; porém, como estamos interessados na análise de séries temporais, nos referiremos sempre a processos que evoluem no tempo.

O conceito de estacionariedade para processos estocásticos admite duas formas: estacionariedade fraca (ou no sentido amplo) e estacionariedade forte (ou no sentido estrito). Para a extensão do conceito de estacionariedade que se vai fazer para o caso de séries temporais, basta-nos a definição da estacionariedade em sentido amplo [37]:

Definição 1: Um processo estocástico $Z = \{z(t), t \in T\}$ é dito estacionário em sentido amplo se e somente se:

i $E\{z(t)\} = \mu(t) = \mu$, constante, para todo $t \in T$;

ii $Var\{z(t)\} = \sigma^2(t) = \sigma^2$, para todo $t \in T$;

iii $\gamma(t_1, t_2) = Cov\{z(t_1), z(t_2)\}$ é uma função de $|t_1 - t_2|$.

A partir de agora, chamaremos simplesmente de *estacionários* os processos que obedeçam a essas três condições, ou seja, que tenham média e variância iguais para todos os instantes de tempo, e em que a covariância entre duas variáveis quaisquer do processo seja função apenas da distância temporal entre elas.

Para que se possa estender o conceito de estacionariedade para séries temporais, consideramos que cada série temporal em particular é uma *realização* de

um dado processo estocástico. Uma *realização* de um processo estocástico é obtida quando cada uma das variáveis aleatórias $Z(t)$ que compõem o processo assume um determinado valor, seguindo a sua função densidade de probabilidade própria.

Sendo assim, diz-se que uma série temporal é estacionária quando pode ser considerada a realização de um processo estocástico estacionário. Neste momento, nos deparamos com a seguinte dificuldade: quase sempre, é impossível gerar outras realizações para o mesmo processo estocástico que gerou a série temporal que se tem em mãos, e esta constitui então a única informação disponível para fazermos inferências sobre o processo gerador, que por sua vez precisa ter a sua estacionariedade assegurada.

O procedimento adotado busca então encontrar algum fator de não-estacionariedade neste processo, a partir de testes estatísticos pertinentes. Se pelo menos algum destes testes indicar não-estacionariedade, faz-se necessário aplicar as transformações convenientes para tornar a série estacionária, como será visto a seguir. Caso nenhum teste aponte não-estacionariedade, a série (assim como a série convenientemente transformada), poderá ser sempre considerada como a realização de um processo estocástico estacionário, porque, usando o conceito de *ergodicidade*, pode-se garantir que existe um processo estocástico estacionário capaz de gerar a série temporal analisada [36].

3.1.1.2 Detecção da Não-Estacionariedade

Conforme visto na subseção anterior, o procedimento adotado para verificar a estacionariedade de uma série temporal consiste na aplicação de testes que detectem alguma não-estacionariedade, ou seja, que mostrem que a série não atende a pelo menos uma das três condições expostas em 3.1.1.1.

A seguir, veremos os principais testes comumente adotados para séries temporais univariadas.

Detecção de Tendência Primeiramente, verifica-se se a série temporal analisada possui alguma *tendência*, o que significa (no caso de tendências determinísticas, conforme veremos adiante) um efeito de longo prazo na média, anulando a condição **(i)** de média constante para todos os instantes de tempo.

Essa tendência pode estar presente de duas maneiras: *estocástica* ou *determinística* [36]. A correta detecção do tipo de tendência é bastante importante para o sucesso da modelagem de uma série temporal. Por exemplo, tomar como determinística uma tendência que, na verdade, é estocástica, leva em geral a resultados espúrios na predição dos valores futuros da série.

A presença de tendência estocástica é detectada através dos chamados *testes de raiz unitária*, conforme será visto a seguir. Na seqüência, abordaremos o tema da remoção das tendências determinísticas.

Testes de Raiz Unitária Suponhamos um processo AR (Auto-Regressivo) de ordem 1, que pode ser modelado na forma $\rho(t) = \phi_0 + \phi_1\rho(t-1) + e(t)$. Se o valor de ϕ_1 for igual a 1, então diz-se que esse processo possui uma *raiz unitária*. A presença de uma raiz unitária gera uma tendência estocástica de grau 1 (diz-se, nesse caso, que o processo tem ordem de integração igual a 1), somente removida através do operador diferença $\nabla = 1 - B$, sendo B o operador atraso (isto é, $B[y(t)] = y(t-1)$). Assim, $\nabla[\rho(t)] = \rho(t) - \rho(t-1) = \phi_0 + e(t)$ para os casos em que $\phi_1 = 1$.

Um processo com raiz unitária em que $\phi_0 \neq 0$ nunca é estacionário, porque sua média não pode ser constante devido à presença de $\rho(t-1)$ na formação de $\rho(t)$. A amostra atual é sempre igual à anterior acrescentada de outro termo. Nesse caso, diz-se haver uma "tração" na série e, em geral, ela evolui para $\pm\infty$ (de acordo com o sinal de ϕ_0). A presença de uma raiz unitária insere uma forte *memória* no processo, fazendo com que a sua função de autocorrelação permaneça constante.

Um processo do tipo $y(t) = \mu + y(t-1) + r(t)$ é chamado de *passeio aleatório* nos casos em que $r(t)$ for, por exemplo, um ruído branco [36]. Um passeio aleatório é o caso típico de processo não-estacionário que se faz estacionário através da aplicação do operador ∇.

As Figuras 3.1 a 3.4 ilustram exemplos de passeios aleatórios, sem e com tração, e antes e depois de se aplicar o operador ∇, juntamente com as respectivas funções de autocorrelação amostrais (FAC). Pode-se observar que, para as séries não-estacionárias, ou seja, antes da aplicação de ∇, o correlograma decai muito lentamente, ao passo que, para as séries já estacionárias, todas as correlações de defasagem nula (lag=0) recaem dentro dos limites do intervalo de confiabilidade de 95% (representado pelas linhas horizontais no gráfico). Essas características serão

importantes no Capítulo 5, quando tivermos necessidade de verificar se algumas séries específicas podem ser modeladas como passeios aleatórios.

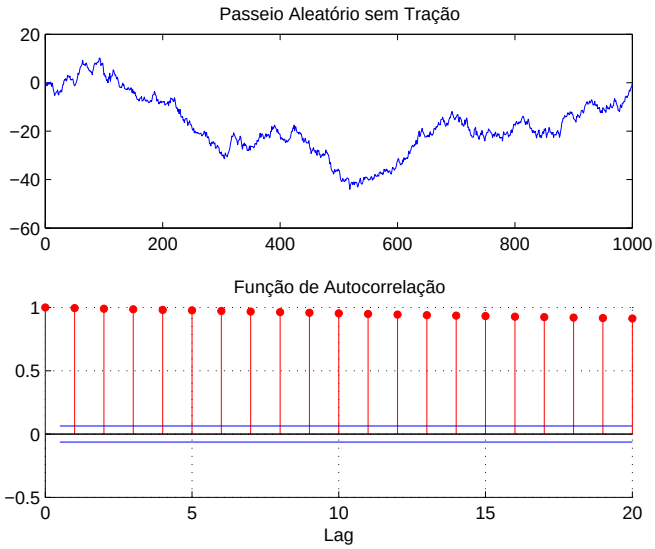

Figura 3.1: Passeio aleatório sem "tração" (acima) e FAC (abaixo).

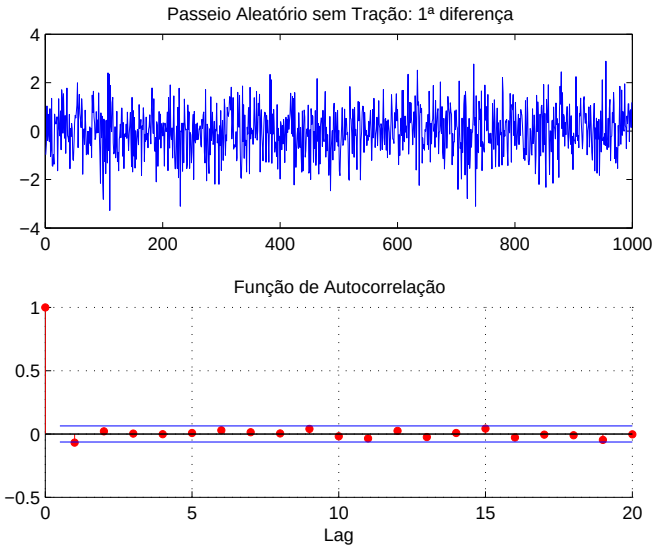

Figura 3.2: Primeira diferença do passeio aleatório sem "tração" (acima) e FAC (abaixo).

Serão agora brevemente descritos dois importantes testes para a detecção de raízes unitárias em séries temporais.

Teste ADF (Augmented Dickey-Fuller) O teste de Dickey-Fuller [38] é um teste de hipóteses econométrico que verifica a probabilidade de que haja uma

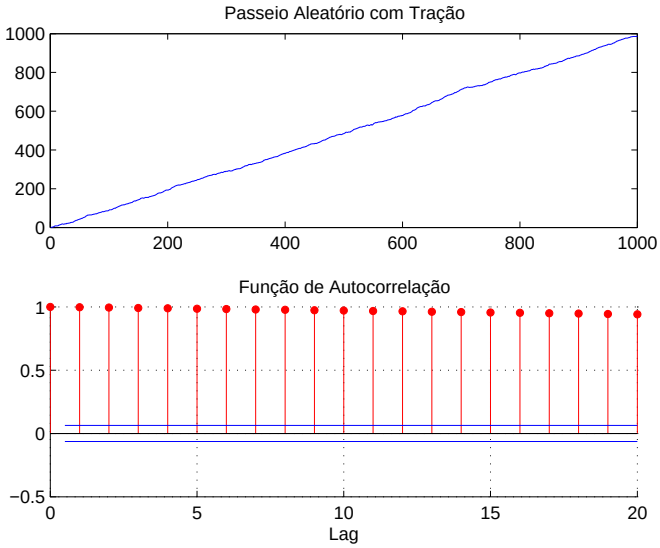

Figura 3.3: Passeio aleatório com "tração" (acima) e FAC (abaixo).

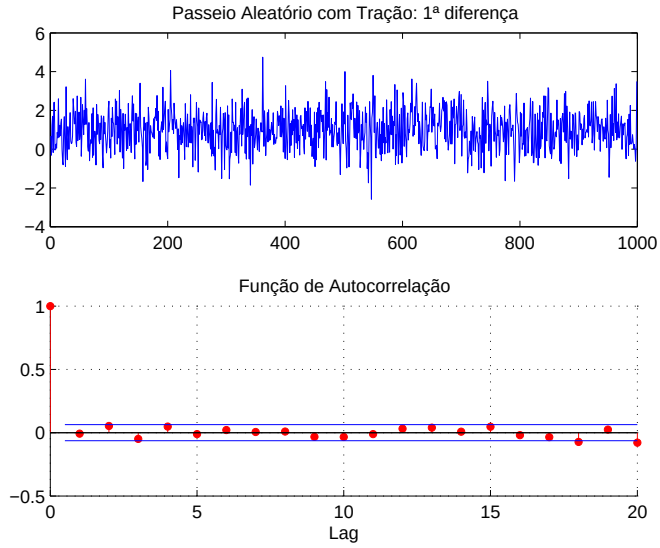

Figura 3.4: Primeira diferença do passeio aleatório com "tração" (acima) e FAC (abaixo).

raiz unitária (hipótese nula, ou H_0) na formação de uma dada série $\rho(t)$, isto é, se $\rho(t) = \phi_0 + \phi_1\rho(t-1) + e(t)$ tem $\phi_1 = 1$ a um determinado nível de confiabilidade. O teste aumentado (ADF) estende-se a modelos de séries temporais mais complexos, retirando a estrutura de autocorrelação da série antes de proceder o teste DF.

Vimos que um processo AR(1) genérico (com tração) pode ser escrito como $\rho(t) = \phi_0 + \phi_1\rho(t-1) + e(t)$. Subtraindo-se $\rho(t-1)$ de ambos os lados da equação chega-se a $\nabla[\rho(t)] = \phi_0 + (\phi_1 - 1)\rho(t-1) + e(t)$, ou ainda $\nabla[\rho(t)] = \phi_0 + \gamma\rho(t-1) + e(t)$ considerando-se $\gamma = \phi_1 - 1$.

O teste (A)DF tem como hipótese nula (H_0) que $\gamma = 0$, ou seja, que $\phi_1 = 1$. Nesse caso, o processo possui uma raiz unitária, tendo ordem de integração igual a 1 (processo $I(1)$). A hipóstese alternativa (H_a) restringe-se a que $\gamma < 0$, ou seja, o teste é unilateral à esquerda. Neste caso, os valores críticos (limiares de decisão para aceitação de H_0) são todos negativos. Se H_a fosse $\gamma \neq 0$, estaríamos permitindo $\gamma > 0$ ou $\phi_1 > 1$, o que levaria a um processo explosivo.

Este teste de hipóteses é então realizado a partir de uma regressão dos dados. A grande contribuição de Dickey e Fuller foi realizar a tabulação dos valores críticos para a distribuição da estatística τ.

O teste de Phillips-Perron [39][40] é bastante semelhante ao (A)DF, com a diferença de que inclui correções para permitir a presença de correlação e heteroscedasticidade na série sob teste.

Teste Combinado Nesta tese, utilizamos um teste combinado ADF-PP para a detecção de raízes unitárias. Através desse teste, calcula-se a ordem de integração n da série analisada. Dessa forma, é possível saber que precisamos tomar n diferenças da série para remover a sua tendência estocástica. Na quase totalidade dos casos, observa-se que $n \leq 1$. Note-se que, ao se tomar n diferenças de uma séries, reduz-se o tamanho da série em n amostras, que são posteriormente recuperadas quando da reconstrução da série ao término do seu processamento.

Este teste combinado[2] resulta na rejeição da hipótese de raiz unitária (hipótese nula) caso seja verificada pelos menos uma das duas seguintes condições:

[2]Agradecemos a Ludwig Kanzler (Boston College Department of Economics) a disponibilização de algumas funções em MATLAB® para a implementação do teste.

- a estatística τ do teste ADF tem significância estatística ($\sigma_\tau \leq 0.1$) e os resíduos da regressão não são correlacionados (caso contrário, a estatística é ineficiente);

- a estatística τ do teste PP tem significância estatística ($\sigma_\tau \leq 0.1$). Aqui não é necessário verificar que os resíduos não são correlacionados porque o teste PP já faz as correções necessárias.

Remoção da Tendência Determinística Se os testes de raiz unitária dão resultado negativo, podemos considerar que a tendência, se houver, é de natureza determinística.

A presença de uma tendência determinística significa que a série evolui sobre uma curva, quase sempre polinomial ou exponencial. Remove-se uma tendência deste tipo através de um ajuste dos dados a uma determinada curva parametrizada. A própria observação dos dados e o conhecimento prévio do fenômeno que gerou a série indicam a forma da possível tendência, que é ratificada pela qualidade do ajuste feito. Também, nesse caso, são raras as vezes em que nos deparamos com tendências não lineares.

Se uma dada série, para a qual já foi verificado que não há raízes unitárias, também não aparenta possuir nenhuma tendência determinística, não há inconveniente em considerar que há uma tendência linear, fazendo-se um ajuste dos dados da série para um polinômio de grau 1 (que não afetaria a série caso não houvesse nem mesmo uma tendência linear residual). A remoção desta reta terá, ao menos, o efeito de levar a média da série a zero, o que é conveniente para o posterior processamento da série, como será visto mais adiante.

Cabe agora comentar um interessante fenômeno que ocorre em algumas séries particulares, e que deve ser levado em conta à hora de fazer sua modelagem. Nessas séries, a aplicação do teste de raiz unitária dá resultado positivo (ordem de integração igual a 1). Porém, se aplicamos o teste após a remoção do suporte linear da série (possível tendência determinística linear), o resultado aponta para a ausência de raízes unitárias. Nestes casos, pode-se afirmar que a tendência da série é de fato determinística, conforme comprovado em [41][42]. A série de demanda de energia elétrica, ilustrada na Figura 3.5, é um exemplo desse fenômeno. O teste de raiz

unitária dá resultado positivo para a série original, mas negativo para a série após a remoção da reta suporte, indicando que a tendência é determinística.

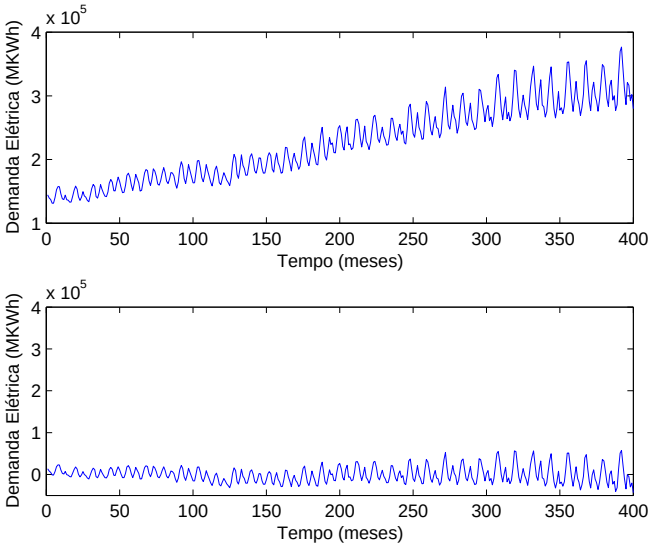

Figura 3.5: Série de demanda de energia elétrica nos EUA. Série original (acima) e após a remoção da tendência determinística (abaixo).

Homoscedasticidade vs. Heteroscedasticidade A condição **(ii)**, descrita na Definição 1, diz que um processo estacionário deve possuir variância constante e finita. Esta característica é chamada de *homoscedasticidade*, também conhecida como homogeneidade da variância. Por outro lado, um processo é dito *heteroscedástico* quando as variâncias das variáveis aleatórias que o compõem não são iguais.

Dessa maneira, a primeira verificação que se deve realizar quando se está comprovando a estacionariedade de uma série é a presença de heteroscedasticidade, antes mesmo da verificação da tendência. Esta pode se apresentar na forma aditiva ou multiplicativa, e há uma grande variedade de testes para detectá-la, como, por exemplo, os testes de Goldfeld-Quandt (1965), Glejser (1969), Szroeter (1978) e White (1980). Porém, para a maior parte das séries, a própria inspeção visual da série e do diagrama de dispersão dos resíduos constitui um procedimento suficiente para decidir sobre a presença da heteroscedasticidade [43]. Em geral, séries que representam evolução de demanda de alguns tipos de bens, como por exemplo energia elétrica, possuem heteroscedasticidade.

Uma vez detectada, a heteroscedasticidade pode ser removida a partir de

transformações como a função logarítmica ou a raiz quadrada, que tornam homogênea a variância ao longo da série. A Figura 3.6 ilustra a série mensal de demanda de energia elétrica nos EUA antes e depois da transformação logarítmica, que homogeneíza a variância da série ao longo do tempo.

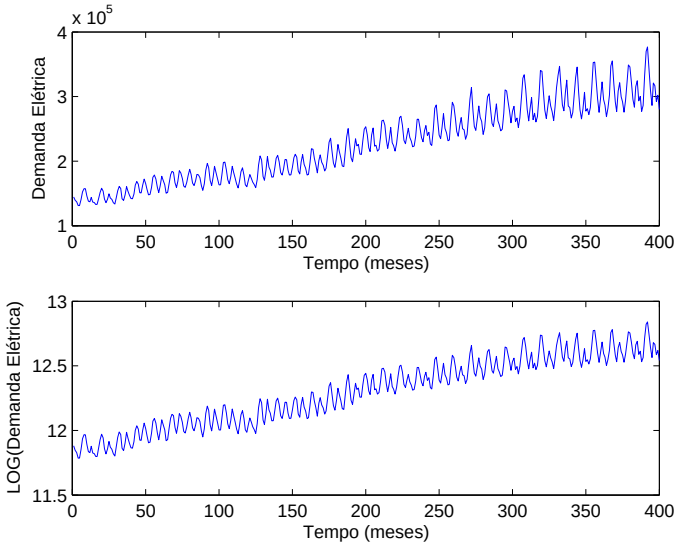

Figura 3.6: Série de demanda de energia elétrica nos EUA (acima) e após a transformação logarítmica (abaixo).

Presença e Remoção da Sazonalidade Outro fator que leva à não-estacionariedade de uma série temporal é a presença de sazonalidade, que também invalida a condição **(i)** de média constante. Boa parte dos fenômenos físicos e econômicos apresenta algum tipo de sazonalidade, ou seja, de variação periódica.

No processo de estacionarização da série, toda sazonalidade deve ser removida[3]. Uma maneira simples e eficiente de detectar a sazonalidade (determinística) de uma série é a análise da sua função de autocorrelação. Por exemplo, uma série de freqüência diária que apresente sazonalidade semanal contém uma correlação acentuada no lag=7. Esta análise é facilitada quando feita após a remoção da tendência da série. Normalmente, o conhecimento do fenômeno que originou a série permite saber *a priori* da existência de algum tipo de sazonalidade. Também a inspeção visual dos dados da série permite detectar algumas sazonalidades, sobretudo as de período menor, que ficam mais visíveis no gráfico da série.

[3]Vamos nos concentrar nos casos em que a sazonalidade é determinística.

Uma vez detectada uma determinada sazonalidade com o respectivo período s, a maneira mais eficaz de removê-la da série é aplicar o operador $(1 - B)^s$. Este operador transforma a série $\{x_1, x_2, ..., x_s, ..., x_N\}$ em $\{x_{s+1} - x_1, x_{s+2} - x_2, x_{s+3} - x_3, ..., x_N - x_{N-s}\}$, ou seja, diminui o tamanho da série de N para $N - s$. As s primeiras amostras são posteriormente devolvidas à série quando do processo de reconstrução da série ao término do seu processamento.

Para a série de demanda mensal de energia elétrica que temos analisado, podemos observar uma forte sazonalidade anual, ilustrada na Figura 3.7(b) pela componente acentuada no *lag*=12 da função de autocorrelação da série já sem tendência (determinística e linear). A série dessazonalizada, cuja FAC está ilustrada na parte (c) da figura, foi obtida com a aplicação do operador $(1 - B)^{12}$ sobre a série sem a tendência.

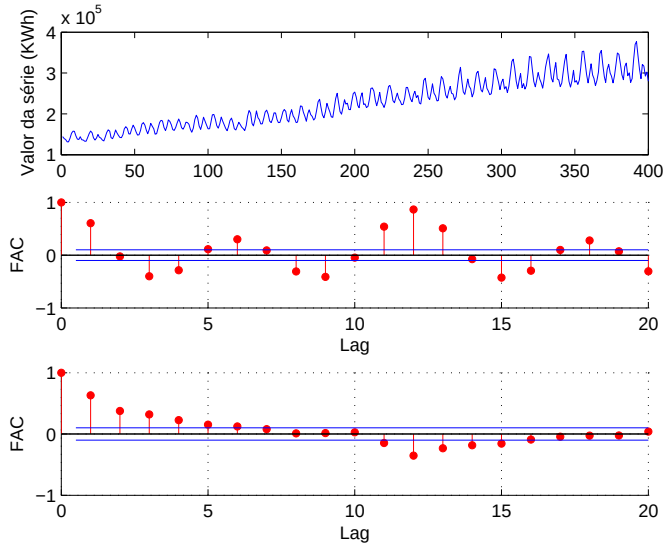

Figura 3.7: Série de demanda de energia elétrica (a); FAC da série diferenciada (b); e FAC da série dessazonalizada (c).

Variações Cíclicas Convém ressaltar aqui a diferença entre *sazonalidade* e *ciclicidade*. A sazonalidade refere-se sempre a fenômenos *periódicos*, com período constante e de fácil compreensão, como por exemplo a sazonalidade climática ou de calendário (por exemplo, fins de semana). Já os *ciclos* representam variações em que os períodos estão sujeitos a pequenas mudanças e refletem, em geral, fenômenos de duração mais longa, como os ciclos econômicos (5 a 7 anos), ciclos epidêmicos

(variam o período conforme a doença) e outros.

Para a detecção e remoção de um ciclo, a análise da função de autocorrelação e o operador $(1-B)^s$ podem não ser mais convenientes. O *lag* de correlação significativa para apontar o ciclo seria muito alto e, dependendo do tamanho total da série, nem seria visto no gráfico de autocorrelação. Além disso, a aplicação de $(1-B)^s$ com s alto implicaria na perda da quase totalidade das amostras da série.

Dessa forma, realizar a análise espectral da série através da Transformada de Fourier surge como uma boa alternativa para detecção e remoção de ciclos de longa duração. Em [44] encontra-se descrita uma metodologia para a utilização da análise de Fourier na remoção de ciclos senoidais. A estratégia consiste em detectar componentes com alta energia no espectro da série $s(t)$. Para cada componente de freqüência f_i com alta energia, calculam-se α_i e β_i (respectivamente, as componentes do cosseno e seno), e se faz a retirada no domínio do tempo através de:

$$s'(t) = s(t) - \left(\sum_{i=1}^{NC} \alpha_i \cos(2\pi f_i t) + \beta_i \sin(2\pi f_i t)\right) \tag{3.1}$$

onde $s'(t)$ é a série resultante após a retirada de NC componentes da série $s(t)$. NC equivale ao número de ciclos presentes na série, e é determinado através da inspeção do seu espectro e também do conhecimento prévio que se tenha da série.

A Figura 3.8 traz a série do número de nascimentos na cidade canadense de Quebec entre 01/01/1977 e 31/12/1990. A grande quantidade de pontos da série não nos permite ver algo que, no gráfico da FAC (Figura 3.9), fica evidente: a forte sazonalidade semanal existente devido ao baixo número de nascimentos nos fins de semana (o que é natural de se esperar, tendo-se em conta que boa parte dos partos é programada). Esta sazonalidade reflete-se nos *lags* múltiplos de 7.

Porém, alguns ciclos de período mais longo não ficam visíveis nesta análise da FAC. A Figura 3.10 mostra a análise espectral da série (com a freqüência normalizada entre 0 e 1). As componentes destacadas acusam a presença dos ciclos e sazonalidades. Ficam claros a sazonalidade semanal (com mais 2 harmônicos) e um grupo de componentes que se destacam na região de baixa freqüência.

A Figura 3.11 traz em detalhes a região de baixa freqüência, em que se distinguem claramente a sazonalidade anual (com um respectivo harmônico) e uma forte componente de mais baixa freqüência, correspondendo a um período de aproxima-

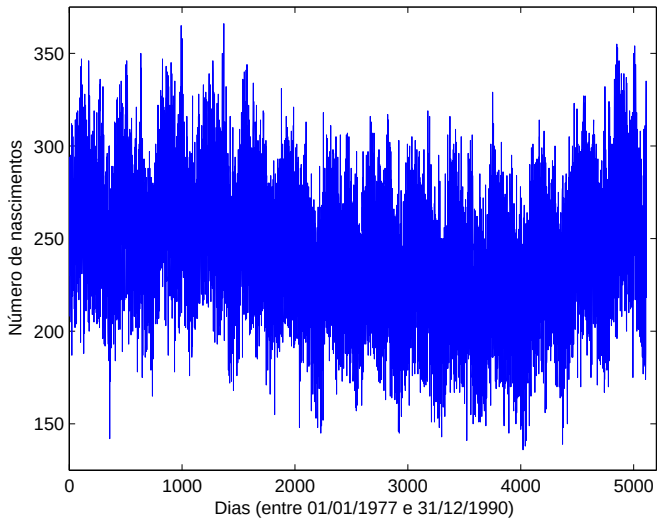

Figura 3.8: Número de nascimentos na cidade de Quebec entre 01/01/1977 e 31/12/1990.

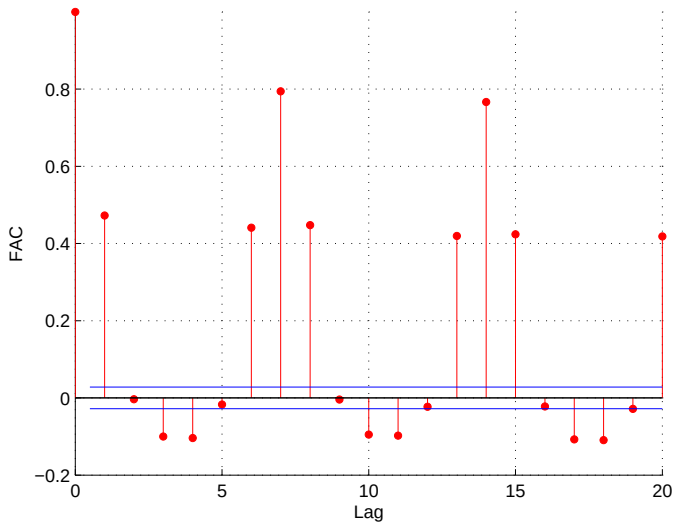

Figura 3.9: Função de autocorrelação para a série de Quebec.

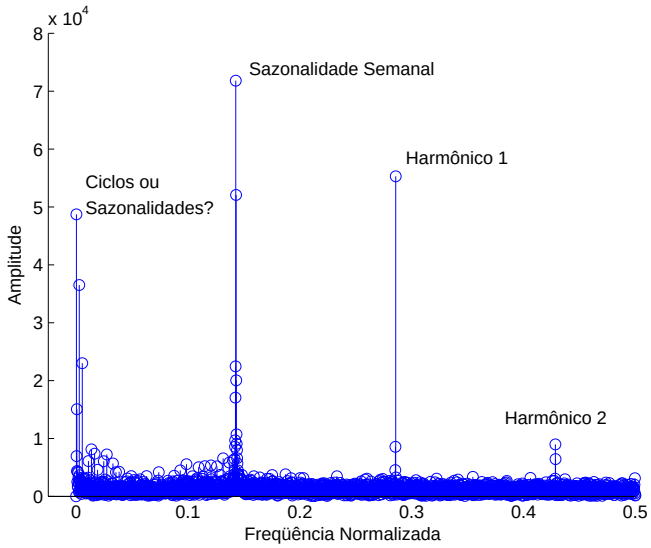

Figura 3.10: Espectro de Fourier para a série de Quebec, com a freqüência normalizada.

damente 15 anos. Esta componente deve ser classificada como um ciclo: possui alto período, não se deve a nenhum fenômeno sazonal (repare na ausência de harmônicos para esta componente).

Este ciclo, juntamente com a variação anual, estão ilustrados na Figura 3.12.

Desta forma, a remoção da sazonalidade/ciclicidade para esta série se dá através da aplicação da Equação 3.1 (com $NC = 1$ e f_i, α_i e β_i relativos ao ciclo de 15 anos) e da posterior aplicação do operador $(1 - B)^7(1 - B)^{365}$, relativos às sazonalidades semanal e anual.

Para evitar a perda de 1 ano de dados com a aplicação do operador $(1-B)^{365}$, poderíamos também ter considerado a sazonalidade anual como um ciclo, fazendo a remoção através da remoção dos pares seno/cosseno correspondentes (considerando os harmônicos).

3.1.2 Pré-processamento de Séries Temporais

Com base nas considerações feitas nas páginas anteriores, é possível agora formalizar uma metodologia de ação para o pré-processamento de séries temporais. O objetivo deste pré-processamento é tornar a série estacionária, apta portanto para modelagem segundo a metodologia que estamos propondo.

A Figura 3.13 mostra de forma esquemática o procedimento de estaciona-

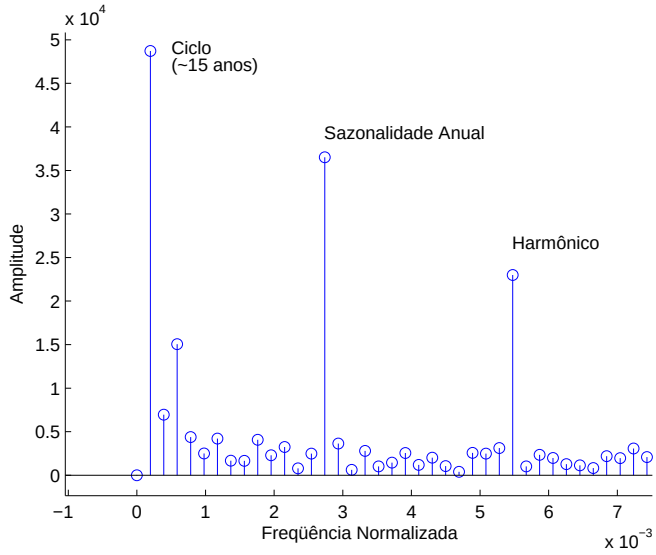

Figura 3.11: Espectro de Fourier para a série de Quebec, com a freqüência normalizada. Detalhe da região de baixa freqüência.

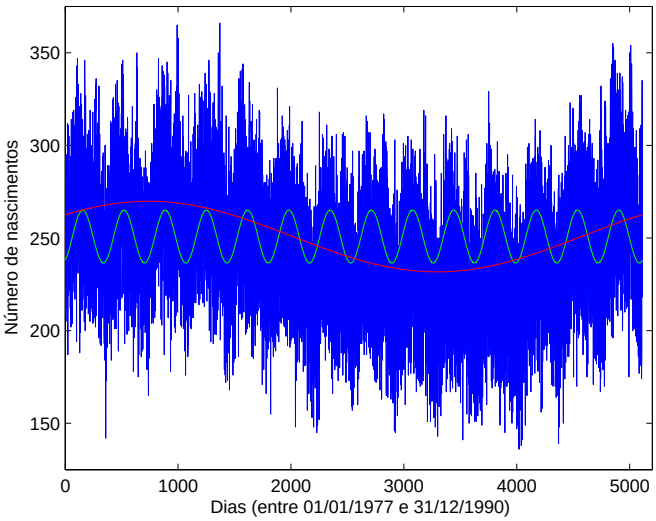

Figura 3.12: Ciclo de 15 anos e variação anual para a série de Quebec.

rização adotado neste pré-processamento. A primeira verificação que se deve fazer diz respeito à homoscedasticidade da série. Caso ela seja heteroscedástica, deve passar pela transformação logarítmica (ou raiz quadrada) para que sua variância seja uniformizada ao longo do tempo.

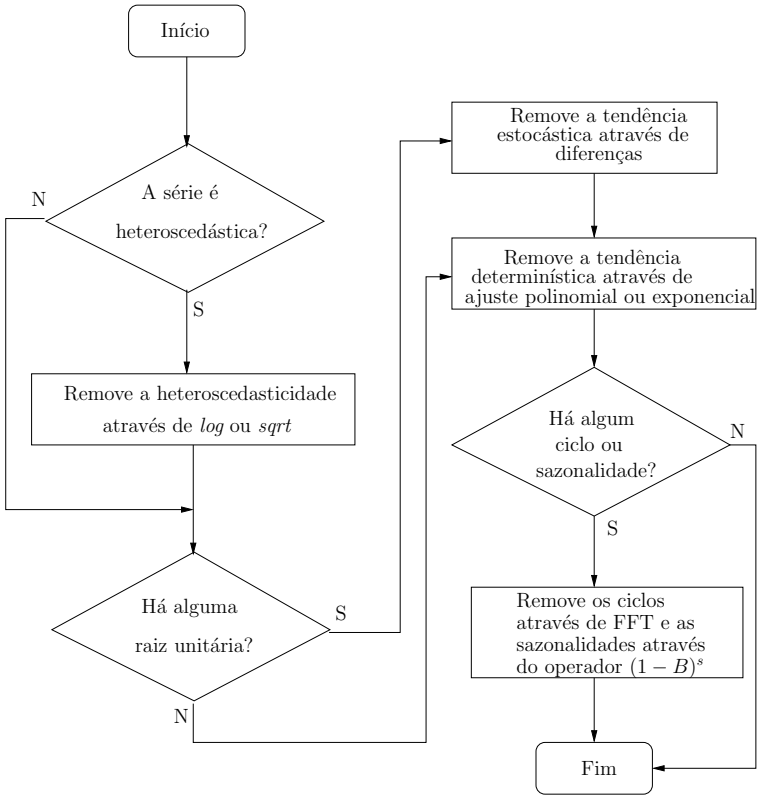

Figura 3.13: Procedimento para a estacionarização de séries temporais.

O passo seguinte verifica se há a presença de raízes unitárias na série já homoscedástica. Com isso, calcula-se a ordem de integração n da série. Caso n seja maior que zero, deve-se aplicar n vezes o operador $\nabla = (1 - B)$ à série, para retirar a sua tendência estocástica. A ausência de raízes unitárias ($n = 0$) indica que a tendência da série é determinística. Nesse caso, sua remoção deve ser feita através de um ajuste polinomial ou exponencial[4]. Sendo polinomial, são raros os casos em que a tendência tem grau maior que 1. Por isso, a tendência linear é assumida *a priori* para as séries que não apresentem nenhuma outra tendência significativa.

Por fim, deve-se verificar a presença de ciclos e sazonalidades. Os primeiros

[4]Há alguns poucos casos em que a tendência determinística tem forma exponencial. Uma tendência desta forma é visível a partir da inspeção da série. Isto direciona a natureza do ajuste a ser feito.

são detectados através da análise de Fourier e retirados com o procedimento descrito em 3.1.1.2. As sazonalidades são detectadas através da análise da FAC e retiradas com o operador $(1 - B)^s$, onde s assume os valores dos períodos das sazonalidades encontradas.

Neste momento, a série está estacionarizada e apta para ser modelada. O tipo de modelagem a ser feita depende de alguns fatores; dependendo também da aplicação e do tipo de série, pode ser conveniente aplicar à série estacionária algum tipo de filtragem, como por exemplo a de Kalman[5] [45]. Voltaremos a este assunto no Capítulo 4.

3.1.3 Exemplos de Séries Temporais Utilizadas

Após expor os aspectos teóricos relacionados às séries temporais, descreveremos agora algumas das séries utilizadas neste trabalho. Inicialmente, será introduzida a problemática da Qualidade de Dados para séries temporais reais, alvo da monitoração do sistema a ser descrito no Capítulo 4 e cuja implementação será detalhada no Capítulo 5.

3.1.3.1 Comentários Iniciais

Existem basicamente duas formas de se ter acesso a grandes massas de dados. Uma das alternativas é comprá-las, visto que muitos dados, devido à natureza da informação contida, não são disponibilizados de forma gratuita. Em geral, bases de dados financeiras se encontram nessa situação, sobretudo se correspondem a séries com alta freqüência (por exemplo, períodos de uma hora ou um minuto) e entregues em tempo real. Costumam ser bases caras, mas que oferecem maior probabilidade de conter dados "limpos", isto é, livres de erros.

Outra alternativa é recorrer a bases gratuitas, também disponíveis em alguns sítios especializados da *Internet*. Estas têm como desvantagem a falta de garantia da pureza dos dados, conforme podemos depreender do alerta colocado ao final da página principal do sítio "Yahoo!Finance" [46], um dos principais provedores de dados gratuitos na área financeira:

[5]Ainda que, para a aplicação da filtragem de Kalman, a série não precise ser estacionária.

"Os dados e as informações são disponibilizados somente para propósitos de informação, não tendo finalidades comerciais. Nem Yahoo! nem qualquer outro de seus provedores de dados ou conteúdo (como Reuters e CSI) poderão ser responsabilizados por quaisquer erros ou atrasos no conteúdo, ou por quaisquer danos causados a partir da utilização dessas informações."

Para este trabalho, todos os conjuntos de dados (sejam séries temporais ou não) foram conseguidos de forma gratuita. Para o caso das séries temporais financeiras, obtivemos também, além de séries "sujas" (oriundas de [46]), dados de origem "intermediária": são dados que acompanham a versão de demonstração gratuita de um *software* [47] que, quando pago, dá a garantia de fornecer dados limpos. Para a formação do nosso modelo, admitimos que os dados provenientes dessa base são totalmente corretos.

3.1.3.2 Séries Temporais Financeiras

Dentro do grande grupo das séries temporais, um importante segmento com o qual trabalhamos foi o das séries temporais financeiras. A análise de QD para dados desse tipo tem grande importância para diversas tarefas comumente realizadas nas áreas econômicas e empresariais, como, por exemplo, a estimação da volatilidade[6] para o dia seguinte, a análise de risco, otimização de carteiras de ações, etc.

Convém dizer novamente que o foco do nosso trabalho não está em desenvolver mecanismos preditores para séries temporais, mas em gerar um sistema de monitoração que faça a análise da qualidade de dados para bases grandes e dinâmicas[7]. Desta forma, para o caso concreto das séries temporais financeiras, ao desempenharmos a análise de QD de suas amostras, estaremos possibilitando aos usuários da base a extração de modelos mais acurados a partir dos dados analisados.

Séries do S&P 500 De forma a ter um conjunto que representasse de forma razoável as séries do mercado financeiro, e levando em conta também a facilidade de aquisição de tais séries, optamos por trabalhar com ações presentes no índice do

[6]Medida da taxa de variação do valor de uma ação durante um dia.

[7]Naturalmente, as tarefas de predição e de monitoração da qualidade dos dados estão intimamente ligadas.

mercado americano "Standard & Poor's 500" [48]. Trata-se de um índice que contém as ações das principais empresas americanas, de maneira que se pode tomá-lo como bastante representativo do mercado de ações como um todo.

A partir do sítio "*Stockwiz*" [47], tivemos acesso às séries diárias, de janeiro de 1998 (ou 1996/1997, dependendo da série) a agosto de 2006, de quase cem empresas do índice S&P 500, como a IBM, a Microsoft, a AOL e a Sun. Como já dissemos, admitimos (de forma razoável) que tais dados têm garantia de qualidade. Para cada série, constam os valores de abertura, máximo, mínimo e fechamento do valor da ação em cada dia (*Open-High-Low-Close* ou formato OHLC), além do valor do volume diário negociado. Na Figura 3.14, vemos a evolução da série diária de fechamento da IBM para o período analisado.

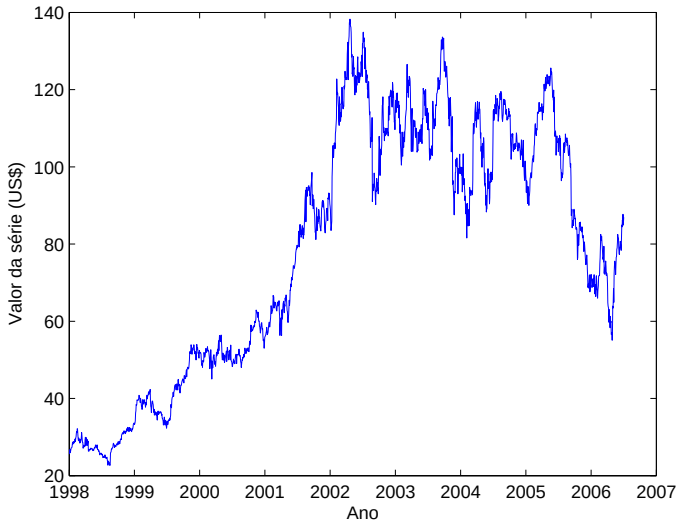

Figura 3.14: Valores (certificados) de fechamento para a série da IBM de 01/1998 a 06/2006.

De modo a ilustrar a diferença existente entre bases certificadas e não certificadas, fizemos a comparação, para um trecho de cerca de dez anos da série da AMD (Advanced Micro Devices Inc.), entre a série obtida no sítio "*Stockwiz*" (certificada) e a série obtida no sítio "*Yahoo!Finance*" (não-certificada). O resultado é mostrado na Figura 3.15. Algo bastante interessante que se pode notar é a quase eliminação total dos erros na série não-certificada a partir do ano 2001, o que indica uma clara mudança de postura da *Yahoo!Finance* com relação à qualidade dos dados disponibilizados em seu sítio. Naturalmente, esse evento reforça a importância da pesquisa

em sistemas de monitoração de qualidade de dados, uma vez que é reflexo de uma preocupação crescente com a confiabilidade da informação disseminada, sobretudo, através da *Internet*.

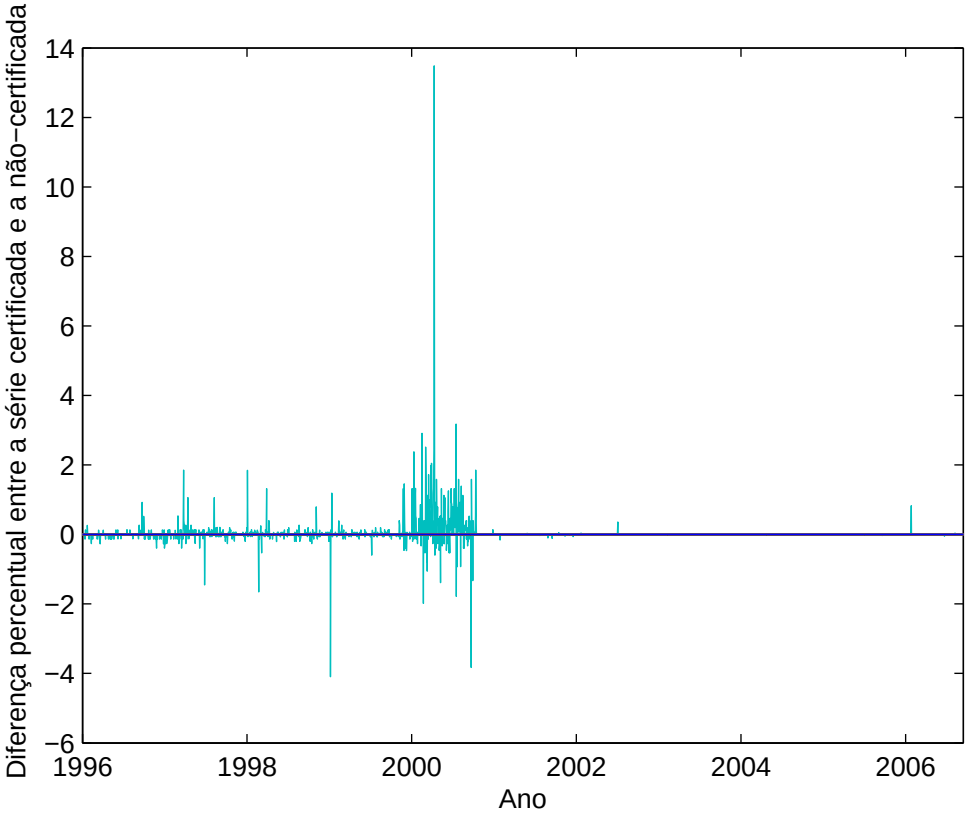

Figura 3.15: Comparação entre duas séries para a AMD: diferença percentual entre as séries certificada e não-certificada.

3.1.3.3 Outras Séries Temporais

As séries temporais financeiras descritas acima constituem um grupo importante de dados devido às peculiaridades deste tipo de série. De fato, são crescentes as pesquisas relacionadas à modelagem dos mercados financeiros.

Porém, este não é o único tipo de série com que trabalhamos nesta tese, mesmo porque o sistema de monitoração tem por objetivo ser o mais abrangente possível. Procuramos formar um banco de séries que representasse uma gama diversa de fenômenos físicos, econômicos e sociais, de modo que pudéssemos testar a eficácia do sistema de monitoração em diferentes cenários.

Diferentemente do que acontecia para as séries financeiras, em que dispúnha-

Série	Freqüência	Intervalo	Fonte
Demanda de Energia Elétrica (EUA)	Mensal	jan/1973 - abr/2006	[49]
Nascimentos em Quebec	Diária	jan/1977 - dez/1990	[50]
População EUA	Mensal	jan/1952 - jun/2006	[49]
Inflação EUA	Mensal	fev/1913 - jul/2006	[49]
Chuva Melbourne	Diária	jan/1981 - dez/1990	[51]
Casas EUA	Mensal	jan/1963 - jul/2006	[49]
Desemprego EUA	Mensal	jan/1948 - jun/2006	[49]
Material Particulado PM10 (SP/Brasil)	Diária	jan/1997 - dez/1997	[36]
Emissão de CO (SP/Brasil)	Diária	jan/1997 - dez/1997	[36]
IPI - Produtos Alimentares (Brasil)	Mensal	jan/1985 - jul/2000	[36]

Tabela 3.1: Listagem de algumas das séries temporais não-financeiras utilizadas neste trabalho.

mos de séries certificadas e não-certificadas, as séries não-financeiras estão apresentadas em suas fontes como dados certificados. Isto é interessante no que se refere à modelagem e à utilização do sistema de monitoração como mecanismos de certificação de tais dados, sobretudo quando uma mesma série for obtida a partir de duas fontes distintas.

A Tabela 3.1 traz um resumo de algumas das principais séries utilizadas neste trabalho, conforme será visto no Capítulo 5.

3.2 Tópicos em Registros Multivariados

Além das séries temporais, um grande grupo de dados com o qual trabalhamos nesta tese são os chamados *registros multivariados*, normalmente armazenados em tabelas multidimensionais. Estes dados geralmente são provenientes de bases cadastrais, informações de usuários de um determinado produto/serviço, resultados de entrevistas/pesquisas, observações espaciais de algum fenômeno físico etc. Se alguma variável de uma base multivariada evolui temporalmente, esta pode ser tratada separadamente como uma série temporal.

3.2.1 Exemplos de Registros Multivariados Utilizados

Este tipo de dados, que aparece com muita freqüência nas grandes bases de dados empresariais, são geralmente armazenados em tabelas multidimensionais. Tais estruturas têm capacidade de registrar dados de fenômenos multivariáveis de origens bastante diversas. Além disso, fornecem um grande campo de ação para as ferramentas de Mineração de Dados (*Data Mining*), visto que, quanto maior o tamanho da tabela, maior a chance de se obter conhecimento novo a partir da análise dos dados nela contidos.

Tais bases de dados são bastante adequadas para realizarmos testes de detecção e correção de *outliers* e também a substituição de dados faltantes (*missing values*). Outra possível informação que pode ser extraída de tais dados diz respeito à identificação dos modelos probabilísticos para as variáveis envolvidas. A modelagem das densidades de probabilidade dessas variáveis leva-nos a desenvolver testes de validação estatística para novas amostras, além de representar uma descoberta de conhecimento importante.

No caso dos registros multivariados, há uma falta intrínseca de qualidade dos dados, causada por diversos fatores (desde o método de aquisição até problemas na manutenção dos dados). Normalmente, as bases vêm acompanhadas de relatórios que descrevem estes problemas na aquisição dos dados. Dessa forma, o fato de estarmos buscando a implementação de um sistema que atue sobre dados com falhas e os corrija faz com que o uso de bases naturalmente ruidosas nos ajude a medir a eficiência da atuação do SMQD.

3.2.1.1 Os dados do ICPSR

O ICPSR (*Inter-university Consortium for Political and Social Research*) [52], da Universidade de Michigan/EUA, foi criado em 1962 com o objetivo de prover o acesso a amplas bases de dados sociais, políticas e econômicas, todas provenientes de pesquisas de campo e contendo, portanto, dados reais. Trata-se de um imenso repositório de bases de dados que é colocado à disposição da comunidade acadêmica. No Brasil, as instituições de ensino universitário têm acesso aos dados do ICPSR através do CESOP (Centro de EStudos da Opinião Pública) [53], da Unicamp.

A base de dados do ICPSR é composta por milhares de pesquisas que lhe

Números tentados	3280	Dentre os lares	2073
Não eram lares	1019	Entrevistas	**1151**
Nunca responderam	188	Recusas	541
Lares	2073	Não houve contato	283
		Outros problemas	97

Tabela 3.2: Descrição do resultados das ligações telefônicas para a base FISCT.

são enviadas pelos membros cooperadores, tornando-a dinâmica e atualizada. Cada pesquisa está pormenorizadamente documentada, o que permite ao usuário ter um amplo domínio sobre os dados, que muitas vezes têm estrutura complexa. Além de explicar a metodologia utilizada na pesquisa, a documentação lista as possíveis fontes de erros para a base, fornecendo alguns campos em potencial para testes com QD.

3.2.1.2 A Base de dados FISCT

Solicitamos ao CESOP o envio de cerca de trinta bases de dados relativas a temas que tratam, por exemplo, do histórico da ação governamental, da economia doméstica, de hábitos de consumo, de áreas da educação etc. Dentre todas, a base que mais utilizamos foi a relativa ao projeto FISCT (*Family Interaction, Social Capital, and Trends in time use*), realizado entre os anos de 1998 e 1999.

A base FISCT [54] é o resultado de um projeto que procurou traçar um perfil do comportamento diário do cidadão americano. Para tanto, a metodologia adotada foi a de sortear 3280 números telefônicos proporcionalmente espalhados por todos os estados dos EUA com o objetivo de entrevistar o atendente. Desse total de ligações, obteve-se o resultado descrito na Tabela 3.2, onde se destaca o total de 1151 entrevistas realizadas com sucesso.

Ao entrevistado pedia-se que dissesse quais haviam sido todas as suas atividades do dia anterior, de 0h às 24h. Deveria informar, percorrendo o dia, os horários de início e fim da atividade, se havia (e quais eram) atividades paralelas secundárias, o local onde a atividade foi realizada, e também se havia (e quem eram) pessoas que o acompanhavam na atividade. Um resultado típico de uma entrevista do FISCT pode ser encontrado na Tabela 3.3, cujas colunas estão descritas na Ta-

A	B	C	D	E	F	G	H	I	J	L	M	N	O
11002	10	Dormindo/cochilando	45			360	32598	0	600	1	3	1	
11002	11	Vestindo-se	47	0	0	30	32598	600	630	1	3	1	9
11002	12	Deslocamento	9	0	0	20	32598	630	650	11	3	1	9
11002	13	Reunião de trabalho	1	0	0	100	32598	650	830	4	3	5	9
11002	14	Deslocamento	9	0	0	20	32598	830	850	11	3	1	9
11002	15	Trabalhando	1	0	0	190	32598	850	1200	4	3	5	9
11002	16	Refeição/lanche	43	0	0	60	32598	1200	1300	4	3	5	9
11002	17	Trabalhando	1	0	0	330	32598	1300	1830	4	3	5	9
11002	18	Refeição/lanche	43	0	0	60	32598	1830	1930	4	3	5	9
11002	19	Trabalhando	1	0	0	105	32598	1930	2115	4	3	1	9
11002	20	Deslocamento	9	0	0	15	32598	2115	2130	11	3	1	9
11002	21	Banho/ducha	40			30	32598	2130	2200	1	3	1	
11002	22	Assistindo TV	91	96	0	30	32598	2200	2230	1	3	2	9
11002	23	Dormindo/cochilando	45			90	32598	2230	2400	1	3	1	

Tabela 3.3: Exemplo de resultado de entrevista para a base FISCT.

bela 3.4. A descrição detalhada dos códigos utilizados e a metodologia usada para o agrupamento de atividades encontram-se em [54].

Para os testes com o SMQD, dispusemos então de 1151 eventos, como o ilustrado na Tabela 3.3. Conforme veremos no Capítulo 5, os testes com estes dados incluíram a análise de dados faltantes e de *outliers*, além da modelagem de distribuições para as variáveis envolvidas.

A	Código do entrevistado	**H**	Data da entrevista
B	Número da atividade	**I**	Hora de início
C	Atividade relatada	**J**	Hora de término
D	Código da atividade	**L**	Local da atividade
E	Atividade secundária (1)	**M**	Dia da semana
F	Atividade secundária (2)	**N**	Acompanhante? (1)
G	Tempo de duração (min)	**O**	Acompanhante? (2)

Tabela 3.4: Descrição das colunas da Tabela 3.3.

Capítulo 4

O Sistema de Monitoração de Qualidade de Dados (SMQD)

A metodologia desenvolvida para o sistema de monitoração é o tema deste capítulo. Com efeito, o núcleo desta tese consiste neste sistema, que atua desde a chegada de um dado novo até o seu carregamento na base de dados após as devidas verificações quanto às dimensões da Qualidade de Dados.

4.1 Modelos Propostos na Literatura

Quando se percorre a literatura da área de Qualidade de Dados, percebe-se que grande parte dos textos ainda está voltada para a busca de conceitos e definições mais precisos sobre os fundamentos da QD. De certa forma, é natural que seja assim, já que se trata de uma área de pesquisa ainda embrionária.

Quanto ao desenvolvimento de mecanismos de medição e melhoria da QD em bases de dados, encontram-se algumas propostas de procedimentos que lidam com dimensões específicas da QD, e que em geral resultam em soluções *ad hoc* para contextos mais específicos [55][56]. Tais procedimentos normalmente são implementados de maneira *offline*, ou seja, sua ação não se dá de forma contínua. Isto é algo contrário ao pressuposto de que um sistema de monitoração é tanto mais eficiente quanto mais imediata é a sua ação. De fato, valem mais as pequenas e freqüentes intervenções do que grandes ações esparsadas no tempo, já que dessa forma busca-se uma base de dados sempre atualizada e evita-se a propagação de eventuais

incorreções para dentro e para fora da base: “Prevenir é melhor que remediar”.

Esta dinâmica, de pequenas e constantes modificações, dá origem a um ciclo que deve reger a estrutura de todo o sistema de monitoração de qualidade de dados [57]. Tal ciclo é ilustrado na Figura 4.1, refletindo um processo contínuo na monitoração da qualidade de dados. Portanto, “monitorar” significa “... → medir → analisar → melhorar → definir → medir → ...”.

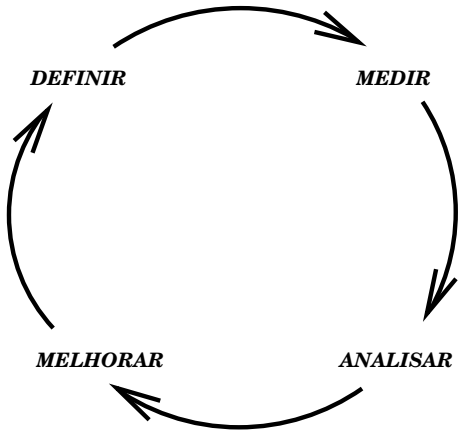

Figura 4.1: Ciclo para a monitoração da Qualidade de Dados.

A seguir, vamos analisar um modelo de monitoração genérico proposto em [28]. Este modelo serviu de inspiração inicial para o desenvolvimento do sistema de monitoração apresentado nesta tese.

4.1.1 Modelo Baseado na Teoria de Controle

O modelo proposto em [28] serviu-nos de ponto de partida para o desenvolvimento de um sistema adequado para os nossos objetivos. O paradigma desse modelo baseia-se na definição de um fluxo de dados existente entre as fontes geradoras de informação e os usuários finais, que desejam receber dados formatados para fins específicos, com a devida qualidade.

A definição desse fluxo de dados segue um consenso entre os analistas de QD. Conforme ilustrado na Figura 4.2, nos extremos do fluxo encontram-se as fontes dos dados, onde ocorre a sua geração, e os usuários finais dos dados. Entre esses dois extremos, deve estar presente todo o processamento do dado, que tem por objetivo monitorar a sua qualidade, de maneira que o usuário possa utilizá-lo com maior rendimento.

Este processamento é conhecido por ETL (*Extraction, Transformation and Load*) [58], que consiste nas tarefas de extração dos dados contidos nas fontes, na sua transformação (isto é, "limpeza" e reformatação, conforme a necessidade) e no seu carregamento nas respectivas bases de dados. A parte de ETL[1] é bastante importante para o sucesso de um sistema de monitoração de QD. É dentro dela em que se devem inserir boa parte das "ferramentas" de análise de QD, desde a detecção de falhas nos dados até a escolha do melhor método de correção disponível.

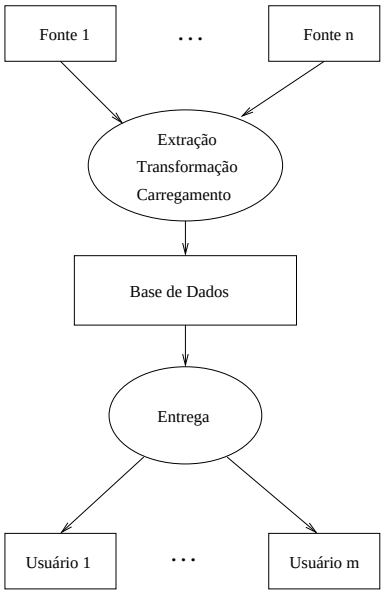

Figura 4.2: Fluxo de dados fonte-usuário.

Com base neste fluxo geral, o passo seguinte é o desenvolvimento de algum mecanismo de monitoração da QD da base. Em [28], sugere-se um modelo baseado na Teoria de Controle. A Figura 4.3 ilustra esse modelo, destacando a analogia existente entre a sua estrutura e a de um sistema de controle de processos.

A tarefa de monitoração da Qualidade de Dados de uma base *dinâmica*, isto é, cujo conteúdo se modifique continuamente, encaixa-se muito bem dentro da perspectiva de um sistema de controle. De fato, o modelo mostrado na Figura 4.3 pode ser comparado ao modelo clássico de Teoria de Controle ilustrado na Figura 4.4 [59]. Tanto em um caso quanto no outro, busca-se minimizar um erro associado à planta (representada pela própria base de dados) através da realimentação de sua saída.

[1]Em português, ETC. Preferimos manter a sigla na língua inglesa, por se tratar de termo já consolidado nos meios acadêmicos e comerciais.

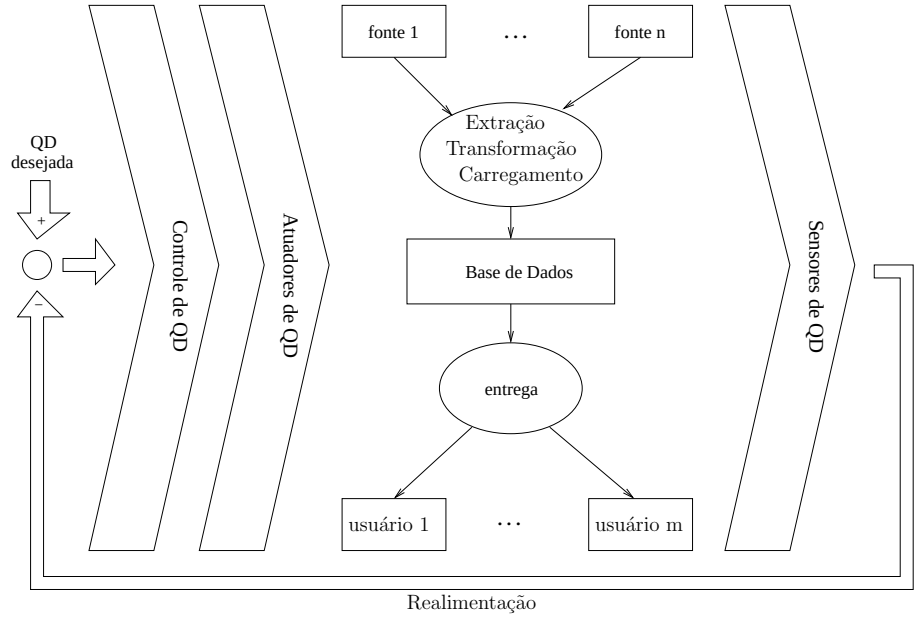

Figura 4.3: Sistema de controle de qualidade de dados.

No caso do sistema da Figura 4.3, os Sensores de QD geram as "saídas" das plantas (equivalentes ao próprio vetor de métrica com as dimensões de QD, conforme o exposto no Capítulo 2), que é comparado ao valor desejado (de referência). A diferença (o "erro") serve de entrada para o Controlador de QD, que deve definir as modificações a serem implementadas pelos Atuadores de QD.

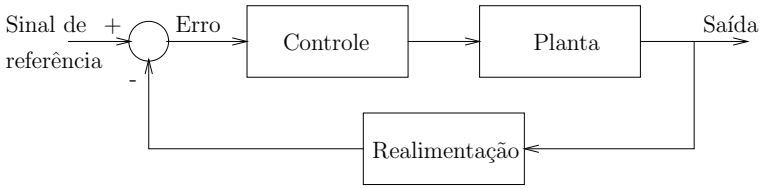

Figura 4.4: Controle clássico de uma planta.

Esta analogia com os sistemas de controle está ratificada em [60]. Neste artigo, os sistemas de informação (por exemplo, bases de dados dinâmicas) são tratados como "Sistemas de Controle Realimentados" (*Feedback-Control System* ou FCS). Dentro desta perspectiva, derivam-se 6 regras para o gerenciamento da QD:

1. Dados não-utilizados não se mantêm corretos por longo tempo;

2. A QD de um sistema de informação é função de seu uso e manutenção, e não da aquisição dos dados;

3. A QD máxima de uma base será atingida com sua utilização mais rigorosa;

4. Problemas com a QD tendem a crescer com o envelhecimento do sistema;

5. Quanto menos provável for uma mudança em um atributo de um dado, mais traumático será quando ele finalmente mudar;

6. As Leis para QD aplicam-se tanto aos dados quanto aos meta-dados (dados sobre os dados).

Estas 6 regras serão levadas em consideração à hora de especificarmos o SMQD.

4.2 O SMQD

Tendo já apresentado um modelo genérico para o fluxo de dados entre a sua fonte e o usuário final, além de uma proposta baseada na teoria de controle para a monitoração da QD de grandes bases de dados, estamos aptos para descrever o Sistema de Monitoração de Qualidade de Dados (SMQD) proposto nesta tese. Nesta seção será exposta a metodologia do SMQD, incluindo as suas especificações, o processo de validação de novos dados e as técnicas para monitoração das dimensões de QD envolvidas.

4.2.1 Metodologia

Ao iniciarmos o desenvolvimento da metodologia do sistema de monitoração, devemos ter bem claras as funcionalidades que desejamos incorporar a ele. Constituem as *especificações* do SMQD, que devem refletir as regras descritas na seção anterior.

A Figura 4.5 ilustra com uma estrutura hierárquica de camadas a organização do SMQD. Note-se que neste esquema as flechas não representam o fluxo de dados, servindo apenas para definir os níveis hierárquicos. Estes níveis encontram-se listados abaixo:

- 1º nível: Interface do Sistema. Neste nível se dá a interação do usuário com o SMQD;

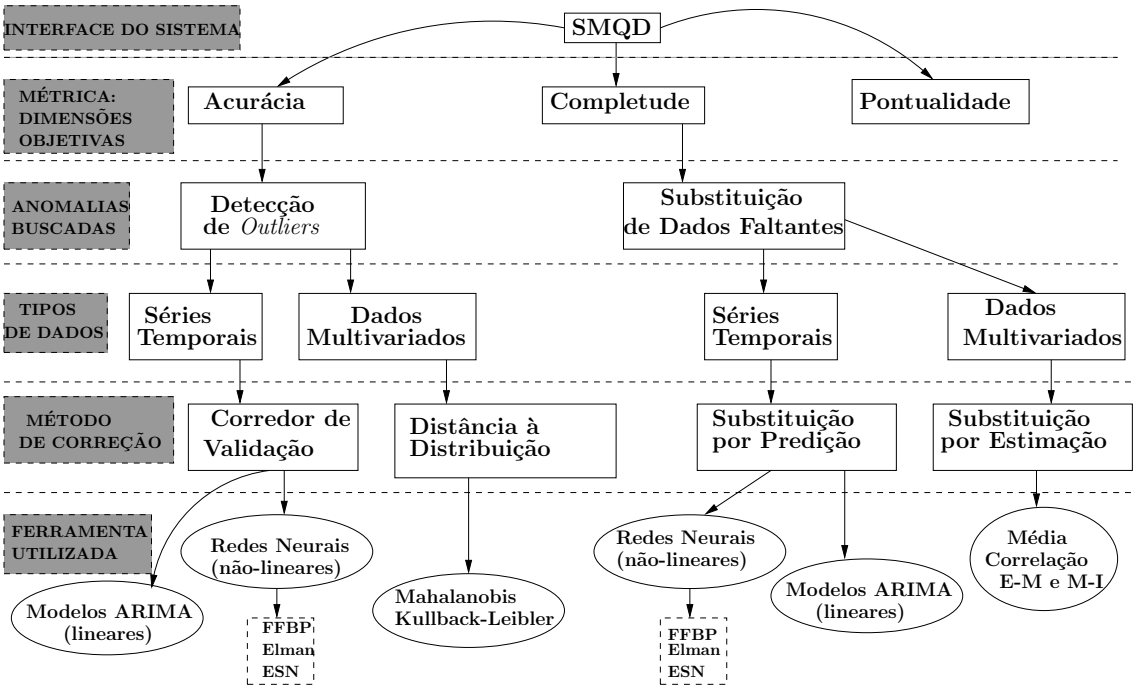

Figura 4.5: Estrutura hierárquica em camadas para o SMQD.

- 2º nível: Dimensões da QD que compõem a métrica. Aqui devemos definir quais atributos serão monitorados pelo sistema;
- 3º nível: Anomalias buscadas. Derivam-se das dimensões a serem monitoradas;
- 4º nível: Tipos de dados. Define-se aqui sobre que tipos de dados serão buscadas as anomalias;
- 5º nível: Método de correção. Aqui se definem as estratégias de atuação para corrigir as anomalias encontradas em cada tipo de dados;
- 6º nível: Ferramentas utilizadas para implementar os métodos de correção definidos.

A seguir, analisaremos cada um destes níveis, mostrando como estão concretizados no SMQD proposto, o que define suas especificações.

4.2.1.1 Interface do Sistema

A interface do sistema é o canal de comunicação do usuário com a base de dados. De maneira especial, destaca-se o papel do supervisor, especialista que tem a faculdade de validar ou não as intervenções propostas pelos mecanismos do sistema.

O papel do supervisor é fundamental, visto que ele possui um tipo de conhecimento que não está contido na parte automática do sistema. Por exemplo, só o supervisor pode informar ao sistema que um determinado dado, detectado como *outlier*, está isento de erro e portanto deve ser incorporado normalmente à base.

4.2.1.2 Dimensões Monitoradas

No projeto do SMQD, uma etapa importante é a definição de quais dimensões da QD, dentre as descritas no Capítulo 2, serão objeto de monitoração.

Para este trabalho, decidimos trabalhar apenas com dimensões *objetivas*, ou seja, os chamados *indicadores* da QD de uma determinada base. Todavia, ao projetar a estrutura do SMQD tivemos a preocupação de deixá-lo apto para a incorporação, no futuro, de dimensões *subjetivas* (*parâmetros*).

Dentre as dimensões objetivas, as mais importantes e utilizadas para a medição da QD são a *acurácia*, a *completude* e a *pontualidade*. Além disso, são as que permitem uma medição mais automática.

A seguir, descreveremos as dimensões contempladas no SMQD, definindo, para cada uma delas, uma forma de cálculo que as limite entre 0 e 1.

Acurácia Esta dimensão mede o grau de concordância entre os valores dos dados coletados e os fornecidos por uma fonte que supre valores certificados. Em outras palavras, mede o quão perto dos valores de referência estão os dados adquiridos.

Em [61], este atributo é quantificado como:

$$Ac = 1 - \frac{NI}{NT} \tag{4.1}$$

onde definem-se NI e NT como o número de dados incorretos e o número total de dados requeridos pela aplicação, respectivamente.

Notemos que esta maneira de quantificar a acurácia não leva em consideração a magnitude do erro em cada dado específico, distinguindo apenas entre dados corretos e incorretos. Isto está em conformidade com a idéia do *corredor de validação* para os dados, intervalo dentro do qual um dado é considerado correto e fora do qual é rejeitado. A formação desse corredor será descrita ainda nesse capítulo.

Completude Na área de QD, esta dimensão é utilizada para medir o grau em que os dados estão presentes. Em um contexto mais dependente do usuário (e não tanto da base), pode-se defini-la como o grau em que os dados são suficientemente disponíveis onde são necessários, isto é, para a tarefa que se quer realizar [18].

Tem como dimensão associada a *quantidade apropriada de dados* [62], definida como $\min(D_p/D_r, D_r/D_p)$, sendo D_p e D_r, respectivamente, o número de dados providos pela base e o número de dados necessários.

Sendo NF como o número de dados faltantes, podemos quantificar a completude através da seguinte equação:

$$Cp = 1 - \frac{NF}{NT} \tag{4.2}$$

Pontualidade Do inglês *timeliness*, significa, em Qualidade de Dados, o quão rapidamente um dado está disponível para o usuário da base de dados. Mede, em outras palavras, o grau de atualização da base de dados.

Reparemos que, para uma série temporal, podemos reduzir esta dimensão à própria completude. Neste caso, a *pontualidade* equivale à *completude temporal*.

Para dados que não evoluem de forma direta com o tempo, o tratamento deve ser diferente. Segundo [30], três aspectos devem ser levados em consideração neste caso:

1. Sobre que período os dados foram coletados?
2. Quando foi feita a última atualização da base, de maneira a refletir as mudanças no mundo real?
3. Por quanto tempo espera-se que os dados mantenham-se em vigor, atuais?

Assim sendo, podemos quantificar a pontualidade através da seguinte equação, que leva em consideração as questões colocadas acima:

$$Pt = \max(0; 1 - \frac{TD}{TE})^s \tag{4.3}$$

onde TD e TE são, respectivamente, o *tempo para disponibilização* (intervalo de tempo entre a geração do dado e sua disponibilização para o usuário) e o *tempo*

de expiração do dado. O fato de TD ser diferente de zero indica que a chegada de um dado à base não é imediata, e TE indica que o dado possui um "prazo de validade" [63][64]. O valor de s é deixado à escolha do supervisor, de maneira a regular a sensibilidade desta dimensão à relação TD/TE para cada grupo de dados. Enquanto TE costuma ser uma medida estática, TD geralmente varia no tempo.

Quando tratarmos da monitoração dessa dimensão dentro do SMQD, será explicado como determinar TD e TE para cada tipo de dados monitorado.

4.2.1.3 Anomalias Buscadas

Tendo definido as dimensões da QD a serem monitoradas pelo sistema, podemos especificar que tipo de anomalias serão alvo da detecção e correção por parte do SMQD.

Diretamente associado à acurácia está o conceito de *outlier*. Com efeito, a presença de dados com essa característica faz com que a acurácia da base como um todo diminua. Assim, ainda que a monitoração da acurácia não tenha como objetivo apenas a detecção de *outliers*, a identificação destes é fundamental para o controle dessa dimensão.

Outro tipo de anomalia que deverá ser monitorada pelo SMQD é a existência de dados faltantes (em inglês, *missing data* ou *missing values*). Esta anomalia afeta diretamente a completude dos dados e também a sua pontualidade, sobretudo no caso de séries temporais. Portanto, o correto gerenciamento de dados faltantes (em geral, através de sua substituição por outro valor, conforme veremos adiante) é o meio ordinário para o controle da completude da base de dados.

Discutiremos em detalhes o tratamento dessas anomalias nas subseções 4.3.1 e 4.3.2, quando tratarmos da monitoração da acurácia e da completude.

4.2.1.4 Tipos de Dados

Seguindo a estrutura hierárquica proposta na Figura 4.5, cabe-nos especificar agora os tipos de dados que serão tratados pelo SMQD. Trata-se principalmente de uma questão organizacional, visto que, ao menos idealmente, o sistema deve ser capaz de atender quaisquer tipos de dados.

No caso presente, classificamos o universo de dados atendidos em dois grandes

grupos, conforme já havia sido delineado no Capítulo 3: o grupo das *Séries Temporais* e o grupo dos *Registros Multivariados*. Trata-se de duas classes que englobam a quase totalidade dos dados presentes nos sistemas de informação de empresas e universidades.

4.2.1.5 Método de Correção

Para cada par *anomalia/tipo de dado* existe um método de correção correpondente, projetado para detectar e corrigir o problema. Dessa forma, a detecção/correção de *outliers* em séries temporais se dá através do *Corredor de Validação*, e em registros multivariados se utilizam técnicas de *Distância à Distribuição*.

Quanto ao gerenciamento de dados faltantes, utilizamos a *Substituição por Predição* e a *Substituição por Estimação* para tratar, respectivamente, séries temporais e registros multivariados.

4.2.1.6 Ferramentas Utilizadas

Cada método de correção listado acima é implementado através de ferramentas estatísticas e de processamento de sinais. Enumeramos aqui tais ferramentas, que serão descritas com detalhes posteriormente. Para os métodos associados à tarefa de predição (séries temporais), as ferramentas utilizadas são os *Modelos ARIMA* (lineares) e os *Modelos Baseados em Redes Neurais* (não-lineares), com diferentes arquiteturas.

Para os métodos associados à tarefa de estimação de distribuições (registros multivariados), utilizam-se as distâncias de *Mahalanobis* e *Kullback-Leibler*, além de algoritmos de *Expectation-Maximization* e *Imputação Múltipla*.

4.2.2 Dinâmica de Atuação do SQMD

Tendo visto já a estrutura geral do sistema de monitoração, com suas especificações e funcionalidades, convém agora que nos detenhamos na sua dinâmica de atuação, isto é, a maneira pela qual os dados fluem dentro do sistema, desde a sua chegada até a sua validação e incorporação à base de dados (ou, eventualmente, rejeição e descarte).

A Figura 4.6 ilustra o fluxo de dados dentro do SQMD que propomos nesta tese. Diferentemente do esquema mostrado na Figura 4.5, aqui as setas representam efetivamente o sentido deste fluxo de dados. Reparemos que este fluxo tem origem no tipo do dado que é recebido e termina na atualização das dimensões de QD monitoradas.

No caso de séries temporais, a verificação de que a nova observação é um dado faltante exclui a necessidade de se detectar um *outlier*, visto que a monitoração se dá a cada novo dado recebido (as séries são sempre modeladas de forma univariada). Ou seja, ou o dado existe e deve-se verificar se é um *outlier*, ou não existe e deve ser tratado como dada faltante. No caso de registros multivariados, como cada novo dado é, em geral, multidimensional (por exemplo, linha ou coluna de uma tabela, representando um novo evento), faz-se necessária a verificação paralela de dados faltantes e *outliers*, que podem ocorrer simultaneamente entre os valores de um determinado evento. Apesar de não estar explicitamente indicado no esquema da Figura 4.6 (para simplificar o fluxograma), a detecção dos possíveis *outliers* deve ser feita antes da substituição dos dados faltantes, para evitar a introdução de tendência na estimação destes.

Notemos que a monitoração da dimensão da pontualidade é feita após a monitoração das outras duas dimensões consideradas, conforme será explicado em 4.3.3.

Uma etapa fundamental deste fluxo é a *validação* do dado recebido. Este processo de validação é feito com base nos resultados da parte automática do sistema e numa possível intervenção do supervisor. Dada a importância deste processo, continuamos o texto com uma discussão sobre o assunto.

4.2.2.1 Validação de Novos Dados

O procedimento de validação de novos dados pode ser considerado o núcleo do SMQD, pois permite a monitoração em tempo real das bases de dados envolvidas. Isso reflete o princípio de que a monitoração deve ser feita de forma contínua e através de pequenas ações.

O processo de validação dos dados (dentro da fase de ETL), segundo o modelo da Figura 4.3, se dá através do cruzamento entre a informação a ser validada e a informação de referência [28]. Este cruzamento, tal como está ilustrado na Figura 4.7,

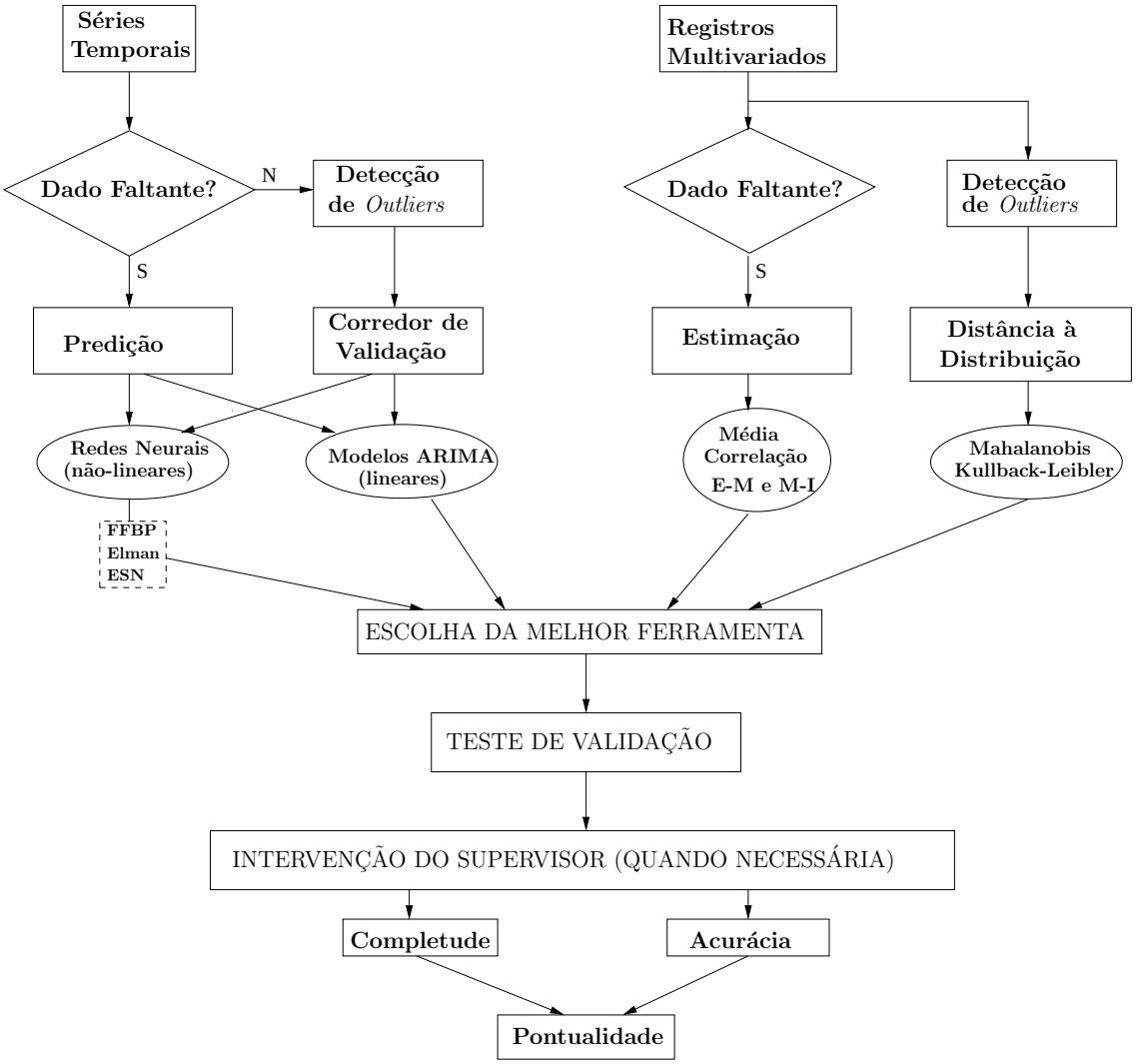

Figura 4.6: Fluxo de dados dentro do SMQD.

leva à classificação do novo dado como "validado" ou "não-validado".

A informação de referência pode ser de 2 tipos, a saber:

- *Meta-informação.* Trata-se de informação independente dos dados, ou seja, regras que devem ser satisfeitas quaisquer que sejam os valores dos dados. Exemplos são o formato dos dados, a validade do campo (para marcar campos em branco), relações entre variáveis etc. O teste feito com meta-informação é chamado de teste *técnico*;

- *Modelos estatísticos.* São obtidos a partir de dados já validados, e têm a forma de relações aproximadas. Tais testes estatísticos incluem o cálculo da correlação entre variáveis, análise de séries temporais, distância entre distribuições etc.

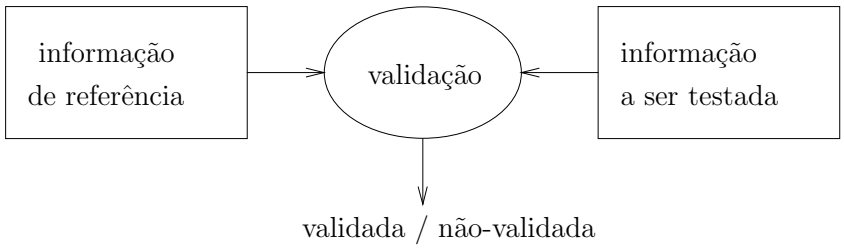

Figura 4.7: Validação de um dado novo.

Os testes que usamos no sistema desenvolvido neste trabalho encontram-se principalmente no segundo tipo. Neles, usamos informação retirada da própria base de dados para formar as métricas para a validação de dados novos. Isto permite ao SMQD trabalhar de forma semi-automática. Por exemplo, o histórico de uma série temporal é utilizado para validar os seus valores futuros, assim como para traçar as tendências de evolução da série. Analogamente, amostras de determinadas variáveis são usadas para modelar as suas distribuições; tais modelos podem então ser usados para validar novos dados.

Estes testes estão implementados através das ferramentas listadas anteriormente e descritas mais adiante, na Seção 4.3.

A seguir, explicaremos como o SMQD atua na monitoração de cada uma das dimensões de QD envolvidas, seguindo o esquema hierárquico ilustrado na Figura 4.5.

4.3 Monitoração das Dimensões Objetivas

Já vimos que são três as dimensões a serem monitoradas pelo SMQD, todas elas objetivas: Acurácia, Completude e Pontualidade. Esta seção é dedicada à descrição de como se faz a monitoração de cada uma dessas dimensões a partir das técnicas estatísticas e de processamento de sinais empregadas.

4.3.1 Monitoração da Acurácia

A monitoração da acurácia tem por objetivo tornar a base de dados o mais livre de erros possível. No SMQD, isto é implementado através da detecção e remoção (com posterior substituição) dos *outliers*.

Começamos esta discussão com algumas considerações gerais a respeito da incidência de *outliers*.

4.3.1.1 Detecção de *Outliers*

Na maioria dos conjuntos de dados, alguns pontos apresentam valores muito distantes do que se esperaria segundo a distribuição dos demais pontos. Isto pode acontecer devido a erros sistemáticos, falhas na teoria que gerou os valores esperados ou simplesmente dá-se o fato de que algumas observações encontram-se bastante afastadas do centro do conjunto de dados. Os assim chamados *outliers* podem estar indicando, portanto, a presença de dados defeituosos, procedimentos errôneos ou áreas em que a teoria usada para descrever os dados não é válida. Por isso, costuma-se dizer que os "*outliers* não devem ser eliminados, mas compreendidos"[2]. De fato, a metodologia do SMQD prevê que o supervisor seja sempre consultado quanto ao descarte de um dado que se encontre fora do seu *corredor de validação*, conforme veremos mais adiante.

Em distribuições normais, um pequeno número de *outliers* é esperado. Além disso, costumam ser classificados em dois grupos [65]:

1. *Outliers* suaves: Definindo Q_1 e Q_3 como o primeiro e terceiro quartis, res-

[2]No SMQD, ainda que um dado seja classificado como *outlier* e, portanto, não seja validado, fica sempre guardado de forma separada para uma posterior análise *offline*, quando outras informações podem vir a elucidar a sua origem.

pectivamente, e $IIQ = (Q_3 - Q_1)$ como o intervalo interquartil, chamamos de *outliers* suaves os dados d que satisfazem:

$$d < Q_1 - 1.5IIQ \text{ ou } d > Q_3 + 1.5IIQ$$

2. *Outliers* extremos: Analogamente, são aqueles dados que satisfazem:

$$d < Q_1 - 3IIQ \text{ ou } d > Q_3 + 3IIQ$$

Segundo essas definições, em uma distribuição normal (gaussiana) encontra-se 1 *outlier* suave para cada 150 dados, e 1 *outlier* extremo para cada 425.000 dados [65][66].

Todavia, esta é apenas uma maneira de se definir um *outlier* em termos de distância à distribuição. O mais importante aqui é que a intervenção humana (de um especialista) é necessária. Por exemplo, novas observações podem ser excluídas pela ação direta de um supervisor, que percebe que tais observações possuem algum viés devido a algum fator temporário de que ele tem conhecimento (e o qual o sistema não seria capaz de detectar a tempo). Por outro lado, o sistema é desenvolvido para fazer a triagem automática de todos os dados, buscando identificar os que parecem ser *outliers*. Cada um desses *outliers* deve ser encaminhado ao supervisor apropriado para julgar a conveniência de se incluir ou não tal amostra no modelo dos dados [67]. De fato, cada *outlier* pode possuir uma origem razoável, ou simplesmente ser resultado de erro.

Vejamos agora como a problemática dos *outliers* é tratada pelo SMQD para o caso de séries temporais e de registros multivariados.

Validação em Séries Temporais A monitoração da acurácia em séries temporais se dá, no SMQD, a cada novo dado recebido. Isto garante uma ação imediata do sistema, impedindo a propagação de erros, uma vez que, quando recebido e validado, o novo dado é incorporado à base de dados. Se o sistema permite que um dado errado seja integrado ao modelo e participe da validação da amostra seguinte, rapidamente deixa-se de rastrear a série monitorada devido à realimentação do erro.

Este processo de utilização das amostras anteriores na validação da amostra recém-chegada se dá através do corredor de validação, descrito a seguir.

Corredor de Validação As ferramentas utilizadas para implementar o corredor de validação para séries temporais utilizam amostras passadas da série para gerar um intervalo dentro do qual se espera que recaia a amostra seguinte. Ou seja, em uma série $\{X(t)\}$, os valores de $x(t)$, $x(t-1)$, $x(t-2)$, ..., $x(t-N+1)$, são usados para estabelecer os limites mínimo e máximo para o valor de $x(t+1)$. O valor de N é otimizado para cada série e ferramenta utilizada.

A Figura 4.8 ilustra esse processo. O primeiro intervalo do corredor de validação representado no gráfico, correspondente ao instante $t+1$, é usado para validar o dado $x(t+1)$. Este valor é incorporado à base de dados e ao modelo da série caso esteja dentro do intervalo $[x_I(t+1), x_S(t+1)]$. Os demais intervalos do corredor, representados com linhas tracejadas, vão sendo gerados nos instantes de tempo posteriores.

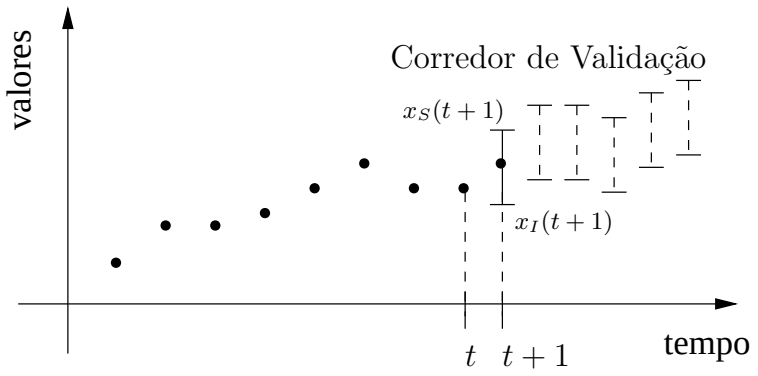

Figura 4.8: Corredor de validação para novas amostras.

Caso o valor recebido para $x(t+1)$ esteja fora do corredor de validação, o caso é submetido à análise do supervisor, que decide pela incorporação, ou não, do dado à base de dados, ainda que não necessariamente o integre ao modelo da série monitorada.

Entendamos melhor esta problemática através de um exemplo. Na Figura 4.9 vemos a evolução do valor diário das ações da Hawaiian Airlines em torno ao atentado de 11 de setembro de 2001 [47]. Entre os dias 10 e 17 (dentro desse período a bolsa de valores não fez operações) houve uma queda abrupta no valor da ação ("quebra estrutural"), impossível de prever através dos valores anteriores da série. Dessa forma, o sistema de monitoração rejeitaria o valor da série no dia 17, indicando haver erro na sua aquisição. Neste caso, porém, o supervisor tem subsídios para

julgar que este valor, apesar de não-usual, está correto. É um *outlier* que tem explicação, chegando inclusive a introduzir uma mudança de regime na série (que passa a evoluir em torno a uma nova média).

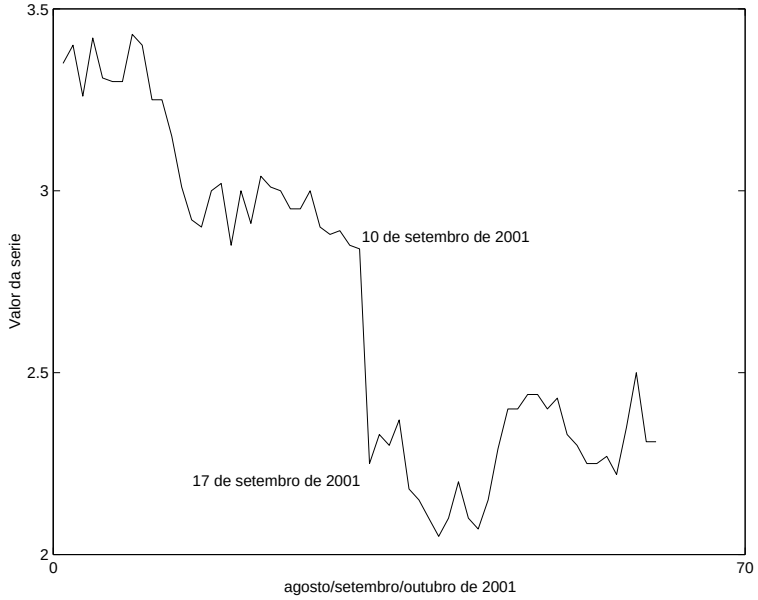

Figura 4.9: Descontinuidade ("quebra estrutural") no valor das ações da Hawaiian Airlines em torno ao 11 de setembro de 2001 (o mercado de ações ficou fechado por uma semana).

Este exemplo constitui um típico caso em que a ação do supervisor é essencial. A nova amostra deve ser incluída na base mesmo não tendo sido validada e, aos poucos, o modelo da série deve se adaptar ao novo regime.

Existem, portanto, três possibilidades para a procedência das amostras que estejam fora do corredor de validação (e que, portanto, classificamos como *outliers*), conforme já mencionamos anteriormente:

- *Irregularidades.* São dados corretos (por exemplo, o ocorrido na série da Figura 4.9), mas que não são incorporados imediatamente ao modelo preditor, já que não têm comportamento regular. Uma vez detectada a irregularidade (a partir de análise posterior, *offline*), o dado pode ser carregado normalmente na base e, se a mudança de regime é confirmada (pela evolução das observações seguintes), pode ser inserido ao modelo preditor. Caso contrário, é substituído pelo valor predito pelo modelo;

- *Dados falsos.* Trata-se aqui de observações efetivamente com erro. O SMQD age corretamente em rejeitá-las, substituindo-as pelo valor predito e, quando disponível, pelo valor real (que seria obtido posteriormente, a partir de verificação na fonte por parte do supervisor);

- *Dados corretos.* Neste caso, é o corredor quem está incorreto. Quando se descobre o erro posteriormente, retifica-se o valor carregado na base, atualizando o modelo. É necessário, aqui, que o sistema refaça, através do modelo atualizado, a validação das observações posteriores à que foi rejeitada pelo corredor incorreto.

Como são gerados os limites x_I e x_S do corredor de validação? Para sua formação, partimos do princípio de que os *outliers* podem ser identificados através da análise do erro de previsão $e(T) = x(T) - \tilde{x}(T)$, onde $\tilde{x}(T)$ é o valor predito pelo sistema para o instante T e $x(T)$ é o valor coletado pelo sistema nesse instante, e que deve passar pelo processo de validação. Intuitivamente, sabemos que, se este erro é grande o suficiente, podemos concluir que a observação $x(T)$ contém erro ou foi gerada por um processo diferente. Dessa forma, o teste de verificação de *outliers* assume naturalmente a forma:

$$\left|\frac{e(T)}{\hat{\sigma}_e}\right| > k \tag{4.4}$$

onde tipicamente $4 < k < 5$, tendo seu valor ajustado de acordo com a série e com o modelo desenvolvido [67]. O valor $\hat{\sigma}_e$ corresponde ao valor RMS da distribuição dos valores do erro $e(T)$ acumulados até o instante $(t-1)$. Assim, a largura total do corredor vale $2k\hat{\sigma}_e$, e os limites são $x_I(T) = \tilde{x}(T) - k\hat{\sigma}_e$ e $x_S(T) = \tilde{x}(T) + k\hat{\sigma}_e$. Caso o dado $x(T)$ não seja validado, o valor estimado $\tilde{x}(T)$ é incorporado ao modelo em seu lugar.

O ajuste no valor de k é o que regula as probabilidades de "perda" (P_M) e "falso alarme" (P_F), tal como estão definidas na teoria de detecção de sinais [68]. A probabilidade P_M mede a chance de validarmos um *outlier*, e P_F mede a chance de classificarmos como *outlier* um dado que faz parte do modelo e deveria ser validado. A filosofia do SMQD privilegia diminuir o valor de P_F (com um possível aumento em P_M), ou seja, deseja-se validar todos os dados reais, ainda que com isso não se

detectem alguns *outliers*. Idealmente, a qualidade do modelo deve ser tal que ambas as probabilidades tendam a ser minimizadas.

Reparemos que esta metodologia para a detecção de *outliers* atua de forma independente da ferramenta utilizada para modelar a série. Ou seja, cabe à ferramenta apenas gerar os valores de $x_I(T)$ e $x_S(T)$, a partir do valor de $\tilde{x}(T)$ e de $\hat{\sigma}_e$.

Continuemos a descrição do processo da monitoração da Acurácia em séries temporais comentando o procedimento de pré-processamento das séries.

Pré-processamento das séries Conforme visto no Capítulo 3, procedemos a estacionarização das séries analisadas antes de passá-las por qualquer processo de modelagem.

Dessa forma, estabelecemos um procedimento de estacionarização que deve ser aplicado em todas as séries que são monitoradas pelo SMQD. Este procedimento está descrito na Figura 4.10, apresentada no Capítulo 3 e repetida aqui.

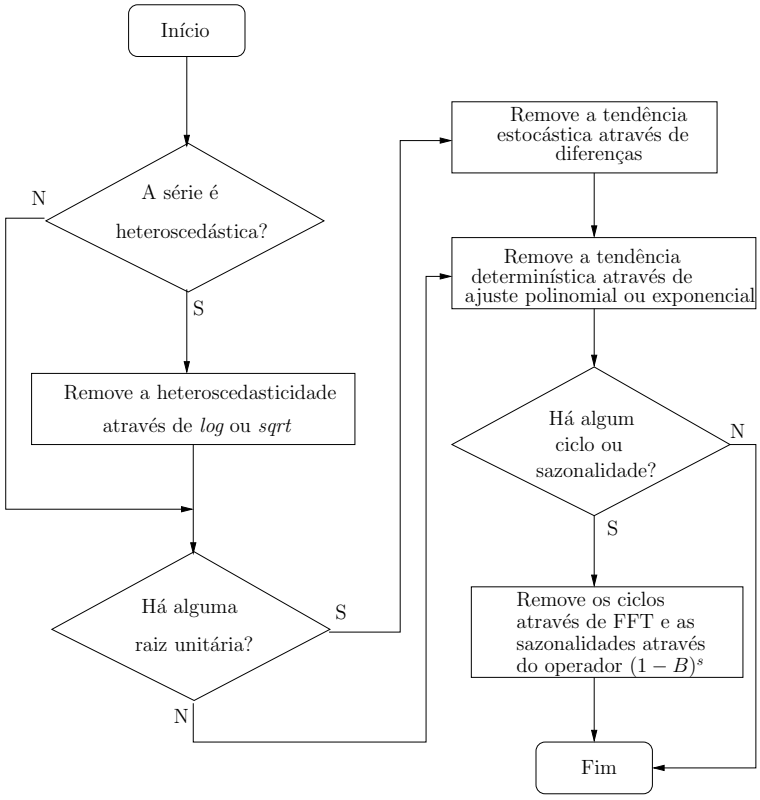

Figura 4.10: Procedimento para a estacionarização de séries temporais.

Este procedimento garante que a série é estacionarizada e está pronta para

ser modelada; porém, algumas vezes é conveniente acrescentar uma nova etapa à fase de pré-processamento, de acordo, sobretudo, com o tipo de série monitorada. Esta última etapa consiste, em geral, na filtragem da série residual (já estacionária). Um exemplo de filtragem utilizada é a de Kalman [45][69], que em alguns casos contribui para o aumento da eficácia do modelo preditor. Esta questão voltará a ser comentada com mais detalhes no Capítulo 5, quando apresentarmos os estudos de caso realizados com o SMQD.

Também no Capítulo 5 veremos casos concretos de utilização dessa dinâmica de pré-processamento, incluindo o procedimento de atualização a ser adotado quando da chegada de novos dados. A seguir, descreveremos as ferramentas utilizadas para a modelagem de séries temporais.

Modelo Linear: ARIMA Por se tratar de uma metodologia consolidada na área de modelagem de séries temporais, usamos os modelos ARIMA como uma das alternativas para a modelagem das séries monitoradas pelo SMQD.

O método ARIMA (*Auto-Regressive Integrated Moving Average*) constitui uma extensão do modelo ARMA para séries temporais não-estacionárias [36][69][70]. Os modelos ARMA combinam um processo de média móvel (MA):

$$x_t = \mu + \theta_0\varepsilon_t + ... + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q} \tag{4.5}$$

(onde ε é um processo de ruído branco com média zero e variância σ_ε^2) com um processo auto-regressivo (AR):

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + \varepsilon_t \tag{4.6}$$

Combinando as equações acima, vemos que um processo x_t, ARMA(p, q) e com média zero pode ser escrito da forma:

$$x_t - \phi_1 x_{t-1} - ... - \phi_p x_{t-p} = \varepsilon_t - \theta_1\varepsilon_{t-1} - ... - \theta_q\varepsilon_{t-q} \tag{4.7}$$

(aqui, ε pode ser chamado de *inovação* [69], que assumimos ser gaussiana e descorrelacionada).

Definindo B como o operador de atraso (ou seja, $B(x_t) = x_{t-1}$), temos:

$$\phi(B) = 1 - \phi_1 B - ... - \phi_p B^p \tag{4.8}$$

$$\theta(B) = 1 - \theta_1 B - ... - \theta_q B^q \tag{4.9}$$

Sabendo que os modelos ARMA são inadequados para séries não-estacionárias, devemos usar séries de diferenças, já que essas são estacionárias. Ajustando os modelos ARMA para séries de diferenças, chegamos aos modelos ARIMA(p, d, q), dados por:

$$\phi(B)\nabla^d(x_t) = \theta(B)\varepsilon(t) \tag{4.10}$$

onde $\nabla = 1 - B$ é o operador de primeira diferenciação com $\nabla^d = (1 - B)^d$ sendo o d-ésimo operador de diferenciação. O cálculo dos valores para os coeficientes do modelo, geralmente através de autocorrelação e análise do espectro em freqüência, nos dá as estimativas dos valores futuros da série x_t. Problemas com a metodologia ARIMA aparecem quando se trabalha com séries temporais de variância alta ou quando as séries representam processos não-lineares. Nesses casos, o uso de outros modelos, como os neurais, pode proporcionar maior eficácia.

Modelos Neurais Não-Lineares O uso de Redes Neurais Artificiais (ou RNAs) na modelagem de séries temporais tem aumentado bastante nos últimos anos. Algumas referências sobre o assunto podem ser encontradas em [71]–[82].

Há diferentes arquiteturas de RNAs propostas na literatura. Nós desenvolvemos ferramentas baseadas em 3 modelos não-lineares diferentes, os quais descreveremos a seguir. Todos eles são usados para modelar as séries já pré-processadas pelo procedimento descrito anteriormente.

Rede Não-Recorrente (FFBP) A arquitetura mais tradicional no uso de RNAs na modelagem de séries temporais é a estrutura que propaga o sinal de entrada (*feedforward*, FF) através de camadas de neurônios e retropropaga o erro de saída (*backpropagation*, BP) para proceder o ajuste dos pesos sinápticos [35]. Dessa forma, constitui a nossa primeira opção dentro do SMQD.

A Figura 4.11 ilustra o método adotado para treinar uma rede neural para a predição em séries temporais. Para a série $\{x_t\}$, fornece $\tilde{x}_{t+1}$, a estimativa de x_{t+1},

com base nos valores de x_t, x_{t-1}, ..., x_{t-N+1}.

Para cada neurônio desta rede, a sua saída y é calculada a partir das suas entradas x_i (que por sua vez podem ser as saídas dos neurônios da camada anterior ou propriamente as entradas da rede neural) da seguinte forma: $y = f(\sum w_i x_i)$, onde $f(\cdot)$ é a chamada função de ativação do neurônio (podendo ser linear ou não) e w_i são os pesos sinápticos entre cada entrada e o neurônio.

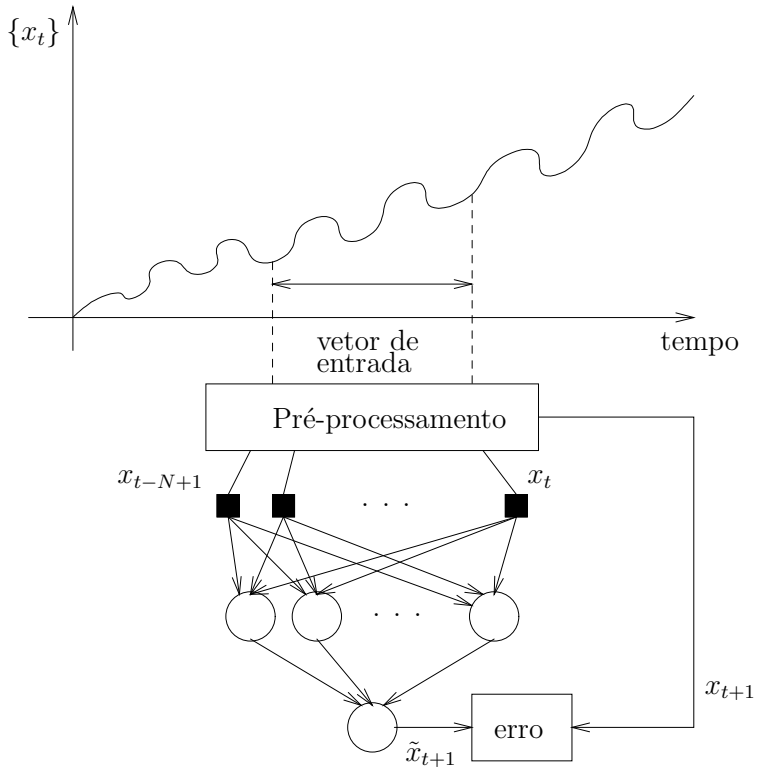

Figura 4.11: Modelo básico para o estimador neural de séries temporais.

Uma variante do modelo descrito acima é a rede neural de base N. O que se pretende neste novo modelo é não permitir que os vetores de entrada da rede neural estimadora tenham amostras repetidas, nem mesmo em posições diferentes. Isto é, se o primeiro vetor de entrada é formado por $\{x_1, x_2, ..., x_N\}$ visando a estimar x_{N+1}, o próximo deve ser $\{x_{N+1}, x_{N+2}, ..., x_{2N}\}$ visando à estimação de x_{2N+1}. Desta forma, reduz-se N vezes o tamanho do conjunto de entrada da rede neural, além de só estimarmos os valores da série a cada N amostras. Para alguns tipos de série, esta estrutura retorna estimações mais acuradas que o modelo anterior (onde $N = 1$). Este esquema é ilustrado na Figura 4.12, para o caso particular de $N = 3$. Reparemos que nesse caso são necessárias N redes para fazer a estimação contínua da série.

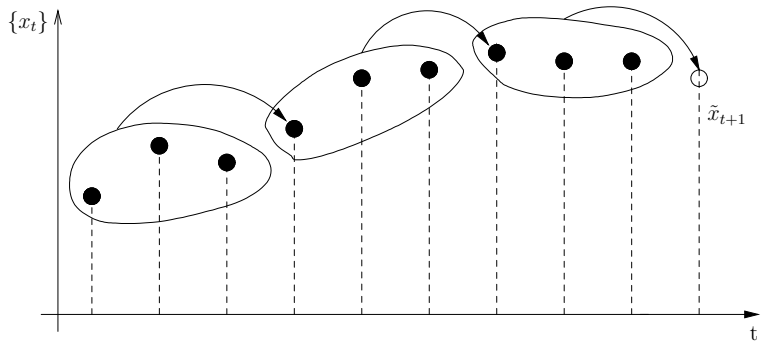

Figura 4.12: Esquema entrada-saída para o estimador de base $N = 3$.

Comentário: Validação Cruzada Com o objetivo de aumentar a capacidade de generalização das redes neurais envolvidas na modelagem de séries temporais, utilizamos no seu treinamento o procedimento de validação cruzada com k dobras, em que o conjunto de dados original é particionado em k sub-conjuntos, dos quais um é reservado como conjunto de validação para testar o modelo. Os demais $(k-1)$ sub-conjuntos são usados durante o treinamento. Neste caso, o processo de validação é repetido k vezes, usando, em cada uma delas, um dos k sub-conjuntos para validação. Os k resultados obtidos podem ser combinados (através, por exemplo, de média) para produzir uma estimação única [83].

Rede Recorrente 1: Elman Descrita em [84] por Jeffrey L. Elman, a rede de Elman é uma das mais utilizadas arquiteturas recorrentes em redes neurais [35]. A Figura 4.13 ilustra a arquitetura genérica de uma rede neural recorrente de Elman. A principal diferença com relação à estrutura FFBP é o fato de as redes de Elman possuírem realimentação da camada escondida, o que permite que todas as observações passadas contribuam para a predição das observações futuras.

Esta característica torna a rede de Elman particularmente capaz de detectar estruturas no tempo e, portanto, modelar a predição em séries temporais. Devido à realimentação na camada escondida, o problema da estabilidade em redes recorrentes tem sido bastante pesquisado e as condições para evitar a divergência no treinamento são já conhecidas [85].

Também com a rede de Elman se pode utilizar a estrutura de base N descrita para a rede FFBP.

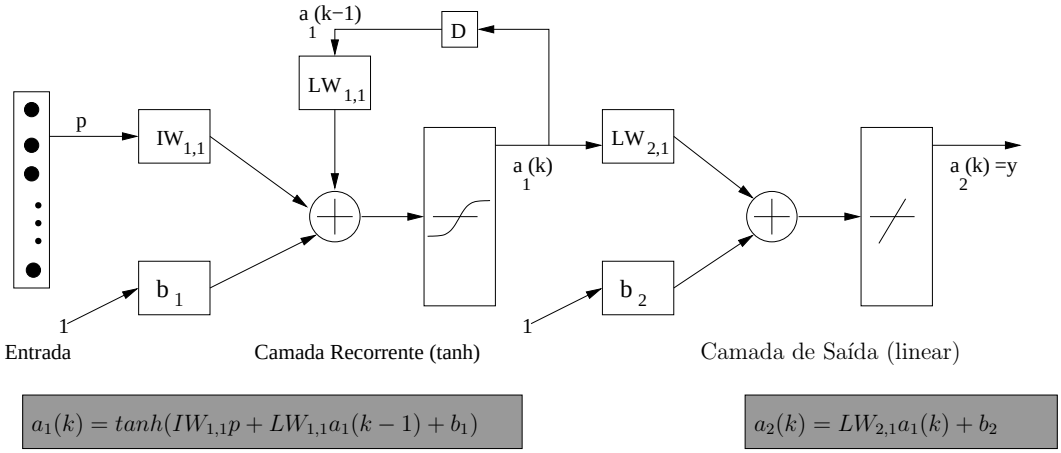

Figura 4.13: Esquema geral para a arquitetura de Elman.

Rede Recorrente 2: ESN Há alguns tipos de série, particularmente as de comportamento caótico [86], cuja modelagem através de redes neurais do tipo *feed-forward*, como a FFBP e a rede de Elman, não é bem sucedida.

Nos últimos 5 anos, uma nova arquitetura para redes recorrentes, as *Echo State Networks* (ESN) [87]–[90], tem demonstrado uma surpreendente capacidade de modelar o comportamento caótico.

Dessa forma, implementamos um preditor para séries temporais usando a arquitetura ESN, cujo esquema geral é mostrado na Figura 4.14. As séries modeladas pelo esquema baseado em ESNs passam pelo mesmo pré-processamento imposto às séries que são tratadas com os demais tipos de rede.

As redes ESN trabalham com pesos sinápticos fixos (representados pelas linhas cheias na Figura 4.14) para a parte recorrente da rede, sendo treinados apenas os pesos que ligam o seu "reservatório dinâmico" à saída (linhas tracejadas na figura). Este reservatório de neurônios deve ser grande o suficiente[3] para conseguir armazenar os "estados de eco" que fazem o mapeamento entrada-saída da rede.

O treinamento dos pesos da saída podem ser feitos de maneira *o*nline (a cada par entrada-saída que é apresentado à rede) ou de maneira *offline*, através da solução de Wiener [87]. É através do treinamento desses pesos que o erro entre a saída da rede e o sinal de referência é minimizado (de forma análoga ao algoritmo BP). Alguns parâmetros, como o tamanho do reservatório dinâmico (número de neurônios) e o grau de interconexão entre esses neurônios (variando entre 0 e 1,

[3]O número ideal varia conforme o caso e em geral é alcançado a partir de múltiplas tentativas

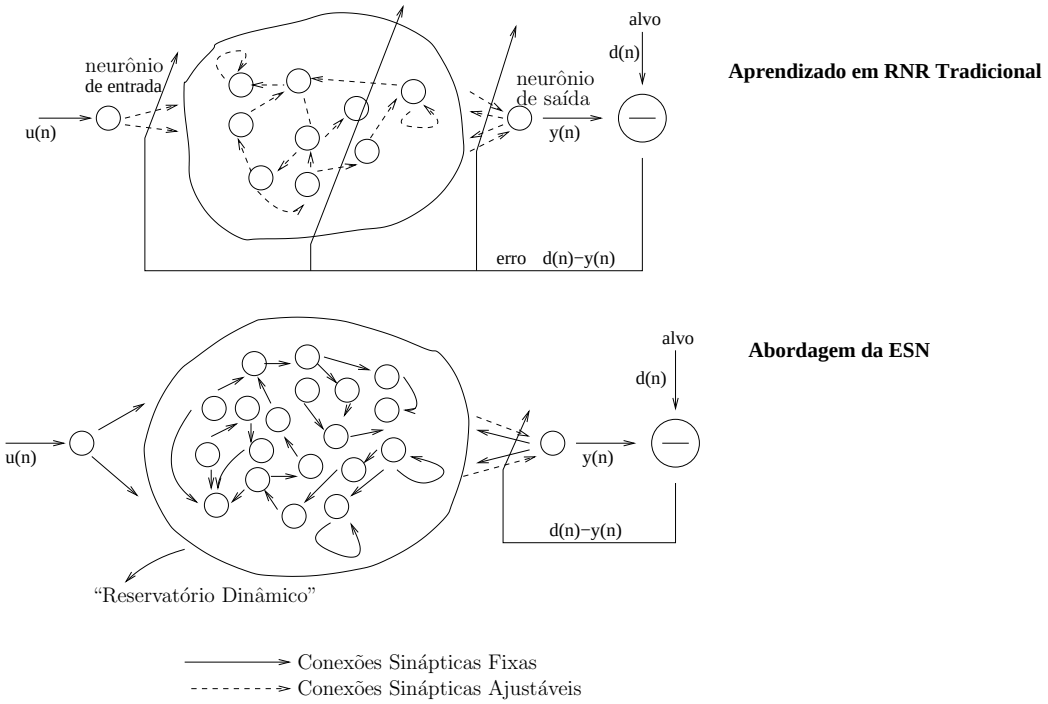

Figura 4.14: Estrutura das redes ESN em comparação com o aprendizado em Redes Neurais Recorrentes (RNRs) tradicionais.

sendo 1 a configuração em que todos os neurônios do reservatório estão conectados entre si), devem ser ajustados. Não há regras fixas para a determinação destes parâmetros, de maneira que o projetista deve testar, a cada caso, uma quantidade razoável de valores, para então escolher a melhor configuração. Também a questão da estabilidade das redes ESN tem sido discutida, e as condições para assegurá-la encontram-se descritas em [89].

O treinamento das redes ESN consiste em formar uma representação espacial (conjunto de bases) da relação entrada-saída através dos "estados de eco" que se formam no reservatório dinâmico. Dessa forma, quando se coloca na entrada da rede um novo evento (do conjunto de teste, por exemplo), a função da rede consiste em encontrar, dentre os "estados de eco", o que mais se assemelha à entrada, gerando assim a resposta mais próxima à desejada. No Capítulo 5 veremos os casos em que o uso de redes desse tipo pode trazer benefícios à modelagem e, conseqüentemente, à monitoração da QD.

Redes Classificadoras Os modelos neurais MLP (*Multilayer Perceptron*) descritos acima (FFBP, Elman) podem servir tanto como estimadores para os valores

das observações futuras da série analisada, bem como classificadores da *tendência imediata* da série, isto é, podem ser treinados para indicar se a amostra seguinte será maior ou menor que a atual.

Em geral, as redes neurais classificadoras atuam de forma a auxiliar a estimativa feita pelo modelo preditor, confirmando ou não a subida ou descida de valor com relação à amostra anterior. Para o supervisor, é uma informação a mais que o ajuda a decidir pela aceitação ou não do teste de validação feito, quando é chamado a intervir.

A rede classificadora opera com a mesma topologia mostrada na Figura 4.11. Neste caso, o único neurônio de saída tem como função ativadora a tangente hiperbólica. Sendo assim, a rede conseguiria, a princípio, distinguir entre duas classes. Os primeiros resultados com esse tipo de rede demonstraram que, pelo fato de que boa parte das flutuações no valor da série é muito pequena (variação próxima de zero), havia um risco grande de erro na saída da rede, fazendo com que a rede mais prejudicasse que ajudasse. A solução adotada para esse problema foi a criação de uma terceira classe, em que se considera que a série permanece estável se o seu valor oscila dentro de uma pequena faixa pré-estabelecida.

Para a implementação desta versão da rede classificadora, testaram-se duas alternativas. A primeira consiste no caminho natural da inclusão de mais um neurônio na camada de saída, permitindo à rede a distinção entre três classes através de codificação binária. Na formação dos pares entrada-saída para o conjunto de treino, os alvos formados eram 00 (para a classe de descida), 11 (para a classe de subida) e 01 (para a classe de estabilidade, definida a partir do parâmetro P_E).

A outra maneira, que se mostrou mais eficiente e vantajosa computacionalmente, consistiu em manter apenas um neurônio na camada de saída, definindo como alvos os valores -1 e 1 (para descida e subida) e 0, para a estabilidade. Estabeleceram-se então dois patamares de corte (ao invés de apenas um, no valor 0) para criar a região da terceira classe. Tais patamares são $\frac{-P_E}{100}$ e $\frac{P_E}{100}$. O intervalo entre $\frac{-P_E}{100}$ e $\frac{P_E}{100}$ forma a região de estabilidade. Dessa forma, a saída da rede (y_t) é interpretada da seguinte forma:

- Se $y_t < \frac{-P_E}{100}$, considera-se que a observação seguinte será menor que a atual;
- Se $\frac{-P_E}{100} < y_t < \frac{P_E}{100}$, considera-se que a observação seguinte não se moverá

significativamente (o que significa que a rede "abstém-se" de opinar pela subida ou descida do valor da observação seguinte);

- Se $y_t > \frac{P_E}{100}$, considera-se que a observação seguinte será maior que a atual.

Existem, resumidamente, dois tipos de erro que podem acontecer na saída do classificador: indicação de estabilidade (dentro do patamar) quando há subida ou descida no valor da série, e a indicação de variação (seja subida ou descida) quando há estabilidade. No nosso sistema, devemos definir a faixa de estabilidade de maneira tal que se minimize a probabilidade de ocorrência de erros do segundo tipo, ainda que, com isso, a maior parte dos eventos seja classificada dentro da faixa de estabilidade. Como o classificador é responsável por auxiliar o supervisor em caso de dúvida na validação da amostra, é importante haver uma razoável prudência para se evitar que este teste falhe. Uma busca rigorosa por pequenas variações de tendência, logo no início da validação, poderia ocasionar a rejeição de amostras válidas.

O valor do patamar depende da série que está sendo analisada. Para séries financeiras, onde a variação entre observações consecutivas é bem pequena (e, por isso, é mais difícil detectar corretamente o sentido da variação), a tendência é que a região de estabilidade tenha que ser maior, para evitar erros na classificação. Para a série da IBM, por exemplo foi definido o patamar de 6% para a região de estabilidade. A esse valor se chegou através de testes com uma rede classificadora (FFBP): foi o mínimo valor para o qual a rede consegue atingir 100% de acerto na classificação[4]. A Figura 4.15 ilustra a variação percentual entre os valores de observações consecutivas para essa série, juntamente com o patamar de 6% para a região de estabilidade. Em torno de 5% das amostras variam (em módulo) mais de 6% em relação à amostra anterior. São exatamente essas amostras que a rede classificadora pretende rastrear, uma vez que são as que mais provavelmente representam um *outlier* (seja oriundo de irregularidade, dado verdadeiro que introduza mudança de regime na série ou erro na fonte ou aquisição do dado). Quando a variação percentual entre amostras consecutivas é menor que 6%, a rede classificadora se abstém de julgar a amostra, deixando a tarefa para o corredor de validação estabelecido pela rede estimadora.

[4]Como eram 100 eventos de teste, admite-se que precisão da eficiência de acerto é de 1%

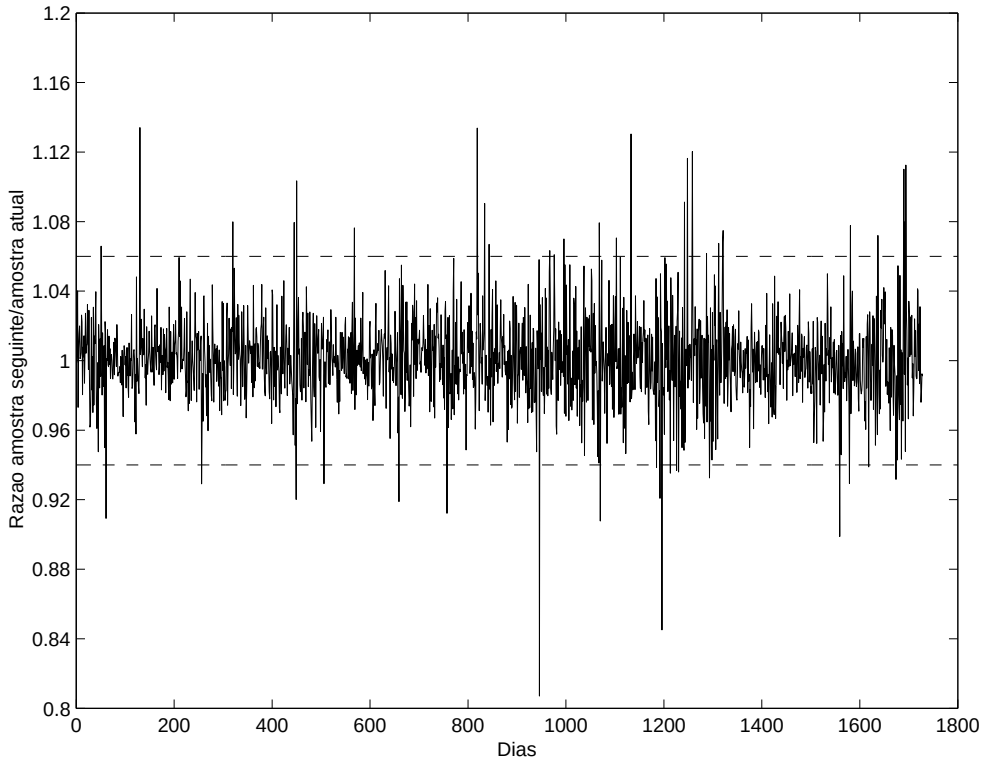

Figura 4.15: Razão entre observações consecutivas para a série da IBM. Patamar em 6%.

A atualização (retreino) do modelo neural Um último aspecto que deve ser contemplado, dentro do contexto da modelagem de séries temporais, é a questão da atualização dos modelos neurais desenvolvidos. Estes modelos devem ser atualizados periodicamente, visando a incorporar no modelo as observações mais recentes, ao mesmo tempo em que se descartam as amostras mais antigas da série. O SMQD prevê a opção de retreino da rede, deixando ao supervisor a decisão de escolher o momento de retreinar os modelos para cada série. Quando se solicita a atualização de um modelo neural para uma dada série, deve-se retreinar tanto as redes estimadoras quanto as classificadoras envolvidas.

Para tanto, utiliza-se o procedimento de Janela-Móvel (JM) [78], que consiste em retreinar periodicamente a rede neural com dados atualizados. Geralmente, é aplicado de maneira *offline* a um conjunto de dados, de maneira a simular as condições reais da evolução de uma série temporal e a testar a robustez do modelo através do freqüente retreino. No nosso caso, como o sistema está continuamente recebendo novos dados, o procedimento de JM é implementado através da substituição das amostras mais antigas pelas mais recentes, o que permite o retreino da

rede com a conseqüente incorporação ao modelo das novas características da série. A Figura 4.16 ilustra este processo[5].

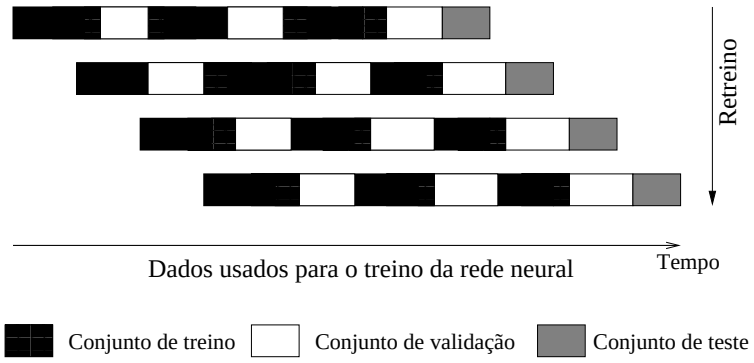

Figura 4.16: Esquema genérico para o procedimento de Janela-Móvel.

Quando opta pela atualização dos modelos, o usuário tem a possibilidade de alterar os valores dos parâmetros utilizados no treinamento das redes envolvidas. Entre esses parâmetros, destacam-se o número de meses que formam a janela de treino para a rede, assim como o número de meses ao término dos quais se faz novamente a verificação de "quebra de tendência". Consideramos haver uma quebra quando a média dos valores de um intervalo de um mês da série é maior (ou menor) que as médias dos meses adjacentes (imediatamente anterior e imediatamente posterior), ou seja, quando há formações de "picos" ou "vales". Tomando-se o intervalo de um mês e calculando a sua média, evita-se considerar como quebra de tendência o que, na verdade, são apenas flutuações naturais da série, seja em um período de subida/descida ou de estabilidade prolongada[6]. Reparemos que buscam-se quebras no *sentido* da tendência: de subida para descida ou vice-versa. A Figura 4.17 ilustra dois exemplos de quebra de tendência.

Quando se detectam quebras de tendência, passa-se a considerar, para o pré-processamento das séries que estão sendo modeladas, apenas o trecho posterior à última quebra detectada. Com isso, garante-se que os modelos de estimação e classificação estarão gerando saídas segundo a estatística atual da série.

[5]Na figura, a divisão entre treino e validação é apenas simbólica.

[6]Além disso, a média de um mês de pico deve ser, na verdade, maior que (1+Q) vezes a média dos meses adjacentes, ou menor que (1-Q) vezes as médias adjacentes. Isso também evita a detecção de quebras em períodos longos de estabilidade. Um valor típico para Q é 0,1, podendo ser ajustado pelo supervisor para cada série.

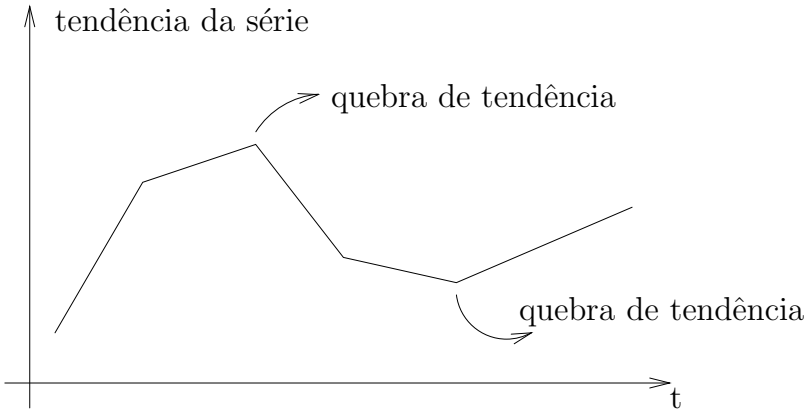

Figura 4.17: Ilustrando quebras de tendência.

A Figura 4.18 ilustra uma etapa deste processo de retreino. Nela, o trecho da série correspondente à janela de treinamento é mostrado juntamente com a série de média zero (obtida a partir da retirada da "tendência global" do trecho da série analisado). A parte de baixo da figura traz o espectro da série de média zero, onde podemos ver claramente a existência das componentes senoidais (sazonalidades) que devem ser retiradas.

Nesta figura, pode-se observar também a existência de quebra de tendência em torno da amostra 125. O SMQD detecta esta quebra (segundo o patamar estabelecido pelo usuário) e mostra posteriormente o trecho da série após a quebra e já normalizado para ter o valor máximo igual a 1 (ver Figura 4.19). O próximo passo é a retirada das componentes senoidais.

Após a retirada do número oportuno de componentes senoidais (ou diferenciação sazonal através de $(1-B)^s$, conforme o caso), o sistema prepara os conjuntos de treino, validação e teste para a atualização do modelo neural. Termina a normalização do conjunto de dados (subtrai de todos os eventos a média dos eventos do conjunto de treino, divindo-os por um múltiplo de seu desvio padrão). As funções de treinamento e aprendizado utilizadas na atualização do modelo neural podem ser escolhidas pelo usuário.

Finalmente, o usuário acompanha, através do sistema, a evolução do treinamento, escolhendo o melhor momento de terminá-lo. Neste ponto, o modelo neural já se encontra atualizado e apto para a validação de novas amostras. À medida que estas vão chegando, passam pelo mesmo pré-processamento (estacionarização e normalização) aplicado sobre a série durante o retreino, uma vez que os modelos

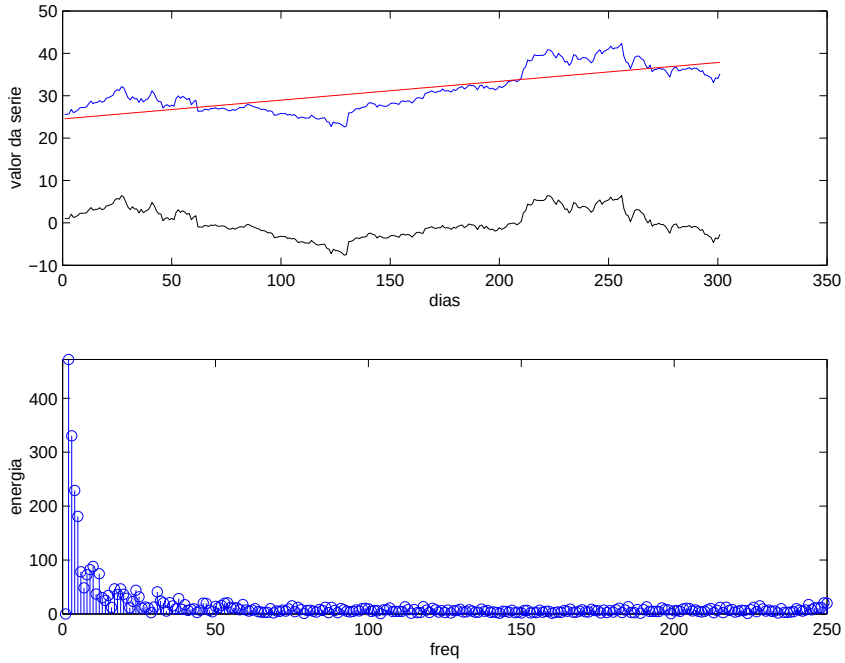

Figura 4.18: Exemplo de operação do SMQD durante a fase de pré-processamento. Acima, o trecho da série analisado juntamente com a retirada da tendência global. Abaixo, o espectro da série de média zero (com freqüência não normalizada).

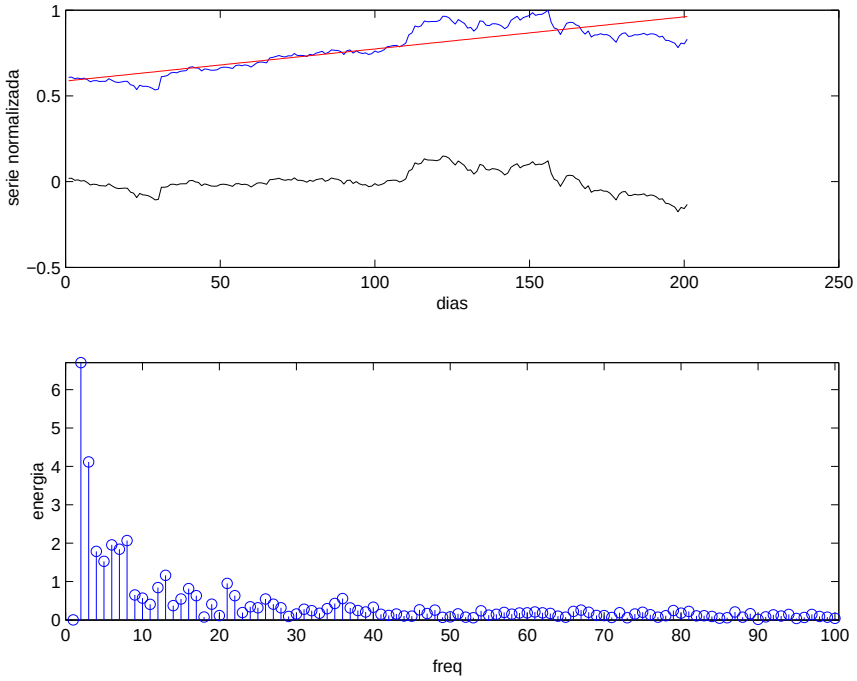

Figura 4.19: Trecho da série já normalizada e após a quebra de tendência, com o espectro da série de média zero.

são treinados sobre a série residual (o erro de predição, porém, é calculado sobre a série predita reconstruída, com a reposição do fator de normalização e dos ciclos/sazonalidades, tendência e heteroscedasticidade).

Voltemo-nos agora para o processo de monitoração da acurácia em registros multivariados.

Validação em Registros Multivariados A monitoração da acurácia em dados desse tipo se dá através do uso de algumas métricas que verificam a distância do novo dado (ou de um conjunto de dados, ou seja, uma amostra) à distribuição presente na base de dados e, portanto, já validada.

Para medir a distância de um novo dado (evento) singular à base de dados, usamos uma ferramenta baseada na *Distância de Mahalanobis*; para medir a distância entre a distribuição validada e uma nova amostra, utilizamos a *Distância*[7] *de Kullback-Leibler*.

Uso da Distância de Mahalanobis na Detecção de *Outliers* Introduzida por P. C. Mahalanobis em [91], a métrica de Mahalanobis é uma distância estatística que pode ser utilizada para a detecção de *outliers* em dados multivariados. O principal objetivo com esta métrica é a verificação da pertença ou não de um determinado ponto, em particular, a um grupo de dados. A distância de Mahalanobis (DM) tem a vantagem de utilizar, além das médias e variâncias para cada variável, a covariância entre as medidas. Daí a sua vantagem com relação, por exemplo, à distância euclidiana.

A distância D_g^2 de um ponto x a um grupo de dados g é igual a:

$$D_g^2(x) = (x - m_g)^T S_g^{-1} (x - m_g) \tag{4.11}$$

onde S_g representa a matriz de covariância do grupo g e m_g é o vetor de médias das variáveis envolvidas. Repare que se a matriz S_g for a identidade, a distância de Mahalanobis toma a forma da distância euclideana (caso particular).

Já podemos perceber, desde já, que a DM de um *outlier* ao restante da distribuição será bem maior que as DMs dos pontos realmente pertencentes à distribuição.

[7]Neste caso, o termo mais correto é "divergência".

Desta forma, uma das maneiras de se detectar *outliers* é calcular a DM de todos os pontos de um dado conjunto com relação ao próprio conjunto (repare que a DM é sempre calculada entre um ponto e uma distribuição; a DM entre dois pontos equivale à distância euclideana). A Figura 4.20 ilustra um grupo de dados em que alguns pontos têm junto a si o valor de sua DM. O ponto cuja DM vale 13,89 é claramente um *outlier*, e aquele cuja DM vale 0,33 certamente pertence à distribuição. O problema se coloca em pontos como o que se destaca na figura com DM igual a 7,49: pode ser considerado um *outlier*?

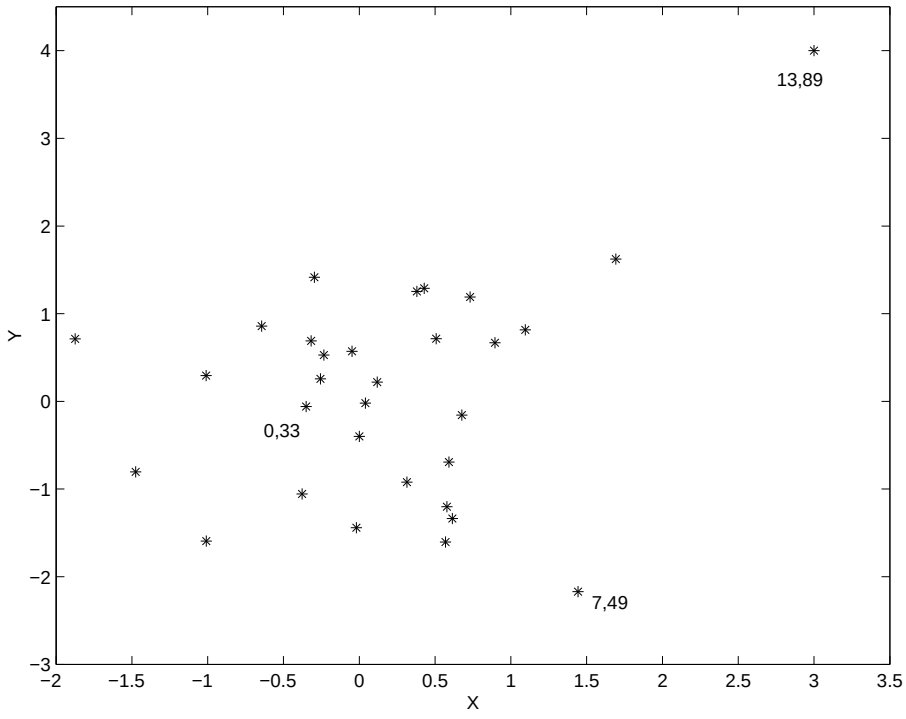

Figura 4.20: Detectando *outliers* através da distância de Mahalanobis.

Vemos que, em uma amostra de tamanho pequeno, este problema fica mais acentuado, visto que a distribuição ainda não se encontra totalmente caracterizada. No caso do SMQD, a tendência é que as distribuições cresçam rapidamente, devido ao grande volume de dados recebido pelo sistema. Assim sendo, o SMQD pode aguardar até que os conjuntos de dados monitorados atinjam um determinado tamanho mínimo e, a partir daí, começar a busca por *outliers*. Posteriormente, as novas amostras são testadas a partir de confrontamento com o conjunto já validado.

Com base nas dificuldades expostas até agora, vemos que o algoritmo de detecção de *outliers* não pode simplesmente afirmar que os pontos que tiverem maiores

DMs devem ser descartados. Como, para nós, descartar um dado válido é pior que não descartar um *outlier*, o SMQD deve primar pela prudência à hora de classificar um dado como *outlier*. Desta forma, propomos o seguinte algoritmo:

1. Recebe uma nova amostra $\{x\}$;
2. Calcula a DM_i para todos os pontos dessa nova distribuição[8];
3. Calcula a média μ e o desvio σ dos valores de DM_i;
4. Se a DM de algum dado específico da nova amostra for maior que $\mu + n\sigma$, classifique tal ponto como *outlier* (removendo-o da distribuição);
5. Atualiza o conjunto de dados com a nova amostra já verificada.

A Figura 4.21 ilustra um caso de remoção de *outliers* bastante claros através do algoritmo acima.

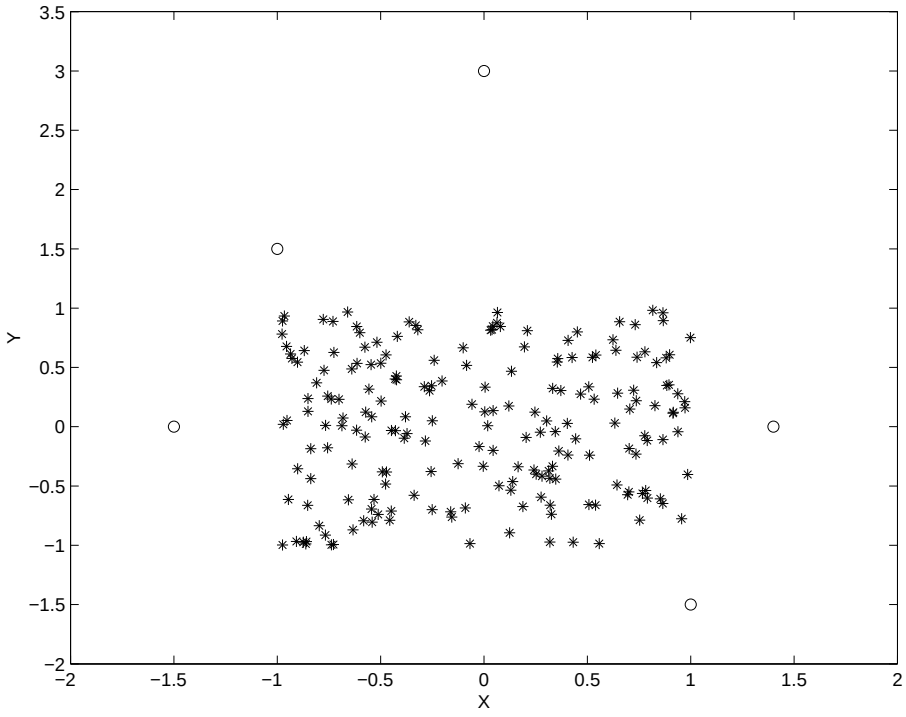

Figura 4.21: Remoção de *outliers* através do algoritmo proposto. Os pontos removidos estão representados por "o".

[8]Para evitar que, à medida que o conjunto de dados aumente, o cálculo de todas as DM_i fique computacionalmente muito custoso, pode-se fazer uma amostragem da distribuição

O valor de n em $\mu + n\sigma$ deve ser determinado experimentalmente, para cada distribuição específica. Através de testes com dados simulados em que sabíamos de antemão haver ou não *outliers*, definimos $n = 3$ como sendo um valor padrão. O critério usado foi encontrar um patamar que não permitisse a retirada de dados corretos, ainda que permitisse a permanência de alguns *outliers*. Esse valor é a sugestão inicial do SMQD, devendo ser ajustado para cada conjunto de dados analisado.

Dentro do SMQD, o uso da Distância de Mahalanobis é auxiliado pela de Kullback-Leibler, descrita a seguir.

Uso da Distância de Kullback-Leibler na Detecção de *Outliers* Introduzida por S. Kullback e R. A. Leibler em [92], a distância ou divergência de Kullback-Leibler fornece uma medida da diferença entre duas distribuições. Tem muitas aplicações no campo da teoria da informação, onde é utilizada para calcular, por exemplo, a quantidade extra de *bits* que deve ser utilizada para transmitir uma informação quando se está usando uma codificação não ótima.

A ferramenta baseada nesta medida que desenvolvemos para o SMQD procura auxiliar o uso da distância de Mahalanobis: enquanto esta última fornece a probabilidade de um determinado ponto pertencer à distribuição analisada, a distância de Kullback-Leibler mede a divergência entre duas distribuições.

Matematicamente, a distância de Kullback-Leibler (D_{KL}) entre duas distribuições P e Q está descrita como:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \text{ no caso discreto} \tag{4.12}$$

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \text{ no caso contínuo} \tag{4.13}$$

Assim, uma maneira de aproveitar esta medida[9] na validação de dados em registros multivariados é verificar, através dela, se a distância entre a distribuição anterior e a distribuição após a inserção de novos dados é significativa (ou seja,

[9]No sentido estrito, a medida de Kullback-Leibler não pode ser considerada uma métrica, já que $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$, além de não satisfazer a desigualdade triangular . Por isso o mais correto é chamá-la de "divergência", e não de "distância".

maior que um determinado patamar d_{kl}). Em caso afirmativo, é sinal de que entre os novos dados deve haver pontos não pertencentes à distribuição original.

No Capítulo 5, analisaremos em cada caso como regular o patamar d_{kl} e utilizá-lo em conjunto com a DM na detecção de *outliers* em registros multivariados.

4.3.2 Monitoração da Completude

Tratemos agora das técnicas utilizadas pelo SMQD na monitoração da Completude. Esta dimensão é uma das mais importantes dentre os atributos de QD, e está sempre presente quando se trata de avaliar a QD geral de uma base de dados.

4.3.2.1 Completando Dados Faltantes

Na prática, monitorar a dimensão da completude significa detectar os campos em branco (dados faltantes) e, salvo indicação em contrário do supervisor, completá-los com valores estimados por um modelo cada vez mais acurado.

O princípio que norteia a ação do SMQD com relação aos dados faltantes é o de que se deve evitar ao máximo o "desperdício" de dados reais, uma vez que, num cenário em que a informação desempenha um papel fundamental na economia e na sociedade, os dados têm adquirido um valor cada vez maior. O impacto causado pela ausência de dados foi primeiramente analisado em [93], trabalho que desencadeou o surgimento de técnicas de substituição de dados faltantes.

A pior forma de se tratar um dado faltante é fazer o descarte do evento (no caso de um registro multivariado) que o contém, pois isso acarreta sempre em perda de informação, muitas vezes imprescindível para a correta caracterização do conjunto de dados. Com efeito, em muitas análises a solução adotada é o descarte de dados, o que leva a um grande desperdício por desconhecimento. A Tabela 4.1 ilustra um caso em que, devido à existência de apenas 2 dados faltantes (NA, *Not Available*) de um total de 16 (12,5%), o descarte dos eventos levaria à eliminação de 8 (50%) dados.

A seguir, descreveremos os mecanismos para a substituição de dados faltantes em séries temporais e em registros multivariados.

	Var 1	Var 2	Var 3	Var 4
Evento 1	35	NA	2,6	300
Evento 2	11	188	0	541
Evento 3	23	207	1,5	283
Evento 4	NA	99	-0,7	97

Tabela 4.1: Amostra com dados incompletos.

Validação em Séries Temporais As ferramentas aqui utilizadas são as mesmas descritas na Subseção 4.3.1 para a detecção de *outliers*. Aqui, o valor $\tilde{x}(T)$, estimado pelo modelo linear ou neural, é assumido como melhor aproximação para o valor faltante $x(T)$. Denominamos esta técnica de "substituição por predição".

No caso de séries temporais, ocorre muitas vezes que a ausência de um dado se dá por atraso no sistema de coleta dos dados. Dessa forma, um dado $x(T)$ faltante pode estar presente no instante $(T + n)$. Nesse momento, o próprio sistema (ou o supervisor, caso essa funcionalidade não esteja implementada de forma automática) tem que proceder a validação do dado $x(T)$. Caso seja validado, ele substitui o valor estimado $\tilde{x}(T)$; caso contrário, é classificado como dado não-validado e não é incorporado ao modelo.

Validação em Registros Multivariados A monitoração da completude nesse tipo de dados dentro do SMQD pode ser feita através de três métodos, descritos a seguir.

Substituição pela média Dentre os métodos mais comumente utilizados para fazer a estimação de dados faltantes, o mais simples consiste em substituir os campos em branco pela média dos valores existentes para as respectivas variáveis.

Apesar de ser uma primeira aproximação para o valor faltante, este método tem sérios inconvenientes. Primeiramente, não leva em consideração a existência de correlação entre as diferentes variáveis do registro multivariado, fazendo com que os eventos reconstituídos não tenham significado prático e possam até mesmo gerar *outliers*. Além disso, a substituição de valores faltantes pela média leva, ao longo do tempo, à formação de picos nas distribuições das variáveis. A Figura 4.22 ilustra

um caso de substituição pela média para uma distribuição com 3 variáveis. Na primeira linha de gráficos, ilustram-se as distribuições originais, completas. Neste exemplo, os dados faltantes são representados pelo valor zero (usualmente, a maioria dos sistemas utiliza um símbolo como NA – *Not Available* – ou NaN – *Not a Number* – para representar tais dados), daí o pico em zero na segunda linha. A terceira linha mostra as distribuições recuperadas através de substituição pela média. Repare que o valor da média original é recuperado, mas a distribuição adquire forma bem diferente da original, o que se reflete em distorção no valor do novo desvio padrão.

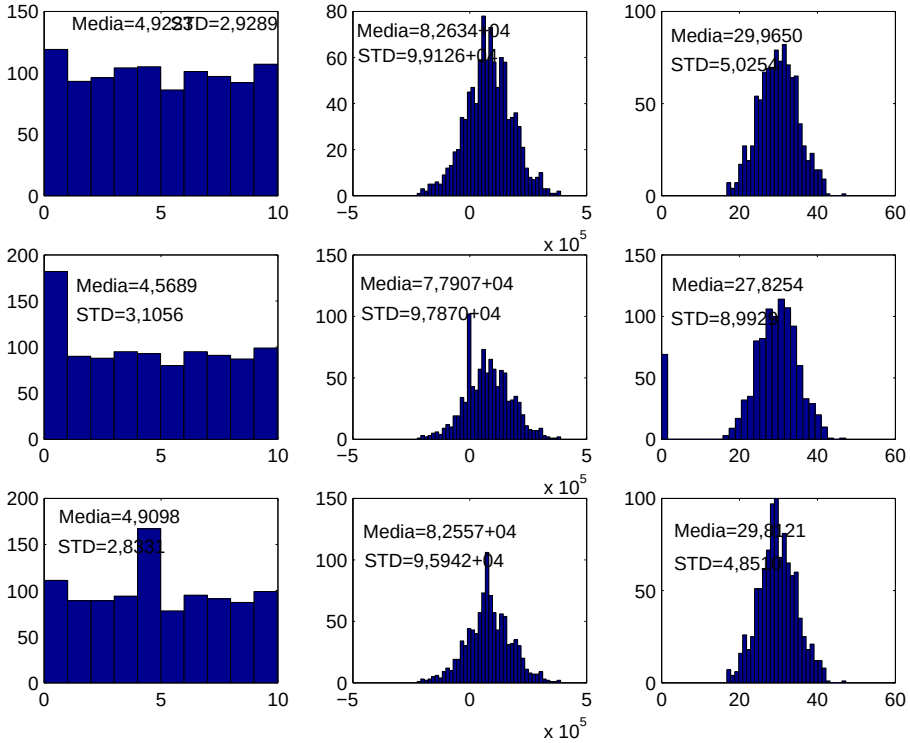

Figura 4.22: Exemplo de substituição de valores faltantes pela média.

Algoritmo EM Regularizado Descrito pela primeira vez em [94], o algoritmo *Expectation-Maximization*, EM, é um estimador de máxima verossimilhança que determina os valores dos parâmetros de um modelo probabilístico sobre distribuições com dados faltantes. Aplicações deste algoritmo encontram-se em diferentes áreas, como por exemplo agrupamento de dados e aprendizado de máquina [95].

Trata-se de um algoritmo iterativo (uma descrição passo-a-passo do algoritmo é encontrada em [96]), em que os passos E e M se sucedem. Nos passos E, estimam-se os valores para os dados faltantes com base nos dados disponíveis e partindo-se

dos valores dados pelo último passo M; neste, busca-se maximizar a verossimilhança do evento conjunto (isto é, dados disponíveis mais os dados faltantes estimados).

O procedimento adotado no SMQD utiliza o algoritmo EM Regularizado, uma variação do algoritmo original [97][98][99].

Imputação Múltipla Descrita por J. L. Schafer em [100], a técnica de imputação múltipla surge como uma das mais eficientes técnicas para a reposição de valores faltantes em dados multivariados com dimensão e tamanho de amostra muito grandes, pois torna desnecessário o cálculo das matrizes de covariância, presente nos métodos baseados em correlação. Encontra aplicações em diversas áreas, desde estudos médicos até pesquisas demográficas.

O algoritmo de imputação múltipla (em sua forma genérica) é dividido em três etapas, a saber:

- Fase de imputação: geram-se m conjuntos completos a partir do conjunto inicial (usando-se modelos paramétricos ou não-paramétricos);
- Fase de análise: analisam-se, através de testes estatísticos, as m distribuições geradas na etapa anterior;
- Fase de combinação: finalmente, os m conjuntos gerados são combinados de modo a formar um único conjunto com dados inferidos.

Com base nestas três etapas, diversas variantes do método podem ser geradas. A implementação de cada etapa (quais testes e modelos usar) varia conforme cada caso. A Figura 4.23 traz os resultados do uso de imputação múltipla a partir de regressão, com os mesmos dados usados no exemplo da Figura 4.22. Nela, podemos ver a recuperação dos valores de média e desvio padrão (o pequeno erro existente deve-se, sobretudo, ao fato de que dados efetivamente nulos (valor igual a zero) foram considerados como dados faltantes e, portanto, tiveram seus valores alterados), mantendo-se a forma da distribuição original.

4.3.3 Monitoração da Pontualidade

Em 4.2.1.2, definimos a pontualidade como $Pt = \max(0; 1 - \frac{TD}{TE})^s$. Na Figura 4.6, vimos que a sua monitoração sucede a da completude e a da acurácia.

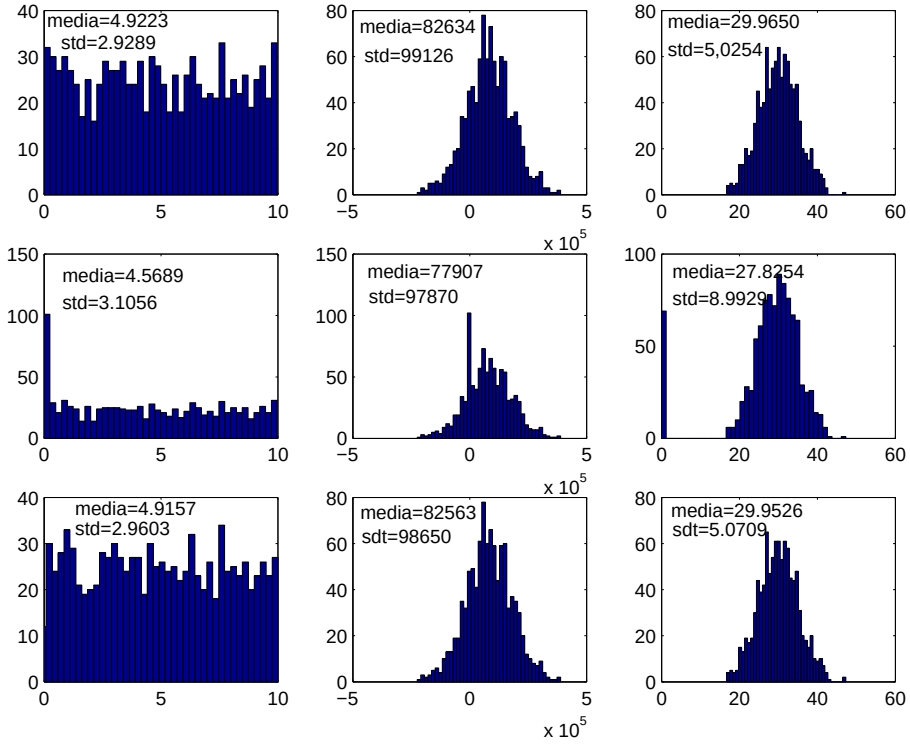

Figura 4.23: Exemplo de substituição de valores faltantes por imputação múltipla.

Para séries temporais, a pontualidade equivale à completude temporal. Para quantificar o seu valor segundo a expressão acima, usamos a seguinte metodologia:

1. Sendo T o instante de tempo atual e $x(T-n)$ o último dado (real, não estimado) que temos disponível, $TD = n$;

2. TE é uma medida estática e deve ser definida pelo supervisor do sistema para cada série que começa a ser monitorada. Deve refletir o intervalo máximo de tempo que se considera possível para o modelo da série continuar rastreando os seus valores reais, sem os estar recebendo. Ou seja, mede o tempo que a série estimada pelo modelo leva para divergir da série real;

3. O valor de s é ajustado para regular a sensibilidade à relação TD/TE. Varia para cada série. Tem como valor padrão $s = 1$.

Segundo a definição acima, TD pode superar o valor de TE; nesse instante, o valor de Pt estabiliza-se em 0.

Para registros multivariados, muitas vezes a dimensão temporal não está explicitamente envolvida com os dados. Dessa forma, o valor de TD é determinado

para cada novo dado recebido como sendo $t_r - t_g$, onde t_r e t_g são, respectivamente, o instante de recebimento do dado e o instante de sua geração.

Em dados desse tipo, convém calcular o valor da pontualidade para cada dimensão da tabela, e tomar a média desses valores como resultado final para Pt. Assim, o valor de TE deve ser estabelecido pelo supervisor para cada coluna (que representa uma variável). Um valor padrão para TE equivale a 10 vezes o tempo médio de atualização para a respectiva variável (segundo as atualizações mais recentes). O supervisor pode então arbitrar valores em torno do valor padrão. Para s, vale o mesmo procedimento usado no caso das séries temporais.

Capítulo 5

Testes com o SMQD

Este capítulo é dedicado à implementação do SMQD sobre dados reais. Trata-se da aplicação do sistema descrito no Capítulo 4 a alguns dos conjuntos de dados apresentados no Capítulo 3, entre outros, com o objetivo de quantificar a eficácia deste sistema e de exemplificar a sua atuação em situações concretas.

Faremos a análise do desempenho do sistema à guisa de estudos de casos, de acordo com o grupo de dados monitorados, visando a cobrir os diferentes campos de atuação do SMQD. Assim, também este capítulo estará dividido em duas grandes seções: na Seção 5.1, contemplaremos a ação do SMQD sobre diferentes tipos de séries temporais; já na Seção 5.2, analisaremos a sua atuação sobre conjuntos de dados multivariados.

Salvo dito o contrário, os algoritmos para a metodologia de ação que compõem o SMQD foram programados utilizando o *software* MATLAB®.

5.1 Monitoração de Séries Temporais

Conforme visto no capítulo precedente, a primeira verificação que se deve fazer, quando da monitoração da QD de séries temporais, diz respeito à presença do novo dado. De fato, como já se sabe a freqüência (diária, semanal, mensal etc) com que uma série deve ser atualizada, sempre se pode verificar, no momento correto, a chegada ou não do novo dado.

Caso o dado esteja presente, a atuação do SMQD é direcionada para a validação deste dado (teste de *outlier*), monitorando-se diretamente a dimensão da

acurácia. Em caso contrário, o sistema é direcionado para a monitoração da completude, fazendo a estimação do melhor valor para substituir o dado faltante.

Em ambos os casos, o SMQD dispõe de um modelo linear (ARIMA) e de alguns modelos neurais para fazer a estimação do novo dado (x_{t+1}), gerando assim o corredor de validação (no caso de monitoração da acurácia) ou um valor substituto $\tilde{x}_{t+1}$ (na monitoração da completude).

Nas subseções seguintes, serão feitos estudos de casos para alguns grupos de séries temporais, de acordo com o fenômeno associado à sua origem.

5.1.1 Séries Temporais Financeiras

Um grupo de séries temporais que tem sido alvo de intenso estudo por parte de economistas, estatísticos, matemáticos, engenheiros e outros profissionais são as séries financeiras. Este grupo de séries é composto principalmente pelas séries de cotações em mercados de ações, mas engloba também séries de taxas de câmbio e outras cotações.

Sem dúvida, trata-se do grupo que apresenta maior dificuldade de modelagem. Com efeito, a previsão em séries financeiras é bastante difícil porque os sistemas que geram os valores das séries são extraordinariamente complexos e a informação disponível para realizar as previsões é sempre muito limitada [101]. Por outro lado, esta dificuldade de previsão vai ao encontro de algumas teorias, como a EMH (*Efficient Market Hypothesis*, Hipótese dos Mercados Eficientes[1]), que defendem a impossibilidade de realizar previsões úteis para este tipo de séries temporais [102].

5.1.1.1 Fenômeno do "Atraso" na Predição de Séries Financeiras

Antes de empreendermos os estudos de casos para séries financeiras específicas, convém que nos detenhamos um pouco a analisar um fenômeno que acontece sistematicamente quando das tentativas de previsão desse tipo de série.

Consideremos o caso em que pretendemos utilizar as últimas n observações

[1]Segundo a EMH, o preço dos ativos negociados em mercados financeiros reflete toda a informação conhecida pelos membros do mercado e todas as expectativas dos investidores com relação ao futuro, de maneira que seria impossível conseguir, de forma consistente, prever os resultados do mercado, a não ser através de sorte (acaso) ou informação privilegiada.

para a predição da observação seguinte, isto é, buscamos fazer o seguinte mapeamento: $x_{t+1} = f(x_t, x_{t-1}, ..., x_{t-N+1})$. A função $f(\cdot)$, caso exista, tem uma estrutura não conhecida, e pode ser aproximada, por exemplo, por um modelo linear (do tipo ARIMA) ou por redes neurais multi-camadas (do tipo FFBP), que são conhecidas como aproximadores universais de funções [35]. Ou seja, é lícito esperar que, caso $f(\cdot)$ exista, uma rede neuronal convenientemente projetada e treinada será capaz de aproximá-la com erro arbitrariamente pequeno.

Após realizar o pré-processamento conveniente (que será ilustrado mais adiante, quando da análise da atuação do SMQD através dos estudos de casos), treinamos algumas arquiteturas neurais segundo o esquema da Figura 5.1 (minimizando o erro médio quadrático, EMQ) e também um modelo linear ARIMA na tarefa de predição de algumas séries temporais financeiras.

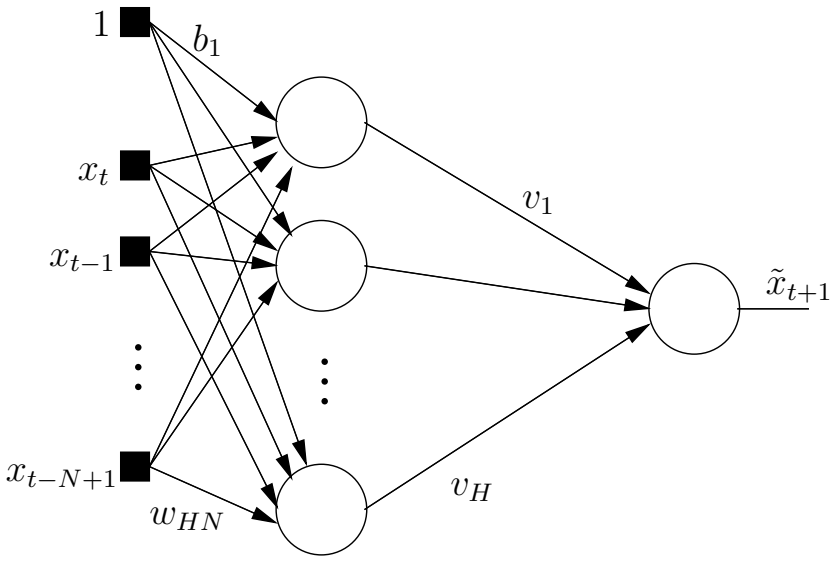

Figura 5.1: Arquitetura para a rede neural treinada com o algoritmo FFBP.

Em todos os casos, com algumas pequenas variações, a série predita (conjunto de teste) aproxima-se muito de uma versão atrasada da série alvo ($\tilde{x}_{t+1} \simeq x_t$), conforme ilustrado na Figura 5.2 para o caso da série com as cotações da SUN (valor do fechamento diário da ação). Para melhor comparação, vemos na figura o confronto do alvo com a série predita e também com a série predita deslocada, em que fica bastante claro que a saída da rede neural equivale a uma versão atrasada da série alvo, somada a um erro. Este "erro" (diferença entre a série alvo e a série deslocada) está ilustrado na Figura 5.3 em termos de sua distribuição (histograma)

e da sua função de autocorrelação.

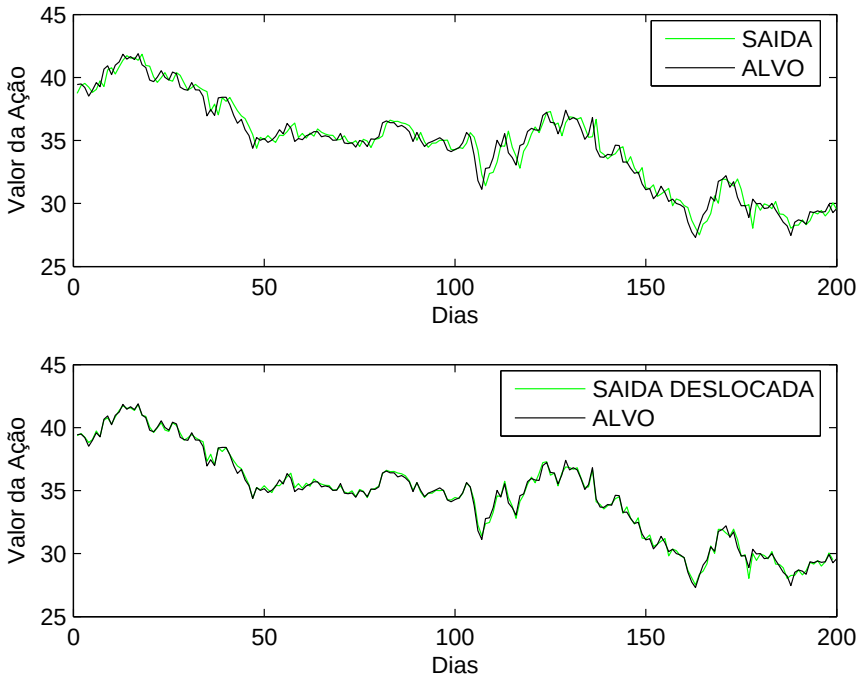

Figura 5.2: Série da SUN predita pela rede neural vs. alvo (acima) e com deslocamento da saída da rede (abaixo).

A Figura 5.4 ilustra a curva do erro médio quadrático dos conjunto de treino e validação, durante o processo de treinamento. Podemos observar que o treinamento converge em um pequeno número de épocas, ou seja, a rede está efetivamente minimizando o EMQ entre a saída e o alvo. Isto, porém, não impede o aparecimento do atraso na série predita. Este mesmo fenômeno acontece para outras séries financeiras, conforme ilustra a Figura 5.5. Para efeito de comparação, todas foram modeladas com a mesma rede neural: arquitetura FFBP, N=20 na entrada, 8 neurônios na camada escondida (tangente hiperbólica como função de ativação) e 1 neurônio (linear) na camada de saída.

Por que acontece este fenômeno? Podemos começar a entendê-lo analisando o comportamento da mesma rede neural usada para gerar o resultado da Figura 5.2 quando se coloca, em sua entrada, uma amostra de ruído branco gaussiano. O resultado está ilustrado na Figura 5.6: a saída da rede não consegue acompanhar a excursão do alvo, oscilando em torno de zero com baixa amplitude. Este comportamento é natural e esperado uma vez que, sendo a entrada composta por um

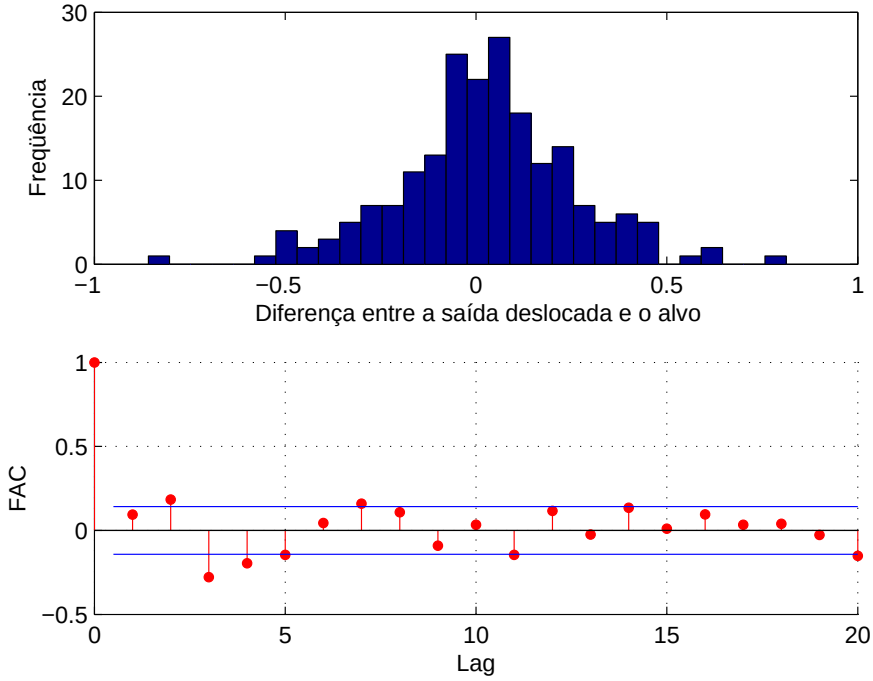

Figura 5.3: Histograma da diferença entre a série alvo e a série deslocada (acima) e FAC da diferença (abaixo).

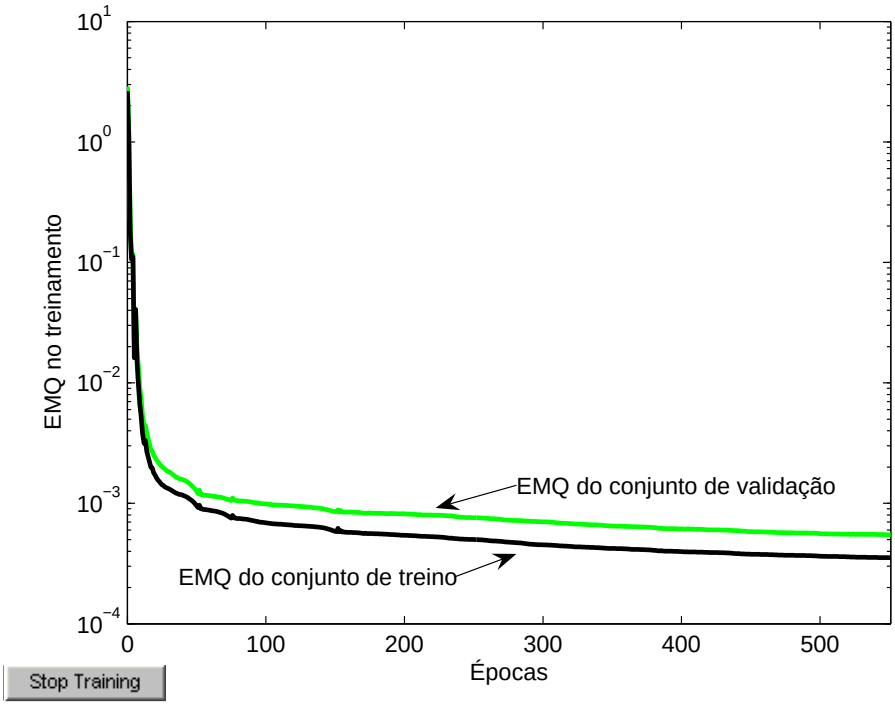

Figura 5.4: Erro médio quadrático durante o treinamento, para os conjunto de treino e validação.

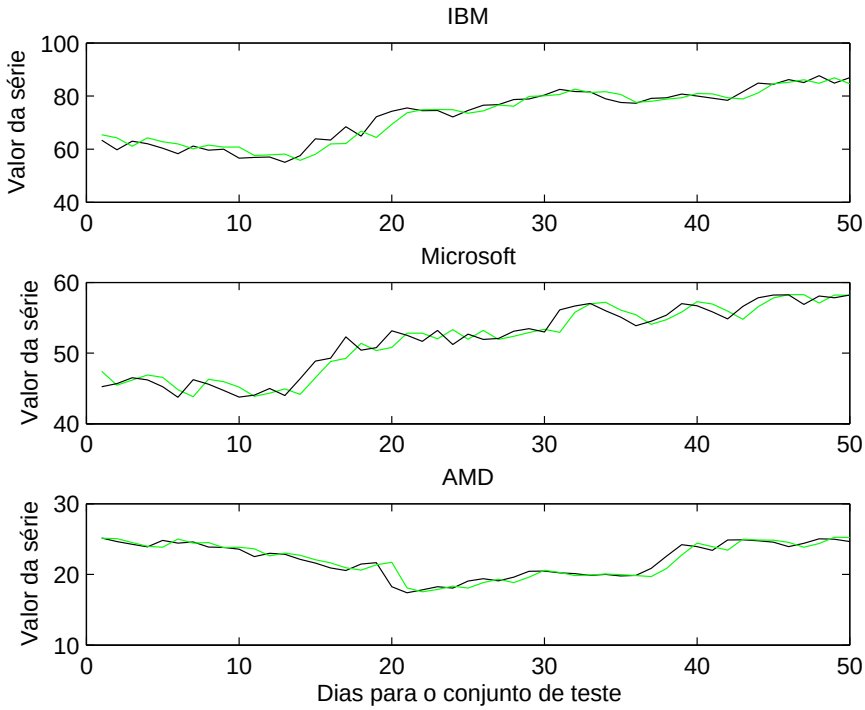

Figura 5.5: Detalhes do "fenômeno do atraso" para algumas séries financeiras (a saída da rede está representada pela linha mais clara).

ruído de média zero e variância σ^2 (no caso deste exemplo, $N(0,1)$), e sendo a rede neural treinada para minimizar o EMQ, a melhor resposta que ela pode fornecer é um valor próximo a zero. Além disso, a função de autocorrelação de uma amostra de ruído branco aproxima-se da função impulso[2] e seu espectro em freqüência é homogêneo[3] (conforme ilustra a Figura 5.7), ou seja, a série a ser modelada pela rede é altamente descorrelacionada, o que faz com que um mapeamento do tipo $x_{t+1} = f(x_t, x_{t-1}, ..., x_{t-N+1})$ seja praticamente impossível de ser alcançado. De fato, sabemos que é absolutamente impossível fazer modelos de predição para um ruído branco perfeito (FAC igual à função impulso e espectro de freqüência constante), por definição, já que seria prever a aleatoriedade.

Façamos agora a análise espectral e de correlação para uma série do tipo passeio aleatório, conforme o exposto no Capítulo 3. Em concreto, observemos a série $y_t = y_{t-1} + r_t$ (onde $y_0 = 0,2$ e $r_t = N(0,1)$). A Figura 5.8 ilustra esta série juntamente com a sua FAC (cujo decaimento bastante lento evidencia a não-

[2]Seria exatamente a função impulso se a série fosse infinita.

[3]Seria uma constante caso o ruído fosse perfeitamente branco e a série infinita.

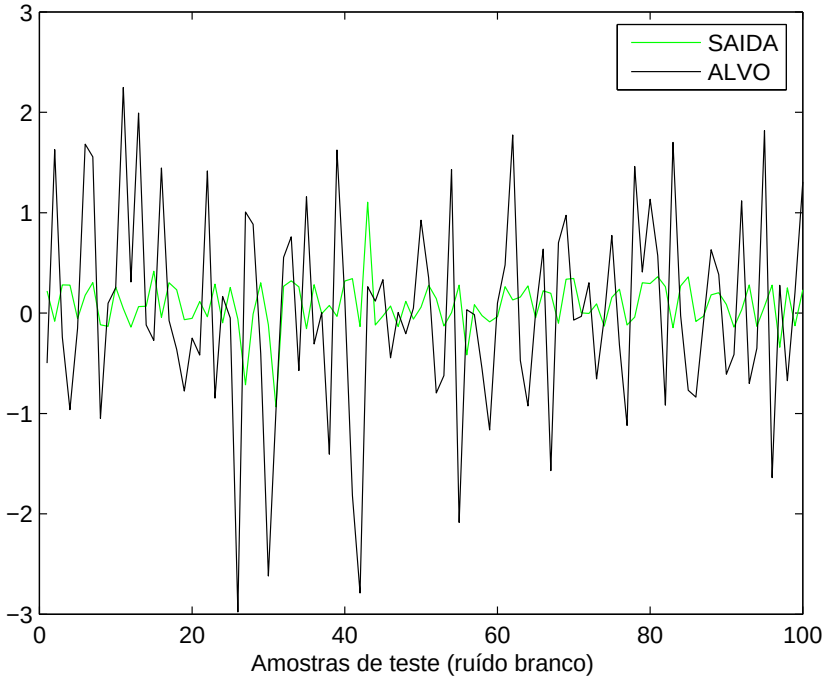

Figura 5.6: Tentativa de predição de ruído branco através de rede neural minimizando o EMQ.

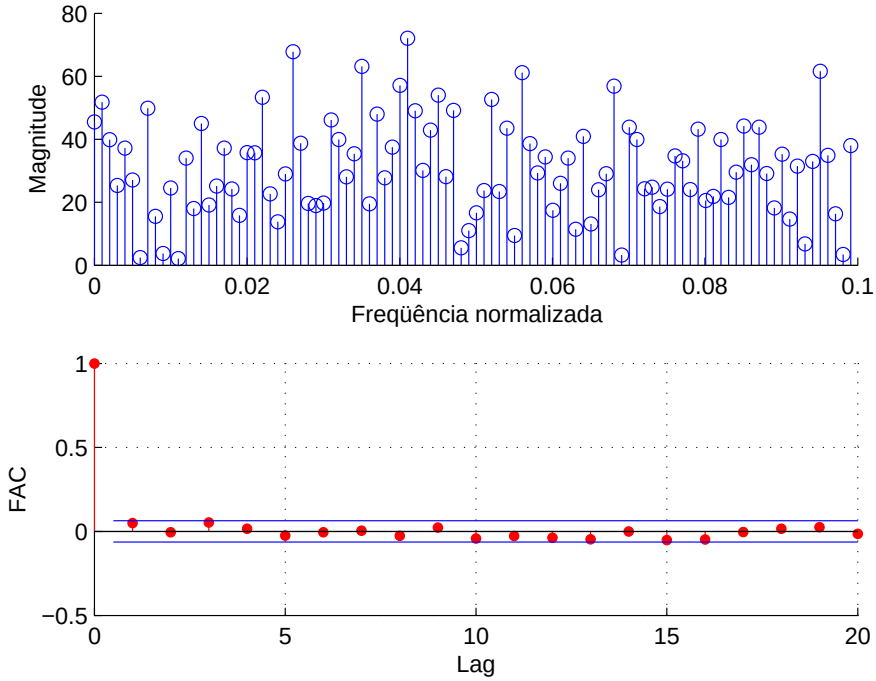

Figura 5.7: Espectro de freqüência (acima) e função de autocorrelação para a amostra de ruído branco colocado na entrada da rede neural (abaixo).

estacionariedade da série).

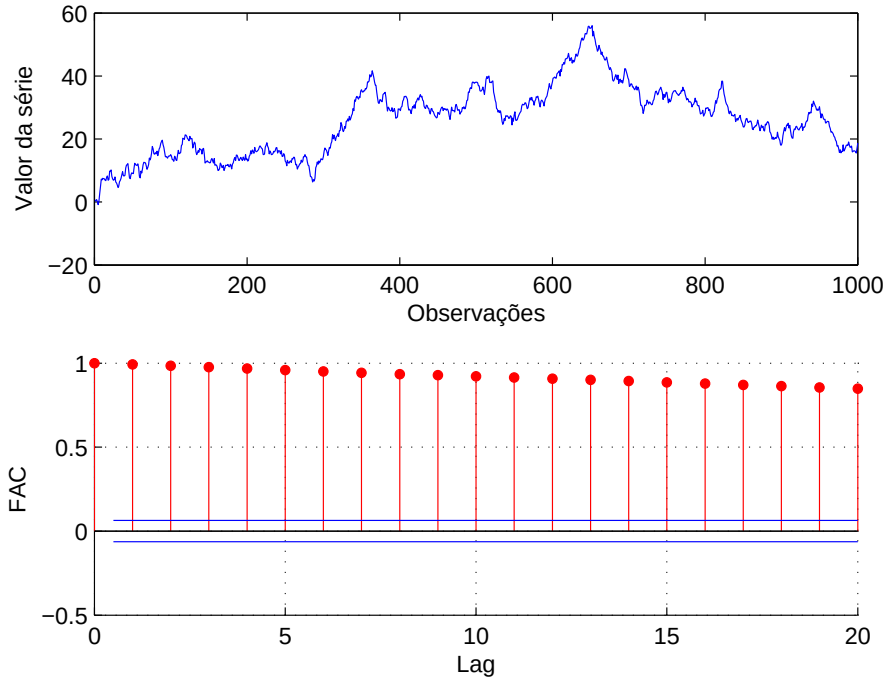

Figura 5.8: Série do passeio aleatório (acima) e sua FAC (abaixo).

A série $\nabla(y_t)$, porém, já é estacionária (equivale a ruído branco), e está ilustrada na Figura 5.9, juntamente com a sua função de autocorrelação. Com a mesma rede neural utilizada para modelar a série da SUN (cujo resultado está na Figura 5.2), fizemos a modelagem da série do passeio aleatório. Utilizamos 900 observações para formar os conjuntos de treino e validação e 100 para testar a rede, obtendo o resultado ilustrado na Figura 5.10. Também nesse caso pode-se observar o fenômeno do atraso.

Quando comparamos as funções de autocorrelação da amostra de ruído branco (Figura 5.7) e do passeio aleatório estacionarizado (Figura 5.9), vemos que são essencialmente iguais. Neste momento, perguntamo-nos se o fenômeno do atraso causado na tentativa de predição em séries financeiras pode ser devido ao fato de que estas séries são muito próximas a passeios aleatórios, podendo mesmo ser modeladas como tais. Para verificar essa hipótese, façamos a análise FAC e do espectro em freqüência de séries financeiras.

As Figuras 5.11 e 5.12 ilustram, respectivamente, as funções de autocorrelação e os espectros de freqüência das séries de fechamento no mercado de ações para a

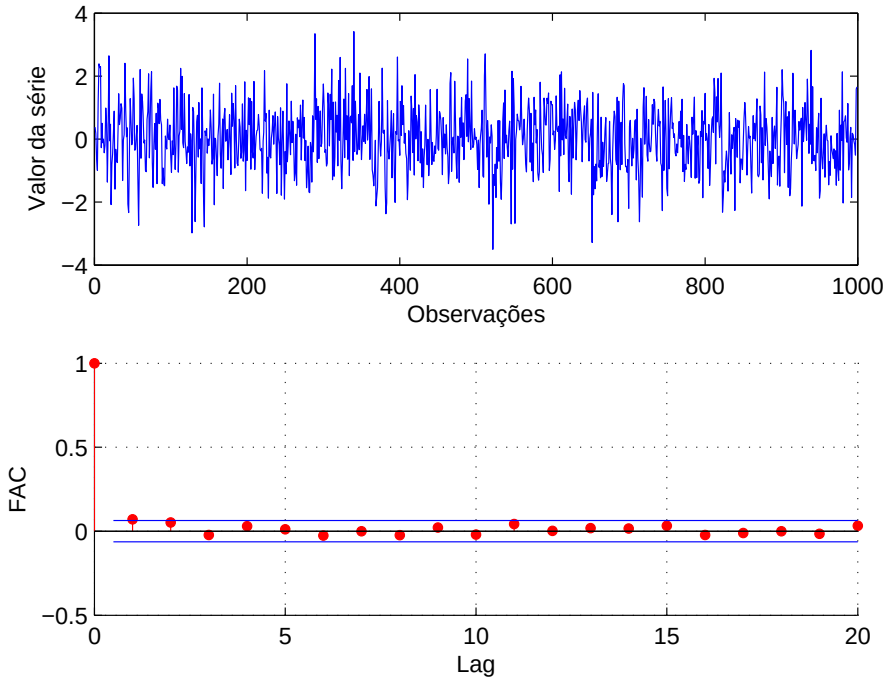

Figura 5.9: Série do passeio aleatório estacionarizado (acima) e sua FAC (abaixo).

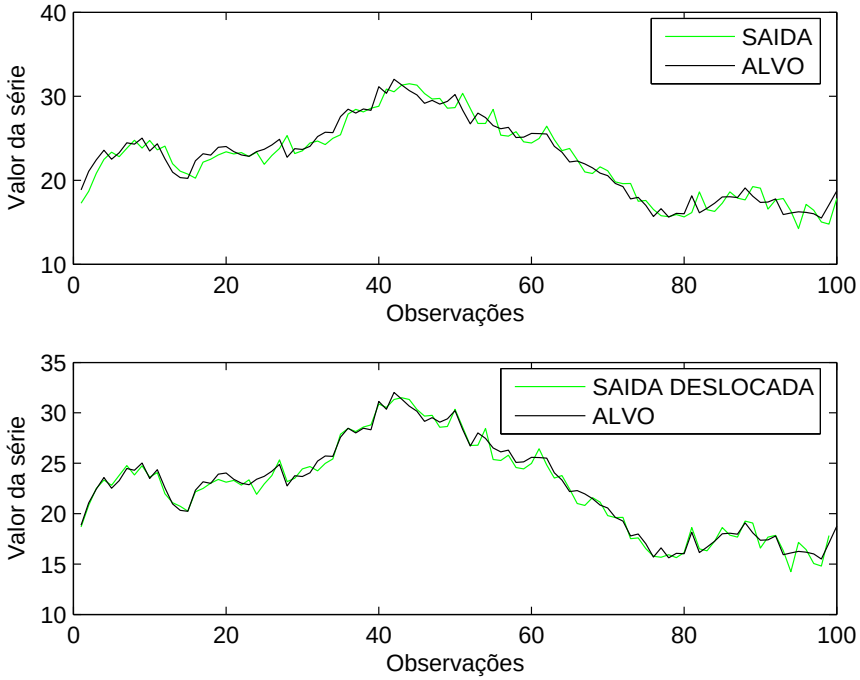

Figura 5.10: Série do passeio aleatório predito pela rede neural vs. alvo (acima) e com deslocamento da saída da rede (abaixo).

SUN, a Microsoft e a IBM (estacionarizadas pela operador ∇). De fato, as funções de autocorrelação e os espectros de freqüência para essas séries financeiras permitem-nos concluir que são séries altamente descorrelacionadas. O fato de que, para as três séries, a correlação de lag=1 seja sempre negativa e um pouco além do intervalo de confiança de 95% reflete apenas a tendência que tais séries têm de inverter o sentido de variação, ou seja, caso a última observação tenha representado uma subida, a tendência é que a seguinte represente uma descida no valor da série, fazendo com que a série estacionarizada (1ª diferença) tenda a ter sinais trocados entre dois valores consecutivos, gerando esta componente negativa na FAC.

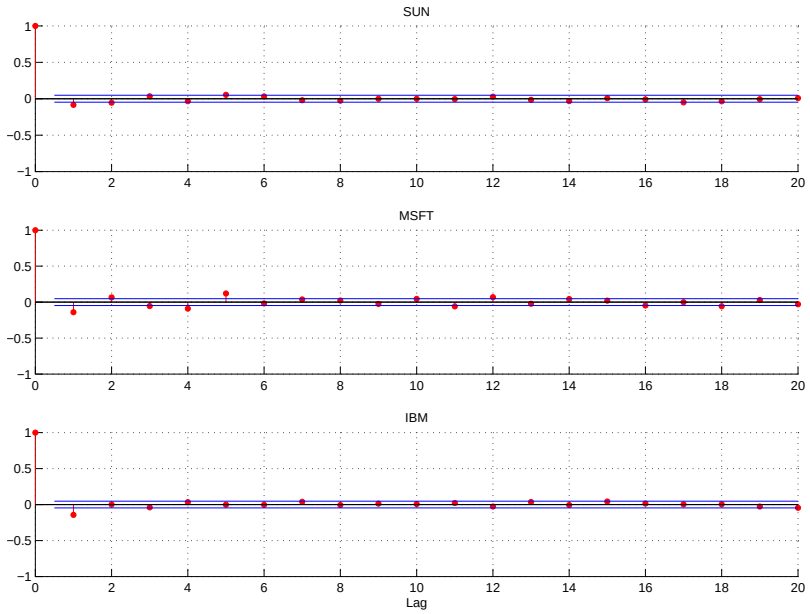

Figura 5.11: Função de autocorrelação para 3 séries de mercado financeiro (séries estacionarizadas pela operador ∇).

A Figura 5.13 ilustra de forma conjunta as funções de autocorrelação para a série estacionarizada da SUN e de uma série de ruído branco com o mesmo tamanho. Pode-se observar que não há diferença significativa.

Isso nos leva a concluir que séries temporais de origem financeira assemelham-se fortemente a passeios aleatórios, o que torna a tarefa de predição extremamente intrincada. De fato, o trabalho publicado em [82] mostrou que redes FFBP ou RNN (Elman) não são capazes de modelar séries financeiras sem incorrer no problema do atraso, indicando que tais séries não têm correlação suficiente entre as amos-

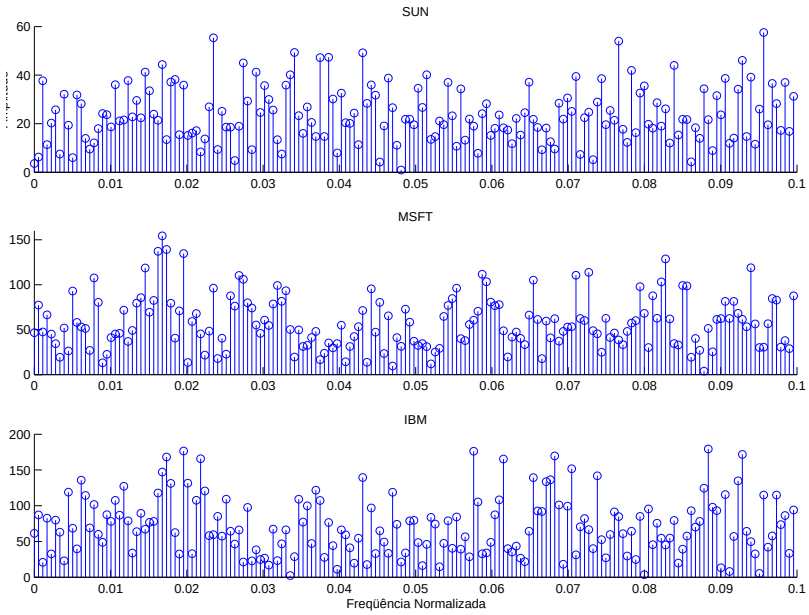

Figura 5.12: Espectro em freqüência para as 3 séries analisadas.

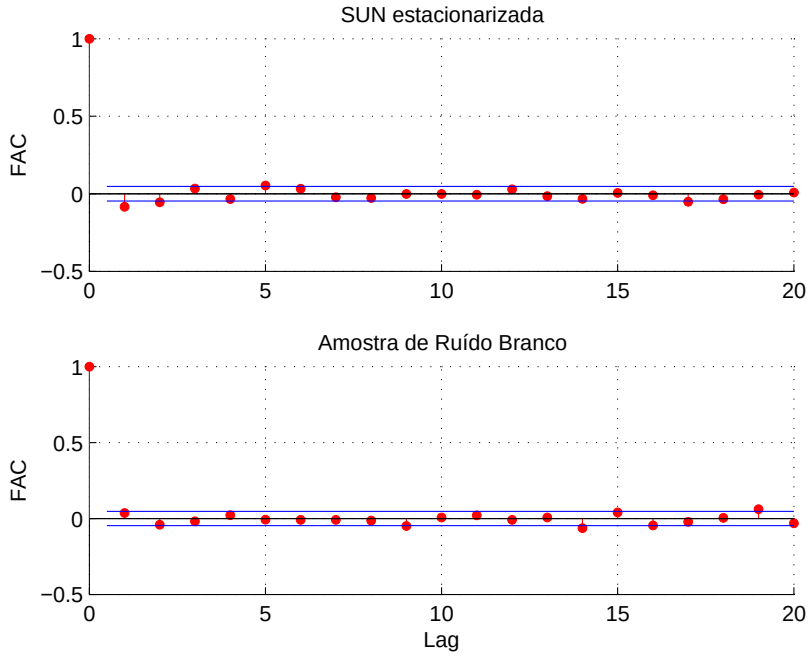

Figura 5.13: Comparação da FAC da série estacionarizada da SUN com a de uma amostra de ruído branco do mesmo tamanho.

tras. Para ilustrar de outro modo a proximidade das séries financeiras com passeios aleatórios, a Figura 5.14 mostra 4 séries, sendo 3 delas séries financeiras reais e a outra, um passeio aleatório. Visualmente, é impossível distinguir qual delas é o passeio aleatório simulado[4].

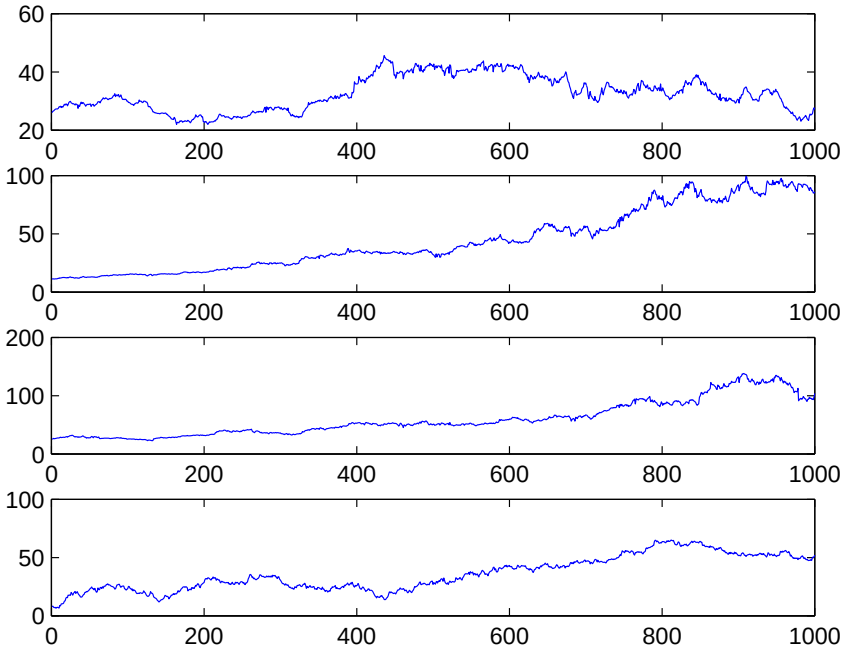

Figura 5.14: Qual destas 4 séries é o passeio aleatório?

Resumindo: quando se coloca uma série com tais características na entrada de uma rede neural, esta logo aprende que o valor de x_{t+1} pode ser modelado como $x_t + \Delta$, sendo Δ uma variável aleatória com média igual a zero. Assim sendo, como o treinamento da rede a leva a minimizar o erro médio quadrático (EMQ), ela acaba por fazer Δ muito próximo a zero, fornecendo um valor muito próximo a x_t como melhor estimativa para x_{t+1}. De fato, é conhecido pelas ciências econômicas o resultado de que "em geral, a melhor previsão para o valor do dólar amanhã é o valor do dólar hoje" (pois é a previsão que retorna o menor erro médio quadrático a longo prazo).

Este "fenômeno do atraso", que acontece em grau elevado nas séries financeiras, se dá também, em menor grau, em outras séries, quando o nível de correlação

[4]O passeio é a quarta série, sendo as 3 primeiras as séries da SUN, MSFT e IBM. Chama a atenção de modo especial a semelhança do passeio aleatório com a série da IBM.

entre as amostras, embora baixo, seja ainda suficiente para uma modelagem razoável da série. Nesses casos, obtém-se alguma vantagem recorrendo às redes recorrentes (Elman), conforme veremos adiante.

Para as séries financeiras, o uso das redes de Elman ameniza o problema, mas não o resolve. Para quantificar essa melhora, calculamos a função de autocorrelação entre a saída da rede para o conjunto de teste e o respectivo alvo. Quando há o fenômeno do atraso, a correlação mais significativa tende a estar no lag=1. Sem o atraso, a maior correlação deveria dar-se no lag=0. Uma maneira que desenvolvemos para medir o "nível de atraso" da série é calcular a razão R entre a correlação de lag=0 e a correlação de lag=1. Podemos dizer que há uma redução no fenômeno do atraso quando R (que chamamos de "fator de sincronismo") aumenta com a mudança da arquitetura da rede neural utilizada. No caso de séries financeiras, como as amostras consecutivas são muito parecidas (as variações bruscas são muito raras), as correlações de lag=0 e lag=1 são muito próximas, de maneira que, mesmo havendo o atraso, R costuma ser próximo a 1. Em todo caso, ilustramos nas Figuras 5.15 e 5.16 o detalhe da FAC em torno ao lag=0 para a predição da série da Microsoft utilizando redes FFBP e Elman, respectivamente. Comprova-se que a variação de R (0,96 para 0,99) com a utilização da rede de Elman não é muito significativa.

Alguns trabalhos publicados nos últimos anos têm proposto novas metodologias para contornar o problema do atraso (também chamado de "eco" na predição de séries temporais). Em [103], sugere-se inserir uma fator de penalidade para o atraso dentro do algoritmo de treinamento da rede neural. Os resultados encontrados, porém, mostram que, além do treinamento ficar extremamente lento e custoso computacionalmente, o erro de predição aumenta, ainda que o fenômeno do atraso diminua. Também neste trabalho, sugere-se a utilização de algoritmos genéticos na seleção e evolução de um grupo de redes neurais geradas aleatoriamente, até alcançar uma topologia ótima. Resultados satisfatórios, porém, ainda não foram encontrados.

Em [104], é sugerido o uso de redes neurais não-supervisionadas juntamente com redes treinadas com algoritmos evolucionários. A implementação dessas arquiteturas, porém, foge ao escopo desta tese.

Finalmente, pode-se também focar no pré-processamento das séries, e não na

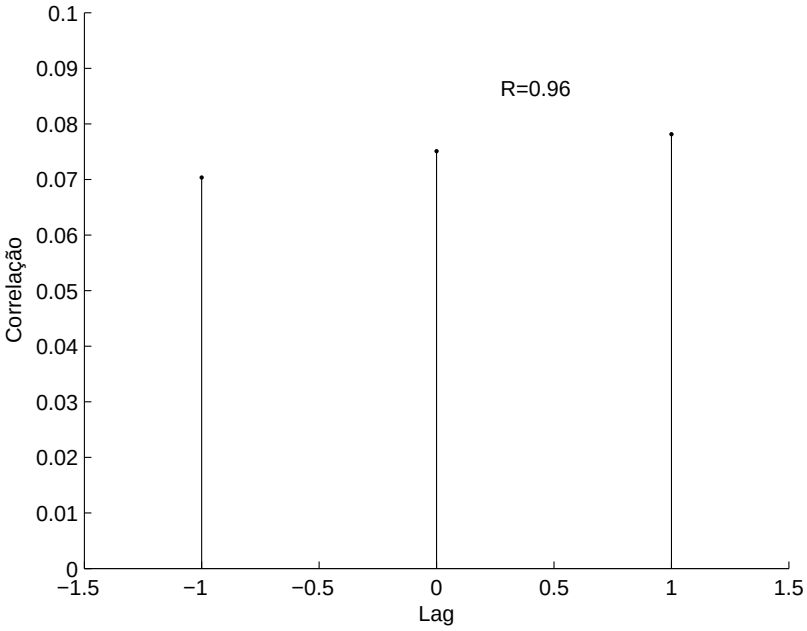

Figura 5.15: Correlação entre a série real e a série predita (Microsoft), usando-se uma rede FFBP.

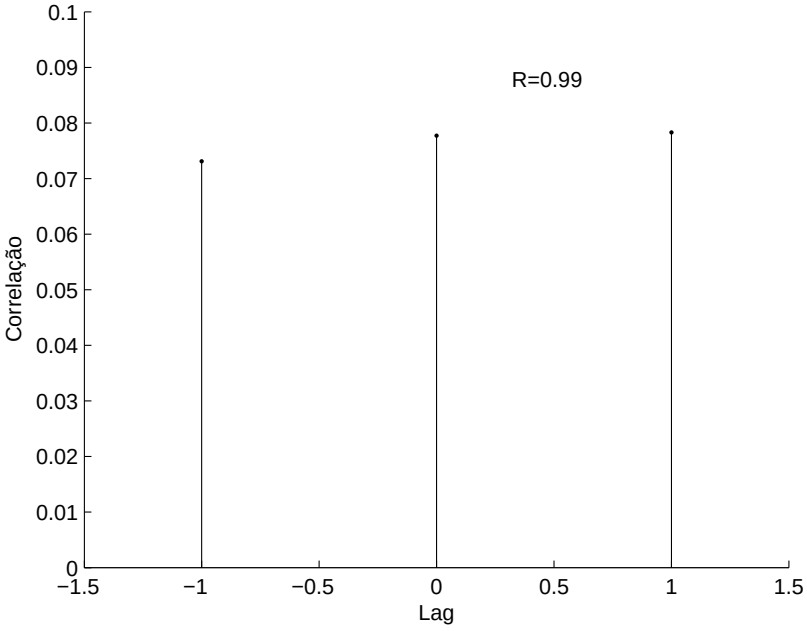

Figura 5.16: Correlação entre a série real e a série predita (Microsoft), usando-se uma rede Elman.

mudança da arquitetura da rede utilizada. Em [105], por exemplo, sugere-se aplicar a filtragem de Kalman ou FNF (*Feedback Neural Filters*) à série estacionarizada como última etapa do pré-processamento. Testamos a filtragem de Kalman [45], sem obter melhorias na modelagem de séries financeiras. Por fim, testamos também o uso de filtros de média-móvel, o que não ocasionou diminuição no fenômeno do atraso.

Com relação ao modelo linear (ARIMA), o mesmo fenômeno ocorre. Utilizamos um modelo ARIMA(p=1,d=1,q=1), já que vimos que a série estacionarizada ($d = 1$) não possui correlação significativa para *lag* maior que 1 (por isso $p = 1$ na parte auto-regressiva do modelo). Observamos que o modelo era praticamente insensível à variação do valor de q. A Figura 5.17 ilustra o modelo de predição para a série da IBM, exibindo claramente o fenômeno do atraso.

De fato, o estado da arte da pesquisa em séries financeiras não contém métodos eficazes para a predição dos valores futuros. Em contrapartida, dispõe-se de uma metodologia já consolidada para a modelagem da variância condicional da série. Trata-se do modelo GARCH (*General Auto-Regressive Conditional Heteroscedasticity*), que possui diversas aplicações em modelagem de mercados financeiros, sobretudo na previsão da volatilidade (e conseqüentemente o valor de risco) de séries de mercados de ações [106][107].

Diante desse panorama, como deve então agir o SMQD na monitoração de séries temporais financeiras? Devemos relembrar aqui que o objetivo final do sistema não é realizar a predição de séries e sim monitorar a sua qualidade, detectando *outliers* e substituindo valores faltantes com um valor estimado próximo do real. Assim sendo, valemo-nos do fato de que as séries financeiras não apresentam variações bruscas. Com isso, o modelo conseguido através de redes recorrentes de Elman é suficiente para gerar um corredor de validação (conforme detalhado no Capítulo 4) que detecte *outliers* verdadeiros.

A Figura 5.18 ilustra o processo de validação da série da AMD a partir da formação do corredor de validação pelo SMQD. Os limites do corredor, para este exemplo, são $x_I(T) = \tilde{x}(T) - 4\hat{\sigma}_e$ e $x_S(T) = \tilde{x}(T) + 4\hat{\sigma}_e$, sendo σ_e o desvio da distribuição do erro de estimação acumulada a cada instante. Observamos que, como a variação entre as amostras consecutivas da série é baixa, a largura escolhida

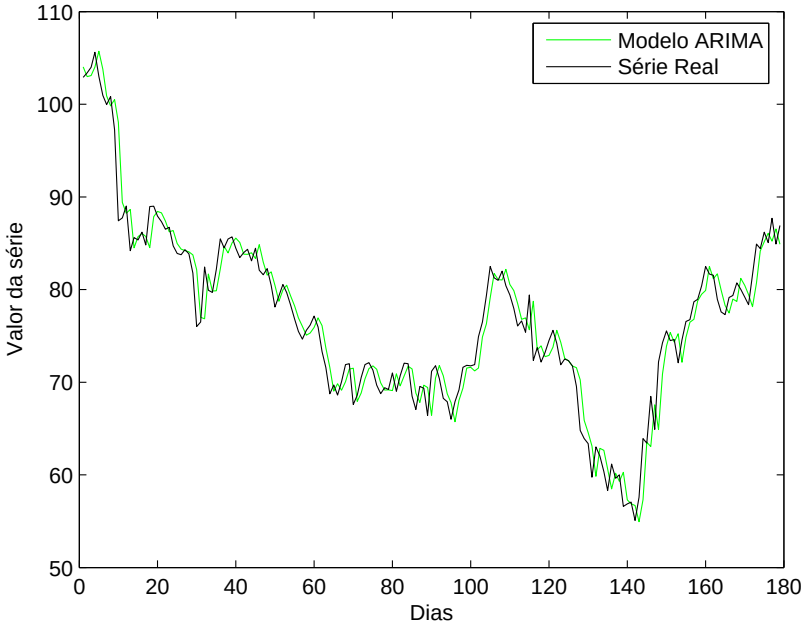

Figura 5.17: Predição da série da IBM através de modelo ARIMA(1,1,1).

para o corredor é grande o suficiente para validar todas as amostras corretas. Neste exemplo, simulamos a existência de dois *outliers* que, detectados pelo corredor, são substituídos pelo valor da série predita. Na Subseção 5.1.3, veremos em detalhes o processo de formação do corredor na detecção de *outliers*.

Quando uma nova observação x_t da série é validada (a partir do confronto com o valor estimado $\tilde{x}_t$), ela é incorporada ao modelo. Caso seja rejeitada, conforme os dois *outliers* da série recebida, ilustrada na Figura 5.18, o valor incorporado ao modelo é o de $\tilde{x}_t$. O SMQD permite que, uma vez detectado um *outlier* que tenha sido gerado por um erro, o seu valor correto seja posteriormente incorporado ao modelo, caso haja a possibilidade de adquiri-lo.

5.1.2 Séries Temporais Caóticas

Na subseção anterior, vimos as dificuldades encontradas pelos modelos neurais e lineares na tarefa de predizer, de forma acurada, séries temporais financeiras. Conforme visto no Capítulo 4, foi desenvolvida nos últimos anos uma nova arquitetura para redes neurais recorrentes, as chamadas Redes de Estado de Eco (*Echo State Networks* ou ESN). Esse tipo de rede tem particular habilidade em modelar fenômenos caóticos, pois consegue mapear, nos estados de eco armazenados no seu

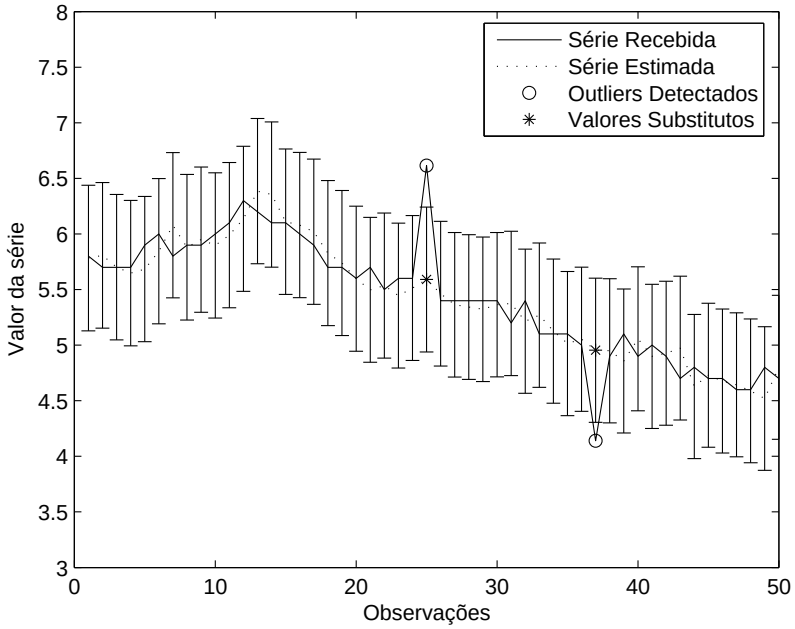

Figura 5.18: Processo de validação da série da AMD.

reservatório dinâmico, a relação entre as últimas observações da série modelada e a observação seguinte.

São raros os fenômenos conhecidamente caóticos que podem ser de interesse para a monitoração do SMQD. Entretanto, decidimos investigar a possibilidade de as séries financeiras serem modeladas como séries caóticas e, portanto, preditas com maior acurácia por redes ESN.

Fenômenos caóticos são caracterizados por possuírem um comportamento com grande sensibilidade a variações nos valores dos parâmetros que os regem (nos pontos de bifurcação) e também às coordenadas do ponto inicial onde se começa a observar o fenômeno. Porém, não se deve confundir o fenômeno aleatório com o fenômeno caótico: enquanto o primeiro não permite conhecer o seu estado seguinte, mesmo sabendo-se o estado atual, para um fenômeno caótico é sempre possível saber o estado seguinte quando se conhece o estado atual, já que se trata de um fenômeno determinístico.

Por essas considerações, já se antevê que as séries financeiras não são manifestações de fenômenos caóticos, mas sim de processos mais assemelhados a fenômenos aleatórios, no sentido de que, ainda que se possam mapear algumas relações do tipo causa-efeito no comportamento de tais séries, é extremamente complexo prever os

seus valores seguintes, mesmo conhecendo os valores passados[5].

A modelagem neural de séries caóticas já havia sido discutida no início da década de 90 em [110]. Porém, o surgimento das ESNs trouxe um ganho de qualidade a essa tarefa, através de maior acurácia na predição.

A Figura 5.19 ilustra a predição de uma série caótica bastante conhecida, a série de Lorenz [111], por meio de uma ESN (com 300 neurônios no reservatório dinâmico e grau de interconexão 0,5). Pode-se perceber que o erro de predição é praticamente zero, porque a ESN consegue guardar, no seu reservatório dinâmico, estados de eco que mapeiam a observação seguinte, dadas as observações anteriores. Ao mesmo tempo, podemos observar, na Figura 5.20, que a série (mesmo estacionarizada) é bastante correlacionada e que o seu espectro em freqüência cai a zero bastante rapidamente (em torno da freqüência normalizada 0,03). Isto mostra o quão diferente é esta série de uma série financeira.

Com efeito, a aplicação de ESN na predição de séries financeiras também resulta no fenômeno do atraso, como ilustra a Figura 5.21.

5.1.3 Outras Séries Temporais

Tendo já visto os casos mais complicados, em que as séries temporais monitoradas possuem complexidades que dificultam sobremaneira a sua modelagem, dedicamo-nos agora à análise da atuação do SMQD sobre outros tipos de séries temporais.

A primeira série, cuja monitoração pelo SMQD é aqui analisada, é a série de consumo mensal de energia elétrica nos EUA (a base para o estudo consiste nos anos de 1973 a 2006, conforme visto no Capítulo 3) [49]. A série ilustrada na Figura 5.22 é claramente não-estacionária, possuindo tendência, heteroscedasticidade e sazonalidade. A figura traz também, na parte superior, os valores da variância da

[5]Por exemplo: é sabido que a compra e venda de dólares por parte do Banco Central afeta diretamente a sua cotação. Todavia, são tantas as variáveis envolvidas na determinação do seu valor (sendo muitas delas guiadas por agentes humanos, cuja atividade é muitas vezes imprevisível) que se torna inviável a modelagem dos valores futuros nesse tipo de séries. Uma abordagem físico-estatística do problema, dentro do que se conhece por Econofísica [108][109], tem tentado modelar tais séries através da analogia com a termodinâmica, onde também há uma grande quantidade de agentes com atuação quase imprevisível.

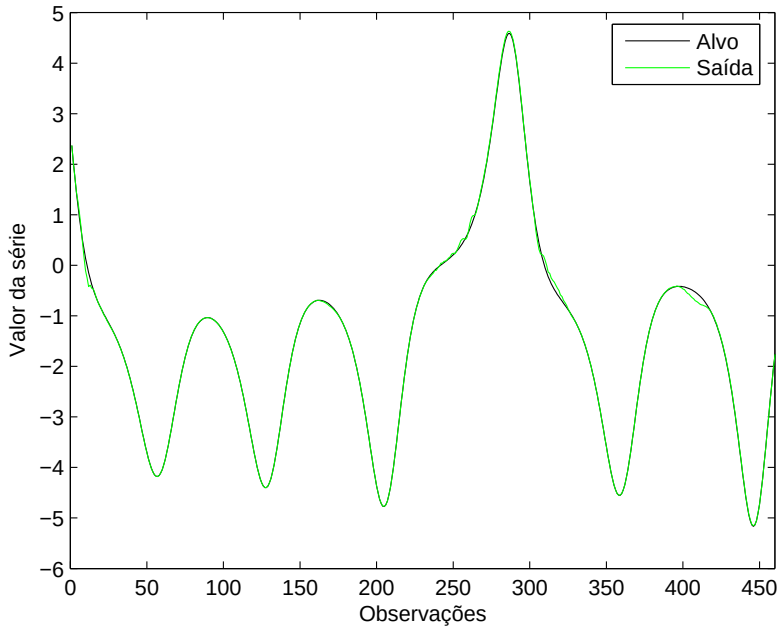

Figura 5.19: Modelagem de uma série de Lorenz através de ESN.

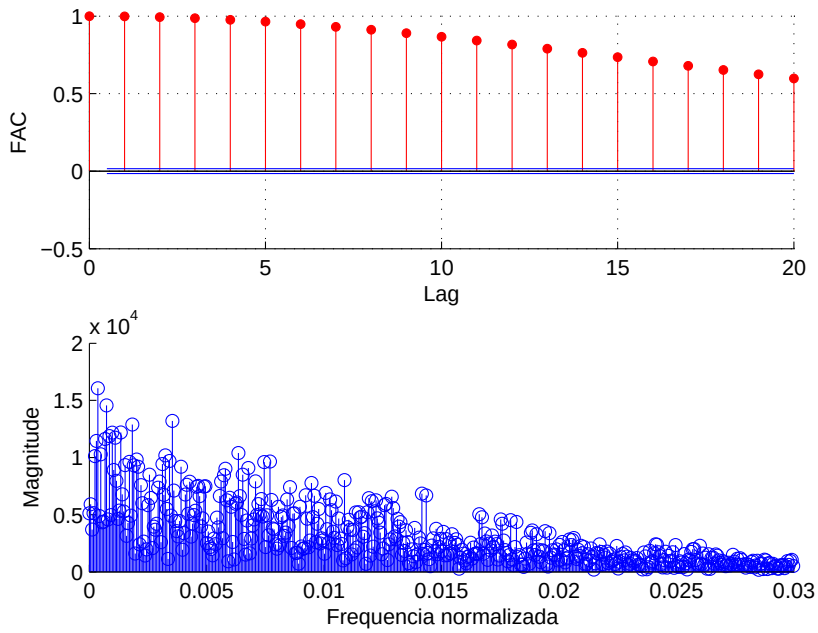

Figura 5.20: Função de autocorrelação para a série de Lorenz (acima) e seu espectro de freqüência (abaixo).

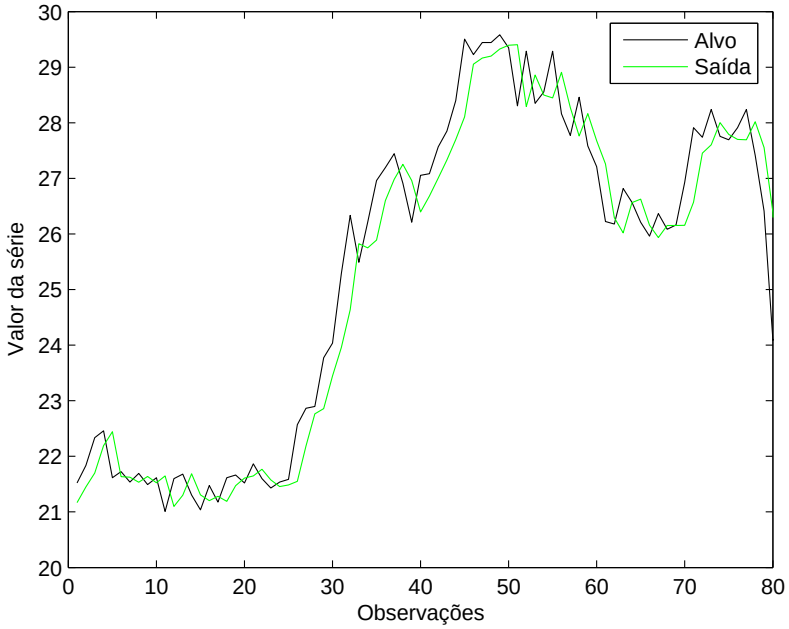

Figura 5.21: Modelagem da série da Microsoft por uma rede ESN.

série a cada grupo de 40 observações[6]. A variância é crescente, o que é característico de séries de consumo de bens necessários, em função do aumento da população e do desenvolvimento da economia.

A homoscedasticidade da série é alcançada com a aplicação da função logarítmica. O resultado é mostrado na Figura 5.23. Conforme discutido no Capítulo 3, a tendência desta série é determinística e linear (ou seja, não possui raiz unitária). Após a sua remoção, resta apenas retirar a sazonalidade da série para torná-la estacionária. Devido ao tipo de fenômeno que origina a série (consumo *mensal* de energia), espera-se que ela apresente uma sazonalidade anual, o que é confirmado pela função de autocorrelação ilustrada na Figura 5.24, onde a correlação de lag=12 e seus múltiplos estão bastantes destacados.

Dessa forma, o processo de estacionarização leva à série mostrada na Figura 5.25, juntamente com o seu espectro em freqüência (não foi necessária a remoção de componentes senoidais através da análise de Fourier). A respectiva função de autocorrelação encontra-se na Figura 5.26.

Para esta série, testamos o desempenho de uma rede não-recorrente (FFBP)

[6]Número suficiente para dar algum significado à variância calculada e permitir observar a sua evolução ao longo de 10 períodos da série.

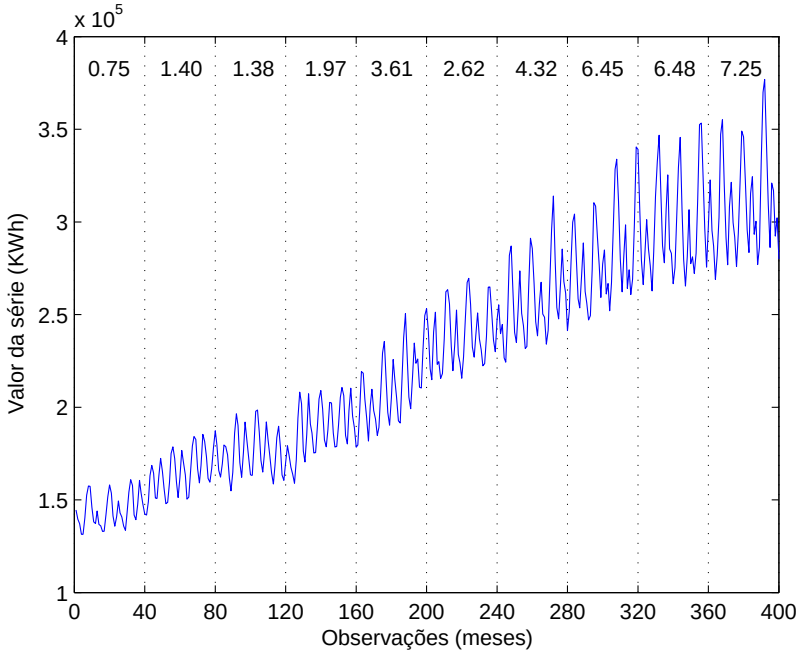

Figura 5.22: Consumo mensal de energia elétrica nos EUA (1973 a 2006) com as variâncias a cada trecho.

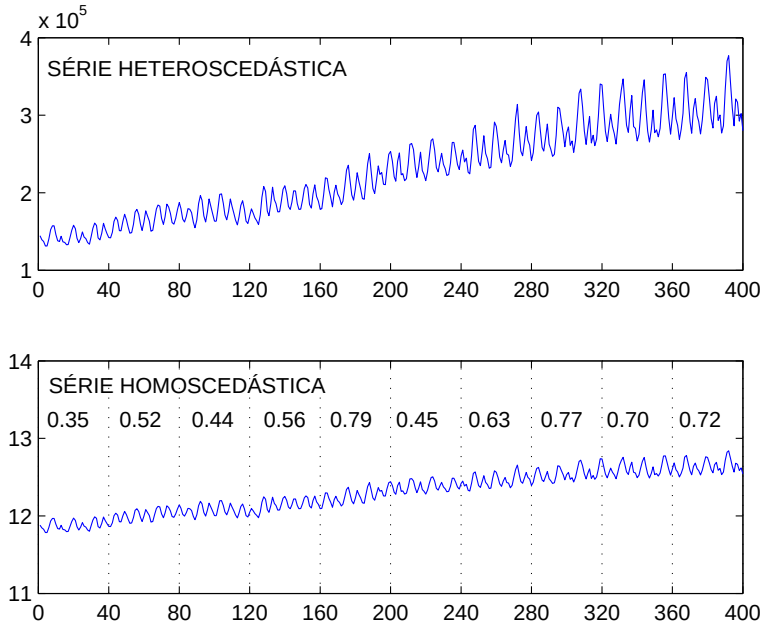

Figura 5.23: Homogeneização da variância da série de energia elétrica.

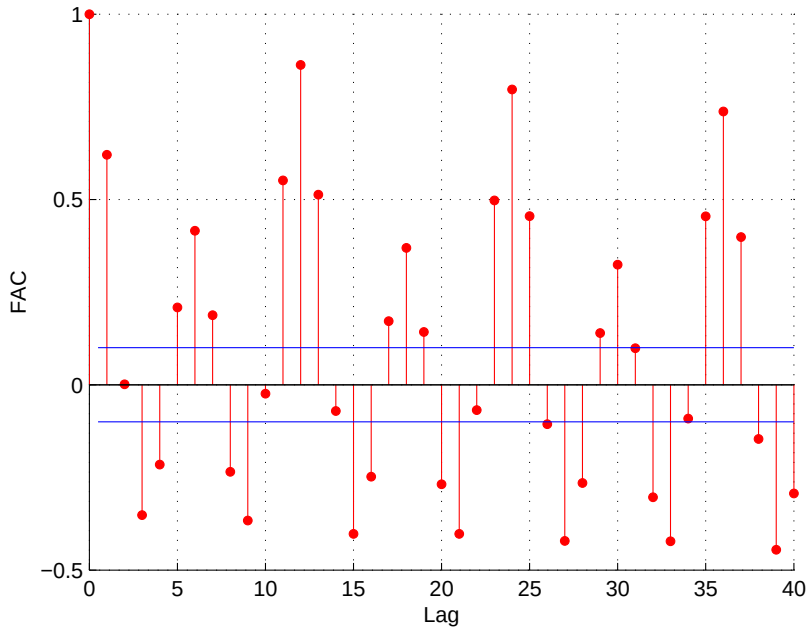

Figura 5.24: Função de autocorrelação para a série de energia elétrica já homoscedástica e sem tendência, mas ainda com sazonalidade.

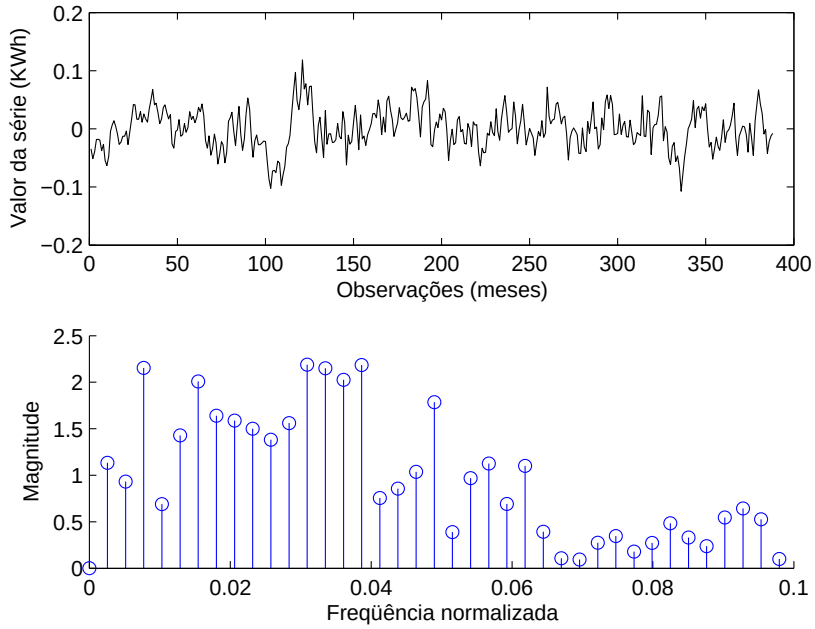

Figura 5.25: Série do consumo elétrico estacionarizada (acima) e espectro em freqüência (abaixo).

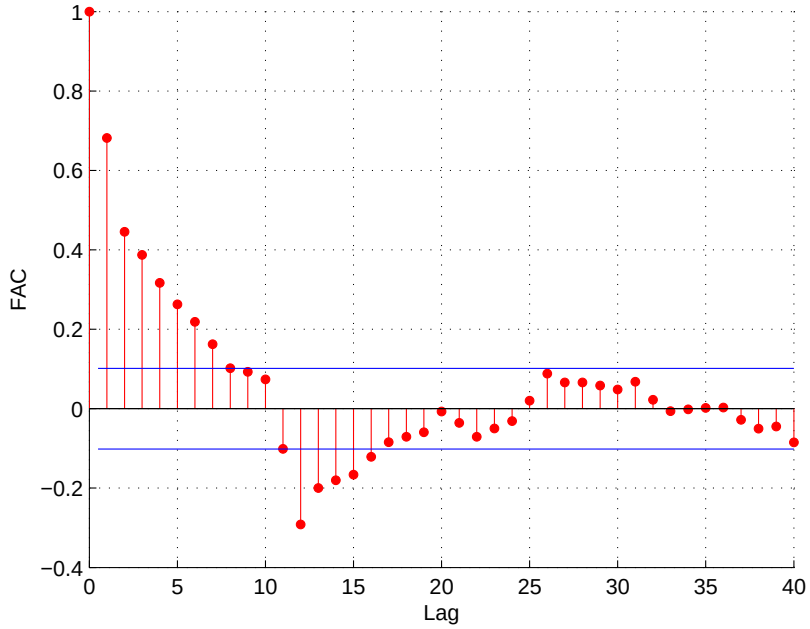

Figura 5.26: Função de autocorrelação para a série de energia elétrica estacionarizada.

e outra recorrente (Elman). Primeiramente, com a rede FFBP, buscamos a topologia ótima para as camadas de entrada e escondida. Com relação à entrada, fizemos N=12, aproveitando a correlação existente entre cada observação com as 11 anteriores (até chegar ao mesmo mês do ano anterior). Para determinar o número ótimo de neurônios na camada escondida, testamos o desempenho da rede para diversas configurações: com 2, 4, 6, 8, 10 e 12 neurônios. A rede com 8 neurônios é a que apresenta a melhor relação custo-benefício, visto que fornece o mesmo erro que a rede com 10 neurônios (4,1%), sendo mais parcimoniosa e vantajosa computacionalmente. A medida de erro utilizada para quantificar a qualidade da estimação é o MAPE (*Mean Absolute Percentage Error*):

$$MAPE = \frac{\sum_{i=1}^{P} \mid \frac{\tilde{x}_i - x_i}{x_i} \mid}{P} \tag{5.1}$$

onde P é o tamanho do conjunto de teste[7]. A Figura 5.27 ilustra a modelagem da série de eletricidade com a rede FFBP 12-8-1, cujas especificações estão listadas na

[7]Durante o treinamento, utilizamos o trecho final da série (últimos 30%) para o conjunto de teste; o trecho inicial (70%) é dividido entre os conjuntos de treino e validação, sorteando-se 50% para cada um.

Arquitetura	FFBP 12-8-1 com *bias*
Funções de ativação	tanh e linear (saída)
Função de treinamento	BP Resiliente
Função de aprendizado	Gradiente descendente com momento
Função de desempenho	Erro Médio Quadrático
Método	Batelada & Validação Cruzada

Tabela 5.1: Especificações para a rede FFBP usada para gerar a Figura 5.27.

Tabela 5.1. A largura total $(x_S - x_I)$ do corredor gerado nesta figura é $8\hat{\sigma}_e$, ou seja, k está no valor padrão de 4.

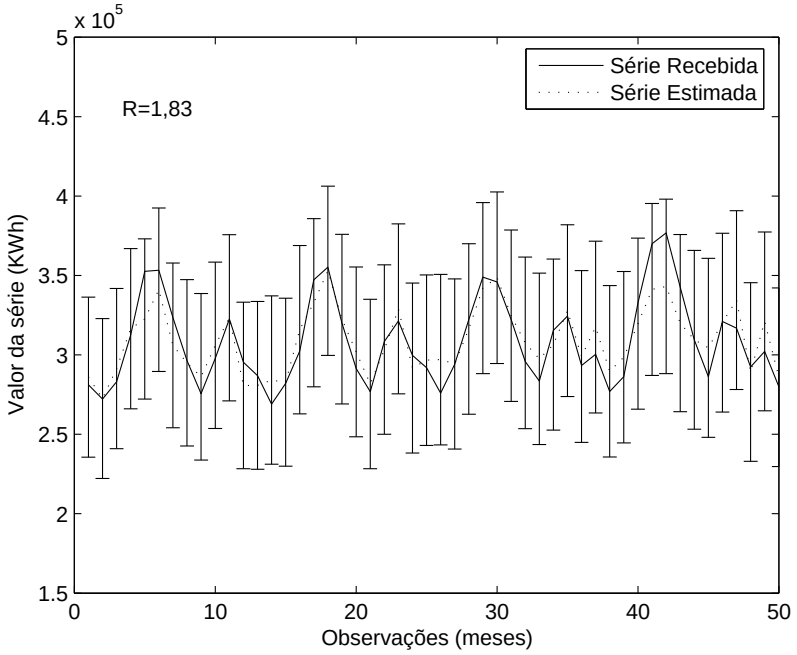

Figura 5.27: Predição da série de eletricidade pela rede FFBP.

Na própria Figura 5.27 consta o valor do fator R (1,83) de sincronismo entre a série real e a predita, conceito que definimos quando estudamos o atraso nas séries financeiras. Este valor bem acima de 1 mostra que o fenômeno do atraso não ocorreu na predição desta série. O MAPE de predição para o conjunto teste foi de 4,1%.

Mesmo tendo já obtido um valor de R alto com a rede FFBP, verificamos o desempenho de uma rede de Elman (com as mesmas características da rede FFBP, conforme listadas na Tabela 5.1) na modelagem desta série. Os resultados ilustrados

na Figura 5.28 mostram que o valor de R se mantém (1,8), mas o valor do MAPE diminui para 3,8%. Aqui, as características próprias da rede de Elman (realimentação na camada escondida) fazem-se valer.

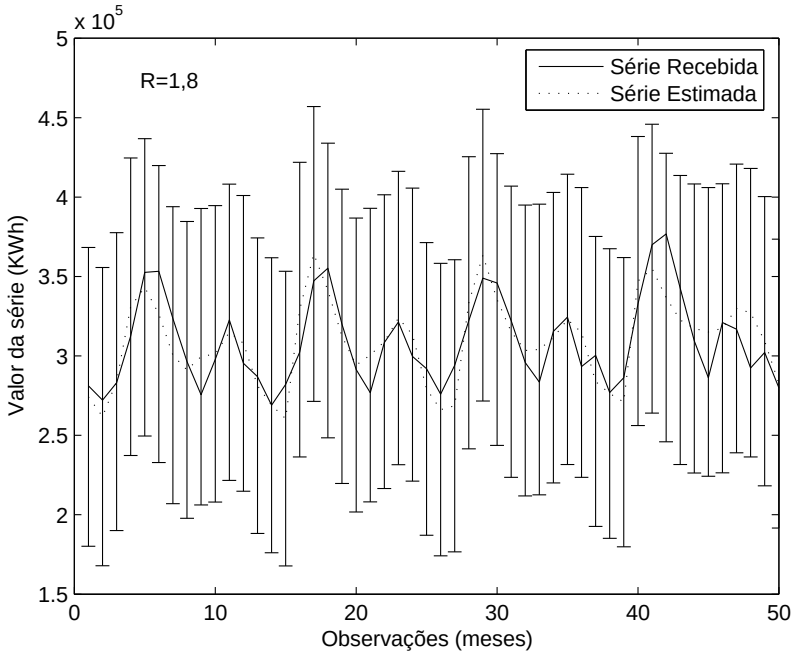

Figura 5.28: Modelagem da série de eletricidade pela rede de Elman.

Dessa forma, adotamos esta rede de Elman como o modelo gerador do corredor de validação para a série de eletricidade. A desvantagem da rede de Elman com relação à rede FFBP – o seu alto custo computacional no treinamento – em geral não é um empecilho, visto que não é necessário retreinar a rede a cada nova observação recebida (conforme explicado no Capítulo 4, em 4.3.1.1). Porém, caso essa dificuldade surgisse, o modelo da rede FFBP é uma alternativa viável. Como regra geral, o SMQD fornece os erros de predição de cada modelo neural desenvolvido para uma dada série, e o supervisor escolhe qual adotar como modelo padrão, tendo em conta a precisão e o custo computacional de cada um. Pode até mesmo escolher por formar o corredor a partir de uma combinação dos erros dos modelos FFBP e Elman (*ensemble*). Nesse caso, seria necessário manter ambos atualizados no SMQD.

Para testar o modelo gerado com a rede de Elman, submetemos esta série às duas situações anômalas possíveis para séries temporais: ocorrência de *outliers* e dados faltantes.

Para a simulação de *outliers*, analisamos a distribuição da série histórica (considerando que seus valores estão corretos) e determinamos o intervalo fora do qual uma observação deve ser considerada *outlier*, conforme o estabelecido em 4.3.1.1. Na Figura 5.29, ilustram-se os limites para "*outliers* suaves" e "*outliers* extremos" juntamente com a série estacionarizada[8].

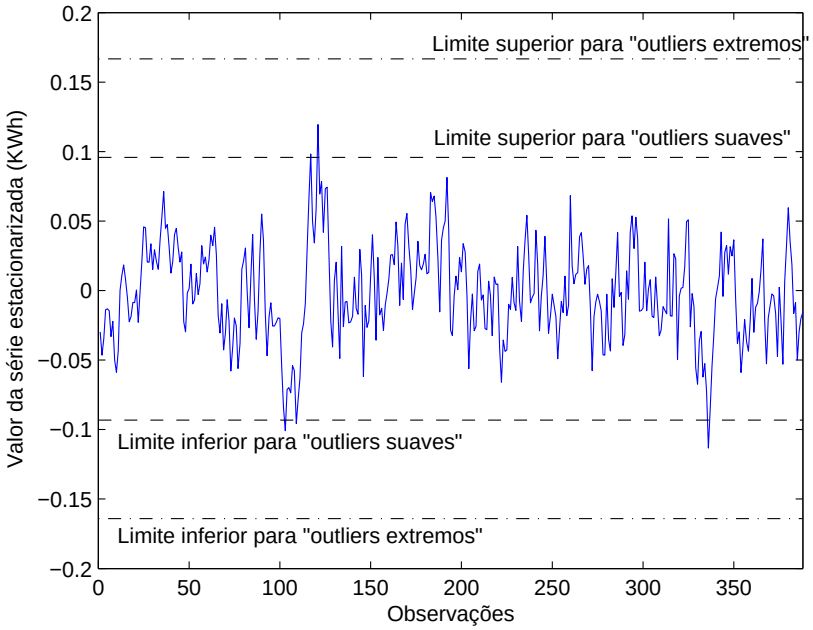

Figura 5.29: Determinação dos limites para *outliers* na série de eletricidade.

Observando a Figura 5.29, decidimos utilizar os limites extremos (e não os suaves) para determinar o intervalo procurado[9]. Dessa forma, simulamos 5 *outliers* (3 exatamente no limite superior extremo e 2 exatamente no limite inferior extremo) ao longo da série de teste, e a passamos pelo SMQD para monitoração. Estes *outliers* são representados na Figura 5.30 pelo símbolo "o".

Nesta figura, a largura do corredor é formada com k=1,5. Este valor permitiu validar todas as observações verdadeiras e detectar os 5 *outliers*, substituindo-os

[8]Como a série possui tendência e heteroscedasticidade, tais limites devem ser calculados sobre a série estacionarizada, e posteriormente transformados para se adaptarem à série original.

[9]Notemos que, usando esse intervalo, nenhuma observação da série original será classificada como *outlier*, ratificando nossa suposição inicial, ao passo que o uso dos limites suaves implicaria em classificar algumas observações como *outliers*. Como temos a garantia, por parte da fonte ([49]), de que tais dados são corretos, e o SMQD baseia-se no princípio de não rejeitar dados verdadeiros, usamos os limites extremos.

Número da observação	2	12	25	37	46
Valor Original	2,72	2,95	2,92	3,00	3,21
***outlier* (no limite)**	3,24 (sup.)	2,42 (inf.)	3,48 (sup.)	2,54 (inf.)	3,75 (sup.)
Limite do corredor	3,20	2,67	3,39	2,80	3,47
Valor Substituto	2,85	3,01	3,05	3,14	3,13

Tabela 5.2: Detecção e substituição de *outliers* na série de eletricidade. Os valores estão em 10^5 KWh.

pelos valores estimados pela rede neural. Caso o valor de k fosse menor, algumas observações reais seriam classificadas como *outlier* ("falso alarme"), e caso fosse maior, o terceiro *outlier* seria considerado como dado válido ("perda"). É assim que se procede a "calibração" da largura do corredor: a partir de *outliers* simulados, ajusta-se o valor de k de maneira a maximizar a detecção destes dados errôneos e a validação dos dados corretos. A Tabela 5.2 resume as informações sobre os *outliers* detectados e substituídos.

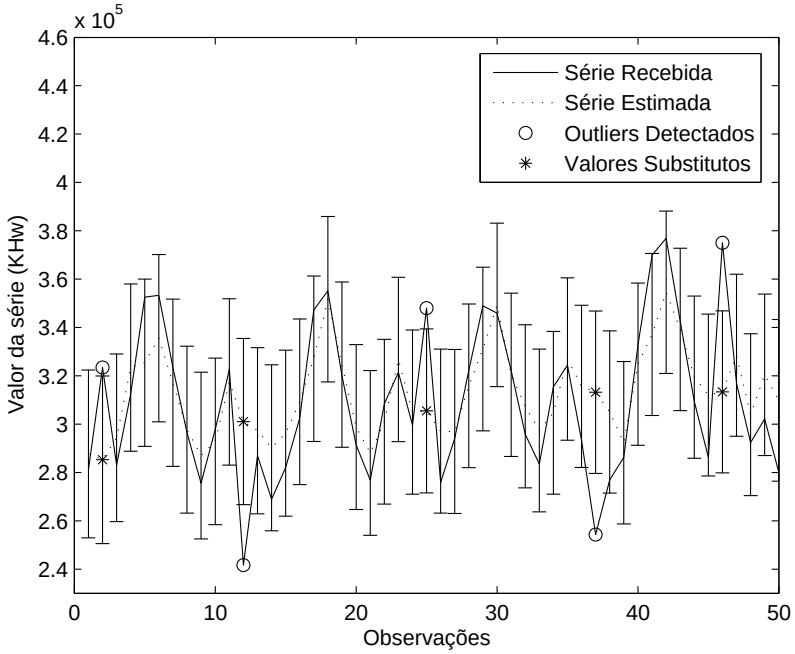

Figura 5.30: Correta remoção e substituição de 5 *outliers* através de corredor com k=1,5 (série da eletricidade).

A simulação seguinte diz respeito à substituição de dados faltantes, conforme

ilustra a Figura 5.31. Nesse caso, foram removidos 3 dados da série real (cujo gráfico fica interrompido nos trechos em torno dos dados faltantes). Aqui, a tarefa do SMQD consiste em repor tais dados através das predições do modelo neural.

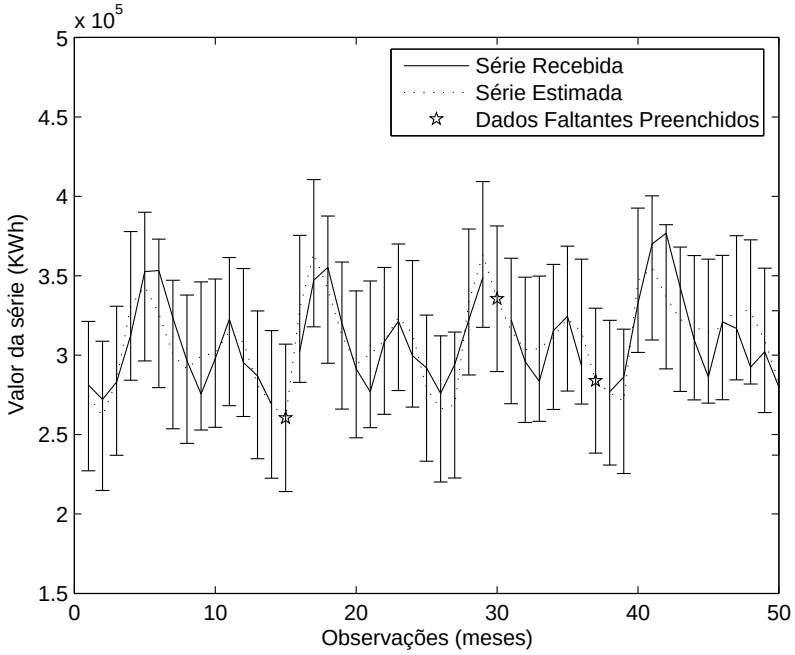

Figura 5.31: Reposição de dados faltantes na série de eletricidade.

A segunda série a ser analisada mede a Produção Física Industrial - Produtos Alimentares (índice IPI, Brasil, entre janeiro de 1985 e julho de 2000) [36]. Os dados têm freqüência mensal, e estão escalados em função do ano 1991 (cuja média foi normalizada para 100). Esta série está ilustrada na Figura 5.32.

A tendência é determinística (o teste ADF-PP indicou a ausência de raízes unitárias) e linear. A Figura 5.33 mostra o correlograma da série antes e após a retirada da tendência. Pode-se vislumbrar, tal como no caso da série da eletricidade, uma forte sazonalidade anual. Porém, não foi detectada heteroscedasticidade significativa. Apenas o que se pode observar é uma pequena estagnação da produção entre os anos de 1990 e 1993 (período em que a média da série fica praticamente constante e sua variância menor que nos períodos anterior e posterior: 234–180–245). A partir de 1994, ocorre uma retomada no crescimento da produção, coincidindo com o advento do Plano Real. É provável que na continuação da série a partir de 2000 (dados de que não dispomos), comece a formar-se um princípio de heteroscedasticidade.

A retirada da sazonalidade anual é feita através da aplicação do operador

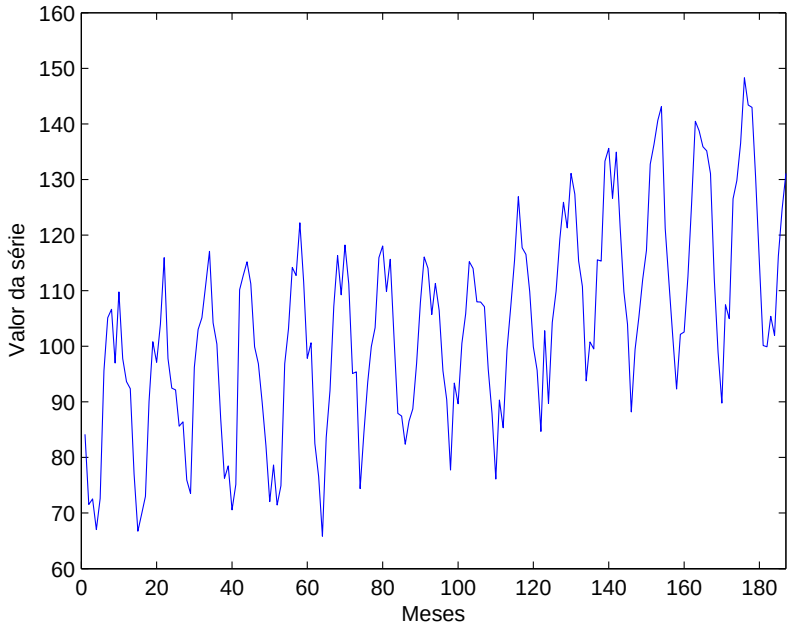

Figura 5.32: Série para a Produção Física Industrial - Produtos Alimentares (índice IPI, base Brasil) entre janeiro de 1985 e julho de 2000.

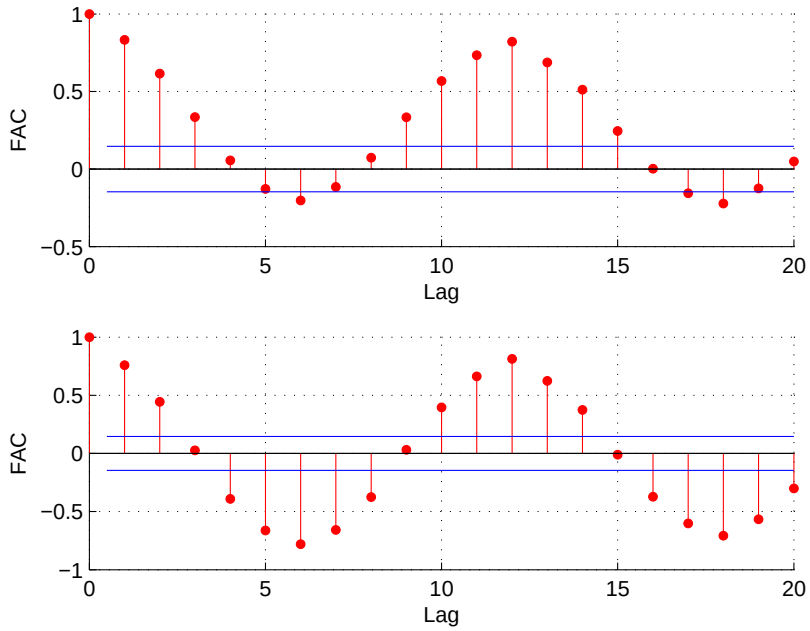

Figura 5.33: Correlograma para a série IPI antes e depois da retirada da tendência linear.

Arquitetura	Elman 12-6-1 com *bias*
Funções de ativação	tanh e linear (saída)
Função de treinamento	BP Resiliente
Função de aprendizado	Gradiente descendente com momento
Função de desempenho	Erro Médio Quadrático
Método	Batelada & Validação Cruzada

Tabela 5.3: Especificações para a rede Elman usada para modelar a série IPI.

$(1 - B)^{12}$. Isto termina o processo de estacionarização da série, conforme ilustrado na Figura 5.34.

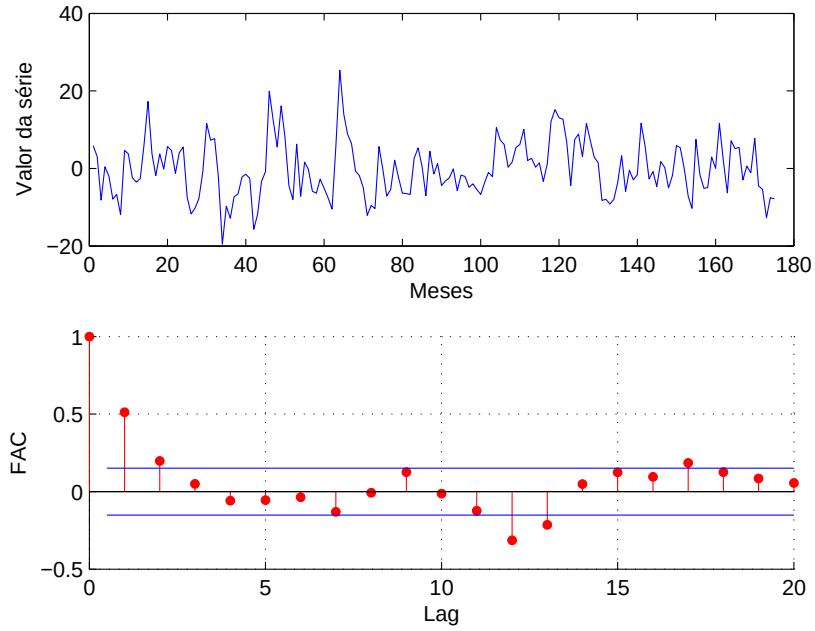

Figura 5.34: Série IPI desazonalizada (acima) e respectivo correlograma (abaixo).

Já sabendo das vantagens da rede de Elman para este tipo de série, fizemos a sua modelagem com a rede caracterizada na Tabela 5.3.

Os testes com a largura do corredor (feitos de forma análoga ao do exemplo anterior) revelaram que k=3 é suficiente para garantir a validação de todas as observações verdadeiras, conforme mostra a Figura 5.35. O fator de sincronismo alcançado foi R=1,22 e o erro de predição em teste é 4,2%.

Para exemplificar a atuação do SMQD, inserimos 2 *outliers* na série, além

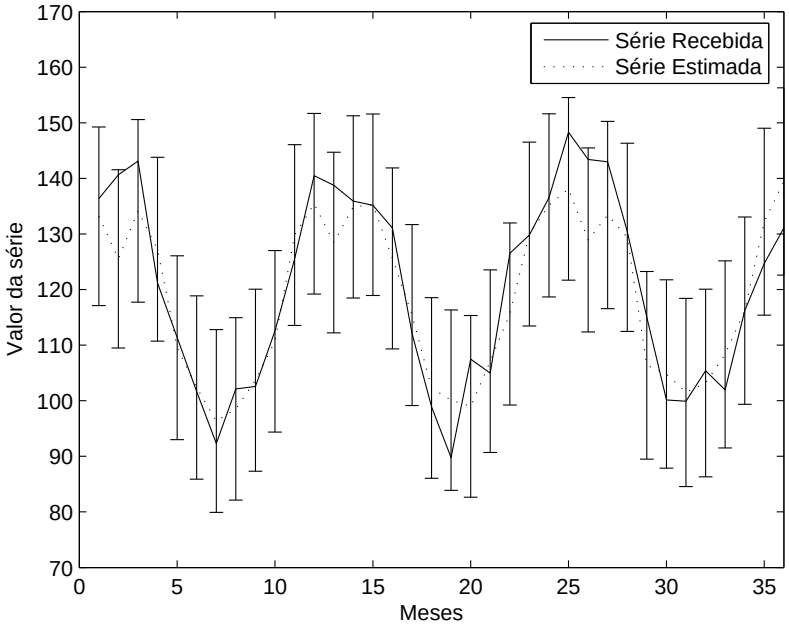

Figura 5.35: Modelagem da série do índice IPI pela rede de Elman, com k=3. O fator de sincronismo é R=1,22, e o erro de predição (conjunto de teste) vale 4,2%.

de simular a existência de 1 dado faltante. Na Figura 5.36, ilustra-se o processo de validação da série com anomalias.

5.2 Monitoração de Registros Multivariados

Depois de termos analisado a atuação do SMQD sobre dados de séries temporais, detemo-nos agora a verificar a sua eficácia na monitoração da QD de dados de registros multivariados.

5.2.1 Dados do ICPSR

A base de dados do projeto FISCT (*Family Interaction, Social Capital, and Trends in time use*) foi apresentada em 3.2.1.2. Vamos utilizá-la aqui para testar a atuação do SMQD na detecção de *outliers*, através da aplicação de ferramentas baseadas nas distâncias de Mahalanobis e Kullback-Leibler.

Com efeito, esse conjunto de dados tem, por natureza, grande probabilidade de conter *outliers*, devido à forma de aquisição dos dados (entrevista telefônica, o que costuma gerar erros de comunicação e posteriormente de digitação dos dados) e

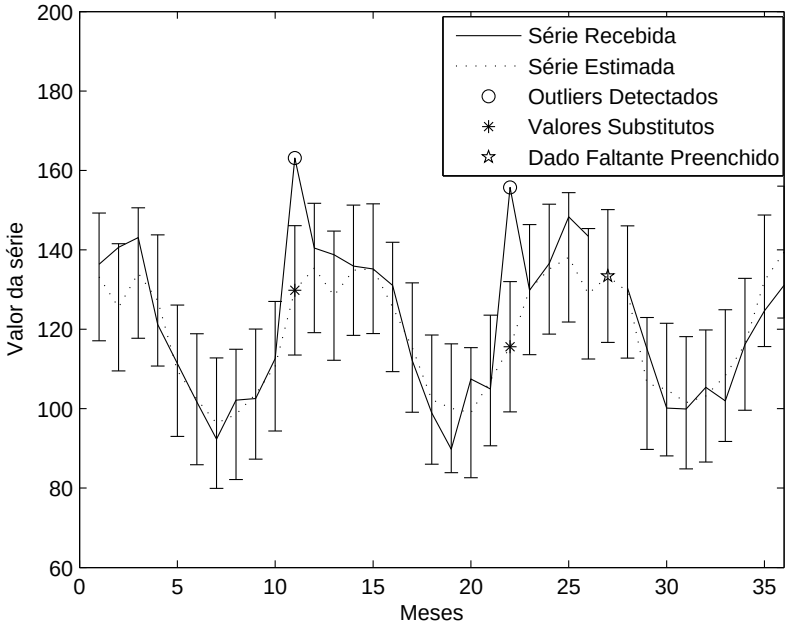

Figura 5.36: Detecção e correção de 2 *outliers* e preenchimento de 1 dado faltante, com k=3 (série IPI).

ao seu próprio conteúdo, que diz respeito a detalhes das atividadades do entrevistado no dia anterior (o que aumenta a chance de enganos e informações erradas).

Conforme exposto no Capítulo 3, essa base é composta por cerca de 2000 entrevistas, cada uma delas gerando um registro de todas as atividades desevolvidas pelo entrevistado no dia anterior. Um exemplo de registro está ilustrado na Tabela 5.4, estando os códigos das colunas listados na Tabela 5.5. A descrição detalhada dos códigos utilizados e a metodologia usada para o agrupamento de atividades encontram-se em [54].

Nosso objetivo é monitorar a base FISCT de maneira a detectar dados errôneos. Para tanto, fazemos a análise das distribuições de cada variável (coluna **C** na Tabela 5.4). Em concreto, seguimos o procedimento descrito abaixo:

1. Recebe uma nova amostra[10] $\{x\}$;

2. Agrega os valores dentro de uma mesma entrevista que pertençam à mesma variável (por exemplo: o tempo total de sono de um mesmo entrevistado, que se alcança somando os intervalos de tempo indicados na coluna **G**);

[10]Cada amostra pode conter uma ou mais entrevistas.

A	B	C	D	E	F	G	H	I	J	L	M	N	O
11002	10	Dormindo/cochilando	45			360	32598	0	600	1	3	1	
11002	11	Vestindo-se	47	0	0	30	32598	600	630	1	3	1	9
11002	12	Deslocamento	9	0	0	20	32598	630	650	11	3	1	9
11002	13	Reunião de trabalho	1	0	0	100	32598	650	830	4	3	5	9
11002	14	Deslocamento	9	0	0	20	32598	830	850	11	3	1	9
11002	15	Trabalhando	1	0	0	190	32598	850	1200	4	3	5	9
11002	16	Refeição/lanche	43	0	0	60	32598	1200	1300	4	3	5	9
11002	17	Trabalhando	1	0	0	330	32598	1300	1830	4	3	5	9
11002	18	Refeição/lanche	43	0	0	60	32598	1830	1930	4	3	5	9
11002	19	Trabalhando	1	0	0	105	32598	1930	2115	4	3	1	9
11002	20	Deslocamento	9	0	0	15	32598	2115	2130	11	3	1	9
11002	21	Banho/ducha	40			30	32598	2130	2200	1	3	1	
11002	22	Assistindo TV	91	96	0	30	32598	2200	2230	1	3	2	9
11002	23	Dormindo/cochilando	45			90	32598	2230	2400	1	3	1	

Tabela 5.4: Exemplo de resultado de entrevista para a base FISCT.

A	Código do entrevistado	**H**	Data da entrevista
B	Número da atividade	**I**	Hora de início
C	Atividade relatada	**J**	Hora de término
D	Código da atividade	**L**	Local da atividade
E	Atividade secundária (1)	**M**	Dia da semana
F	Atividade secundária (2)	**N**	Acompanhante? (1)
G	Tempo de duração (min)	**O**	Acompanhante? (2)

Tabela 5.5: Descrição das colunas da Tabela 5.4.

3. Para cada uma das variáveis, executa os passos seguintes:

4. Calcula a distância $DM1_i$ para todos os pontos pertencentes à distribuição já validada e também a $DM2_i$ dos pontos da amostra que está sendo testada à distribuição já validada[11];

5. Calcula a média μ e o valor RMS σ dos valores de $DM1_i$;

6. Se a $DM2_i$ de algum dado específico da nova amostra for maior que $\mu + n\sigma$, classifique tal dado como *outlier* (removendo-o da distribuição);

7. Calcula a distância de Kullback-Leibler (D_{KL}) entre a distribuição validada anterior e a nova distribuição atualizada com os novos dados validados;

8. Caso D_{KL} esteja acima do patamar estipulado, rejeita a amostra inteira, já que mesmo com a remoção de *outliers* específicos, a introdução da nova amostra alterou significativamente a distribuição já validada (o que pode significar que essa amostra pertence a outro modelo); caso contrário, atualiza o conjunto de dados com a nova amostra já verificada (ou seja, já com a exclusão dos *outliers* específicos).

Uma observação importante: para ser eficaz, o algoritmo acima deve começar a ser aplicado apenas quando já se dispõe de uma distribuição já caracterizada para a variável a ser analisada. Isso implica no risco de se validar, sem fazer o respectivo teste, alguns *outliers*, durante a fase de "construção" da distribuição. De outra forma, os testes de distância realizados perderiam o sentido, já que a distribuição de referência disporia de poucos dados para dar significado aos resultados encontrados. Quanto aos possíveis *outliers* validados inicialmente, podem ser removidos através de análise *offline*, feita especificamente para detectá-los.

Para exemplificar o procedimento descrito acima, ilustramos o caso da detecção de *outliers* na variável "tempo de sono" (correspondente ao total de minutos que cada um dos entrevistados alegou como tempo de sono do dia anterior). A Figura 5.37 mostra a distribuição completa dessa base, sem a remoção de nenhum

[11]Para evitar que, à medida que o conjunto de dados aumente, o cálculo de todas as DM_i fique computacionalmente muito custoso, pode-se optar por fazer uma amostragem da distribuição.

outlier. Pode-se perceber a existência de dados muito provavelmente errôneos, sobretudo ao final da cauda direita da distribuição (tais dados implicariam no fato de o entrevistado ter estado dormindo durante quase todo o dia anterior). De fato, uma análise prévia das entrevistas que geraram esses dados mostram que as respectivas respostas não são confiáveis.

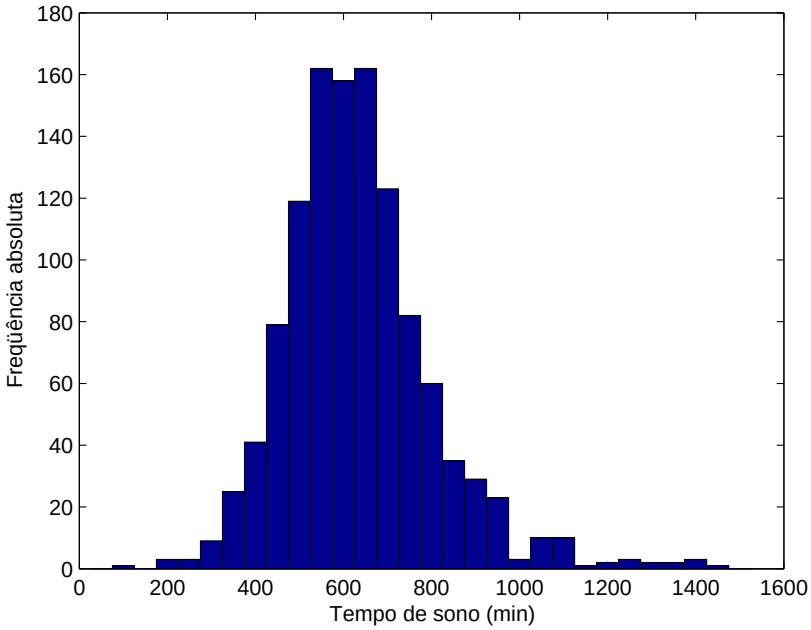

Figura 5.37: Histograma para a variável "tempo de sono" sem a atuação do SMQD.

Implementamos, sobre essa base, o algoritmo descrito anteriormente. Para tanto, construímos uma distribuição inicial que continha 75% do total de registros da base. Posteriormente, simulamos a chegada de novos registros, em amostras contendo TA registros. O valor de TA foi ajustado para TA=10 amostras[12]. A cada chegada de novos registros, agregava-se o tempo total de sono de cada entrevistado. Procedia-se a etapa da Distância de Mahalanobis (passos 4, 5 e 6 do algoritmo), buscando-se identificar *outliers* dentro da nova amostra. Nessa etapa, o valor de n foi igual a 2. A seguir, nos passos 7 e 8, usava-se a Distância de Kullback-Leibler para calcular a divergência entre a distribuição anterior e a acrescida da nova amostra (já sem os *outliers*). Nesse momento, há que se determinar também a granularidade dos histogramas que acumulam a distribuição (valor da "binagem", n_b), já que isto

[12]Esse valor não pode ser nem muito pequeno (para dar significância ao cálculo da D_{KL}) e nem muito grande para não perdermos a sensibilidade aos *outliers*.

influi no cálculo da D_{KL}. No caso deste exemplo, trabalhamos com n_b=30.

Com relação ao patamar estipulado como valor máximo aceitável de D_{KL}, o procedimento padrão que o SMQD adota consiste em calcular essa distância entre duas amostras de distribuição de ruído branco gaussiano que contenham o mesmo tamanho da maior das duas distribuições que se quer comparar (a maior será sempre a distribuição acrescida com a nova amostra). Para evitar flutuação estatística, calcula-se essa distância por 100 vezes e assume-se a média como o valor de referência d_{kl}. Dessa forma, se a D_{KL} entre as duas distribuições que se está comparando for maior que d_{kl}, diz-se que há divergência significativa entre elas.

A Figura 5.38 ilustra a distribuição da variável "tempo de sono" após a monitoração do SMQD. Percebe-se que o principal efeito de sua atuação foi descartar os eventos extremos (ilustrados na Figura 5.39), sobretudo os da cauda direita, tornando a distribuição dessa variável mais próxima à curva gaussiana, que é o modelo tomado para uma variável desse tipo quando o número de eventos é suficientemente grande.

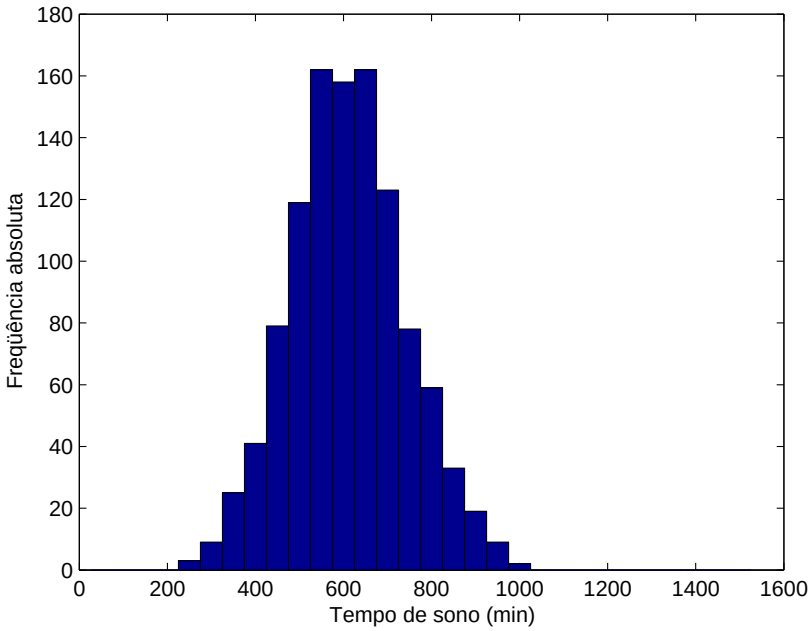

Figura 5.38: Histograma para a variável "tempo de sono" após a atuação do SMQD.

Essa aproximação da curva gaussiana foi quantificada através de um ajuste (*fitting*) da distribuição dos 2 conjuntos de dados (antes e depois da monitoração). Com efeito, conforme ilustrado nas Figuras 5.40 e 5.41, a remoção dos eventos

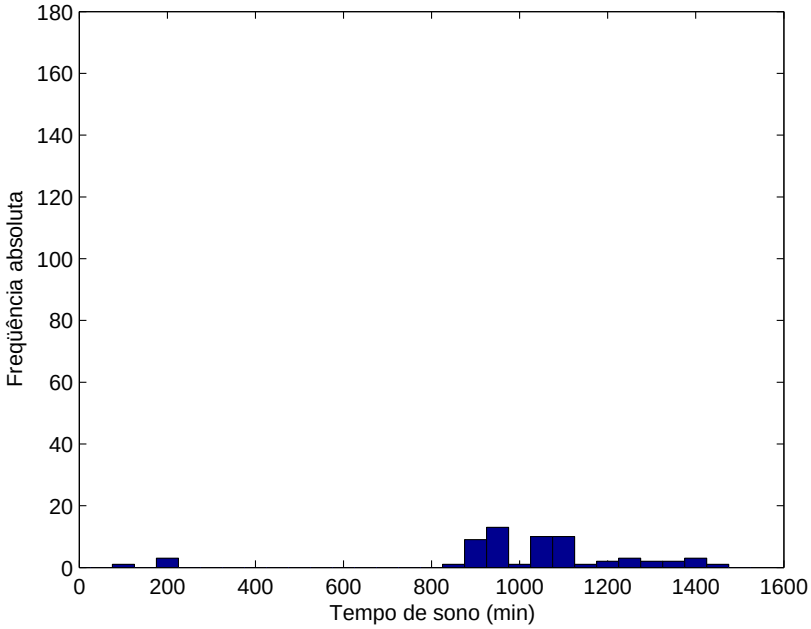

Figura 5.39: *Outliers* removidos pelo SMQD.

extremos ocasionou uma melhora no ajuste, aumentando o valor da estatística R^2 (de 0,985 para 0,994) e sobretudo diminuindo o valor do erro RMSE (de 6,61 para 4,31).

5.2.2 Dados da Área Médica

Procedemos também alguns testes com dados da área médica. Em concreto, os dados utilizados provêm de um estudo feito sobre a incidência da Hepatite A em uma localidade[13] do 2º Distrito do município de Duque de Caxias.

Este estudo tinha por objetivo relacionar a incidência dessa doença com algumas variáveis que mediam as condições sócio-econômicas e sanitárias da população local. De um total de mais de 60 variáveis monitoradas, foram pré-selecionadas 7 (além da variável que indica a presença da hepatite) através de uma análise de relevância (feita em [112]). Estas 8 variáveis estão descritas abaixo.

- **HEP.**: Vale 1 se a pessoa tem o vírus da Hepatite A, e -1 caso contrário;
- **IDADE**: Idade da pessoa;

[13]Setor Parque Fluminense.

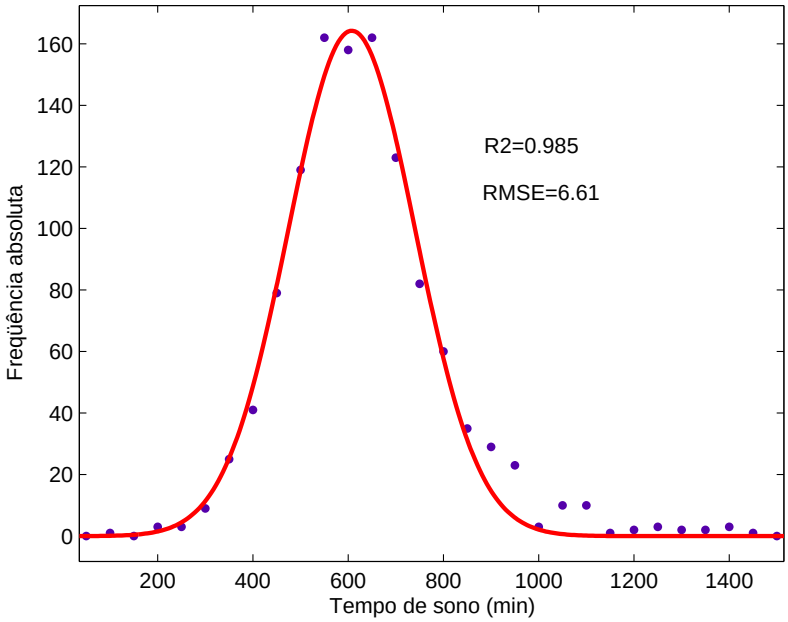

Figura 5.40: Ajuste gaussiano para o conjunto sem monitoração.

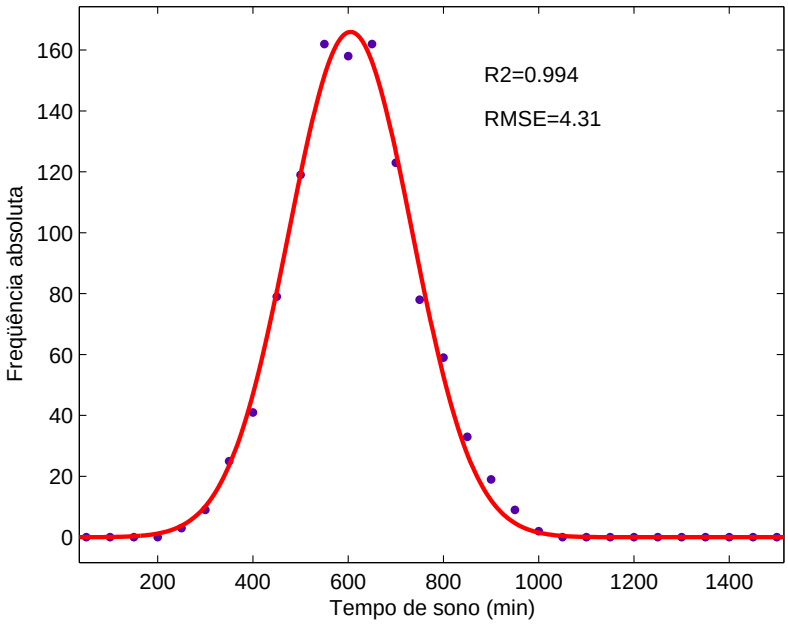

Figura 5.41: Ajuste gaussiano para o conjunto monitorado.

HEP.	IDADE	ÁGUA	FILT.	ESC.	RENDA	VALA	CÔMODOS
1	5	1	1	5	2,0	1	1,8
-1	7	5	1	3	2,6	1	1,5
-1	5	1	-1	5	1,0	1	1,33
-1	2	5	-1	6	2,5	1	1,0
-1	1	6	1	6	0	1	1,33
1	9	2	1	2	3,0	-1	1,25
1	7	0	1	1	1,0	1	2,20
-1	1	4	-1	3	3,3	-1	1,25
...	...	...	...	...	...	...	...

Tabela 5.6: Alguns eventos para o registro de casos de hepatite A.

- **ÁGUA**: Número de pontos de água da casa;
- **FILT.**: Vale 1 se a casa utiliza filtragem de água, e -1 caso contrário;
- **ESC.**: Grau de escolaridade da dona de casa;
- **RENDA**: Renda mensal;
- **VALA**: Vale 1 se a casa está próxima de uma vala negra, e -1 caso contrário;
- **CÔMODOS**: Densidade de cômodos.

A Tabela 5.6 traz alguns eventos (nas linhas) contidos na base de dados com que trabalhamos[14]. Um registro desse tipo pode ser usado para detectar padrões na relação entre as 7 variáveis selecionadas e a presença da doença, o que pode levar a uma maior compreensão dos fatores que favorecem a contaminação pelo vírus da hepatite A e, conseqüentemente, a medidas de prevenção da doença. Daí a importância de que os dados do registro estejam completos e acurados.

Como dispomos, para este estudo, da base completa, podemos simular a existência de dados faltantes, aplicar o SMQD sobre os dados e posteriormente verificar a sua eficácia através de comparação com os dados originais. Dessa forma, desenvolve-se um método para a imputação de valores nos casos em que a base

[14]A base completa possui 2771 eventos.

vier a apresentar dados faltantes, o que evita ter que descartar eventos incompletos. Para esse tipo de base, isto é altamente vantajoso, pois cada evento tem uma aquisição bastante custosa (exige fazer uma entrevista com os moradores, pesquisar suas condições de vida etc).

5.2.2.1 Imputação Usando Algoritmo EM Regularizado

Os valores faltantes são imputados através de um algoritmo EM regularizado[15]. Em cada iteração do algoritmo, os valores estimados para a média de cada variável (colunas do registro) e da matriz de covariância são revisados em 3 passos. Primeiramente, para cada evento X_i (linha do registro) que contenha dado faltante, os parâmetros da regressão das variáveis que também contêm dados faltantes sobre as que não contêm são calculados a partir das estimativas para a média e para a matriz de covariância. Depois, o valor faltante no evento X_i é preenchido com o seu valor esperado condicional, dados os valores disponíveis (não faltantes) e as estimativas para a média e para a matriz de covariância. Este valor esperado condicional é o produto dos valores disponíveis pelos coeficientes de regressão (feita usando o parâmetro de regularização de *ridge* [99]). Por último, a média e a matriz de covariância são reestimadas: a média como sendo a média amostral do registro atualizado com o novo dado estimado, e a matriz de covariância como sendo a soma da matriz de covariância amostral do registro atualizado com a estimativa da matriz de covariância condicional do erro de imputação.

Como condição inicial padrão para o algoritmo de imputação, as médias das variáveis do registro são calculadas a partir dos dados disponíveis e são usadas para preencher os dados faltantes. A matriz de covariância é estimada como sendo a matriz de covariância amostral do registro completo (com as médias no lugar dos dados faltantes).

Sorteamos aleatoriamente 30 valores do registro para hepatite e removemos o seu conteúdo, tornando-os dados faltantes. Em seguida, aplicamos o algoritmo descrito acima para imputar valores aos dados faltantes. Os resultados estão listados na Tabela 5.7

[15]Agradecemos a Tapio Schneider (California Institute of Technology) a disponibilização de uma função em MATLAB® com a implementação do algoritmo EM regularizado.

Variável	**Valor Orig.**	**Valor EM**	**Variável**	**Valor Orig.**	**Valor EM**
IDADE	9,0	8,54	RENDA	2,5	2,79
IDADE	6,0	8,66	RENDA	4,0	3,64
ÁGUA	10,0	8,54	RENDA	2,0	1,58
ÁGUA	4,0	4,58	RENDA	5,0	3,89
ÁGUA	5,0	3,43	RENDA	3,0	3,03
ÁGUA	3,0	3,48	RENDA	5,8	3,49
FILT.	1,0	0,29	RENDA	2,9	3,47
FILT.	1,0	0,01	RENDA	4,0	5,53
FILT.	-1,0	0,38	RENDA	1,5	4,03
FILT.	1,0	0,12	VALA	-1,0	-0,11
FILT.	1,0	0,34	VALA	1,0	-0,02
ESC.	1,0	3,69	VALA	1,0	0,11
ESC.	6,0	5,06	VALA	-1,0	-0,01
ESC.	4,0	4,76	CÔMODOS	1,2	1,57
ESC.	4,0	5,41	CÔMODOS	3,0	1,68

Tabela 5.7: Resultado da imputação via algoritmo EM para dados faltantes do conjunto da hepatite.

Uma dificuldade apresentada pelo algoritmo diz respeito à estimação das variáveis binárias ("FILT." E "VALA"). Estas só assumem os valores -1 ou 1 e têm médias próximas a zero. Por isso, o erro percentual de estimação é alto; porém, caso estabeleçamos um corte em 0, o percentual de acerto é mais alto (para o exemplo acima, 7 acertos e 2 erros). Quando fazemos um teste mais extenso, removendo os dados de 100 campos do registro, o índice de acerto é de 63%.

5.2.2.2 Usando Imputação Múltipla

Quando se dispõe de poucos dados, pode ser vantajoso o uso do método de imputação múltipla[16]. Implementamos neste exemplo uma versão que consiste basicamente em uma junção do algoritmo EM usado no exemplo anterior com um algoritmo do tipo DA (*Data Augmentation*). Este último consiste em uma técnica iterativa (semelhante a MCMC) com 2 passos [113]:

- Passo "I" (*Imputation*): Imputam-se os dados faltantes sorteando-se valores segundo a distribuição condicional dos dados existentes, com os parâmetros atualizados no momento;
- Passo "P" (*Parameter*): Simulam-se novos valores para os parâmetros a partir da distribuição bayesiana "a posteriori", dados os valores existentes e os valores imputados no passo anterior.

As 3 etapas da imputação múltipla (algoritmo combinado EM-DA) são implementadas da seguinte forma[17] (conforme introduzido em 4.3.2.1):

1. Calculam-se, através do estimador de máxima verossimilhança (algoritmo EM, com k iterações), os parâmetros de regressão para as variáveis (colunas) do registro multivariado. Tais parâmetros servem como valores iniciais para o passo seguinte;
2. Executa-se, com mk iterações, o algoritmo DA, produzindo-se imputações a cada k iterações;

[16]Para o conjunto completo da hepatite, usado no exemplo anterior, a eficiência do algoritmo EM é maior que a do método da imputação múltipla.

[17]O cálculo da imputação múltipla foi feito utilizando o *software* "NORM", disponibilizado por J. L. Schafer. A análise dos resultados foi feita em MATLAB®.

3. Toma-se a média das m imputações do passo anterior como resultado para os valores imputados.

Para comparar a eficiência do método de imputação múltipla com a do algoritmo EM em um conjunto com poucos dados, selecionamos cerca de 10% dos dados da base da hepatite e simulamos 10% de dados faltantes nesses eventos. Em seguida, procedemos o preenchimento dos campos em branco através dos 2 métodos. A razão entre os valores do erro (RMSE) com a utilização do método de imputação múltipla e do algoritmo EM foi de 0,98, enquanto que para o caso do conjunto completo a razão calculada foi 2,55.

Capítulo 6

Conclusões e Sugestões para Trabalhos Futuros

Neste último capítulo, expomos as conclusões da pesquisa de tese realizada neste trabalho. Primeiramente, comentamos as contribuições desta tese para a pesquisa na área de Qualidade de Dados. Na seqüência, fazemos a crítica da metodologia criada para o sistema de monitoração, núcleo do trabalho. Os testes implementados com as ferramentas desenvolvidas para o SMQD são também alvo de discussão. Por fim, fazemos alguns comentários sobre possíveis extensões para a pesquisa desenvolvida.

6.1 Estudo dos Tópicos de Qualidade de Dados

A área de Qualidade de Dados possui algumas características bastante motivadoras para a realização de uma pesquisa de tese, a saber: é uma área ainda incipiente, o que torna necessária a abertura de diversas frentes de pesquisa, em que se trabalhe tanto na formalização de conceitos como na geração de ferramentas de desenvolvimento; é uma área multidisciplinar, o que a torna convidativa a pesquisadores de diversos ramos do conhecimento e permite o entrelaçamento de conceitos e métodos de diferentes áreas; e é também uma área com aplicações muito concretas, sobretudo no estágio atual em que tanto instituições comerciais quanto acadêmicas caminham no sentido de tornar este século o "século dos dados", visto que a extração das informações neles contidas afeta cada vez mais a vida da maioria das pessoas,

destinatárias últimas de todo o processo de monitoração de QD.

Nesta tese, procuramos aproveitar um pouco dessas três características da área de Qualidade de Dados. Com efeito, acreditamos que uma primeira contribuição dada por esta pesquisa consistiu na formalização de alguns conceitos envolvidos diretamente na área de Qualidade de Dados, a partir de uma revisão bibliográfica centrada nos principais canais de divulgação da área. Isto é particularmente importante por dois motivos: gera uma base sólida sobre a qual desenvolver a pesquisa e fornece àqueles que queiram iniciar-se no assunto um resumo de sua teoria, com as respectivas referências bibliográficas.

Alguns dos conceitos mais importantes que foram definidos neste trabalho dizem respeito às dimensões da QD e à geração de métricas para a sua monitoração. Também o contexto da área e as motivações para o desenvolvimento da pesquisa em QD foram discutidas, levantando-se argumentos de ordem tecnológica, econômica e social que mostram o impacto do mau gerenciamento da QD nesses diversos âmbitos.

Com relação à multidisciplinariedade desta pesquisa, ela se faz notar não somente na origem dos dados, por si só bastante heterogênea, mas também nas técnicas utilizadas no sistema de monitoração. Em particular, o sistema aqui proposto agrega ferramentas da área de processamento de sinais aos métodos estatísticos mais adequados para tratar cada tipo de falha nos dados.

A aplicabilidade da pesquisa se concretiza na confecção do sistema de monitoração, em que se implementa a metodologia desenvolvida para atuar sobre bases de dados reais. Comentários a este sistema são feitos na próxima seção.

6.2 A Metodologia do Sistema de Monitoração

O desenvolvimento desta metodologia consiste no núcleo desta tese e é sua principal contribuição. De fato, a ausência de sistemas capazes de gerir a qualidade dos dados de sistemas de informação é uma das lacunas mais sentidas na área de Qualidade de Dados.

Vimos no Capítulo 4 a definição da estrutura do SMQD, um sistema integrado de monitoração de Qualidade de Dados. A principal preocupação com relação ao desenvolvimento desse sistema consistia em dar-lhe algumas características essen-

ciais, tais como o dinamismo (capacidade de receber e monitorar dados em tempo real, fazendo as alterações necessárias) e o enfoque *data driven*, isto é, a capacidade de direcionar as suas ações de acordo com o tipo de dados que recebe.

Para tanto, o SMQD conta com uma série de mecanismos de atuação especialmente projetados para monitorar a qualidade dos grupos de dados tratados pelo sistema com a freqüência necessária. Para séries temporais, os modelos de predição dispõem sempre de um valor estimado para a observação seguinte. Esta, quando recebida, passa por um processo de validação que visa à detecção de *outliers*; caso não seja recebida (dado faltante), é substituída pelo valor estimado pelo modelo.

De forma análoga, a monitoração dos registros multivariados é feita através de ferramentas que preenchem os seus dados faltantes e detectam os *outliers* que, neste caso, não necessitam de substituição, sendo removidos da distribuição (são guardados à parte). Diferentemente do caso das séries temporais, aqui a freqüência de monitoração é irregular, determinada pela chegada de novas amostras. Caso estas tardem em chegar, a penalização se dá na dimensão da pontualidade que, no caso das séries temporais, se confunde com a dimensão da completude.

Dessa forma, a dimensão da acurácia é monitorada pela detecção (e substituição, no caso das séries temporais) dos *outliers*, e a dimensão da completude é monitorada pelo correto preenchimento dos dados faltantes.

6.3 Testes Implementados

Conforme dissemos desde o início deste texto, o foco principal deste trabalho é a metodologia do sistema de monitoração. Ou seja, nossa preocupação era deixar estabelecida a estrutura do SMQD, as ferramentas que o compõem, os seus modos de atuação, os tipos de dados com os quais lida e também as dimensões da qualidade dos dados que são por ele monitorados.

Assim sendo, a realização de testes exaustivos com as ferramentas desenvolvidas não era o nosso objetivo. Interessava-nos mais fazer uma seleção das ferramentas a serem utilizadas do que explorar a sua utilização. Com efeito, temos consciência de que cada uma das ferramentas desenvolvidas pode ser melhorada, e mesmo novas ferramentas podem ser incorporadas ao SMQD. Ou seja, a seleção feita não pre-

tende ser definitiva, mas sim um conjunto inicial a partir do qual se pode continuar a evolução do SMQD.

Não obstante, apresentamos no Capítulo 5 alguns testes realizados com o SMQD, com o intuito de ilustrar a sua atuação diante de casos concretos que envolviam as anomalias definidas no Capítulo 4: o tratamento de *outliers* (para monitorar a dimensão da acurácia) e de dados faltantes (para monitorar a *completude*), tanto para séries temporais (de diferentes origens) como para registros multivariados.

De fato, os testes realizados cobriram uma ampla gama de dados, como séries temporais financeiras e de fenômenos físicos e econômicos, além de registros multivariados sociais e da área médica.

6.4 Continuações para esta Pesquisa

Cremos que o trabalho realizado possibilita algumas oportunidades de seguimento, em diversas áreas. Propomos aqui algumas dessas possibilidades, fornecendo também algumas dicas sobre como levar a cabo a implementação de tais sugestões.

- **Extensão do SMQD para as dimensões subjetivas**. Conforme vimos no Capítulo 2, muitas das dimensões da QD são subjetivas. Efetivamente, em muitas bases de dados essas dimensões são bastante relevantes (a acurácia e a completude, dimensões objetivas monitoradas pela versão atual do SMQD, são importantes sempre), como a credibilidade da fonte e a inteligibilidade dos dados. Nesses casos, uma maneira de quantificar tais dimensões é solicitar aos usuários que lhes dêem notas, o que torna possível ao sistema medir a evolução de seus valores ao longo do tempo;

- **Desenvolvimento de *software* com a implementação do SMQD em interface gráfica**. Este trabalho possibilitaria a instalação e a utilização do SMQD em ambientes específicos, além de permitir a interação com o usuário através da interface gráfica;

- **Inserção de processamento paralelo nas ferramentas do SMQD**. Como a tendência é que as bases monitoradas cresçam com o tempo e também o número de ferramentas disponíveis – e conseqüentemente o número de testes a

serem realizados – aumente, o sistema poderia aproveitar-se da implementação de seus métodos de correção em algoritmos paralelos a serem executados em plataformas de computação distribuída como, por exemplo, placas contendo múltiplos DSPs[1];

- **Inclusão de novas ferramentas e melhoria daquelas já desenvolvidas**. É fato que as ferramentas desenvolvidas podem ser melhoradas, e isto representa um amplo campo de trabalho: novas técnicas de pré-processamento dos dados de séries temporais (como ICA (*Independent Component Analysis* [114][115]) ou filtragem via *Wavelet* [116][117]), aperfeiçoamento dos modelos neurais (técnicas não-supervisionadas [104] e mescladas com algoritmos evolucionários [103]). Além disso, novas ferramentas, baseadas em outros modelos, podem ser incorporadas.

Ainda dentro do contexto de sugestões para trabalhos futuros, um campo promissor para a aplicação do SMQD consiste na monitoração da QD de dados da área médica (um exemplo, com os dados da hepatite, foi ilustrado no Capítulo 5). Conforme mencionado no Capítulo 2, o impacto da má qualidade dos dados nesta área é muito grande. Além disso, a implementação de sistemas de QD pode auxiliar na integração das bases de dados de áreas ligadas à saúde.

Pode-se vislumbrar também a aplicação do SMQD na análise de dados intangíveis (como a credibilidade de uma empresa, por exemplo). É sabido, hoje em dia, que a correta valoração de tais fatores é fundamental para uma compreensão profunda do funcionamento da economia e da sociedade como um todo, visto que englobam não só os aspectos técnicos mas também os valores éticos refletidos nas ações humanas e, de alguma forma, nos dados a elas relacionados.

A pesquisa realizada resultou na publicação de 3 artigos: [75][76][77].

[1]Digital Signal Processors.

Referências Bibliográficas

[1] WANG, R. Y., KON, H. B., MADNICK, S. E., "Data Quality Requirements Analysis and Modeling", Ninth International Conference of Data Engineering, April 1993.

[2] ECKERSON, W. W., *Data Quality and the Bottom Line*, Report, The Data Warehousing Institute, 2002.

[3] SCHMIDT, E., "Don't bet against the internet", The Economist (Special Edition, p. 110), December 2006.

[4] DONOHO, D. L., "High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality", Lecture for the American Math. Society "Math Challenges of the 21st Century", August 2000.

[5] "CERN: the world's largest particle physics laboratory", http://www.cern.ch, acessado em abril de 2007.

[6] "How Much Information?", http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm, acessado em fevereiro de 2007.

[7] EPPLER, M. J., HELFERT, M., "A Classification and Analysis of Data Quality Costs". In: *Proceedings of the Ninth International Conference of Information Quality (ICIQ)*, 2004.

[8] PIPINO, L., KOPCSO, D., "Data Mining, Dirty Data and Costs". In: *Proceedings of the Ninth International Conference of Information Quality (ICIQ)*, 2004.

[9] STRONG, D. M., LEE, Y. W., WANG, R. Y., "Data Quality in Context", *Communications of the ACM*, v. 40, n. 5, pp. 103–110, 1997.

[10] SPIEGELHALTER, D., GRIGG, O., KINSMAN, R., *et al.*, "Risk-adjusted sequential probability ratio tests: application to Bristol, Shipman and adult cardiac surgery", *International Journal for Quality in Health Care*, v. 15, n. 1, pp. 7–13, 2003.

[11] GRIGG, O. A., FAREWELL, V. T., SPIEGELHALTER, D. J., "The use of risk-adjusted CUMSUM and RSPRT Charts for Monitoring in Medical Contexts".

[12] "The MIT Total Data Quality Management Program", http://web.mit.edu/tdqm/, acessado em fevereiro de 2007.

[13] DRUCKER, P. F., "The Coming of the New Organization", *Harvard Business Review on Knowlodge Management*, pp. 1–19, 1998.

[14] PIPINO, L. L., LEE, Y. W., WANG, R., "Data Quality Assessment", *Communications of the ACM*, v. 45, n. 4ve, pp. 211–218, April 2002.

[15] LEE, Y. W., STRONG, D. M., KAHN, B. K., *et al.*, "AIMQ: a methodology for information quality assessment", *Information and Management*, v. 40, pp. 133–146, December 2002.

[16] KAHN, B. K., STRONG, D. M., WANG, R. Y., "Information Quality Benchmarks: Product and Service Performance", *Communications of the ACM*, v. 45, n. 4ve, pp. 184–192, April 2002.

[17] CABALLERO, I., GóMEZ, S., PIATTINI, M., "Getting Better Information Quality by Assessing and Improving Information Quality Management". In: *Proceedings of the Ninth International Conference of Information Quality (ICIQ)*, 2004.

[18] WANG, R. Y., STRONG, D. M., "Beyond Accuracy: What data quality means to data consumers", *Journal of Management Information Systems*, v. 12, n. 4, 1996.

[19] CHRISMAN, N. R., "The role of Quality Information in the long-term functioning of a GIS". In: *Proceedings of the AUTOCART06*, v. 2, pp. 303–321, 1983.

[20] OLSON, J. E., *Data Quality: the Accuracy Dimension*. Morgan Kaufmann Publishers, 2003.

[21] WANG, R. Y., ZIAD, M., LEE, Y. W., *Data Quality*. Kluwer, 2001.

[22] LEE, Y. W., STRONG, D. M., "Knowing-Why about Data Processes and Data Quality", *Journal of Management Information Systems*, v. 20, n. 3, pp. 13–39, 2004.

[23] Kotz, S. (ed.), *Encyclopedia of Statistical Sciences*, v. 3, Update. 2 ed. John Wiley and Sons, Inc., 1999.

[24] WAND, Y., WANG, R. Y., "Anchoring data quality dimensions in ontological foundations", *Communications of the ACM*, v. 39(11), pp. 86–95, November 1996.

[25] CHUNG, W. Y., FISHER, C., WANG, R., "What Skills Matter In Data Quality? (Research In-Progress)".

[26] DRAVIS, F., "Data Quality Strategy: A Step-by-Step Approach". In: *Proceedings of the Ninth International Conference of Information Quality (ICIQ)*, 2004.

[27] APLIGRAF APLICATIVOS & GRáFICOS, *Introdução à Análise Técnica*, julho 2003.

[28] REIGROTZKI, M., MILEK, J., BOSCH, H., *et al.*, *A Holistic Approach to Data Quality Management*, Report, Predict AG, 2001.

[29] "Centro de Estudos do Genoma Humano", http://genoma.ib.usp.br, acessado em fevereiro de 2007.

[30] CHAPMAN, A., *Principles of Data Quality*, Report, Global Biodiversity Information Facility, 2005.

[31] SPENCER, N., *SAS programming: the one-day course*. Boca Raton, FL, USA, Chapman and Hall/CRC, 2004.

[32] MONTGOMERY, D. C., *Introduction to Statistical Quality Control*. 4 ed. John Wiley & Sons, Inc., 2001.

[33] LATHI, B. P., *Linear Systems and Signals*. 2 ed. Oxford University Press, August 2001.

[34] DORFFNER, G., "Neural Networks for Time Series Processing", *Neural Network World*, v. 6, n. 4, pp. 447–468, 1996.

[35] HAYKIN, S., *Neural Networks - a Comprehensive Foundation*. 2 ed. Prentice-Hall, 1999.

[36] MORETTIN, P. A., TOLOI, C. M. C., *Análise de Séries Temporais*. Edgard Blücher Ltda, 2004.

[37] PEREIRA, B. B., PAIS, M. B. Z., SALES, P. R. D. H., *Análise Espectral de Séries Temporais*. Arte Final Leasing Editorial Ltda., 1986.

[38] DICKEY, D. A., FULLER, W. A., "Distributions of The Estimators For Autoregressive Time Series With a Unit Root", *Journal of the American Statistical Association*, v. 75, pp. 427–431, 1979.

[39] PHILLIPS, P. C. B., "Time Series Regression with a Unit Root", *Econometrica*, v. 55, pp. 277–301, 1987.

[40] PHILLIPS, P. C. B., PERRON, P., "Testing for a Unit Root in Time Series Regression", *Biometrika*, v. 75, pp. 335–346, June 1988.

[41] MEDEIROS, M. C., SOARES, L. J., "Robust Statistical Methods for Electricity Load Forecasting", RTE-VT Workshop, Paris, May 2006.

[42] SOARES, L. J., MEDEIROS, M. C., "Modeling and Forecasting Short-Term Electric Load Demand: A Two-Step Methodology", Pontifícia Universidade Católica do Rio de Janeiro, Textos para Discussão, junho 2006, A sair no International Journal of Forecasting.

[43] RODRIGUES, S. A., DINIZ, C. A. R., "Modelo de Regressão Heteroscedástico", *Revista de Matemática e Estatística*, v. 24, n. 2, pp. 133–146, 2006.

[44] CALÔBA, L. P., "Introdução ao Uso de Redes Neurais na Modelagem de Sistemas Dinâmicos e Séries Temporais", Livro para mini-curso do Congresso Brasileiro de Automática, 2002.

[45] KALMAN, R. E., "A New Approach to Linear Filtering and Prediction Problems", *Transactions of the ASME - Journal of Basic Engineering*, v. 82, pp. 35–45, 1960.

[46] "Yahoo!Finance", http://finance.yahoo.com, acessado em março de 2007.

[47] "Stockwiz", http://www.stockwiz.com, acessado em fevereiro de 2007.

[48] "S&P 500 Index", http://www.standardandpoors.com, acessado em outubro de 2004.

[49] "Economagic: Economic Time Series Page", http://www.economagic.com, acessado em abril de 2007.

[50] "Statistics Canada", http://www.statcan.ca, acessado em março de 2007.

[51] "Time Series Data Library", http://www-personal.buseco.monash.edu.au/ hyndman/TSDL/, acessado em abril de 2007.

[52] "Inter-university Consortium for Political and Social Research", http://www.icpsr.umich.edu, acessado em outubro de 2004.

[53] "Centro de Estudos da Opinião Pública", http://www.cesop.unicamp.br, acessado em outubro de 2004.

[54] ROBINSON, J. P., BIANCHI, S. M., PRESSER, S., "Family Interaction, Social Capital and Trends in Time Use", Inter-University Consortium for Political and Social Research (ICPSR Study N. 3191), 1998-1999, Distribuído por CESOP/Unicamp, Brasil.

[55] ENGLISH, L., "Data Quality: Standardize, Validate and Improve your Information Assets", White Paper: Information Impact International & DataFlux Corporation, 2004.

[56] LOSHIN, D., INBAR, D., "Integration and the data quality imperative: the data quality monitor", White Paper, Data Junction Corporation, 2001.

[57] WANG, R. Y., "A Product Perspective on Total Data Quality Management", *Communications of the ACM*, v. 41, n. 2, pp. 58–65, 1998.

[58] MONTCHEUIL, Y. D., DUPUPET, C., "Third Generation ETL: Delivering the Best Performance", Sunopsis – White Paper.

[59] KUO, B. C., GOLNARAGHI, F., *Automatic Control Systems*. 8th ed. Wiley, August 2002.

[60] ORR, K., "Data quality and the systems theory", *Communications of the ACM*, v. 41, n. 2, pp. 66–71, 1998.

[61] CAPIELLO, C., *Data Quality and Multichannel Services*. Ph.D. dissertation, Politecnico di Milano - Dipartimento di Elettronica e Informazione, 2005.

[62] NAUMANN, F., ROLKER, C., "Assessment methods for information quality criteria". In: *Proceedings of the Conference on Information Quality*, pp. 148–162, 2000.

[63] BALLOU, D. P., WANG, R. Y., PAZER, H. L., *et al.*, "Modelling information manufacturing systems to determine information product quality", *Management Science*, v. 44, n. 4, pp. 462–533, 1998.

[64] BOVEE, M., SRIVASTAVA, R. P., MAK, B., "A Conceptual Framework and belief-function approach to assessing overall information quality". In: *Proceedings of the Sixth International Conference on Information Quality*, 2001.

[65] MOORE, D. S., MCCABE, G. P., *Introduction to the Practice of Statistics*. 3 ed. W. H. Freeman, 1999.

[66] RENZE, J., "Outlier", MathWorld - A Wolfram Web Resource. http://mathworld.wolfram.com/Outlier.html, acessado em março de 2007.

[67] MONTGOMERY, D. C., JOHNSON, L. A., *Forecasting and Time Series Analysis*. McGraw-Hill, Inc., 1976.

[68] SHANMUGAN, K. S., BREIPOHL, A. M., *Random Signals: Detection, Estimation and Data Analysis*. John Wiley and Sons, 1988.

[69] HAYKIN, S., *Modern Filters*. MacMillan, 1989.

[70] POLLOCK, D. S. G., "Lecture Courses on Time Series", World Wide Web: www.qmw.ac.uk/ugte133, acessado em outubro de 2004.

[71] PODDIG, T., REHKUGLER, H., "A "world" model of integrated financial markets using artificial neural networks", *Neurocomputing*, v. 10, pp. 251–273, 1996.

[72] FRANSES, P. H., *Time Series Models for business and economic forecasting*. Cambridge University Press, 1998.

[73] REFENES, A. N., BENTZ, Y., BUNN, D. W., *et al.*, "Financial time series modelling with discounted least squares backpropagation". In: *Neurocomputing*, n. 14, Elsevier, pp. 123–138, 1997.

[74] HSIEH, W. W., "Nonlinear multivariate and time series analysis by neural network methods", *Review of Geophysics*, v. 42, March 2004.

[75] DANTAS, A. C. H., SEIXAS, J. M. D., "An Adaptive Neural System for Financial Time Series Tracking". In: Ribeiro, B., Albrecht, R. F., Dobnikar, A., *et al.* (eds.), *Adaptive and Natural Computing Algorithms*, Springer, 2005. Proceedings of ICANNGA: International Conference on Adaptive and Natural Computing Algorithms. Coimbra, Portugal, 2005.

[76] DANTAS, A. C. H., SEIXAS, J. M. D., "Neural Networks for Data Quality Monitoring of Time Series". In: *9th International Conference on Enterprise Information Systems*, 2007.

[77] DANTAS, A. C. H., SEIXAS, J. M. D., DINIZ, F. B., *et al.*, "A Statistical and Signal Processing Based System for Data Quality Management". In: Zanasi, A., Ebecken, N. F. F., Brebbia, C. A. (eds.), *Data Mining V: Data Mining, Text Mining and their Business Applications*, v. 10, *WIT Transactions on Information and Communication Technologies*, WITPRESS, pp. 209–218, 2004.

[78] KAASTRA, I., BOYD, M., "Designing a neural network for forecasting financial and economic time series", *Neurocomputing*, v. 10, pp. 215–236, 1996.

[79] ALEKSEEV, K. C. P. G., *Previsão da demanda de passageiros no transporte aéreo doméstico regular: uma modelagem neuronal*. Ph.D. dissertation, Programa de Engenharia de Produção, COPPE/UFRJ, agosto 2002.

[80] MEDEIROS, M. C., TERÄSVIRTA, T., RECH, G., "Building Neural Network Time Series Models: A Statistical Approach", *Journal of Forecasting*, v. 25, pp. 49–75, 2006.

[81] BROWNSTONE, D., "Using percentage accuracy to measure neural network predictions in Stock Market movements", *Neurocomputing*, pp. 237–250, 1996.

[82] SITTE, R., SITTE, J., "Neural Networks Approach to the Random Walk Dilemma of Financial Time Series", *Applied Intelligence*, v. 16, pp. 163–171, 2002.

[83] KOHAVI, R., "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, v. 12, pp. 1137–1143, 1995.

[84] ELMAN, J. L., "Finding Structure in Time", *Cognitive Science*, v. 14, pp. 179–211, 1990.

[85] JIN, L., NIKIFORUK, P. N., GUPTA, M. M., "Absolute Stability Conditions for Discrete-Time Recurrent Neural Networks", *IEEE Transactions on Neural Networks*, v. 5, n. 6, pp. 954–964, November 1994.

[86] ABARBANEL, H. D. I., BROWN, R., SIDOROWICH, J. J., *et al.*, "The analysis of observed chaotic data in physical systems", *Review of Modern Physics*, v. 65, n. 4, pp. 1331–1392, October 1993.

[87] JAEGER, H., *The "echo state" approach to analysing and training recurrent neural networks*, Report 148, German National Research Center for Information Technology, 2001.

[88] JAEGER, H., HAAS, H., "Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication", *Science*, v. 304, April 2004.

[89] JAEGER, H., *A tutorial on training recurrent neural networks, covering BPTT, RTRL, EKF and the "echo state network" approach*, Report 159, German National Research Center for Information Technology, 2005. 2nd revision.

[90] JAEGER, H., "Reservoir Riddles: Suggestions for Echo State Network Research (Extended Abstract)". In: *Proceedings of International Joint Conference on Neural Networks, Montreal, Canada*, 2005.

[91] MAHALANOBIS, P. C., "On the generalised distance in statistics". In: *Proceedings of the National Institute of Science of India*, 1936.

[92] KULLBACK, S., LEIBLER, R. A., "On Information and Sufficiency", *Annals of Mathematical Statistics*, v. 22, pp. 79–86, 1951.

[93] RUBIN, D. B., "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys", *Journal of the American Statistical Association*, v. 72, pp. 538–543, 1977.

[94] DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B., "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society. Series B (Methodological)*, v. 39, n. 1, pp. 1–38, 1977.

[95] MACKAY, D. J. C., *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[96] DELLAERT, F., *The Expectation Maximization Algorithm*, Report GIT-GVU-02-20, College of Computing, Georgia Institute of Technology, February 2002.

[97] SCHNEIDER, T., "Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values", *Journal of Climate*, v. 14, pp. 853–871, 2001.

[98] LITTLE, R. J. A., RUBIN, D. B., *Statistical Analysis with Missing Data*. Wiley Series in Probability and Mathematical Statistics, 1987.

[99] HANSEN, P. C., "Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion", SIAM Monographs on Mathematical Modeling and Computation, 1997.

[100] SCHAFER, J. L., *Analysis of Incomplete Multivariate Data*, n. 72 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 1997.

[101] ALMEIDA, P. D. D., *Previsão do Comportamento de Séries Temporais Financeiras com Apoio de Conhecimento Sobre o Domínio*. Ph.D. dissertation, Universidade da Beira Interior, abril 2003.

[102] FAMA, E. F., "Efficient Capital Markets: A Review of Theory and Empirical Work", *Journal of Finance*, v. 25, n. 2, pp. 383–417, 1970.

[103] CONWAY, A. J., MACPHERSON, K. P., BROWN, J. C., "Delayed Time Series Prediction with Neural Networks", *Neurocomputing*, v. 18, 1998.

[104] PAVLIDIS, N., TASOULIS, D., VRAHATIS, M., "Financial forecasting through unsupervised clustering and evolutionary trained neural networks", 2003.

[105] KONDRATENKO, V. V., KUPERIN, Y. A., "Using Recurrent Neural Networks To Forecasting of Forex", *ArXiv Condensed Matter e-prints*, , April 2003.

[106] HAMILTON, J. D., *Time Series Analysis*. Princeton University Press, 1994.

[107] SADORSKY, P., "Volatility Forecasting and Value at Risk". In: Hamza, M. (ed.), *Proceedings of Financial Engineering and Applications*, 2003.

[108] PLEROU, V., GOPIKRISHNAN, P., ROSENOW, B., *et al.*, "Econophysics: Financial Time Series From a Statistical Physics Point of View", *Physica A*, v. 279, pp. 443–456, 2000.

[109] GOPIKRISHNAN, P., PLEROU, V., Y. LIU, L. A., *et al.*, "Scaling And Correlation In Financial Time Series", *Physica A*, v. 287, pp. 362–373, 2000.

[110] PRÍNCIPE, J. C., RATHIE, A., KUO, J. M., "Prediction of chaotic time series with neural networks and the issue of dynamic modeling", *Int. J. of Bifurcation and Chaos*, v. 2, n. 4, pp. 989–996, 1992.

[111] LORENZ, E. N., "Deterministic Nonperiodic Flow", *Journal of the Atmospheric Sciences*, v. 20, n. 2, pp. 130–141, March 1963.

[112] DINIZ, F. C. C. B., "Detecção de Sinais usando-se Filtros Casados a partir de dados sobre Hepatite A", Relatório de Trabalho para a Disciplina "Detecção e Estimação de Sinais" (PEE-COPPE/UFRJ), outubro 2003.

[113] TANNER, M. A., WONG, W. H., "The calculation of posterior distributions by data augmentation (with discussion)", *Journal of the American Statistical Association*, v. 82, pp. 528–550, 1987.

[114] HYVÄRINEN, A., KARHUNEN, J., OJA, E., *Independent Component Analysis*. John Wiley & Sons, Inc., 2001.

[115] GÓRRIZ, J. M., PUNTONET, C. G., SALMERÓN, M., *et al.*, "Time Series Prediction Using ICA Algorithms". In: *IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, pp. 226–230, September 2003.

[116] DONOHO, D. L., "Nonlinear Wavelet Methods for Recovery of Signals, Densities, and Spectra from Indirect and Noisy Data". In: *Proceedings of Symposia in Applied Mathematics*, v. 00, pp. 173–204, 1993.

[117] STRUZIK, Z. R., "Wavelet methods in (financial) time-series". In: *Physica A*, v. 296, Elsevier, pp. 307–319, 2001.