ANÁLISE DE COMPONENTES INDEPENDENTES PARA A MONITORAÇÃO DA QUALIDADE DE DADOS EM SÉRIES TEMPORAIS

José Márcio Faier

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Elétrica.

Orientador: José Manoel de Seixas

Rio de Janeiro
Junho de 2011

ANÁLISE DE COMPONENTES INDEPENDENTES PARA A MONITORAÇÃO DA QUALIDADE DE DADOS EM SÉRIES TEMPORAIS

José Márcio Faier

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. José Manoel de Seixas, D.Sc.

Prof. Luiz Pereira Calôba, Dr. Ing.

Prof. Mariane Rembold Petraglia, Ph.D.

Prof. André Carlos Ponce de Leon Ferreira de Carvalho, D.Sc.

Prof. Guilherme de Alencar Barreto, D. Sc.

Prof. Vitor Hugo Ferreira, D. Sc.

RIO DE JANEIRO, RJ – BRASIL
JUNHO DE 2011

Faier, José Márcio

Análise de Componentes Independentes para a Monitoração da Qualidade de Dados em Séries Temporais/José Márcio Faier. – Rio de Janeiro: UFRJ/COPPE, 2011.

XVI, 153 p.: il.; 29, 7cm.

Orientador: José Manoel de Seixas

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2011.

Referências Bibliográficas: p. 140 – 150.

1. Qualidade de Dados. 2. Análise de Componentes Independentes. 3. Séries Temporais. I. Seixas, José Manoel de. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Agradecimentos

Primeiramente, eu gostaria de agradecer à minha família por ter me dado suporte todo este tempo, com um agradecimento especial aos meus pais, que me apoiaram incondicionalmente. Um obrigado mais especial ainda para a minha mãe, que sempre deu suporte em outras fases da minha jornada de estudante e que me permitiram chegar até aqui. Agradeço também aos pais de minha namorada, Dona Margarida e Seu Fábio, que tiveram sempre presentes enquanto eu desenvolvia os trabalhos durante os finais de semana. Um obrigado especial pela grande contribuição de Seu Fábio, que sempre oferecia uma cerveja para que meu o pensamento fluísse melhor. Um obrigado com paixão para a minha namorada Mônica, que soube entender a importância deste trabalho, abrindo mão de muitos domingos e feriados e sempre esteve ao meu lado durante todo este tempo. Agradeço ao meu orientador Seixas, que desde o mestrado, transmite o seu conhecimento e me apoia em atividades para o meu crescimento profissional e pessoal. Uma amizade consolidada durante todo este tempo. Agradeço a Augusto Dantas pelos conhecimentos relevantes que ajudaram a construir esta tese e a Frank Block pelo apoio nas atividades profissionais que levam o conhecimento deste trabalho além das fronteiras acadêmicas. Agradeço a Roberto Quinet pela longa estadia que tive no Rio de Janeiro para o desenvolvimento desta tese. Agradeço a todos colegas do LPS que, direta ou indiretamente tiveram alguma participação no meu trabalho. Um obrigado especial aos colegas que tive mais contato durante este tempo: Natanael Moura, um exemplo de vida, seu filho Júnior, Fernando Ferreira, Vitor Cascão, Rodrigo Torres, Carmen Maidantchik, Danton Ferreira, Eduardo Simas, Andressa Sivolella, Jorge Magalhães, Tadeu Ferreira e Danilo. Agradeço a todos os meus amigos Levy, Frank, Moisés, Daniel, Felipe, Genário, Marco Aurélio e Antônio, pelos bons momentos que me faziam esquecer um pouco desta tese, para depois voltar com força total. Por fim, agradeço a todos que fizeram parte da minha vida durante este tempo e que de alguma forma contribuíram para este trabalho.

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.)

ANÁLISE DE COMPONENTES INDEPENDENTES PARA A MONITORAÇÃO DA QUALIDADE DE DADOS EM SÉRIES TEMPORAIS

José Márcio Faier

Junho/2011

Orientador: José Manoel de Seixas

Programa: Engenharia Elétrica

Nesta tese, desenvolve-se um sistema de monitoração da qualidade de dados (QD) para séries temporais, utilizando-se a Análise de Componentes Independentes (ICA). As fontes são extraídas de forma cega e concentram estruturas que permitem estimar padrões de qualidade de dados mais acurados. A ICA é introduzida na cadeia de pré-processamento, na qual se extraem heterocedasticidades, tendências, ciclos e sazonalidades. O resíduo é estimado através de modelos lineares ou não lineares com redes neurais. A partir do erro de estimação, constroem-se corredores de validação dinâmicos que se adaptam às variações estatísticas da série. O corredor pode ser ajustado para estabelecer um compromisso com o usuário e o contexto no qual os dados estão inseridos.

Esta metodologia é desenvolvida e avaliada com séries sintéticas, em condições controladas (não cega), e com séries reais, no contexto de separação cega de fontes. Utilizam-se séries temporais de carga elétrica e séries do mercado financeiro de ações. Os resultados mostram que a metodologia facilita o processamento e produz corredores de validação mais acurados. Os corredores permitem a medição mais precisa da QD, através de indicadores da qualidade, e aumentam a eficiência da monitoração, detectando melhor os problemas e emitindo menos falsos alarmes.

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

INDEPENDENT COMPONENT ANALYSIS FOR DATA QUALITY MONITORING IN TIME SERIES

José Márcio Faier

June/2011

Advisor: José Manoel de Seixas

Department: Electrical Engineering

In this thesis, we propose a time-series data quality monitoring system, using Independent Component Analysis (ICA). The components are obtained in a blind source separation context and concentrate structures which produce accurate data quality patterns. The ICA is used in the preprocessing chain to remove heteroscedasticity, tendency, cycles and seasonality. The residue is estimated through linear or non linear neural networks. From the estimation error, dynamic validation corridors are constructed and adapt to statistical uncertainties. The corridor can be adjusted to include the user and the data context.

The methodology is developed and evaluated for synthetic time-series, in controlled conditions (not blind), and actual time-series, in blind source separation context. Electrical load and stock market time series are used. The methodology improves the processing chain and provide accurate validation corridors. The corridors produce more accurate data quality measurement, through quality indicators, and improve the monitoring performance, detecting problems and reducing false alarms.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Durante os últimos anos, o desenvolvimento mundial tem ocorrido principalmente pela disseminação de conhecimento. Se a informação é considerada a moeda da nova economia, então os dados podem ser entendidos como a matéria-prima dessa economia do século XXI. De fato, devido em grande parte à globalização do conhecimento com a utilização crescente da Internet, deu-se início a uma nova era tecnológica: a "era da informação". Nesse novo mundo, é fundamental conhecer a qualidade dos dados no seu mais amplo sentido; pois, problemas nos dados impactam diretamente no desempenho e competitividade das organizações. Assim, caso não sejam detectados e corrigidos no momento certo, os erros podem propagar-se por todo o "ciclo de vida" da informação e culminar em grandes perdas.

A qualidade de dados (QD) não pode ser entendida apenas do ponto de vista técnico ou apenas como dados livres de erros. A qualidade tem diversas outras dimensões, tais como a interpretabilidade, atualização, valor agregado, forma de representação e credibilidade, dentre outras características. Além disso, o aspecto humano e o contexto nos quais os dados estão inseridos complementam as perspectivas para se abordar o tema. Assim, a QD deve ser definida, medida, analisada, melhorada e monitorada, de acordo com os dados, o contexto e as especificações dos usuários.

Neste trabalho, a ênfase é na monitoração da QD em séries temporais. A monitoração é fundamental para detectar erros precocemente e corrigi-los antes que se propaguem e gerem prejuízos. Para monitorar dimensões-chave nessa classe de dados, propõe-se avaliar a qualidade no início do "ciclo de vida" da informação. A possibilidade de correção deve ser sugerida aos usuários, que decidem se a ação de correção deve ser realizada.

Para detectar apenas os problemas de fato, propõe-se a construção de corredores de validação dinâmicos introduzindo a Análise de Componentes Independentes (*Independent Component Analysis* - ICA) na cadeia de processamento. A técnica proposta extrai fontes independentes de forma cega. Sob estas fontes, um pré-

processamento clássico é aplicado para remover componentes frequentemente encontrados em séries temporais, como heterocedasticidade, tendências, ciclos e sazonalidades, e facilitar a estimação por meio de modelos mais adequados, como redes neurais ou mesmo métodos lineares. A incerteza gerada pelos modelos é utilizada para estimar os limites do corredor e pode ser ajustada para incluir o contexto e as necessidades do usuário.

1.1 Contexto

Um dos desafios desse novo século é a habilidade em lidar com bases gigantescas de dados e informações de diversas naturezas. Com o surgimento de novas fontes de informação e a facilidade de difusão, a quantidade de dados cresce a cada dia com velocidade cada vez maior. Ainda, a necessidade de armazenar o aspecto temporal dos dados contribui substancialmente para este crescimento. Com o aumento no volume de dados e a necessidade de integrar fontes de diversas naturezas, os problemas de QD tornam-se cada vez mais evidentes. Por outro lado, a competitividade entre as empresas tolera cada vez menos qualquer tipo de erro e exige que os processos sejam cada vez mais eficientes.

As séries temporais podem conter padrões fundamentais e não podem estar corrompidas por problemas. A perda de qualidade nesses dados pode impactar fortemente nos processos de tomada de decisão e na gerência da informação de diversos setores. Nesta tese, a QD é abordada no contexto dos setores elétrico e financeiro. No setor de geração, transmissão e distribuição de energia elétrica há a necessidade da oferta de energia acompanhar as variações da demanda ao longo do tempo, não podendo haver risco de "apagão". Se por um lado a demanda exige um serviço de alta qualidade, por outro lado os custos envolvidos para se atender esta expectativa são elevados. Sendo assim, a eficiência é fundamental e decisões corretas são cada vez mais críticas para o sistema como um todo. No setor financeiro, problemas de qualidade nos dados podem acarretar decisões erradas e causar redução dos lucros esperados ou mesmo prejuízos aos investidores.

1.2 Motivação

Apesar da evidente importância da qualidade dos dados, o tema ainda é pouco explorado. Em contextos como o de séries temporais, a qualidade de dados é muitas vezes tratada *ad hoc* e de forma secundária. De fato, uma grande parte do tempo que deveria ser dedicado à aplicação de interesse é gasta na correção de problemas nos dados. De maneira geral, o que se nota é que a QD não é medida e tampouco monitorada de forma sistemática e contínua.

Além de haver espaços inexplorados, outra motivação está na abordagem do tema sob o ponto de vista da inteligência computacional e do processamento de sinais. Em diferentes áreas de aplicação, bons resultados têm sido obtidos em comparação com abordagens tradicionais. Especificamente, a motivação para utilizar ICA está na capacidade da técnica de acessar estruturas ocultas nos dados e, dessa forma, tornarem evidentes os padrões contidos nas séries temporais. Isso facilita o processo de modelagem e a medição mais fidedigna da QD, o que torna a monitoração mais eficiente.

1.3 Contribuições

O trabalho aqui desenvolvido contribui com a especificação e o desenvolvimento de um sistema de monitoração da qualidade de dados para séries temporais. O sistema parte de estudos conceituais iniciados em [1] e propõe novos conceitos, técnicas, métodos e perspectivas para a monitoração da QD neste tipo de dado. A ICA é introduzida na etapa inicial do sistema e impacta em toda a cadeia de processamento. Um conjunto de algoritmos de ICA é proposto para extrair as fontes independentes e os métodos clássicos de pré-processamento são integrados para o tratamento de fontes múltiplas. Para a prospecção das fontes, disponibiliza-se um conjunto de testes e ferramentas de análise. Propõem-se também testes de hipóteses para a seleção automática de parâmetros e modelos (lineares ou não-lineares). Além disso, o sistema permite o tratamento automático do ruído presente nas séries. Após a transformação do espaço ICA para o espaço original, os padrões de QD são utilizados para a construção de corredores de validação. Propôs-se a utilização dos corredores tanto para a monitoração quanto para a formação de indicadores de QD. Assim, dada a quantificação dos problemas, torna-se possível a gestão da QD no contexto de séries temporais. O sistema é testado em condições reais no setor elétrico e financeiro e está disponível para a utilização em diversas outras séries.

1.4 Organização

Além dessa introdução, este texto é organizado em outros sete capítulos. Os dois capítulos seguintes abordam a teoria utilizada na elaboração desta tese. No Capítulo 4, abordam-se a metodologia para a monitoração da QD em séries temporais e os conceitos aplicados. Os resultados obtidos com séries sintéticas são mostrados no Capítulos 5 e, nos Capítulos 6 e 7, os resultados para séries reais. Por fim, no último capítulo, são realizadas as análises e conclusões.

O Capítulo 2 aborda a teoria da qualidade de dados. O tema é inicialmente abordado de forma geral e, na sequência, abordado no contexto de séries temporais.

O capítulo mostra o estado-da-arte da qualidade de dados e apresenta definições e aspectos relevantes, assim como dimensões de qualidade, ambiguidades do tema, prevenção e correção dos problemas, modelagem e medição da qualidade de dados.

O Capítulo 3 trata da teoria de Análise de Componentes Independentes. Inicialmente é dada a definição de ICA e são mostrados os princípios de separação de fontes para ICA básica e ICA para séries temporais. Em seguida, são apresentados algoritmos ICA em contexto de estatística de ordem superior e estatística de segunda ordem. Por fim, são mostradas aplicações recentes de ICA em outros contextos e no contexto de séries temporais.

No Capítulo 4, os conceitos de QD incorporados à metodologia proposta são classificados e organizados. São mostradas os problemas buscados, as dimensões avaliadas e os métodos utilizados. A estrutura do Sistema de Monitoração da Qualidade de Dados para Séries Temporais (SMQD-ST) é mostrada sob forma de um diagrama e explicada na sequência. Por fim, são apresentados os métodos de estudo da QD nas séries temporais. Apresentam-se os critérios de escolha dos contextos utilizados para o estudo, os critérios de escolha da configuração das séries e o método de divisão dos conjuntos de desenvolvimento e teste dos modelos de QD.

No Capítulo 5, são apresentados os resultados com séries sintéticas. Inicialmente, mostra-se o processo de criação das séries sintéticas e das fontes quem compõem estas séries. Em seguida, avalia-se o processo de separação cega em diversos ambientes, assim como ambientes ruidosos e com *outliers*. Nesta análise, diversos algoritmos ICA são testados e avaliados sob diversas perspectivas. Os resultados são mostrados e avaliados em todas as etapas da metodologia, desde o pré-processamento, passando pela modelagem até a medição da QD. Nestas etapas, compara-se o método proposto com o cenário de separação ideal e o cenário sem ICA.

No Capítulo 6, são apresentados os estudos com séries temporais reais de carga elétrica. A mesma metodologia aplicada para séries sintéticas é utilizada para estas séries. A abordagem comparativa é feita apenas para o caso com e sem a aplicação de ICA, dado que não se dispõe das fontes originais. Os resultados obtidos são mostrados para duas configurações dos dados: série de pico diário de consumo de carga elétrica e séries de consumo de carga em horários adjacentes ao horário de pico.

No capítulo 7, são apresentados os resultados para as séries financeiras. São utilizadas séries de preços de ações de companhias de um mesmo setor econômico. Para estas séries, avaliam-se duas configurações distintas: com séries explicativas do mesmo segmento de mercado e com séries múltiplas de uma mesma companhia. Em ambas as configurações são avaliados os impactos da introdução de ICA na monitoração da qualidade de dados em todas as etapas da metodologia proposta.

No capitulo final, os resultados são analisados e a tese é concluída. São mostra-

dos direções futuras para a continuidade da pesquisa. Nas seções complementares, são mostrados os trabalhos publicados durante o desenvolvimento desta tese e a referencia bibliográfica utilizada.

Capítulo 2

Qualidade de Dados

A qualidade de dados (QD) é o principal foco deste trabalho. Neste capítulo, será feita uma revisão bibliográfica e discutiremos conceitos, definições, ambiguidades, o estado-da-arte da pesquisa e alguns outros aspectos importantes relacionados ao tema.

2.1 Visão Geral

O tema qualidade de dados é fundamental para uma ampla faixa de áreas do conhecimento e, desde o início da década de 90, o interesse pelo assunto tem crescido substancialmente. Pode-se fazer uma relação desse interesse com o crescimento do uso da Internet e outras ferramentas da Tecnologia da Informação (TI) que, nos últimos anos, vem possibilitando uma troca de dados jamais vista. Este aumento no volume de informações faz com que problemas na qualidade dos dados se tornem cada vez mais evidentes e surjam iniciativas para tratá-los de forma adequada. Apesar da grande importância do tema, a pesquisa em QD tem áreas ainda pouco exploradas, onde há carência de consenso e de padrões bem definidos [2]. A começar pelo próprio nome do tema de pesquisa, muitos pesquisadores e praticantes utilizam o termo Qualidade da Informação (QI) [2] [3] e Qualidade de Dados [4] indistintamente. De fato, atualmente, as palavras "dado" e "informação" têm sido utilizadas como sinônimas. Assim, é de se esperar que os termos QD e QI sigam a mesma tendência.

Rigorosamente, podemos distinguir ambas as palavras "dado" e "informação". Dados são resultado de um "conjunto de medidas, caracteres ou símbolos" [5] e podem ser entendidos como o mais baixo nível de abstração do qual a informação é derivada. Já a informação pode ser definida como "dados processados" que transmitem uma mensagem. Assim, podemos também diferenciar a QD da QI. A distinção entre os termos é comparada com as diferenças entre sintaxe e semântica [6]. A semântica refere-se ao estudo do significado enquanto a sintaxe refere-se a estrutu-

ras ou padrões formais do modo como algo é expresso [7]. Por exemplo, o valor semântico de "um" pode ser expresso em diferentes sintaxes, como 001, 1,0 ou 1. Assim, a QD está relacionada a estruturas ou padrões encontrados nos dados enquanto a QI relaciona-se ao significado destes dados, ou seja, tem um compromisso maior com o contexto e às necessidades do usuário da informação.

Seguindo a tendência atual, no presente trabalho, não se faz uma distinção rígida entre as definições de dados, informação e, da mesma forma, QD e QI. Ainda, devido à tendência deste trabalho em abstrair fatores subjetivos das análises, optaremos por utilizar o termo "qualidade de dados" ou QD.

Assim, notam-se duas perspectivas na definição de QD: dos usuários e dos dados. Sob a perspectiva dos dados, o termo qualidade tem sido definido como uma conformidade com as especificações ou uma "conformidade ao uso" [8] e tem sido amplamente utilizado [9]. Há autores [10] que argumentam que os usuários não são tão capazes de encontrar erros na informação e alterar a maneira como eles usam a informação. Assim, a partir da perspectiva dos dados, QD pode ser definida como a informação em conformidade com as especificações e requisitos [10]. A partir da perspectiva do usuário, QD tem sido definida como uma "conformidade ao uso pelos usuários da informação" [9]. Os autores argumentam que são os usuários que definem se a informação ajusta-se ao uso ou não e, portanto, são eles que definem o que é a qualidade.

2.2 Estado da Arte

Apesar da aparente falta de consenso e padronização na área, já é possível identificar nichos de pesquisas [11] (veja a Figura 2.1). Assim, pode-se classificar a pesquisa em QD sob três aspectos: avaliação, gestão e contexto da QD, que serão abordados nas seções seguintes.

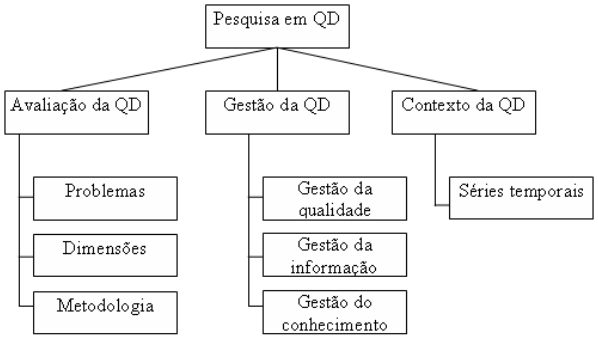

Figura 2.1: Pesquisa em Qualidade de Dados. Adaptada de [11].

2.2.1 Avaliação da QD

Neste nicho de pesquisa, a avaliação da QD é definida como o processo de determinar valores numéricos ou categorias para as dimensões da QD e pode ser concentrado sob três aspectos fundamentais: problemas de QD (representam a definição de métricas), dimensões (representam as características da informação) e metodologia (modelos, panoramas e métodos para integrar as diversas dimensões às métricas) – veja Figura 2.2.

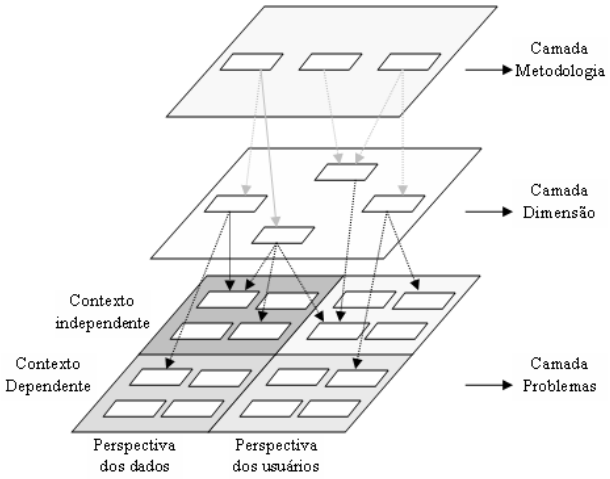

Figura 2.2: Avaliação da Qualidade de Dados. Extraída de [11].

Problemas de QD

Uma grande quantidade de trabalhos tem contribuído para identificação de problemas e definição de métricas de QD. Em [11], os problemas reportados na literatura são classificados segundo a perspectiva dos dados e dos usuários e sob a dependência ou independência do contexto. Sob estas óticas, o autor reporta que a escolha da métrica é uma das partes mais complicadas do estudo, pois a forma de se perceber a qualidade pode variar de um contexto para outro, ou de um usuário para outro.

No exemplo da Tabela 2.1, em uma análise independente do contexto, sob a ótica dos dados, a falta de dados poderia ser relacionada à completude da base de dados. Sob a perspectiva dos usuários, o problema poderia ser visto como uma impossibilidade dos usuários acessarem os dados, refletindo na dimensão acessibilidade. Sob a perspectiva contexto-dependência e do ponto de vista dos dados, o problema poderia significar dados que violem as regras do negócio, refletindo na dimensão da acurácia. Por fim, sob a perspectiva do usuário, o problema poderia estar relacionado a dados irrelevantes para o trabalho em questão, refletindo na dimensão da relevância.

Tabela 2.1: Problema de QD sob várias perspectivas. Adaptada de [11]

Problema: falta de dados		
	Perspectiva dos dados	Perspectiva dos usuários
Independência do contexto	*base de dados incompleta (Completude)*	*informação inacessível para os usuários (Acessibilidade)*
Dependência do contexto	*dados violam as regras do negócio (Acurácia)*	*Informação irrelevante (Relevância)*

Dimensões da QD

Outra categoria identificada na avaliação da QD é o estudo das dimensões da qualidade dos dados. Vários trabalhos confirmam que QD tem características de multi-dimensionalidade [1],[9],[12] e, nas últimas décadas, diferentes conjuntos de dimensões têm sido identificados. Para a identificação de tais dimensões, em [9], três abordagens são propostas: intuitiva, teórica e empírica. A abordagem intuitiva define as dimensões a partir da perspectiva dos dados. Por exemplo, em [1], a completude é definida de forma objetiva como a medida da quantidade de dados faltantes e da representatividade das dimensões do banco de dados (BD). A abordagem teórica define as dimensões a partir da perspectiva de uma situação ideal. Em [13], completude é definida como a habilidade do sistema de informação em representar o um estado real. Já a abordagem empírica define as dimensões a partir da perspectiva do usuário. Por exemplo, em [9], a completude é definida subjetivamente como sendo o nível para qual os dados são suficientes para a tarefa dos usuários.

Com base na identificação e definição de dimensões da QD, pesquisadores têm proposto diferentes tipos de abordagens para classificá-las. Em [11], é apresentado um quadro com as classificações de diversos autores (veja a Figura 2.3). Em [14], as dimensões foram classificadas em intrínseca, contextual, representacional e de acessibilidade. Em [15], classificou-se as dimensões como sendo relacionadas aos dados ou ao sistema e em internas e externas. Em [16], classificou-se em sintática, semântica e representacional. Em [17], agrupou-se em dimensões de acessibilidade, interpretabilidade, relevância e integridade. Em [18], as dimensões foram classificadas como utilizáveis, úteis, dependentes e saudáveis.

Muitas vezes, o que se observa nas diversas classificações é que uma mesma dimensão pode ser classificada de forma distinta e, dependendo da perspectiva, até ambígua. Por exemplo, em [19], a completude é classificada como sendo uma dimensão objetiva e, em [9], como uma dimensão contextual e, portanto, com certo grau de subjetividade.

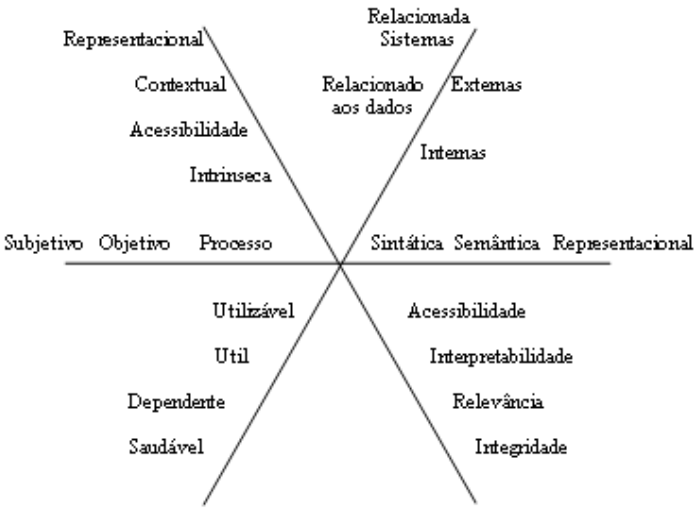

Figura 2.3: Classificações da QD. Extraída de [11].

Metodologia de QD

Outros grupos de trabalhos abordam as metodologias de avaliação da qualidade dos dados. Em [20], as metodologias de avaliação da QD foram categorizadas em objetiva e subjetiva. A avaliação objetiva revela problemas no conjunto de dados e a avaliação subjetiva reflete a necessidade e experiência dos usuários dos dados. Em [11], é apresentado um quadro para indicar as diferenças entre a avaliação objetiva e subjetiva da qualidade (veja a Tabela 2.2).

Tabela 2.2: Diferenças entre a avaliação objetiva e subjetiva da QD. Extraída de [11]

	Objetivo	Subjetivo
Ferramenta	*Software*	*Questionário*
Alvo da medida	*Dado*	*O que a informação representa*
Padrão da medida	*Padrões, regras*	*Satisfação do usuário*
Processo	*Automação*	*Usuário envolvido*
Resultado	*Único*	*Múltiplo*
Local armazenada	*Banco de dados*	*Contexto do negócio*

2.2.2 Gestão da QD

A qualidade de dados abordada sob o ponto de vista de gestão pode ser descrita em três grandes grupos: gestão da qualidade, gestão da informação e gestão do conhecimento. Sob a perspectiva da gestão da qualidade, conceitos como o da Gestão Total da Qualidade de Dados (TDQM) [21] buscam nortear o gerenciamento da qualidade dos dados como o gerenciamento da qualidade de um produto manufaturado. Sob a

perspectiva da gestão da informação, princípios como a integração de dados e contextualização são fundamentais para transformar dados em informações relevantes. Sob a ótica da gestão do conhecimento, princípios como "know-what", "know-how" e "know-why" são abordados com o objetivo de melhora da qualidade da informação e tornar explícito o conhecimento para criar conhecimento organizacional.

2.2.3 Contexto da QD em Séries Temporais

Devido à grande especificidade que o tema pode ter em determinados contextos, a aplicação das técnicas acaba seguindo a sua lógica própria na definição da qualidade de dados. Assim, falar de todos os contextos nos quais a QD está inserida seria uma tarefa quase impossível. Neste trabalho, vamos focar no contexto das séries temporais [22], que são tipos de dados de grande importância para diversas áreas e tema alvo deste trabalho.

Apesar de reconhecer que a análise temporal dos dados é de suma importância para a análise de qualidade dos dados, o estudo das séries temporais neste contexto ainda é pouco explorado [23]. Em uma busca por palavras-chave no contexto da QD, realizado em [2], a expressão "*time-series*" é referenciada apenas em dois trabalhos. No contexto de séries temporais, em [23], é reportada a necessidade da melhoria da qualidade dos dados nas organizações e a necessidade de gerenciar a informação temporal. O autor defende a medição do tempo de duração dos processos e a modelagem da sequência temporal. O autor conclui que modelos temporais levam a representações mais consistentes, tanto dos dados quanto da própria medida de qualidade.

Métodos de QD

Em [4], é apresentado um método geral para medir a QD através da realização de testes de qualidade. O dado testado é comparado a uma informação de referência. O processo genérico é mostrado na Figura 2.4.

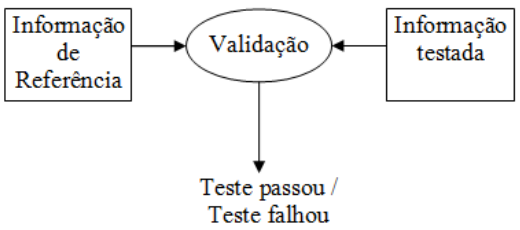

Figura 2.4: Testes de QD. Extraída de [4].

São reportados dois tipos de informação de referência:

- Meta-informação: independente dos dados, mas com relações rígidas sobre eles. Por exemplo, o formato dos campos de uma tabela [24].

- Modelos estatísticos: obtidos a partir de dados livres de erros e tendo a forma de relações aproximadas com os dados como, por exemplo, média e desvio-padrão [23].

Para o caso de modelos estatísticos, técnicas de modelagem com variáveis simples e múltiplas são utilizadas para monitoração da QD. Em [4], sob variáveis únicas, por meio do conhecimento a priori da Função Densidade de Probabilidade, classificam-se as amostras como corretas (prováveis) ou incorretas (improváveis) - veja a Figura 2.5. Em [1], [25], apresentam-se conceitos de monitoração a partir de predições com redes neurais e, em [24], é reportada a utilização de técnicas de reconhecimento de padrões e clusterização. Na Tabela 2.3, são reportado algumas técnicas que modelam padrões de QD de acordo com padrões de QD obtidos no espaço e no tempo.

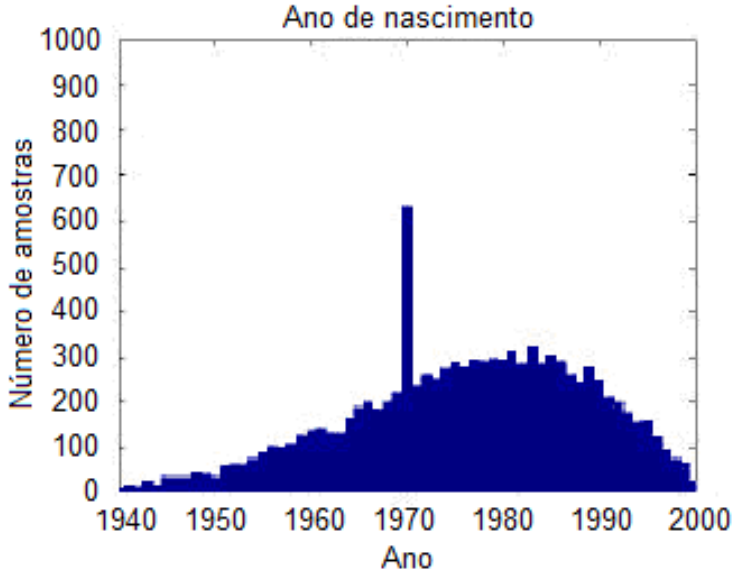

Figura 2.5: Histograma unidimensional. Extraída de [4].

Tabela 2.3: Técnicas utilizadas na monitoração da QD.

Tipo de correlação	Técnicas utilizadas
Temporal	*Redes Neurais [25] e modelos AR e ARIMA [26]*
Espacial	*Correlação estatística [23] e Análise de Componentes Principais [4]*
Espacial e Temporal	*Modelos de séries temporais multivariadas [1]*

Para a correção dos dados, em [25],[26], a previsão temporal das séries é utilizada para preencher dados faltantes e substituir dados suspeitos de erro. Em [27], os

dados falhos são substituídos por estatísticas simples, tais como média e mediana e o modelo deve ser obtido da parte limpa da série.

Métodos de Monitoração da QD

O processo de monitoração consiste em medir a qualidade, acompanhar a sua evolução e corrigir os problemas, caso necessário/solicitado. Em [28], são abordadas dois processos básicos de monitoração. Na primeira abordagem, um gerador de resíduo, assim como os métodos mostrados na Tabela 2.3, é estimado a partir de dados livres de erro (veja Figura 2.6). Então, o gerador é utilizado para testar novos dados. Um alarme é emitido se o resíduo exceder determinados patamares.

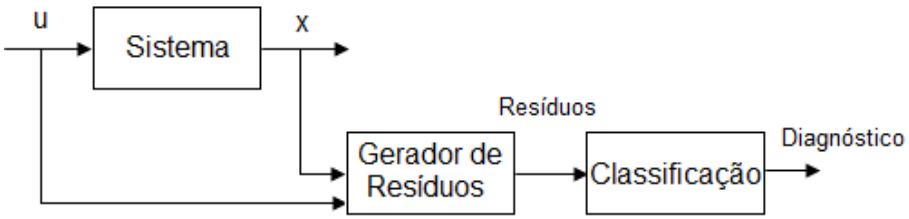

Figura 2.6: Monitoração da QD através de resíduos extraídos dos dados. Extraída de [4]

Na segunda abordagem, não há modelo fixo. Em vez disso, um modelo paramétrico é constantemente estimado a partir dos dados. A monitoração é realizada por meio de testes na variação dos parâmetros do modelo (veja Figura 2.7).

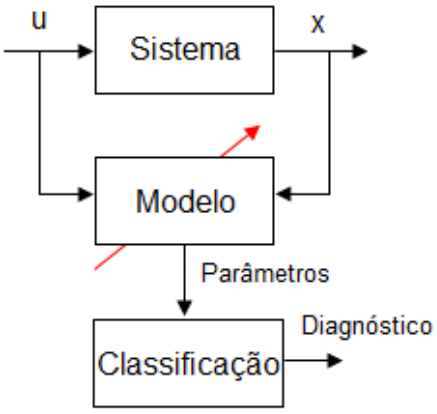

Figura 2.7: Monitoração da QD através de parâmetros de modelos extraídos dos dados. Extraída de [4].

Em [4], o método de monitoração pelos resíduos utiliza a diferenciação das séries temporais. Este método é apropriado para explorar a correlação temporal e testar se as variáveis alteram-se suavemente. O princípio de monitoração consiste em emitir alarmes se os limites que foram determinados para o resíduo são violados. Além

da diferenciação, outros métodos, como os mínimos quadrados recursivos (*Recursive Least Squares* - RLS) [29], têm sido utilizados para monitorar a QD. O método consiste em testar se a quantidade de exemplos de uma partição dos dados, em determinado período de tempo, é proporcional ao conjunto total de dados. O número de exemplos em cada partição x_t^i é modelado em função da média $x_t^i = \hat{\alpha}_t^i \overline{x}$ de todas as outras partições. O coeficiente $\hat{\alpha}_t^i$ é estimado utilizando-se o algoritmo RLS. A Figura 2.8 mostra a contagem real de uma partição, a estimativa do número de exemplos, o erro de previsão e o coeficiente do modelo ($\hat{\alpha}_t^i$). Este coeficiente permanece quase constante e o erro de previsão dentro dos limites. Quando o erro excede limites pré-estabelecidos ou o coeficiente altera-se de forma anormal, alarmes são emitidos.

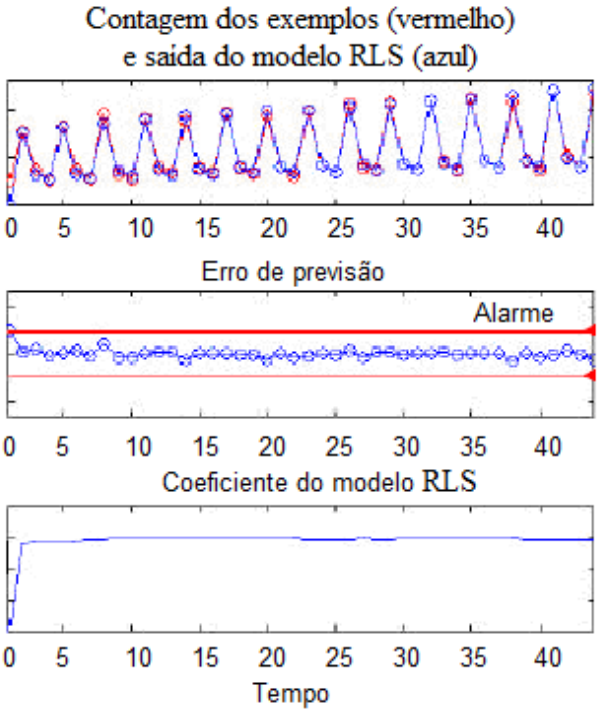

Figura 2.8: Monitoração RLS. Extraído de [4].

A QD também pode ser monitorada através da correlação entre variáveis. Assume-se que os dados de diferentes partições e instantes de tempo devam estar dentro dos limites das curvas de contorno da Função Densidade de Probabilidade (FDP) gaussiana (veja Figura 2.9). Esta estratégia parte da hipótese dos dados serem gaussianos.

Outro método utilizado na monitoração é a Análise de Componentes Principais (PCA) [30]. A PCA realiza uma transformação linear de forma que as amostras transformadas sejam projetadas em componentes ortogonais de acordo com a energia (variância) dos dados (veja a Figura 2.10). A partir de um conjunto livre de defeitos, as componentes principais são obtidas e os novos dados são projetados nes-

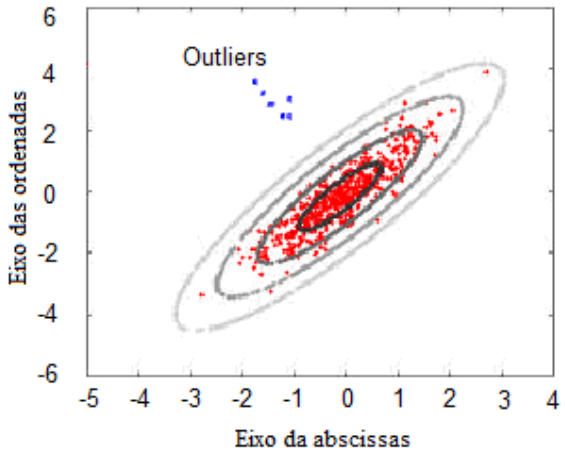

Figura 2.9: Exemplo da monitoração da QD através da correlação entre variáveis e limites de formas elípticas. Extraído de [4].

tas componentes. A partir daí, selecionam-se as projeções de interesse e monitora-se a QD pelo resíduo, que pode ser obtido, por exemplo, pelo método da diferenciação (veja Figura 2.11)

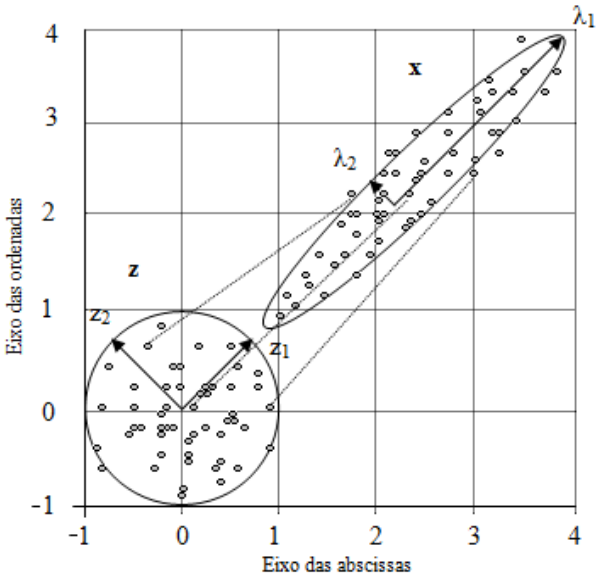

Figura 2.10: Componentes principais extraídas do conjunto de dados.

Em [25], propõe-se um método de monitoração baseado na teoria de controle [31]. Neste método, os erros acumulados são utilizados para construir um modelo dinâmico e monitorar as novas amostras de uma série temporal. Propõe-se a construção de corredores de validação através de redes neurais preditoras. O corredor é construído sob a incerteza gerada pelo modelo e ajustado por uma constante para incluir o papel do usuário. A Figura 2.12 mostra o princípio de monitoração.

Em [26], uma aplicação é desenvolvida a partir dos princípios dos corredores de validação anterior, com um modelo auto-regressivo integrado de média móvel (*Au-*

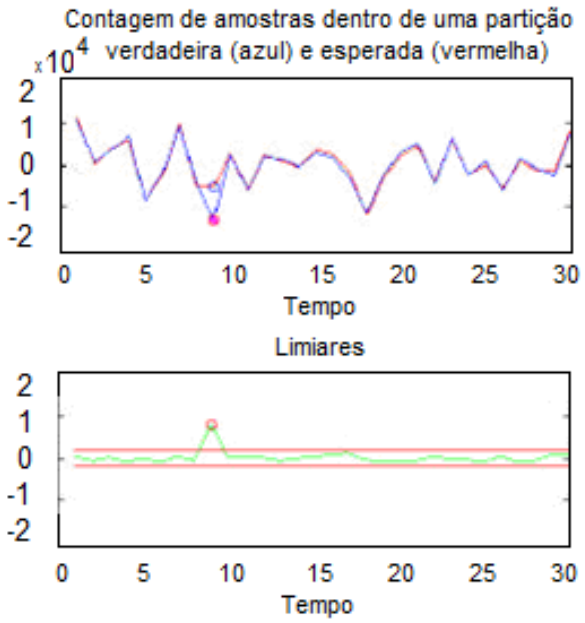

Figura 2.11: Monitoração do resíduo projetado em uma componente principal. Extraído de [4].

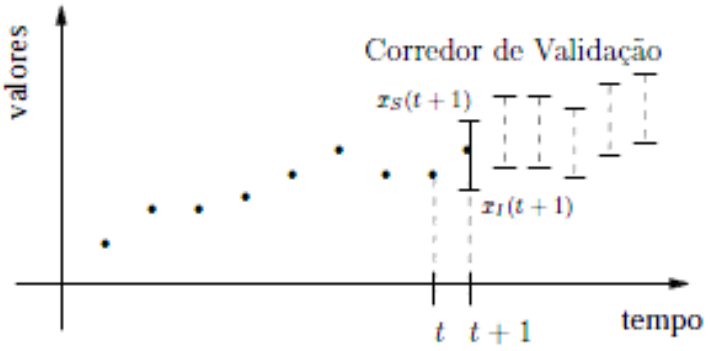

Figura 2.12: Corredor de validação para séries temporais. Extraído de [1].

toregressive Integrated Moving Average - ARIMA) [22] . Os corredores são divididos em patamares para se classificar as amostras entrantes como aceitáveis, possivelmente erradas e provavelmente erradas (veja a Figura 2.13).

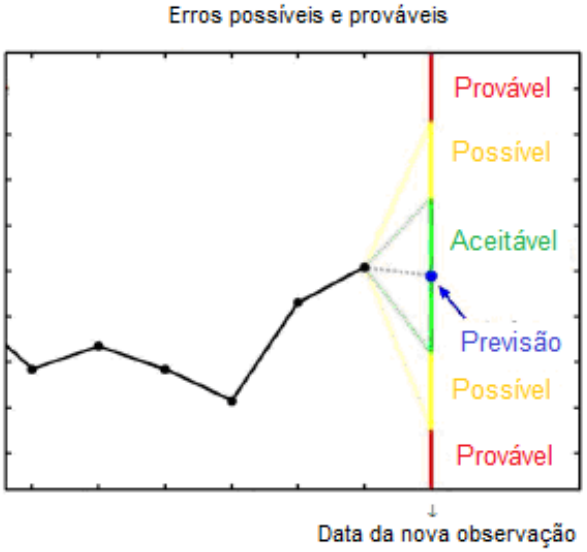

Figura 2.13: Níveis de validação para novas amostras. Extraído de [26].

2.3 Dimensões e Métricas

O aspecto multidimensional da qualidade de dados pode ser constatado fazendo-se uma busca na literatura, onde são reportadas mais de duzentas dimensões relacionadas ao tema [1]. Muitas dessas dimensões são correlacionadas [20] e ambigüidades na terminologia são frequentemente identificadas [17]. Assim, em casos específicos, as dimensões devem ter o seu conceito adaptado e estendido [32] para se adequarem às especificações dos dados e necessidades dos usuários. As dimensões são mensuradas através de indicadores, que definem objetivamente a qualidade do dado [14]. Mesmo as dimensões subjetivas podem ser relacionadas a indicadores para se medir objetivamente a qualidade [14]. Nesta seção, são apresentadas algumas dimensões conforme classificação em [14] e definição de diversos autores.

2.3.1 Dimensões Intrínsecas

Dimensões que avaliam a diferença entre os valores armazenados no sistema de informação e os valores do mundo real [9].

Acurácia

A acurácia é considerada um dos indicadores mais importantes para se medir a QD e é definida como "a medida em que os dados estão corretos, fiáveis e certificados" [9]

ou, ainda, "a correspondência dos dados com os valores do mundo real" [13],[12],[33]. Em [20], a acurácia foi definida como:

$$\text{Acurácia} = \frac{\text{Quantidade de Exemplos Incorretos}}{\text{Quantidade Total de Dados}}. \tag{2.1}$$

Observa-se ainda que é necessário definir o significado do erro. Por exemplo, o erro pode estar associado a um grau de precisão. Mesmo não correspondendo aos valores reais exatos, aceita-se uma faixa como sendo correta. O contexto pode dificultar ainda mais esta definição, pois uma precisão aceita em determinadas situações poderia ser inaceitável em outras. Isto mostra a ambiguidade que pode existir quando se tenta classificar as dimensões. Mesmo sendo classificada como uma dimensão intrínseca, sob outro ponto de vista, a acurácia poderia ser uma dimensão contextual.

Credibilidade

A Credibilidade é definida como "a medida em que os dados são verdadeiros, reais e acreditáveis" [9][33] . Esta dimensão pode ser avaliada através da combinação de três fatores [20]: avaliação da credibilidade da fonte, comparação com um padrão aceito e a experiência. Uma forma de se medir a credibilidade poderia ser associá-la a um valor, por exemplo, entre 0 e 1. Algumas formas de calcular esta dimensão são propostas em [19].

Objetividade

A Objetividade mede a imparcialidade dos dados. Apesar do nome, essa dimensão é afetada pela subjetividade e alguns julgamentos pessoais dos usuários da informação. Questões negativas relacionadas à objetividade podem também refletir em outras dimensões como a reputação e a credibilidade dos dados [20].

Reputação

A Reputação é o grau com que os dados são fiáveis, considerando-se, principalmente, a sua origem e conteúdo [9]. Reputação é, às vezes, usada como sinônimo de confiabilidade ou credibilidade para indicar o nível com que os usuários confiam nos dados [34]. Reputação dos dados é estritamente relacionada com as suas fontes. Por esta razão, muitos autores não distinguem a dimensão reputação de confiabilidade das fontes [32].

2.3.2 Dimensões Contextuais

Dimensões que avaliam a qualidade dentro de um contexto ou processo específico no qual os dados estão envolvidos [9].

Quantidade adequada de dados

Esta dimensão define a quantidade apropriada de dados disponíveis [9], uma vez que podem existir problemas tanto com elevada quantidade de dados quanto com a baixa quantidade. Esta dimensão pode ser definida relacionando-se a quantidade de dados com os resultados [34]. Um indicador para esta dimensão é definido por [20]:

$$\text{Quantidade Adequada de Dados} = \min\left(\frac{D_d}{D_r}, \frac{D_r}{D_d}\right). \tag{2.2}$$

onde, D_d é a quantidade de dados disponível; D_r é a quantidade de dados requerida;

Completude

A definição de dados completos é dada pelo grau de presença de dados para os diversos propósitos dos usuários. Em [9], define-se a completude como o grau de abrangência para a tarefa em questão. A completude pode ser definida pela seguinte expressão [20]:

$$\text{Completude} = \frac{\text{Quantidade de Exemplos Faltantes}}{\text{Quantidade Total de Dados}}. \tag{2.3}$$

Pontualidade

A pontualidade pode ser definida como o quão rapidamente um dado está disponível para o usuário na base de dados. Em [9], o termo é definido como o grau de atualização em relação a sua utilização. Em [34], pontualidade significa a idade dos dados. Em [35], pontualidade é definida como a propriedade da informação de chegar no momento oportuno. Em [36], menciona-se a equivalência entre pontualidade e completude. A Pontualidade também pode ser medida como uma função de duas variáveis elementares: atualidade e volatilidade [37]. A Atualidade é definida como o intervalo de tempo entre o instante em que os dados foram criados e o instante em que os dados estão prontos para serem usados. A Volatilidade é o intervalo de tempo que exprime a validade dos dados num contexto específico. Uma medida para a pontualidade é definida em [37] como:

$$\text{Pontualidade} = \max\left(0, 1 - \left[\frac{Atualidade}{Volatilidade}\right]^s\right) \tag{2.4}$$

onde $s > 0$ é um parâmetro necessário para controlar a sensibilidade para a razão atualidade-volatilidade. O valor deste expoente é, de fato, relacionado com o contexto e absorve a subjetividade introduzida com o julgamento de quem analisa os dados. Com esta definição, o valor da pontualidade varia entre 0 e 1.

Relevância

Relevância considera a adequação dos dados para a tarefa em questão [9] ou como os dados são capazes de satisfazer requisitos dos usuários [34]. Em [17] tem-se uma definição um pouco diferente: os dados são relevantes se forem oportunos e satisfizerem os critérios dos usuários especificados, explicitamente incluindo, assim, o conceito de tempo.

Valor agregado

Esta dimensão mede os benefícios obtidos com o uso dos dados [9],[34]. É possível distinguir duas abordagens para esta avaliação: medidas objetivas ou subjetivas. A abordagem objetiva implica na utilização de técnicas estatísticas para conceber parâmetros de utilidade dos dados. A abordagem subjetiva baseia-se na análise do valor acrescentado percebido pelos usuários.

2.3.3 Dimensões Representativas

Dimensões que julgam a clareza de representação dos dados. Estas dimensões enfatizam importância do papel de sistemas computacionais [9].

Inteligibilidade

Inteligibilidade é "a medida em que os dados são claros e inequívocos e de fácil compreensão"[9]. É uma dimensão subjetiva, uma vez que a sua avaliação inclui as percepções dos usuários em conjunto com outros aspectos, tais como a linguagem dos dados, estrutura e layout gráfico.

Concisão

Concisão é o grau de compactação na representação dos dados [14]. Essa dimensão é também considerada em [38], na qual é chamada minimalidade. Neste estudo, os autores propõem medir a dimensão considerando-se a quantidade registros redundantes (linhas de uma tabela), de relacionamentos e de atributos (colunas).

Consistência

Consistência é a medida em que os dados são sempre apresentados no mesmo formato [9]. O conceito ainda pode ser distinguido em consistência interna e externa [35]. Consistência interna se refere a diferentes valores de dados referentes a uma mesma entidade, por exemplo, nome e sexo de uma mesma pessoa. Já a consistência externa se refere a diferentes valores de dados referentes a diversas entidades.

Interpretabilidade

Interpretabilidade está relacionada com o formato com que os dados são especificados, incluindo a linguagem, símbolos, unidades e etc. Refere-se à clareza (não ambiguidade) das definições de dados [9].

2.3.4 Dimensões de Acessibilidade

Dimensões que enfatizam o papel de garantidor de acessibilidade, disponibilidade e segurança do sistema de informação [9].

Acessibilidade

Acessibilidade exprime o quanto o dado está disponível ou o quão ele é rapidamente recuperável [9]. Em [33] é proposta a seguinte medida:

$$\text{Acessibilidade} = \max\left(0, 1 - \frac{\Delta T_1}{\Delta T_2}\right) \tag{2.5}$$

onde: ΔT_1 é o intervalo de tempo que ocorre entre os dados serem requisitados pelo usuários e efetivamente usados. É associado com o tempo de entrega dos dados. ΔT_2 é o intervalo de tempo que ocorre entre os dados serem requisitados pelos usuários e estarem fora de uso. É o intervalo de tempo relacionado com a necessidade de dados dos usuários

Observa-se que, na expressão 2.5, o foco está somente no aspecto temporal da dimensão acessibilidade. Outras medidas podem ser desenvolvidas considerando-se a estrutura dos dados e o caminho de acesso. Em algumas contribuições, o termo Disponibilidade é utilizado como sinônimo de Acessibilidade [34].

Segurança de acesso

Em [9], segurança de acesso é definida como o grau de limitação ao acesso aos dados pelos usuários. Em [34], o termo é definido como o grau de segurança e privacidade com os quais os dados são transmitidos de uma fonte até os usuários e vice-versa.

2.4 Dependência entre as dimensões

De acordo com o contexto, as dimensões podem ser dependentes umas das outras. As dependências podem ter uma correlação negativa ou correlação positiva. No primeiro caso, a melhora na avaliação de uma dimensão pode levar a uma perda de desempenho em outra dimensão. Por exemplo, introduzir uma nova informação para melhorar a completude pode reduzir a consistência dos dados. Correlações positivas significam que a melhora de desempenho de uma dimensão provoca uma mudança na outra dimensão no mesmo sentido.

2.5 Prevenção × Correção

A perda da qualidade dos dados pode ocorrer em qualquer parte do "ciclo de vida" da informação. Nesse sentido, os princípios da qualidade dos dados devem ser aplicados nas diversas fases do processo (captura, digitalização, armazenagem, análise, apresentação e utilização). Existem dois pontos-chave a se considerar para a melhoria da qualidade dos dados: prevenção e correção. O erro de prevenção está intimamente relacionado com obtenção dos dados e a armazenagem em um banco de dados. Embora um considerável esforço possa e deva ser dado à prevenção do erro [36], a verdade é que os erros em grandes conjuntos de dados continuarão a existir [24]. O que nos leva a dizer que a validação e a correção também não devem ser ignoradas.

A prevenção é mais eficaz do que a detecção do erro, uma vez que os diagnósticos tardios são frequentemente onerosos e nem sempre garantem que a correção seja totalmente bem sucedida [39]. Ao se projetar um sistema de dados, deve-se começar pela definição da visão de dados, desenvolvendo uma política e implementando uma estratégia para os dados - e não pela realização sem planejamento, descoordenada e não sistemática de "limpeza dos dados". A detecção do erro, no entanto, tem um papel particularmente importante quando se lida com bases herdadas [39],[40], como é o caso de grande parte dos dados existentes.

O custo da certificação na fase de entrada dos dados pode ser substancial [41], mas é apenas uma fração do custo de verificação e correção dos dados numa etapa posterior. Fazer correções retroativas também dão margem para que os dados incorretos sejam utilizados em análises anteriores, elevando os riscos. O processo de limpeza é importante para identificar as causas dos erros que já foram incorporados na base de dados, devendo se preocupar com os procedimentos que garantam que esses erros não sejam repetidos. As operações de limpeza de dados e prevenção do erro devem correr paralelamente. Decidir-se por limpar os dados e depois preocupar-se com a prevenção, normalmente significa que a prevenção do erro nunca será satis-

fatoriamente realizada e, entretanto, mais e mais erros serão adicionados à base de dados.

2.6 Da Modelagem de Dados para Modelagem da Qualidade de Dados

Um tipo de modelagem da qualidade de dados pode ser feito através da extensão das tradicionais metodologias de modelagem de dados. Enquanto a modelagem de dados captura a estrutura e a semântica dos dados, a modelagem da qualidade de dados captura as questões implícitas da estrutura e da semântica da qualidade de dados. Para entendermos como a qualidade de dados se encaixa no contexto da modelagem convencional dos dados, é mostrada uma esquematização na Figura 14 [14][42]. O esquema é composto de várias etapas, cada uma com entradas e saídas. O resultado final é o esquema de qualidade. Os detalhes de cada etapa mostrada na Figura 2.14 são abordados na seqüência.

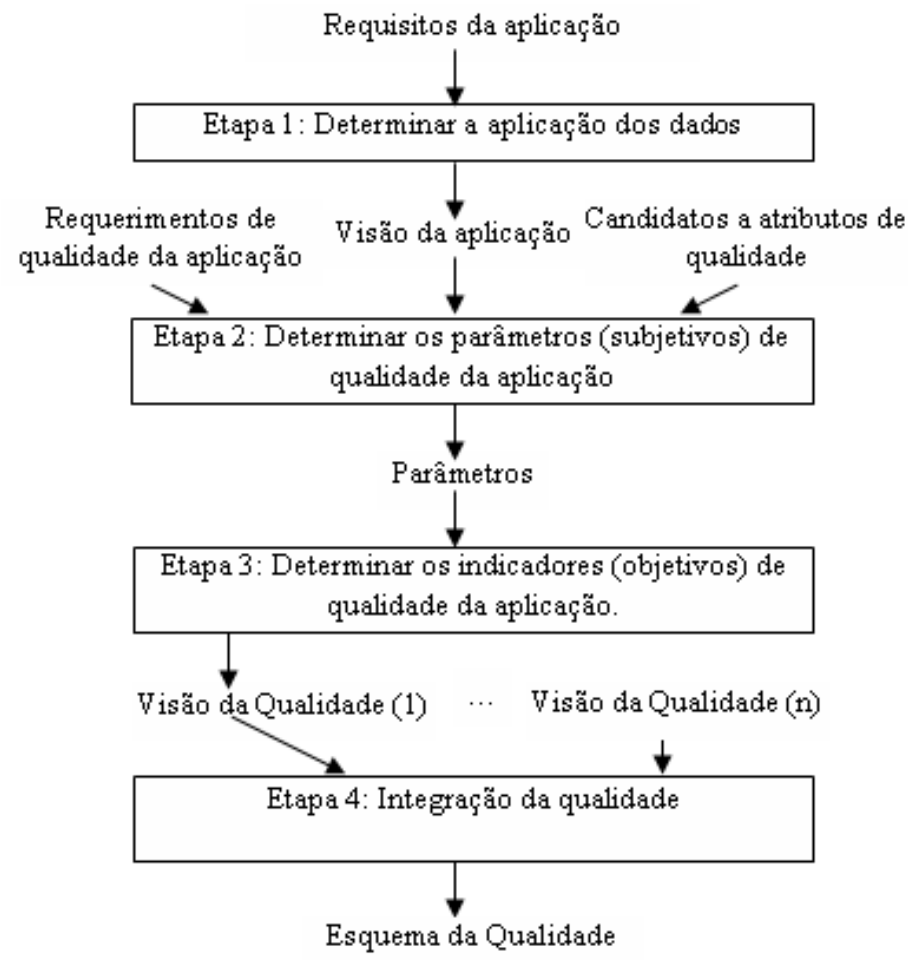

Figura 2.14: Processo de modelagem da qualidade de dados. Extraído de [14].

Etapa 1

Estabelecida a aplicação para a qual a qualidade de dados será implantada, deverão ser avaliados requisitos da aplicação. Esta etapa é semelhante ao processo de modelagem de dados tradicional [43]. Por exemplo, suponhamos que são obtidas informações de ações de empresas negociadas na bolsa de valores. As ações da companhia são identificadas pelo seu código de negociação, que tem o preço de negociação e um relatório periódico associado a ela. Neste caso, a visão da aplicação será semelhante a Figura 2.15. Observa-se que, nesta etapa, a QD ainda não foi incorporada ao modelo de dados.

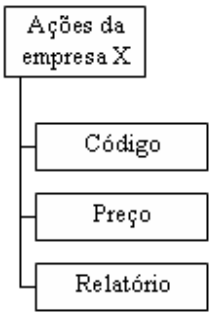

Figura 2.15: Saída da etapa 1: visão da aplicação.

Etapa 2

Na segunda etapa, os parâmetros (subjetivos) de qualidade são definidos. Além da visão da aplicação do passo anterior, requerimentos de qualidade e candidatos a atributos de qualidades devem compor a entrada deste passo. A saída será a visão dos parâmetros. Neste passo, o objetivo é elucidar a necessidade da qualidade de dados. Para cada componente da visão da aplicação, os projetistas devem determinar os parâmetros de qualidade necessários para os requisitos da qualidade de dados (veja a Seção 4.2.2). Um esquema da visão dos parâmetros é mostrado na Figura 2.16. Os parâmetros da qualidade são associados ao modelo de dados. Por exemplo, a atualização e a acurácia são parâmetros de QD para o preço da ação, a interpretabilidade é um parâmetro para o código da ação e a credibilidade e a atualização são parâmetros para o relatório.

Etapa 3

Na terceira etapa, definem-se os indicadores (objetivos) de qualidade. Neste passo, processa-se a visão dos parâmetros e apresenta-se a visão da qualidade como saída. Neste processo, o objetivo é operacionalizar os parâmetros qualitativos em características mensuráveis. Estas características mensuráveis são os indicadores de qua-

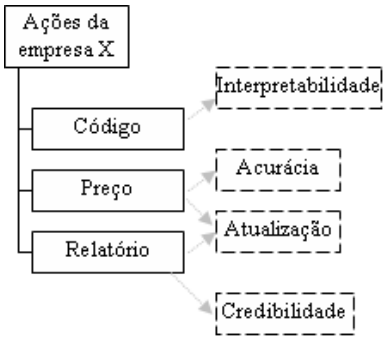

Figura 2.16: Saída da etapa 2: visão dos parâmetros.

lidade. A Figura 2.17 ilustra esta etapa com o exemplo utilizado. Observa-se que o parâmetro subjetivo de atualização do preço foi mapeado para o indicador objetivo data, hora e segundos. A interpretabilidade do código da ação foi mapeada para um campo com o nome da empresa. A credibilidade do relatório foi associado ao nome do analista e a atualização à data de publicação.

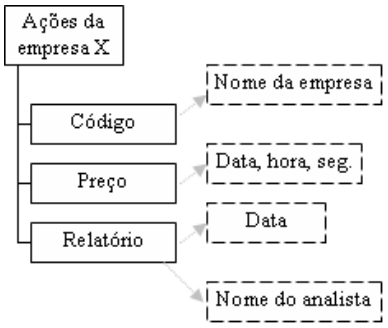

Figura 2.17: Saída da etapa 3: visão da qualidade.

Etapa 4

O último passo da modelagem integra as diversas visões da qualidade. O resultado é um esquema de qualidade. Por exemplo, em outro contexto, o usuário poderia estar interessado somente na acurácia do preço das ações. Assim, esta nova visão integraria as visões existentes e resultaria no esquema da QD.

2.7 Medição da QD

Medir a QD é fundamental para se gerenciar a informação [21]. Em [34], é reportada a necessidade de padronização certificada das medições, para se ter um gerenciamento unificado da QD e elevar a confiança das comparações. Em [44], iniciativas de padronização internacional estão sendo desenvolvidas. Assim, atribuir valores objetivos para a qualidade é fundamental para a monitoração e o gerenciamento da QD. A medição da QD deve ocorrer tanto para dados nos seus níveis

mais elementares, como amostras únicas, quanto para dados agrupados em diferentes formas, como, por exemplo, séries temporais, bancos de dados e repositórios de dados. Tanto as amostras que compõem uma variável (e suas dimensões) quanto as variáveis que compõem um banco de dados deveriam ter associadas medidas de QD individuais. Por exemplo, a acurácia de uma amostra poderia estar associada a um teste de limites, que verifica se amostra está dentro ou fora de padrões estabelecidos como de qualidade (1 se sim e zero se não). Os resultados dos testes de cada amostra poderiam ser agrupados para formarem um indicador de qualidade da amostra. Os indicadores poderiam ser agrupados para formar o indicador da variável e, da mesma forma, os indicadores da variável agrupados para formar o indicador do banco completo. Se este banco estiver em um repositório de dados, essa mesma lógica pode ser seguida para se ter diversas perspectivas da QD. Esse procedimento pode ser realizado para as diversas outras dimensões de QD, como completude, atualização e outras, para se medir o tamanho do problema e compor a métrica de qualidade.

As dimensões que compõem a métrica devem ser ponderadas de acordo com o contexto e com os usuários da informação. Assim, tem-se a perspectiva da QD desde um simples indicador, como, por exemplo, um número que define a QD do banco, até visões mais detalhadas, como indicadores de qualidade de cada série e de cada amostra. No seu nível mais fundamental, deveria ser possível visualizar quais testes passaram ou falharam.

2.8 Discussão

A qualidade de dados (ou qualidade da informação) tem importância fundamental em sistemas de informação. Assim como os dados são modelados, a qualidade dos dados também deve ser modelada. Essa modelagem da qualidade deve ser aplicada já no desenho de esquema dos dados, prevenindo erros origem e elevando a QD desde a origem. Essa abordagem preventiva não elimina todos os problemas, pois mesmo dados corretos podem ter a qualidade degradada ao longo do tempo ou após novas informações serem inseridas no ciclo da informação. Dessa forma, a qualidade deve ser avaliada de forma contínua, antes e depois do dado ter sido inserido no sistema de informação. Monitorar a qualidade permite que erros sejam detectados e corrigidos precocemente e não se propaguem ao longo do ciclo da informação. Qualidade de dados significa ter dados em conformidade com as especificações, seja do ponto de vista dos dados ou dos usuários. Este conceito reflete um aspecto multidimensional da qualidade. Dentre as principais dimensões, destaca-se o conjunto introduzido por [9] (veja a Tabela 2.4), que classifica as dimensões em intrínsecas, contextuais, representacionais e de acesso.

Estas dimensões podem compor métricas que definem objetivamente as di-

Tabela 2.4: Dimensões da Qualidade de Dados

Intrínsecas	Contextuais	Representacionais	Acessibilidade
Acurácia Credibilidade Objetividade Reputação	Valor agregado Relevância Completude Pontualidade Quantidade apropriada de dados	Inteligibilidade Interpretabilidade Consistência Concisão	Acessibilidade Segurança

mensões. Mesmo dimensões subjetivas, podem ser mensuradas pelas métricas e formar indicadores da qualidade dos dados. Por fim, essa medição da qualidade dos dados introduz a perspectiva de monitoração e gerenciamento da qualidade dos dados.

Capítulo 3

Análise de Componentes Independentes

A análise de componentes independentes (ICA) é uma técnica capaz de revelar fatores escondidos em conjuntos de sinais. ICA define um modelo gerador para os dados observados, que são assumidos serem misturas de variáveis ocultas e desconhecidas. As variáveis latentes são assumidas mutuamente independentes, e são chamadas de componentes independentes ou fontes dos dados observados [30].

As raízes de ICA vêm dos trabalhos de Darmois [45] na década de 50 e Kagan et al. [46] na década de 70, caracterizando variáveis aleatórias em estruturas lineares. Os trabalhos pioneiros em análise de componentes independentes foram desenvolvidos por Jutten & Herault [47] na década de 80 e, na década de 90, Comon [48] formalizou e desenvolveu a teoria básica de análise de componentes independentes concentrando os trabalhos nas condições de existência, unicidade e indeterminações da estimação. Utilizando o teorema de Darmois-Skitovitch, Comon demonstrou que a matriz mistura linear pode ser encontrada, exceto pela permutação e escala dos fatores. Neste trabalho, a computação de ICA foi abordada como uma questão de otimização. Durante a década de 90 e até os dias de hoje, diversas aplicações têm sido propostas nos mais variados contextos e bons resultados têm sido demonstrados.

3.1 Definição

Vamos assumir a hipótese de que os dados consistem de m variáveis aleatórias conjuntamente observadas T vezes. Assim, denotaremos os dados por $\mathbf{x}_j(t)$, onde $j = 1, 2, ...m$ e $t = 1, 2, ...T$. A formulação geral para o problema seria encontrar uma função, mapeando-se o espaço m-dimensional para o espaço n-dimensional, de maneira que as variáveis transformadas fornecessem as informações escondidas no espaço original. Ou seja, as variáveis transformadas deveriam ser os componentes

implícitos que descrevessem a estrutura essencial dos dados. É esperado que estes componentes correspondam a alguma causa física envolvida no processo de geração dos dados. Cada componente ($\mathbf{y}_i(t)$) pode ser expresso como uma combinação linear das variáveis observadas ($\mathbf{x}_j(t)$).

$$\mathbf{y}_i(t) = \sum_j b_{ij}\mathbf{x}_j(t) \tag{3.1}$$

onde, $i = 1, 2, ...n$ e $j = 1, 2, ...m$ e b_{ij} são os coeficientes que definem a representação

O problema pode ser resolvido encontrando-se os coeficientes b_{ij}. Com o auxílio da álgebra linear, a transformação linear pode ser representada por:

$$\begin{pmatrix} \mathbf{y}_1(t) \\ \mathbf{y}_2(t) \\ \vdots \\ \mathbf{y}_n(t) \end{pmatrix} = \mathbf{B} \begin{pmatrix} \mathbf{x}_1(t) \\ \mathbf{x}_2(t) \\ \vdots \\ \mathbf{x}_m(t) \end{pmatrix} \tag{3.2}$$

Agora, podemos determinar a matiz $\mathbf{B}$ através das propriedades estatísticas dos componentes transformados $\mathbf{y}_i(t)$, tais como descorrelação não-linear, máxima não-gaussiana e descorrelação no tempo [30].

3.2 Princípios Básicos

Podemos considerar ICA como um passo além da simples descorrelação linear. De fato, a descorrelação linear (ou branqueamento) é utilizada como um pré-processamento para ICA.

Um vetor $\mathbf{z}(t) = (\mathbf{z}_1(t), \mathbf{z}_2(t), ..., \mathbf{z}_n(t))^T$ de média zero é dito branqueado se os elementos $\mathbf{z}_i(t)$ são descorrelacionados entre si e têm variância unitária. Em termos da matriz de covariância, isso significa:

$$\mathrm{E}\{\mathbf{z}(t)\mathbf{z}^T(t)\} = \mathbf{I} \tag{3.3}$$

onde $\mathbf{I}$ é a matriz identidade.

Para se obter as variáveis branqueadas $\mathbf{z}(t)$, aplica-se uma transformação $\mathbf{V}$ na variável observada $\mathbf{x}(t)$:

$$\mathbf{z}(t) = \mathbf{V}\mathbf{x}(t) \tag{3.4}$$

Para solucionar o problema, consideremos a matriz $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n)$, cujas colunas são autovetores de norma unitária da matriz de covariância $\mathbf{C}^x = \mathrm{E}\{\mathbf{x}\mathbf{x}^T\}$, e $\mathbf{D} = \mathrm{diag}(d_1, d_2, ..., d_n)$ a matriz diagonal de autovalores de $\mathbf{C}^x$. A decomposição

em autovetores e autovalores nos dá a matriz de branqueamento:

$$\mathbf{V} = \mathbf{D}^{-1/2}\mathbf{E}^T \tag{3.5}$$

Após o branqueamento, basicamente dois princípios norteiam ICA para acessar as fontes independentes: descorrelação não linear e máxima não-gaussianidade. A descorrelação não-linear encontra uma matriz-separação de maneira que, para qualquer $i \neq j$, os componentes $\mathbf{y}_i(t)$ e $\mathbf{y}_j(t)$ são descorrelacionados e os componentes transformados $g(\mathbf{y}_i(t))$ e $h(\mathbf{y}_j(t))$ são também descorrelacionados, onde $g(.)$ e $h(.)$ são funções não-lineares apropriadas. A Teoria da Estimação e a Teoria da Informação fornecem muitos métodos clássicos para a escolha e estimação das não linearidades $g(.)$ e $h(.)$, tais como a máxima semelhança e a informação mútua [49].

Já o princípio da máxima não-gaussiana busca no teorema do limite central os fundamentos para a separação de fontes independentes. No teorema, a soma de variáveis não-gaussianas é mais próxima de uma gaussiana do que as variáveis originais. Assim, este princípio de separação busca encontrar máximos locais de não-gaussianas de uma combinação linear. Dessa forma, cada máximo local revela um componente independente. Para se medir a não-gaussiana, métodos tais como a assimetria, curtose e cumulantes de ordem elevada são utilizados.

Ainda, em sua formulação básica, o modelo de ICA assume a hipótese (i) das fontes serem estatisticamente independentes entre si, (ii) com distribuições não-gaussianas e (iii) o número de fontes independentes ser igual ao número de misturas observadas. Com base nestas hipóteses, em geral, os algoritmos utilizam estatística de ordem superior para estimar a matriz de separação.

3.3 Princípios para Séries Temporais

Os sinais-mistura podem ser variáveis ordenadas ao invés de variáveis aleatórias. Isto contrasta com a formulação básica de ICA, na qual a sequência das amostras não tem ordem particular. Se os componentes independentes (CIs) são, digamos, séries temporais, eles podem conter muito mais estrutura do que simples variáveis aleatórias. A informação adicional pode tornar possível a estimação dos modelos por meio de estatística de segunda ordem e deixar de assumir a hipótese de fontes não-gaussianas.

Para sinais com estrutura temporal, as t realizações do processo (veja a Equação 3.2) representam a sequência de tempo. Além disso, a estimação do modelo baseia-se em hipóteses alternativas à hipótese de não-gaussiana apresentada em ICA básico: assumir que os CIs têm autocovariâncias diferentes ou assumir que as variâncias dos CIs são não estacionárias. No entanto, apesar de não utilizar a estrutura temporal

das séries, frequentemente a formulação básica também pode ser aplicada em séries temporais.

3.3.1 Hipótese de autocovariâncias diferentes

No caso de assumir a hipótese de autocovariâncias diferentes para cada CI, a forma da estrutura temporal é dada pelas autocovariâncias de cada sinal, $cov(\mathbf{x}_i(t), \mathbf{x}_i(t-\tau))$, e a covariância entre dois sinais, $cov(\mathbf{x}_i(t), \mathbf{x}_j(t-\tau))$. Assim, após a retirada da média de $\mathbf{x}(t)$, as estatísticas necessárias para obter os componentes independentes podem ser agrupadas na Matriz Covariância deslocada no tempo:

$$\mathbf{C}_\tau^x = E\{\mathbf{x}(t)\mathbf{x}^T(t-\tau)\} \tag{3.6}$$

onde, $\mathbf{x}(t)$ é a matriz contendo os sinais-mistura e $\mathbf{x}^T(t-\tau)$ é a matriz transposta contendo os sinais-mistura com atraso τ.

Aqui o ponto chave é que a informação de segunda ordem pode ser usada no lugar da informação de ordem superior para se obter as fontes independentes [50],[51]. Assim, devemos encontrar a matriz-separação $\mathbf{B}$ que faça, além da covariância instantânea ($\tau = 0$), as covariâncias defasadas ($\tau > 0$) serem zero:

$$E\{\mathbf{y}_i(t)\mathbf{y}_j^T(t-\tau)\} = 0 \tag{3.7}$$

A motivação para se igualar a zero todas as covariâncias defasadas é o fato desta característica ser própria da independência. Para melhor compreender esta separação, consideremos apenas uma matriz de covariância atrasada ($\tau = 1$). Após retirar a média e branquear $\mathbf{x}(t)$, tem-se $\mathbf{z}(t)$ e chega-se a matriz de separação ortogonal $\mathbf{W}$:

$$\mathbf{y}(t) = \mathbf{W}\mathbf{z}(t) \tag{3.8}$$

$$\mathbf{y}(t-\tau) = \mathbf{W}\mathbf{z}(t-\tau)) \tag{3.9}$$

Pela linearidade e ortogonalidade, pode-se escrever a matriz covariância atrasada dos sinais branqueados [30]:

$$\overline{\mathbf{C}}_\tau^z = \mathbf{W}^T\overline{\mathbf{C}}_\tau^y\mathbf{W} \tag{3.10}$$

onde,

$$\overline{\mathbf{C}}_\tau^z = \frac{1}{2}[\mathbf{C}_\tau^z + (\mathbf{C}_\tau^z)^T]$$

Observa-se que $\overline{\mathbf{C}}_\tau^y$ é diagonal devido à independência dos vetores de $\mathbf{y}(t)$. O que a equação mostra é que $\mathbf{W}$ deve fazer parte da decomposição de autovalores de $\overline{\mathbf{C}}_\tau^z$ [30].

Estendendo o raciocínio para vários atrasos, é suficiente que apenas a covariância

de um deles seja diferente das demais. Assim, a escolha de τ não seria tão problemática. Em princípio, utilizando vários atrasos no tempo, deseja-se diagonalizar simultaneamente todas as matrizes-covariância correspondentes. No entanto, a diagonalização exata é pouco provável, o que nos leva a formular um indicador para o grau de diagonalização:

$$off(\mathbf{M}) = \sum_{i \neq j} m_{ij}^2 \tag{3.11}$$

onde $\mathbf{M} \in \{\mathbf{C}_1^z, \mathbf{C}_2^z, ..., \mathbf{C}_\tau^z\}$

3.3.2 Hipótese de variâncias não estacionárias

A hipótese anterior pode não ser eficiente quando as componentes independentes têm autocovariâncias iguais (espectro de potência idêntico). Neste caso, uma alternativa é assumir a hipótese de variâncias não estacionárias dos CIs [52]. Assume-se também que a variância se modifica lentamente no tempo. Observa-se que este pressuposto independe das hipóteses mencionadas nas seções anteriores. Dada a hipótese de variâncias não estacionárias, pode-se chegar aos componentes independentes através da análise das autocorrelações locais ou através da análise dos cumulantes cruzados dos CIs.

Autocorrelações locais

Se encontrarmos a matriz $\mathbf{B}$ que produza $\mathbf{y}(t) \in \Re^n$ descorrelacionado a cada instante do tempo, tem-se a independência [52]. Note que, como não é estacionária, a covariância de $\mathbf{y}(t)$ depende do atraso. Assim, se forçarmos os componentes serem descorrelacionados a cada instante, haverá uma condição muito mais forte que o simples branqueamento.

Cumulantes cruzados

Um segundo método, baseado na interpretação das variâncias não estacionárias, é através de cumulantes cruzados de ordem superior. Podemos medir a variância não estacionária do sinal $\mathbf{y}(t)$ usando uma medida baseada na correlação temporal das energias:

$$E\{\mathbf{y}_i^2(t)\mathbf{y}_j^2(t-\tau)\} \tag{3.12}$$

Por uma questão de simplificação matemática, utilizam-se cumulantes. A autocorrelação não-linear é interpretada através do cumulante cruzado de 4ª ordem, correspondente às correlações de energias:

$$cum(\mathbf{y}(t), \mathbf{y}(t), \mathbf{y}(t-\tau), \mathbf{y}(t-\tau)) \tag{3.13}$$

3.3.3 Unificação dos princípios de separação

Os princípios de separação abordados foram unificados em [50],[51], com base no conceito da complexidade de Kolmogoroff [53],[54]. Define-se complexidade de Kolmogoroff de uma "string" $\mathbf{z}$(t) como a descrição mínima de seu comprimento; ou seja, a quantidade mínima de código (bits) necessária para descrever esta variável. Por exemplo, sendo o ruído uma variável não estruturada e não previsível, cada amostra deveria ser codificada. No caso de uma variável estruturada, a redundância pode ser estimada, não precisando ser codificada. Dessa forma, pode se medir a quantidade de estrutura de um sinal $\mathbf{y}(t)$ pela quantidade de compressão possível na codificação do sinal.

Dado o sinal $\mathbf{z}(t)$, o grau de incerteza desta variável pode ser medido pela entropia:

$$\mathrm{H}(\mathbf{z}(t)) = -\mathrm{E}\{\log \mathrm{p}(\mathbf{z}(t))\} \tag{3.14}$$

e, dada uma transformação $\mathbf{y}(t) = \mathbf{W}\mathbf{z}(t)$, a entropia da transformação fica

$$\mathrm{H}(\mathbf{y}(t)) = \mathrm{H}(\mathbf{y}(t)) + \log|\det \mathbf{W}| \tag{3.15}$$

Definindo-se a informação mútua entre as variáveis transformadas $\mathbf{y}_i(t)$

$$\mathrm{IM}(\mathbf{y}_1(t), \mathbf{y}_2(t), ..., \mathbf{y}_n(t)) = \sum_i \mathrm{H}(\mathbf{y}_i(t)) - \mathrm{H}(\mathbf{y}(t)) \tag{3.16}$$

Chega-se a

$$\mathrm{IM}(\mathbf{y}_1(t), \mathbf{y}_2(t), ..., \mathbf{y}_n(t)) = \sum_i \mathrm{H}(\mathbf{y}_i(t)) - \mathrm{H}(\mathbf{z}(t)) - \log|det\mathbf{W}| \tag{3.17}$$

No entanto, dado que a variável $\mathbf{z}(t)$ é conhecida *a priori*, $\mathrm{H}(\mathbf{z}(t)) = 0$.

A entropia pode ser interpretada como o comprimento médio ótimo do código, o que leva à função objetivo:

$$\mathrm{IM}(\mathbf{y}_1(t), \mathbf{y}_2(t), ..., \mathbf{y}_n(t)) = \sum_i \mathrm{K}(\mathbf{b}_i\mathbf{z}(t)) - \log|\det \mathbf{W}| \tag{3.18}$$

onde $\mathrm{K}(\mathbf{b}_i\mathbf{z}(t))$ é a Complexidade de Komolgoroff

Avaliar a complexidade de Komolgoroff dos sinais significa avaliar a correlação entre as amostras - uma vez que códigos mínimos referem-se a variáveis mais correlacionadas. Por outro lado, minimizar a informação mútua equivale a procurar estruturas nos dados - assim como o princípio da não-gaussiana utilizado em ICA básico. Dessa forma, a complexidade de Komolgoroff pode ser interpretada como uma ferramenta que unifica os princípios de correlação temporal e os princípios de

máxima não-gaussiana.

3.4 Algoritmos

Diversos algoritmos têm sido propostos e implementados no contexto de separação cega de sinais. Nesta seção, apresentaremos alguns deles e que compõem a base de algoritmos ICA do sistema de monitoração da QD proposto nesta tese. Basicamente, podemos dividir estes algoritmos em dois grandes grupos: aqueles que utilizam estatística de segunda ordem (ESO) e aqueles que utilizam estatística de ordem superior (EOS).

3.4.1 ESO

Assumindo que as autocovariâncias das fontes são diferentes, um dos mais simples algoritmos desta classe é o AMUSE (*Algorithm for Multiple Unknown Source Extraction*) [55]. Este algoritmo é baseado na decomposição de autovalores de uma simples matriz de covariância atrasada no tempo para dados pré-branqueados [30] e foi implementado em [56]. Assim, baseado na diagonalização de apenas uma matriz de covariância atrasada, AMUSE pode ser escrito como:

1 – Branqueie os dados $\mathbf{x}(t)$ para obter $\mathbf{z}(t)$;

2 – Calcule a decomposição de autovalores de $\overline{\mathbf{C}}_\tau^z = \frac{1}{2}[\mathbf{C}_\tau^z + (\mathbf{C}_\tau^z)^T]$ para um dado atraso τ.

3 – As linhas da matriz separação $\mathbf{W}$ são dadas pelos autovetores obtidos no passo 2.

Apesar de simples e veloz, o algoritmo necessita que os autovalores da matriz $\overline{\mathbf{C}}_\tau^z$ sejam unicamente definidos. Isto ocorreria caso todos os autovalores fossem distintos. No entanto, o que se nota na pratica é que freqüentemente os autovalores não são distintos. Para remediar o problema, deve-se procurar um atraso τ apropriado. No caso da igualdade entre os espectros de potências dos CIs, não há atraso que faça a estimação possível.

Outro algoritmo que aplica estatística de segunda ordem é o EVD2 (*Eigen Value Decomposition*). Este algoritmo faz a separação cega de sinais não branqueados (coloridos), aplicando decomposição de autovalores em uma matriz simétrica que é uma combinação linear de várias matrizes de covariâncias atrasadas no tempo. Esta aproximação foi apresentada em [56],[57].

Outro algoritmo que combina diversas matrizes de covariância atrasadas no tempo é o SOBI (*Second Order Blind Identification*) [58],[59]. O grupo de matrizes pode ser conjuntamente diagonalizado aplicando-se o método de diagonalização conjunta [60],[61]. O objetivo é minimizar $off(\mathbf{M})$ (veja a Equação 3.11) das $\mathbf{M}$

matrizes de covariância atrasada. Assim, substituindo $\mathbf{M}$ pela matriz de covariância atrasada, e realizando as devidas operações algébricas [30], chega-se à função objetivo:

$$\Im_1(\mathbf{W}) = \sum_{\tau \in S} off(\overline{\mathbf{C}}_\tau^y) \tag{3.19}$$

onde $\overline{\mathbf{C}}_\tau^y = \mathbf{W}\overline{C}_\tau^z\mathbf{W}^T$ e S é o conjunto de atrasos

Minimizando $\Im_1(\mathbf{W})$ sob a consideração de que $\mathbf{W}$ é ortogonal, temos o método de estimação. A minimização pode ser realizada pelo gradiente descendente. Uma alternativa seria adaptar os métodos existentes pela decomposição de autovalores, para diagonalização simultânea das várias matrizes. Em [30], desenvolve-se um método de gradiente descendente para atualização de $\mathbf{W}$:

$$\Delta\mathbf{W} = \sum_{\tau \in S} \mathrm{diag}(\mathbf{W}\overline{C}_\tau^z\mathbf{W}^T)^{-1}\mathbf{W}\overline{C}_\tau^z \tag{3.20}$$

sendo que $\mathbf{W}$ deve ser ortogonalizado a cada iteração.

Métodos alternativos para a diagonalização conjunta diversificam a forma de implementação de SOBI, tais como o método de diagonalização conjunta aproximada rápida (FAJD) [61], a diagonalização aproximada seqüencial (SAD) [61], diagonalização conjunta (FFDIAG) [59],[62], otimização quadrática para diagonalização conjunta (QDIAG) [63]. A utilização efetiva de cada tipo de método de diagonalização depende do tipo de dado, nível de ruído aditivo e experiência do usuário. O FAJD utiliza o seguinte critério para a minimização:

$$\min_{\mathbf{V}\in\Re^{N\times N}} \sum_{k=1}^{K} \alpha_k \sum_{i=1}^{N} \sum_{j=1, j\neg i}^{N} |[\mathbf{V}\mathbf{C}_k\mathbf{V}^H]_{ij}|^2 - \beta \ln|\det(\mathbf{V})| \tag{3.21}$$

onde α_k, $1 \leq k \leq K$, e β são ponderadores positivos e $\mathbf{V} \leftarrow (\mathbf{I} + \mathbf{W})\mathbf{V}$.

O FFDIAG:

$$\min_{\mathbf{W}\in\Re^{N\times N}} \sum_{k=1}^{K} \sum_{j\neg i}^{N} |[\mathbf{W}\mathbf{D}_k + \mathbf{D}_k\mathbf{W}^T + \mathbf{E}_k]_j|^2 \tag{3.22}$$

onde $\mathbf{D}_k$ e $\mathbf{E}_k$ denotam os elementos da diagonal e fora dela na matriz $\mathbf{C}_k$, respectivamente.

O QDIAG avalia o problema de otimização:

$$\min_{\mathbf{V}\in\Re^{N\times N}} \sum_{k=1}^{K} \alpha_k \sum_{i=1}^{N} \sum_{j=1, j\neg i}^{N} (\mathbf{w}_j^T\mathbf{C}_k\mathbf{w}_j)^2 \tag{3.23}$$

sujeito a $\mathbf{w}_t^T\mathbf{C}_0\mathbf{w}_t = 1$, onde $\mathbf{w}_i^T$ é uma linha de $\mathbf{W}$, α_k são os fatores de ponderação e $\sum_k a_k = 1$

O SAD [61] considera o problema de otimização restrita:

$$\min_{\mathbf{w},a,d} \sum_{k=0}^{K} ||\mathbf{C}_k\mathbf{w} - d_k a||^2 \tag{3.24}$$

sujeito a $|||[d_0, d_1, ...d_k]||| = 1$, onde $d_0, d_1, ...d_k$ são escalares desconhecidos, a é colinear com um vetor coluna na matriz $\mathbf{A}$ e o vetor $\mathbf{w}^T$ é um vetor linha da matriz $\mathbf{W}$.

Outra variação do SOBI, nomeada de SOBI-RO (*Robust Second Order Blind Identification with Robust Orthogonalization*) [64], aplica como pré-processamento um tipo de ortogonalização robusta proposto em [65]. Para o desenvolvimento do algoritmo, parte-se da hipótese de $\mathbf{x}(t)$ ser uma mistura das fontes independentes e conter um ruído aditivo:

$$\mathbf{x}(t) = \mathbf{A}(\mathbf{y}(t) + \mathbf{n}(t)) \tag{3.25}$$

em que $\mathbf{A}$ é a matriz-mistura não singular, $\mathbf{y}$(t) representa as fontes assumidas mutuamente descorrelacionadas e temporalmente correlacionadas e $\mathbf{n}$(t) representa o ruído aditivo com média zero, independente dos sinais da fonte e com matriz covariância:

$$\mathbf{C}^n = E\{\mathbf{n}(t)\mathbf{n}^T(t)\} \tag{3.26}$$

Dadas as hipóteses acima, as matrizes de correlação dos sinais observados têm a seguinte estrutura:

$$\mathbf{C}_0^x = E\{\mathbf{x}(t)\mathbf{x}^T(t)\} = \mathbf{A}\mathbf{C}_0^y\mathbf{A}^T + \mathbf{C}^n \tag{3.27}$$

$$\mathbf{C}_\tau^x = E\{\mathbf{x}(t)\mathbf{x}^T(t-\tau)\} = \mathbf{A}\mathbf{C}_\tau^y\mathbf{A}^T \tag{3.28}$$

O problema é estimar a matriz mistura $\mathbf{A} = \mathbf{B}^{-1}$ utilizando somente os dados ruidosos observados $\mathbf{x}(t)$. Para resolver este problema, SOBI-RO utiliza uma matriz de branqueamento robusto. Para estimá-la, faz-se a decomposição de autovalores para cada matriz de covariância $\mathbf{C}_\tau^x$. O resultado da decomposição é combinado linearmente para produzir a matriz de branqueamento robusto $\mathbf{F}$. Os coeficientes da combinação linear são calculados por um algoritmo de convergência global de passo finito. Assim, o algoritmo é dado por:

(i) Estime a matriz correlação e calcule uma decomposição de valor singular do conjunto de $n \times nK$ matrizes $\mathbf{C} = [\mathbf{C}_1^x ... \mathbf{C}_{\tau=K}^x]$

$$\mathbf{C} = \mathbf{U}_C\mathbf{D}\ \mathbf{V}^T \tag{3.29}$$

onde, $\mathbf{U}_C \in \Re^{n\times n}$ e $\mathbf{V} \in \Re^{nK\times nK}$ são matrizes ortogonais e $\mathbf{D}$ tem valores não nulos nas posições (i,i), para $1 \leq i \leq n$ e zero nas demais.

(*ii*) Para $i = 1, ...K$, calcular:

$$\mathbf{F}_i = \mathbf{U}_C^T \mathbf{C}_\tau^x \mathbf{U}_R \tag{3.30}$$

(*iii*) Escolha qualquer $\alpha \in \Re^n$ inicial

(*iv*) Calcular:

$$\mathbf{F} = \sum_{i=1}^{K} \alpha_i \mathbf{F}_i \tag{3.31}$$

A característica de SOBI-RO é a sua maior robustez quando há sinais com ruído aditivo com correlação espacial [64]. Esta habilidade do algoritmo pode ser útil em situações reais, nas quais pode haver grande quantidade de ruídos nos dados.

Outra variação no pré-processamento dá origem ao algoritmo SOBI-BPF (*SOBI with bank of Band-Pass Filters*) [66], que utiliza um banco de filtros passa-banda. O algoritmo aplica a diagonalização conjunta sobre dados filtrados em faixas distintas de freqüência. Outras extensões otimizadas de SOBI, tais como WASOBI e EWASABI podem ser encontradas em [67],[68].

Assumindo a hipótese de variâncias não estacionárias, um algoritmo que utiliza apenas estatística de segunda ordem é o SONS (*Second Order Nonstationary Source Separation*) [69]. Este algoritmo permite extrair fontes não estacionárias com estrutura temporal e fontes coloridas com espectros de frequência distintos. Utilizando autocorrelações locais e assumindo que a variância se altera vagarosamente, devemos utilizar uma estimativa local da variância:

$$\hat{\mathrm{E}}_t\{(\mathbf{w}_i^T \mathbf{z}(t))^2\} = \sum_{\tau} \mathrm{h}(\tau)(\mathbf{w}_i^T \mathbf{z}(t-\tau))^2 \tag{3.32}$$

onde, $\mathrm{h}(\tau)$ é o operador de médias móveis normalizado e o subscrito t enfatiza que os sinais são não estacionários.

Assim, chega-se ao algoritmo de atualização da matriz $\mathbf{W}$ [52]:

$$\Delta\mathbf{W} = \sum_{i} \mathrm{diag}(\hat{\mathrm{E}}_t\{\mathbf{w}_i^T \mathbf{z}(t))^2\})^{-1} \mathbf{W}\mathbf{z}(t)\mathbf{z}(t)^T \tag{3.33}$$

3.4.2 EOS

Outra classe de algoritmos ICA refere-se aos métodos que utilizam estatística de ordem superior para extrair as fontes independentes. Um dos mais populares é o algoritmo JADE [60],[70]. O método pode ser entendido como uma generalização da diagonalização de matrizes de covariância, utilizando cumulantes no lugar das matrizes de segunda ordem.

O algoritmo busca uma matriz de separação $\mathbf{W}$ de forma a diagonalizar $\mathbf{F}(\mathbf{M})$

para qualquer matriz $\mathbf{M}$. Em outras palavras, $\mathbf{WF}(\mathbf{M})\mathbf{W}^T$ deve ser diagonal. Assim, dada as matrizes $\mathbf{M}_i$, $i = 1, 2, ...k$, devemos diagonalizar o conjunto de matrizes $\mathbf{Q}_i = \mathbf{WF}(\mathbf{M}_i)\mathbf{W}^T$ tanto quanto possível. Novamente, a diagonalização exata é pouco provável e a Equação 3.11 pode ser utilizada para medir o grau de diagonalização. No entanto, no lugar das matrizes de covariância atrasada, $\mathbf{M}_i$ são as automatrizes do vetor de cumulantes. Assim, pode-se formular a seguinte função objetivo:

$$\Im_{JADE}(\mathbf{W}) = \sum_i ||\text{diag}(\mathbf{Q}_i)||^2 \tag{3.34}$$

onde, $||\text{diag}(.)||^2$ é a soma dos quadrados da diagonal. A maximização de $\Im_{JADE}$ permite a diagonalização conjunta aproximada.

Variações deste algoritmo dão origem a outros métodos, tal como a versão otimizada JADE-op (*Joint Approximate Diagonalization of Eigen matrices with optimized numerical procedures*) [71]. A vantagem do algoritmo é ser livre de parâmetros ajustáveis. Em [72],[73], as matrizes cumulantes são atrasadas no tempo e propõe-se o algoritmo JADETD (*Joint Approximate Diagonalization of Eigen matrices with Time Delays*). Outra variação é o algoritmo FJADE (*Flexible Joint Approximate Diagonalization of Quadricovariance Matrices*). O método é suportado pelos diferentes métodos de diagonalização conjunta: FAJD, FFDIAG, QDIAG e SAD.

Dentre os algoritmos que utilizam estatística de ordem superior, outra classe bastante difundida corresponde aos algoritmos de ponto fixo, conhecidos também como Fast ICA [74],[75],[76]. Em [71], o algoritmo FPICA (*Fixed-Point ICA*) foi implementado para extrair as fontes sequencialmente de forma cega. Após retirar a média e branquear os dados, o algoritmo de ponto fixo ortogonaliza cada um dos vetores $\mathbf{y}_i(t)$ em relação aos demais. Ao ser ortogonalizado, o vetor é fixado e continua-se a iteração. O processo termina quando todos os vetores estão ortogonalizados entre si (não linearmente).

Outras extensões dos algoritmos de ponto fixo são o POWERICA (*Power iteration for ICA*), que foi proposto em [77], e o EFICA (*Efficient Variant of FastICA*) [67],[78]. Outros algoritmos, tais como COMBI & MULCOMBI (*Combination and Multi-combination of WASOBI and EFICA*) [79] combinam estatística de ordem superior com os princípios de SOBI.

Outro grupo de algoritmos que utiliza estatística de ordem superior é o CICA (*Constrained ICA*) e ICA-R (*ICA with Reference*) [80],[81],[82]. Neste grupo, fontes de referência são utilizadas na extração das fontes. Estes algoritmos podem extrair as fontes de forma seqüencial ou simultânea.

Outro grupo importante refere-se aos algoritmos que utilizam o gradiente natural (NG) [30] para extrair as componentes independentes. Em [83], foi proposto

um algoritmo de extração em lote: SANG (*Self Adaptive Natural Gradient algorithm with nonholonomic constraints*). Em [69], propôs-se o algoritmo NG- FlexICA (*Natural Gradient Flexible ICA*) e, em [84], propõe-se o algoritmo Akuzawa, cujo nome é em homenagem ao seu autor. Em [85], propõe-se o algoritmo ThinICA, que extrai simultaneamente um arbitrário número de componentes especificado pelo usuário. O algoritmo é baseado no critério de maximização conjunta do critério de desempenho de cumulantes da saída e/ou matrizes de covariância atrasadas no tempo de segunda ordem. Em [86], o algoritmo ERICA (*Equivariant Robust ICA - based on Cumulants*) utiliza a curtose para separar fontes na presença de ruído gaussiano. O algoritmo é uma iteração quasi-Newton que converge para um ponto com convergência isotrópica local, com relação às distribuições das fontes. O pré-branqueamento não é necessário para o algoritmo convergir. O algoritmo SIMBEC (*Simultaneous Blind Extraction using Cumulants*) [87],[88],[89] otimiza e resolve o problema quando a densidade aproximada dos sinais desejados é conhecida a priori. Este algoritmo estende este resultado para outras funções contraste que não requerem conhecimento explicito da densidade da fonte. A condição necessária e suficiente de estabilidade da densidade local do algoritmo é explorada para obter uma convergência rápida. O algoritmo pode extrair um grupo qualquer de fontes especificadas pelos usuários. O algoritmo UNICA (*Unbiased quasi Newton algorithm for Independent Component Analysis*) [89] é usado para extrair um especifico número de componentes independentes realizando ICA não tendenciosa na presença de ruído gaussiano correlacionado na mistura. O algoritmo realiza uma iteração quasi-Newton para a estimação do sistema de mistura com um critério de resposta da distorção da variância que elimina das saídas os componentes de interferência e todo o ruído que está fora do subespaço de sinal é extraído.

3.5 Aplicações

A ICA tem sido aplicada em diversos contextos. Aplicações para separação de áudio, imagem, mineração de textos e reconhecimento de padrões em séries temporais são alguns exemplos. Em [90], é proposta uma aproximação para a decomposição de vibrações ao longo do tempo. O método se mostrou apto a capturar a estrutura essencial dos dados de vibração medidos. As características da vibração representada fornecem o relacionamento entre a resposta da vibração medida e os componentes independentes. O trabalho reporta a utilidade para aplicações que necessitam de identificação de estruturas, tais como monitoramento médico e detecção de defeitos. O trabalho também destaca a habilidade da técnica de trabalhar com elevado nível de ruído e redução das dimensões. Os resultados mostraram que a decomposição dos sinais de vibração extrai as características repetidas vezes com boa acurácia.

Em [91], ICA é utilizada para extração de ruídos. O trabalho argumenta que devido à independência do ruído com os sinais, o ruído pode ser isolado. Neste trabalho é mostrado que a estrutura sem ruído não foi afetada pela aplicação de ICA.

Em [92], ICA foi utilizada para auxiliar na construção de modelos para séries temporais e a robustez de algoritmos ICA para séries com *outliers* é avaliada. O trabalho avalia a sensibilidade de algoritmos ICA. Em [93], ICA foi utilizada para previsão de séries temporais voláteis. O método introduziu uma rede neural baseada em funções de base radial (RBF), após o pré-processamento com ICA. Em [94], ICA foi utilizada como pré-processamento e Regressão de Vetor Suporte na predição de séries temporais financeiras. Em [95], propôs-se a clusterização de séries temporais baseado em ICA, para analises de séries financeiras. Em [96], uma combinação de ICA e Máquinas de Vetores Suporte (SVM) foram utilizados para a previsão de séries temporais. Em [97], identificou-se sistemas dinâmicos baseado em ICA a partir de séries temporais sintéticas e reais. Em [98], propôs-se um método para a ordenação das fontes obtidas de séries temporais. Em [99] propôs-se um método empírico para selecionar componentes independentes dominantes em análise de séries temporais. Em [100], utilizou ICA em tempo real para séries temporais de imagens de ressonância magnética funcionais (fMRI) e, em [97], realizou-se a identificação de componentes independentes automaticamente para remover ruídos de sinais estruturados em fRMI. Em [101], ICA foi utilizada para previsão de séries temporais multi-variadas. Em [102], técnicas de *data mining* com ICA foram aplicadas para extração de características e, em [103], propôs-se um detector de faltas baseado em técnicas de redução de dimensão.

Um caso especial de ICA ocorre quando há apenas uma única série e múltiplas fontes misturadas. O problema da ICA com apenas uma série-mistura (um canal) é um caso sub-determinado chamado de SCICA (*Single Channel ICA*) [104]. Em [105], SCICA é aplicado em dados de eletro encefalograma e eletrocardiograma. Em [106], esta metodologia, baseada em SCICA é estendida ao problema de múltiplos distúrbios. Mostrou-se que os múltiplos distúrbios se enquadram ao modelo convolutivo SCICA. Análises qualitativas e quantitativas mostraram que, com poucos deslocamentos temporais na mistura monitorada, é possível obter boas estimativas das fontes de distúrbios presentes na mistura. O trabalho mostrou que as linhas da matriz de separação $\mathbf{W}$ constituem vetores de coeficientes de filtros FIR que, ao processar as misturas, levam aos componentes independentes.

3.6 Discussão

No contexto da monitoração da QD em séries temporais, espera-se que as amostras venham dos mais variados contextos. Nesse sentido, ICA pode se revelar como uma

boa solução, uma vez que casos de sucesso têm sido reportados nas mais variadas aplicações. Especificamente, a técnica pode contribuir ao permitir a redução do espaço de variáveis monitoradas e a extração de ruídos. Além disso, espera-se que contribuição fundamental esteja na capacidade de acessar a estrutura essencial dos dados. Assim, utilizando ICA na cadeia de processamento, esperam-se estimativas mais acuradas dos padrões de qualidade dos dados. Alguns trabalhos também reportam a robustez da técnica, que é fundamental para a proposta de um sistema geral para séries temporais.

Neste ponto, uma questão deve ser levantada: em quais situações se deve usar determinada hipótese de separação de fontes? Os diferentes critérios de separação assumem diferentes hipóteses nos dados, sendo assim, a escolha do critério/algoritmo deve depender dos dados analisados. Como uma orientação geral, quando os dados não têm estrutura temporal (ordem arbitrária e sem importância), ICA básico é apropriado. Muitas vezes trabalham bem mesmo quando há estrutura temporal, mas os resultados podem ficar distantes do ótimo. Já dados com clara autocorrelação temporal, no caso de CIs com diferentes autocorrelações, o método mais apropriado seria baseado na autocovariância. No caso de CIs com mesma auto-correlação, os métodos deveriam ser baseados na hipótese de variâncias não-estacionárias. Por fim, a utilização de ICA ainda é inexplorada no contexto da QD, o que nos motiva explorar as particularidades ainda desconhecidas nesta nova situação.

Capítulo 4

Sistema de Monitoração da Qualidade de Dados

Neste capítulo, apresentam-se a metodologia e os conceitos aplicados na especificação/desenvolvimento do Sistema de Monitoração da Qualidade de Dados para Séries Temporais (SMQD-ST). Esse trabalho introduz novos métodos, técnicas e perspectivas de monitoração. Um conjunto de técnicas estatísticas e de processamento de sinais é proposto para a construção de modelos, medição e monitoração da QD. Aplicam-se diversas teorias em torno do conceito de qualidade de dados para propor um sistema robusto e capaz de adaptar-se a diferentes cenários propostos. Como plataforma de desenvolvimento do SMQD-ST, utiliza-se o *software* MatLab®.

4.1 Monitoração da QD

O SMQD-ST é construído com base na necessidade de monitoração de um fluxo de dados entre as fontes geradoras e os usuários dos dados, que desejam receber informações em conformidade com os seus propósitos [31]. No fluxo mostrado na Figura 4.1, os dados são extraídos das fontes, transformados conforme as especificações e armazenados em alguma base de dados. Este conjunto de ações é conhecido como fase *ETL* (*Extraction, Transformation and Load*). Os problemas de QD podem ocorrer em qualquer fase deste fluxo, mas é principalmente na fase *ETL* que ocorrem grande parte dos problemas de qualidade de dados [1]. Ainda, como os dados podem ser utilizados por usuários distintos em contextos variados, a mesma amostra pode ser classificada, ao mesmo tempo, como de alta e baixa qualidade. Assim, o sistema de monitoração deve também permitir que a figura do usuário seja incluída nos modelos de qualidade.

O SMQD-ST pode ser inserido nas diversas fases do fluxo fonte-usuário. No entanto, o quanto antes os problemas forem corrigidos, menor será o impacto da

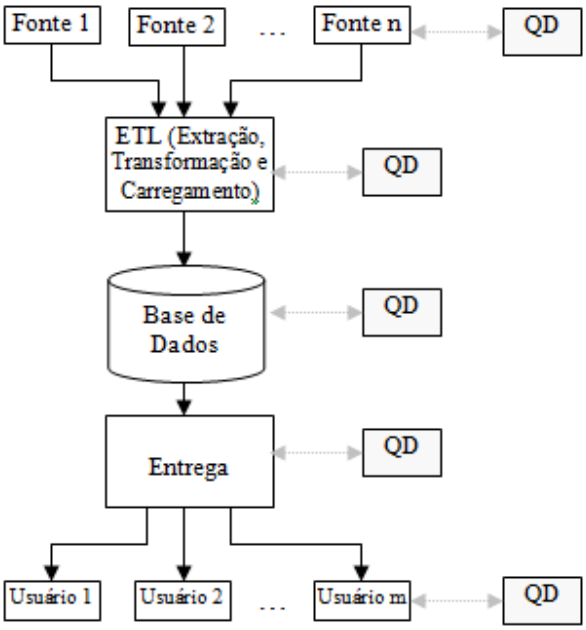

Figura 4.1: Fluxo de dados fonte-usuário.

propagação do erro. Assim, é fundamental monitorar cada amostra entrante o mais cedo possível (*on-line*). Isto não significa que a qualidade dos dados também não deva ser monitorada em lote (*off-line*), quando as amostras formarem uma base de dados. De fato, novas informações podem revelar outras perspectivas que permitem re-classificar a qualidade da amostra.

Para monitorar os dados em tempo real, aplicam-se os fundamentos da teoria de controle para o desenvolvimento do SMQD-ST [36]. Na Figura 4.2, modelo de controle é concebido com base no erro entre a qualidade medida e um padrão de qualidade (referência). A diferença é utilizada pelo sistema para aperfeiçoar os modelos que, dinamicamente, adaptam-se às variações dos dados em tempo real.

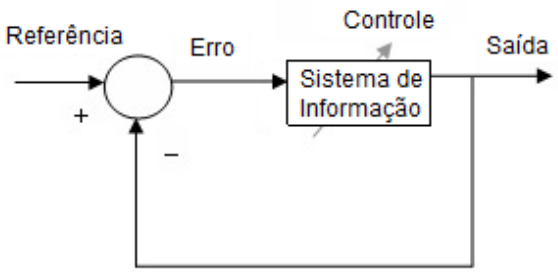

Figura 4.2: Controle clássico.

Os fundamentos da teoria de controle e *ETL* são aplicados no contexto de QD. Na Figura 4.3, a monitoração ocorre em todo o fluxo de dados (*on-line* e *off-line*), das fontes aos usuários finais. Os sensores medem a qualidade com base em uma métrica (veja a Seção 4.2.2) previamente definida. Assim, o nível de qualidade é

mensurado objetivamente e comparado a um valor de referência (qualidade desejada). Havendo diferença (erro) entre os valores desejados e o valor de referência, ações são realizadas no sentido de redução deste erro como, por exemplo, correção ou validação da amostra identificada como suspeita. Assim, a QD é controlada e o ciclo de monitoração se repete.

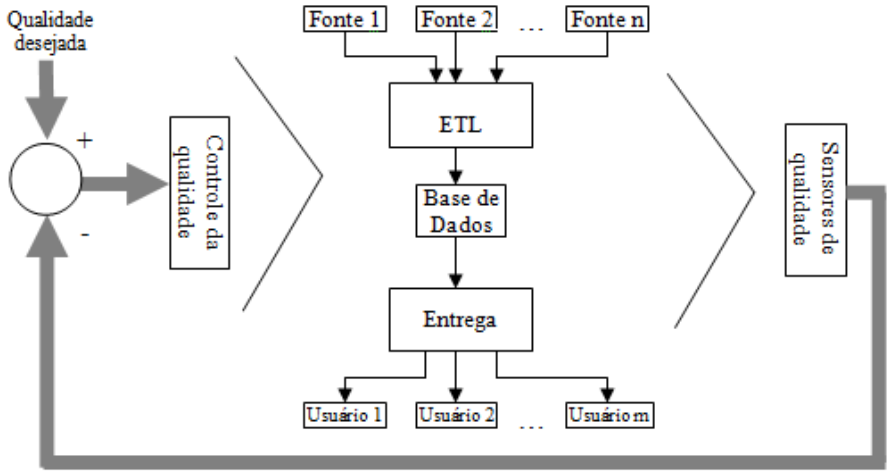

Figura 4.3: Sistema de controle de qualidade de dados.

4.2 Organização do Sistema de Monitoração

Os níveis de organização do sistema de monitoração da qualidade de dados são mostrados na Figura 4.4. No nível mais elevado, o usuário interage com a ferramenta aqui desenvolvida (SMQD-ST). O sistema permite a configuração de parâmetros e a criação de modelos de qualidade de acordo com as especificações dos dados e necessidades do usuário. No segundo nível, são mapeados os problemas de qualidade de dados que serão associados a dimensões e monitorados pelo sistema. No terceiro nível, são definidas as dimensões que comporão a métrica de QD. Um conjunto de dimensões objetivas e subjetivas é definido de acordo com o contexto de séries temporais. Os métodos para medir a qualidade e corrigir as anomalias são agrupados no quinto nível e, no último nível, são mostradas as diversas técnicas utilizadas. Nas seções seguintes, as diversas camadas do sistema de monitoração são abordadas em detalhes.

4.2.1 Problemas

Em séries temporais, frequentemente são encontrados problemas como dados faltantes, erros, valores fora do padrão e valores extremos. O dado faltante pode estar relacionado à inexistência de medição na posição esperada, a perda deste valor em

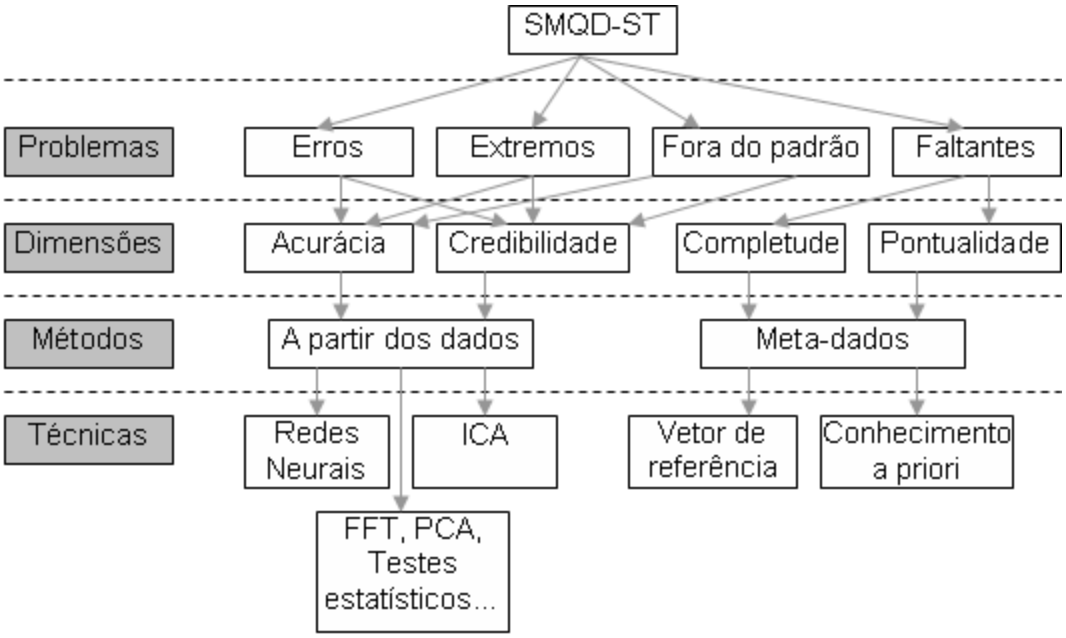

Figura 4.4: Estrutura em camadas para o SMQD-ST.

algum momento do "ciclo de vida" da informação ou a desatualização dos dados. Já os dados fora do padrão e valores extremos, também chamados de *outliers*, podem estar ligados a uma ampla faixa de possibilidades. Estes dados podem representar erros ou mesmo dados corretos que são atípicos. No entanto, mesmo sendo corretos, em muitos casos é preciso detectá-los para evitar problemas. Por exemplo, a presença de valores extremos nos dados poderia prejudicar a construção de modelos. Erros de fato significam que as amostras não deveriam pertencer às séries. No entanto, muitas vezes, estes erros podem seguir o mesmo padrão dos dados e serem difíceis de detectar sob determinadas perspectivas. Dessa forma, múltiplas perspectivas da qualidade dos dados são fundamentais para detectar todo o tipo de problema nos dados. Por fim, nota-se que a dúvida na classificação dos *outliers*, erros e outros problemas nas séries sempre pode existir. Assim, a presença do usuário é fundamental para auxiliar o SMQD-ST na classificação destes problemas.

4.2.2 Dimensões

As dimensões para a medição da QD foram propostas com base nos problemas anteriormente identificados e foram calculadas objetivamente. Mesmo as dimensões classificadas como subjetivas são medidas através de indicadores objetivos. Foram mapeadas 4 dimensões para se compor a métrica: acurácia, credibilidade, completude e pontualidade.

A acurácia mede o grau de aproximação entre os valores amostrados e o padrão de qualidade estimado. A grande dificuldade está em determinar se o valor amostrado

é o valor real, pois muitas incertezas podem estar embutidas na amostra avaliada e na própria referência. Dessa forma, assume-se que o valor de referência tem uma incerteza intrínseca e definem-se limites para determinar o padrão de qualidade. Surge então uma nova questão: qual é a incerteza que define os limites da acurácia? Para responder esta questão, poderíamos recorrer às estatísticas dos dados, assumindo a hipótese de modelos paramétricos para a distribuição do erro produzido pelos modelos de QD. Surge agora a dificuldade de se definir uma função de distribuição para este erro. A distribuição normal poderia ser uma alternativa, no entanto, nem sempre esta hipótese ocorre no mundo real. Já os modelos não paramétricos como, por exemplo, o percentil, podem ser preferíveis nestas situações práticas.

Já a credibilidade tem um viés ainda mais subjetivo do que a acurácia. Por exemplo, em [21], a credibilidade é classificada como uma dimensão contextual. No entanto, sem entrar nesta discussão do nível da subjetividade, podemos medir objetivamente esta dimensão. O mesmo modelo utilizado para medir a acurácia pode ser utilizado para medir a credibilidade da amostra. O que difere uma medida da outra são os níveis de confiança da medida. Por exemplo, amostras cuja diferença em relação ao modelo de qualidade são acima do desvio-padrão até aquele momento podem não estar erradas, no entanto, são menos prováveis. Assim, poderiam ser consideradas suspeitas ou com credibilidade menor do que as amostras dentro do desvio padrão. De fato, definir estes limites cai novamente na subjetividade.

Assim, assumindo-se o papel do usuário dos dados, convencionou-se utilizar apenas o termo acurácia e definiu-se níveis de precisão (alto, médio e baixo) para classificá-la. No entanto, o SMQD-ST ainda permite que estes valores possam ser alterados de acordo com as especificações dos dados, o contexto e a aplicação dos usuários. Caso a distribuição do erro seja uma normal, definem-se a precisão pelos níveis de confiança correspondentes a esta distribuição. Caso seja uma distribuição desconhecida, os níveis de precisão podem ser definidos empiricamente pelos usuários dos dados ou por meio das especificações do contexto.

A completude foi definida apenas pela presença ou não de uma amostra no intervalo de tempo esperado. Para se definir este intervalo, duas alternativas foram utilizadas. A primeira é a partir de conhecimento a priori do comportamento temporal da série. Neste caso define-se um vetor com os períodos de tempo esperados para determinada série (meta-informação). O segundo método utiliza o período de uma série paralela no tempo assumida como completa. Assim, comparam-se os intervalos de tempo da série avaliada com os intervalos da série de referência. Por exemplo, em séries de ações, espera-se que haja amostras somente quando houver pregão na bolsa de valores na qual a ação é negociada. Assim, os valores com as datas de referência do pregão são utilizadas como meta-informação. Já em séries de carga elétrica, espera-se que todos os dias haja consumo de energia. Assim, a

meta-informação é o vetor de datas com todos os dias do ano.

Por fim, a pontualidade pode ser definida simplesmente como a presença da amostra em determinado instante de tempo. Assim, esta dimensão poderia ser avaliada em séries temporais como sendo a presença ou não da amostra referente ao fenômeno físico recente. No entanto, como essa forma de medida também é abrangida pela completude, o conceito de pontualidade poderia ser entendido como estando embutido nesta outra dimensão. Assim, para simplificar as análises aqui realizadas, convencionou-se utilizar apenas a dimensão completude. Ainda, a pontualidade poderia ser medida pela Equação 2.4. No entanto, alguns conceitos como a volatilidade têm um grande viés subjetivo relacionado ao contexto das aplicações. Dessa forma, esta expressão não foi utilizada para a medição da pontualidade.

4.2.3 Métodos

Basicamente, dois métodos podem ser utilizados para se avaliar a qualidade dos dados. Através de modelos extraídos dos próprios dados ou através de informações disponíveis sobre os dados (meta-informação). A meta-informação é independente dos dados e pode ser vista como um conjunto de regras que devem ser satisfeitas quaisquer que sejam os valores das amostras. Já os modelos estatísticos são obtidos a partir dos próprios dados. É desejável que estes modelos sejam também de boa qualidade e obtidos de dados previamente certificados.

Para o sistema de monitoração proposto, os dois métodos foram utilizados. Para a medição da acurácia e da credibilidade, os modelos foram extraídos dos dados. Para a medição da completude e pontualidade, utilizou-se a meta-informação. No caso da completude, para algumas séries estudadas, avaliou-se a presença dos dados através do conhecimento a priori. Para outras séries, avaliou-se a sequência temporal de séries de referência.

As amostras com problemas de QD não devem, a princípio, ser utilizadas para ajustar o modelo dinâmico, pois podem acarretar uma perda na confiança dos modelos utilizados na medição da QD e na própria QD medida. Tampouco, estes dados deveriam ser utilizados para a tomada de decisão. Dessa forma, é preciso criar um banco de dados para o qual as amostras são certificadas. Além disso, a própria informação da QD medida deve ser armazenada. Assim, é possível criar uma série temporal com o próprio índice de qualidade, armazenando evolução histórica da QD.

Corredores de Validação

O principal método de monitoração da QD neste trabalho são os corredores de validação. Estes corredores podem ser utilizados para detectar problemas na acurácia e credibilidade das séries temporais e ainda serem utilizados para preencher valo-

res faltantes ou corrigir amostras identificadas como problemáticas, sempre com a supervisão do usuário.

O corredor de validação é construído dinamicamente a partir da incerteza entre a série amostrada e a série modelada (veja a Figura 4.5). O desvio-padrão do erro (σ_{erro}) entre as séries amostrada e modelada, até o instante de tempo $t-1$, é ajustado por uma constante k que define um *trade-off* entre a probabilidade de falso alarme e a probabilidade de detecção dos problemas. Corredores mais estreitos podem detectar vários tipos de erros, no entanto, tendem a emitir mais falsos alarmes. Por outro lado, corredores mais largos emitem menos falsos alarmes, mas tendem a não detectar todo o tipo de problema. Assim, para adequar as necessidades dos usuários e o contexto onde os dados estão inseridos, são introduzidas as constantes k_1 e k_2 na determinação do corredor (Equação 4.1).

$$Corredor(t) = \begin{cases} x_S(t) = \hat{x}(t) + k_1\sigma_{erro} \\ x_I(t) = \hat{x}(t) - k_2\sigma_{erro} \end{cases} \tag{4.1}$$

O intervalo entre o limite superior $x_S(t)$ e o limite inferior $x_I(t)$ é utilizado para validar a amostra entrante. Os corredores vão sendo gerados dinamicamente à medida que novas amostras validadas são introduzidas na série temporal. Caso um exemplo entrante $x(t)$ esteja fora dos limites de validação (o valor seja faltante), sugere-se a substituição da amostra (ou o preenchimento do valor faltante) pelo padrão de QD ($\hat{x}(t)$), após a análise do usuário.

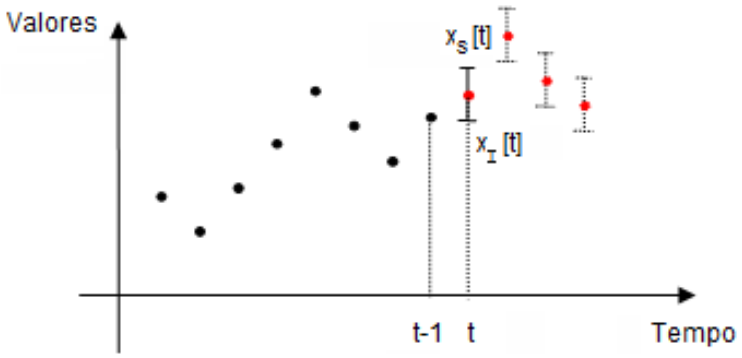

Figura 4.5: Corredor de validação do sistema de monitoração da qualidade de dados.

Existem três possibilidades para as amostras estarem fora do corredor. A primeira possibilidade é o dado estar correto e o seu valor estar fora de um padrão regular. Neste caso, as ações do sistema permitem a supervisão de um especialista que irá decidir o destino das amostras. O sistema de monitoração fornece alternativas de descarte ou reintegração à base de dados. A segunda possibilidade trata de observações efetivamente com erro. O sistema age corretamente ao rejeitá-la e fornece um valor para a substituição da amostra defeituosa. O procedimento também é

acompanhado pelo supervisor, que pode, posteriormente, optar por buscar a amostra na fonte. A última possibilidade é a amostra ficar fora do corredor, apesar de correta. Neste caso, o corredor estaria definido de forma incorreta. Dessa forma, a amostra é reintegrada a base e o modelo deve ser reavaliado e atualizado.

Para se definir os limites dos corredores para a monitoração da QD, voltamos a mesma discussão levantada na definição das dimensões acurácia e credibilidade. Assim, empiricamente, definiram-se valores simétricos de k para monitorar a acurácia nos níveis baixo ($k_1 = k_2 = 3$), médio ($k_1 = k_2 = 2$) e alto ($k_1 = k_2 = 1$). Fixados estes parâmetros, é possível comparar as diversas configurações do sistema.

4.3 Sistema de Monitoração da Qualidade de Dados

A dinâmica de atuação do SMQD-ST avalia a amostra desde a sua chegada até a sua validação/correção e incorporação à base de dados. O primeiro procedimento é a verificação da existência da amostra. Não havendo amostra, cessa a necessidade de se detectar outros problemas. Ou seja, ou o dado existe e deve-se verificar se há alguma anomalia, ou não existe e deve ser tratado como dado faltante ou não atualizado. O procedimento permite a monitoração em tempo real (*on-line*) das amostras entrantes, o que é fundamental para evitar propagação de erros.

A estrutura do sistema é mostrada na Figura 4.6. Inicialmente, avalia-se a presença de *outliers* nas séries $\mathbf{x}(t)$ (bloco OUT - Figura 4.6). Estas amostras podem ser retiradas da série e substituídas por uma estimativa. Esta etapa será explicada em maiores detalhes na seção seguinte. Na sequência, as séries têm a média retirada e são branqueadas (bloco Bran - Figura 4.6), projetando-se as séries em eixos ortogonais. Nesta fase, é possível também considerar apenas as projeções de maior energia (componentes principais). Na metodologia proposta, apenas o branqueamento é realizado.

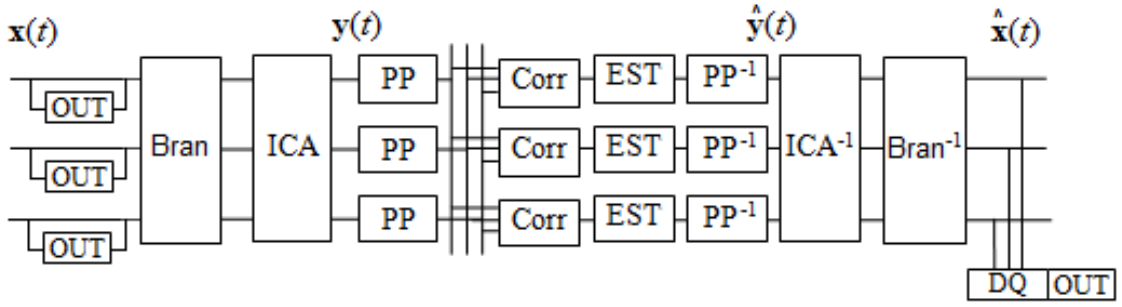

Figura 4.6: Estrutura do Sistema de Monitoração da Qualidade de Dados.

Na etapa seguinte, aplica-se a Análise de Componentes Independentes (bloco

ICA - Figura 4.6). O sistema tem à disposição um banco de algoritmos para a extração de fontes independentes. Dentre eles, destacam-se algoritmos que utilizam estatística de ordem superior como, por exemplo, FastICA e JADE e algoritmos que utilizam estatística de segunda ordem para séries temporais como, por exemplo, AMUSE, SOBI e SOBI-RO. Os sinais apresentados na entrada do bloco ICA são assumidos misturas de fontes independentes. Dessa forma, o processamento ICA separa as fontes $\mathbf{y}(t)$ ocultas. Nesta etapa, além das fontes independentes, ruídos podem ser isolados e extraídos (deflação).

Após ICA, realiza-se um pré-processamento clássico nas fontes independentes (bloco PP - Figura 4.6). O pré-processamento busca normalizar, tornar estacionária e encontrar componentes frequentemente presentes nas séries temporais [1], para se tornarem adequadas aos métodos de modelagem. Primeiramente, é verificado se há variação no comportamento da variância ao longo do tempo (heteroscedasticidade) [107]. Em caso afirmativo, uma ação apropriada é considerada, como a aplicação da função logarítmica, função quadrática ou modelando o crescimento da variância por um polinômio ou uma função exponencial. Com a série homoscedástica, verifica-se a presença de raízes unitárias através de uma combinação dos testes de Dickey-Fuller aumentado (ADF) [108] e Phillips-Perron [109]. O objetivo é identificar a presença de tendência estocástica. Se raízes unitárias forem detectadas, significa que a tendência é estocástica e diferenças sucessivas devem ser aplicadas n vezes, para torná-la estacionária (onde n é o número de raízes unitárias ou ordem de integração do processo). Se o teste não detecta raízes unitárias, avalia-se a presença de tendência determinística. A análise também pode ser feita visualmente ou a partir de um conhecimento a priori dos dados. Caso este componente seja identificado, ajusta-se uma função polinomial ou exponencial nos dados. Após a retirada das tendências, verifica-se a presença de sazonalidades e de ciclos, através da análise visual do especto de Fourier e da função de autocorrelação da série [1]. Detectados ciclos, estes podem ser removidos anulando-se a componentes de frequência do espectro de Fourier. Sazonalidades são extraídas pelo operador de diferença para o período de sazonalidade encontrado ou pela retirada das componentes de frequência da transformada de Fourier.

Outros testes também podem ser aplicados, como o teste de Goldfeld-Quandt[110] para se detectar a heteroscedasticidade, o teste de hipótese no erro produzido pelo *fitting* de tendência determinística e o teste de hipótese para identificação de ciclos relevantes no espectro de Fourier [111]. No entanto, os testes devem ter a metodologia avaliada para cada tipo de série e os resultados devem ser comparados com a análise do comportamento temporal observado. Assim, apesar de fornecerem um parâmetro objetivo, para alguns casos, acabam confundindo análise do usuário dos dados. Dessa forma, para o sistema proposto, optou-se por

não incluir estes testes no pré-processamento.

A Figura 4.7 mostra o fluxo do pré-processamento. Este procedimento foi proposto em [1] e foi implementado de forma sistemática no SMQD-ST. O sistema aqui proposto permite que a ordem de cada etapa seja também alterada de acordo com as análises do usuário. Isso dá uma maior flexibilidade para tratar de forma adequada diferentes tipos de séries temporais.

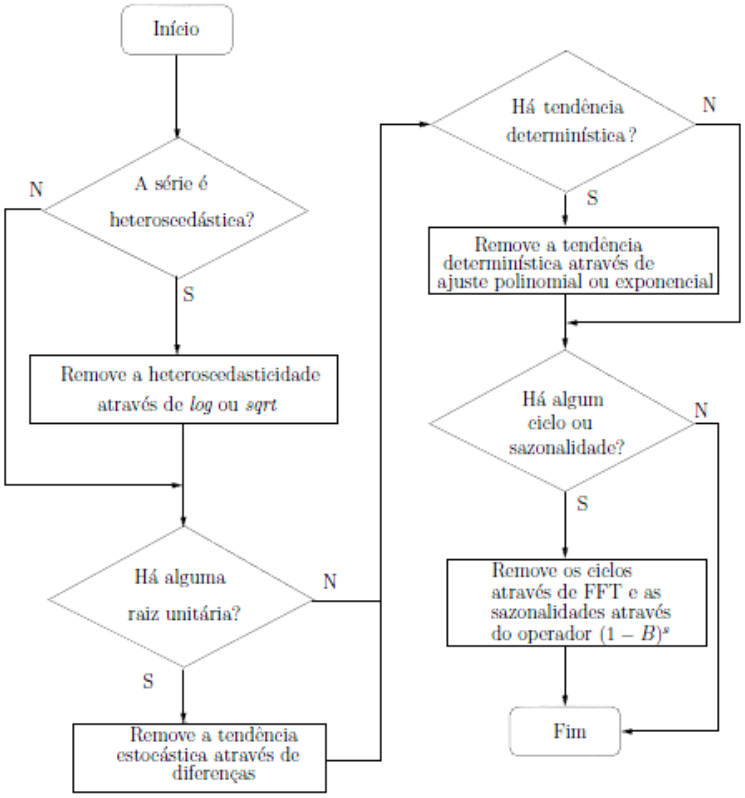

Figura 4.7: Pré-processamento clássico para tornarem estacionárias as séries temporais. Adaptado de [1]

Na sequência, a correlação entre os resíduos do pré-processamento e o alvo da modelagem são avaliados (bloco Corr - Figura 4.6). Ambas as funções de correlação linear e não linear devem ser analisadas. Quando ICA é introduzida, o bloco Corr deveria tornar-se irrelevante, uma vez que ICA obtém componentes descorrelacionadas e independentes. No entanto, em situações não ideais, as hipóteses assumidas para a obtenção de fontes independentes podem não ser satisfeitas. Isso pode ocorrer, por exemplo, quando a hipótese de quantidades iguais de fontes e séries misturas não for verdadeira. Assim, tanto os atrasos relevantes do alvo quanto os atrasos relevantes das outras fontes podem ser utilizados para estimação do modelo. Para tal, aplica-se um teste de hipótese para se verificar valores relevantes na função de correlação (linear e não-linear).

O estimador (bloco EST - Figura 4.6) é composto por diversos métodos que

podem ser ajustados às especificações e características de cada série. Dado que a proposta do SMQD-ST é ser utilizado por uma ampla faixa de séries temporais, modelos parcimoniosos são propostos e, criteriosamente, a complexidade é elevada. Se o teste da correlação (linear e não linear) não detectar nenhum atraso relevante ou série explicativa, assume-se que não há estruturas definidas na série e nenhum modelo mais sofisticado é elaborado. Neste caso, a série poderia ser classificada como um ruído. Assim, ou se assume o melhor estimador de caminho aleatório (*The Best Random Walk Estimator*) ou estima-se o modelo simplesmente pela média das amostras anteriores. No caso de haver alguma correlação significativa na entrada, estimam-se modelos lineares ou não lineares. Os modelos são implementados por uma rede neural de múltiplas camadas (MLP) [112],[113] (veja a Figura 4.8). No caso de modelos lineares, utiliza-se apenas um ponderador linear no lugar do neurônio.

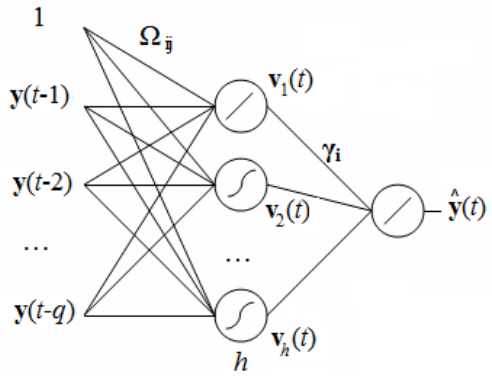

Figura 4.8: Rede Neural MLP

Matematicamente,

$$\hat{\mathbf{y}}(t) = \gamma_1 \mathbf{v}_1(t) + \sum_{i=2}^{h-1} \gamma_i \mathbf{v}_i(t) + \gamma_h \mathbf{v}_h(t) \tag{4.2}$$

onde,

$$\mathbf{v}_1(t) = \sum_{j=1}^{q} \Omega_{1j} \mathbf{y}(t-j) + \Omega_{10} \tag{4.3}$$

$$\mathbf{v}_i(t) = \mathrm{f}\left(\sum_{j=1}^{q} \Omega_{ij} \mathbf{y}(t-j) + \Omega_{i0} \right) \tag{4.4}$$

onde h é o número de neurônios escondidos, q é o atraso e f(.) é uma função não-linear, tipicamente a tangente hiperbólica.

O teste para a seleção do modelo [113] avalia a hipótese $\mathrm{H}_0 : \gamma_i = 0$, onde, $i = 1, 2, ...h$ e $j = 0, 1, ...q$. Para se chegar ao modelo mais adequado, parte-se novamente da hipótese de modelos parcimoniosos. Inicialmente, estima-se um simples modelo

linear e um modelo não linear com um neurônio na camada escondida e verificam-se as hipóteses:

$$\mathrm{H}_0 : \text{Modelo Linear } (\gamma_2 = 0) \tag{4.5}$$

$$\mathrm{H}_1 : \text{Modelo Não Linear } (\gamma_2 \neq 0) \tag{4.6}$$

Para a verificação, avalia-se a razão de semelhança do erro médio quadrático (EMQ) produzido por cada um dos modelos.

$$r = \frac{EMQ_{\mathrm{H}_0}}{EMQ_{\mathrm{H}_1}} \tag{4.7}$$

Se a razão for igual a 1, para um dado nível de significância (tipicamente 95%) do teste estatístico (chi-quadrado [113]), rejeita-se H_1, pois o erro não se altera quando a estimação é feita pelo modelo mais complexo. Assim, o teste é imediatamente paralisado e o modelo parcimonioso é utilizado para estimar os padrões de QD. Caso contrário, assume-se que o modelo é não linear com um neurônio na camada escondida e aplica-se o segundo teste de hipótese (veja as expressões 4.8 e 4.9). O número de neurônios é incrementado até que a hipótese H_0 não seja mais rejeitada.

$$\mathrm{H}_0 : h \text{ neurônios } (\gamma_{h+1} = 0) \tag{4.8}$$

$$\mathrm{H}_1 : h + 1 \text{ neurônios } (\gamma_{h+1} \neq 0) \tag{4.9}$$

Após a modelagem do resíduo, todos os componentes retirados no pré-processamento são adicionados novamente nas séries (bloco PP^{-1} - Figura 4.6), com exceção dos *outliers*. Em seguida, os padrões de qualidade são estimados e mapeados para o espaço original, com a aplicação de ICA e branqueamento inversos (blocos ICA^{-1} e Bran^{-1} - Figura 4.6). Assim, definem-se os padrões de QD no espaço das séries originais. Os erros entre os padrões de QD e os valores amostrados são utilizados na construção dos corredores de validação.

Neste ponto, nota-se que é possível monitorar a qualidade da série nas diversas etapas do SMQD-ST (antes de PP^{-1}, antes de ICA^{-1} ou antes de Bran^{-1} - Figura 4.6), criando diferentes perspectivas para a qualidade dos dados. No entanto, para manter o entendimento físico dos padrões de QD, as análises deste trabalho são realizadas principalmente no espaço original das séries monitoradas (após Bran^{-1} - Figura 4.6).

4.3.1 Detecção de *Outliers*

Na estrutura do SMQD-ST (veja a Figura 4.6), é mostrado que, antes da amostra ser utilizada na construção do padrão de QD, verifica-se se a amostra não se trata

de um *outlier*. Esta identificação pode também ser realizada sem este bloco, pelo próprio corredor produzido para a monitoração das séries. Para este fim, o corredor é configurado de forma a detectar, além dos problemas de QD, também os *outliers* nas séries. Os corredores podem ser configurados para encontrar valores extremos na séries, ajustando-se as constantes k_1 e k_2. Tipicamente, este valor é ajustado para se detectar amostras afastadas três desvios padrão do modelo de QD.

O método anterior busca *outliers* a partir do conhecimento do padrão de QD. Propõe-se também outra metodologia para detectar *outliers* a partir de fontes desestruturadas . Para este fim, utiliza-se uma adaptação de ICA para obtenção de fontes em canais únicos (*Single Channel ICA* - SCICA). O método insere atrasos nas séries individuais (bloco $t-1$ - veja a Figura 4.9) e aplica os algoritmos de ICA (bloco ICA). Com apenas dois atrasos, obtêm-se duas fontes com ICA: fonte estruturada e fonte desestruturada. A primeira fonte concentra os componentes estruturados da série e é descartada pelo método (deflação - bloco Def). Apenas a segunda fonte, que contém ruídos e *outliers*, é avaliada. Nesta fonte, constroem-se os corredores de validação, após o pré-processamento (Bloco PP) e a estimação dos modelos (bloco EST). Agora, os corredores são construídos no espaço de fontes independentes e não se faz a transformação inversa de ICA.

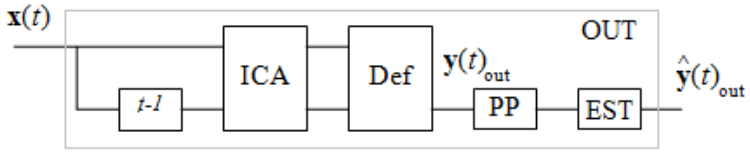

Figura 4.9: Método de detecção de *outliers* com SCICA

4.4 Algoritmos ICA

As fontes foram separadas utilizando-se os algoritmos AMUSE [59], SOBI [114], SOBI-RO [70],[114], FastICA [74],[75],[76] e JADE [55]. Os três primeiros são testados por utilizarem estatística de segunda ordem e se valerem da informação sequencial entre as amostras para obter os componentes independentes. Os dois últimos são incluídos nas análises por serem algoritmos básicos de ICA largamente utilizados, incluírem estatística de ordem superior na separação de fontes e não considerarem a sequência temporal. O objetivo é avaliar o comportamento destes métodos na separação das fontes independentes e analisar a contribuição para a monitoração da qualidade de dados.

Os algoritmos SOBI, SOBI-RO e AMUSE estimam as componentes independentes diagonalizando matrizes de covariância atrasadas no tempo. Para o método mais

simples (AMUSE), utiliza-se apenas uma matriz atrasada. O parâmetro de atrasos (τ) é definido a partir de um indicador de separação, após avaliar-se todos os atrasos possíveis, dado o número de amostras da série. O mesmo critério foi adotado para se escolher o conjunto de matrizes atrasadas para SOBI e SOBI-RO.

Para o algoritmo JADE, estimam-se matrizes cumulantes de quarta ordem que, em seguida, são diagonalizadas com rotações de Jacobi. Além disso, não se espera que o modelo ICA se mantenha exatamente constante ao longo do tempo. Poderia ser uma perda de foco tratar precisamente o problema, pois é pouco provável que o tensor cumulante possa ser acuradamente resumido pelas auto-matrizes, principalmente quando há ruídos nas séries. Então, considera-se a diagonalização simultânea de todo o conjunto de auto-matrizes. No entanto, não é necessário calcular o conjunto completo de auto matrizes porque se pode utilizar equivalentemente frações do tensor cumulante. Exploram-se as simetrias dos cumulantes para mais adiante reduzir o número de matrizes a serem diagonalizadas. Estas considerações levam à versão otimizada do algoritmo JADE utilizado nos nossos testes.

Já o algoritmo fastICA utilizado estima cada componente independente de forma sequencial. Um vetor inicial é definido aleatoriamente e ortogonalizado em relação aos outros. São estimadas tantas fontes quanto são as séries misturas. A não linearidade utilizada para extração de fontes é a potência cúbica $g(u) = u^3$.

4.5 Separação dos conjuntos de desenvolvimento e teste

O método para a formação dos conjuntos para desenvolvimento (treino e validação) e teste do modelo baseou-se no procedimento de janela-móvel [115]. Este procedimento consiste em re-treinar a rede com os dados mais atualizados da série. Uma vez que o sistema está continuamente recebendo novas amostras, o re-treino ocorre com estas amostras mais recentes.

Consequentemente, novas características são incorporadas ao modelo. A Figura 4.10 ilustra este processo. Na janela superior, a série é dividida em vários conjuntos de desenvolvimento. Em cada um desses conjuntos, para o caso da utilização de redes neurais, divide-se uma parte das amostras para treino e a outra para validação da rede neural. Apenas o último conjunto de amostras é utilizado para testar o modelo gerado. Em um momento posterior, havendo a necessidade de re-treino do modelo, a janela é deslocada no tempo para incorporar as amostras mais recentes. Assim, o mesmo procedimento para desenvolvimento/teste é realizado novamente nesta nova janela.

O SMQD-ST deixa para o supervisor a decisão da escolha do melhor momento

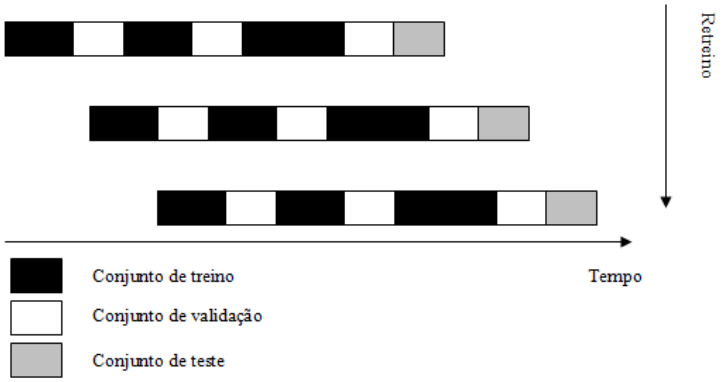

Figura 4.10: Esquema para o procedimento de janelamento móvel.

de re-treino. Já a proporção dos conjuntos de treino e validação são fixados em, respectivamente, 70% e 30% das amostras disponíveis de desenvolvimento. Optou-se por fixar esta proporção em valores típicos devido ao grande número de variáveis a serem definidas no SMQD-ST. Já o conjunto de testes é configurado caso a caso de acordo com a necessidade de análise para cada uma das séries aqui estudadas. Utilizam-se as amostras mais recentes para comporem este conjunto.

4.6 Metodologia de Teste

O sistema de monitoração da qualidade de dados foi inicialmente desenvolvido e avaliado com séries sintetizadas. A avaliação dos métodos em condições controladas permite conhecer detalhes que dificilmente poderiam ser observados em situações reais. Neste ambiente simulado, as fontes independentes são conhecidas e é possível avaliar o desempenho da ICA e dos demais métodos utilizados, em uma situação ideal de extração (não cega) de fontes.

As fontes independentes foram extraídas através de métodos frequentemente utilizados para séries temporais e através de métodos ICA de uso geral. Esta avaliação busca estabelecer métodos ICA mais apropriados para a monitoração da qualidade de dados em séries temporais. Espera-se que o sistema recupere as estruturas das fontes originais e facilite a monitoração da qualidade de dados, obtendo modelos de qualidade mais acurados. Os testes também foram realizados em ambiente ruidoso e com *outliers*, para avaliar o comportamento dos métodos em um cenário com padrões pouco definidos e complexidade elevada. Foram avaliadas matriz de separação, as fontes no tempo e as distribuições de probabilidade das fontes.

No espaço ICA, as fontes são pré-processadas e o padrão de qualidade é obtido. As fontes são obtidas com os algoritmos mais promissores encontrados na análise de separação de fontes. Para efeitos de comparação, a metodologia é aplicada também sem o bloco ICA. Nestes cenários, o método de monitoração é avaliado e a qualidade

das séries e a qualidade dos modelos são medidos e comparados.

O sistema de monitoração da qualidade de dados foi desenvolvido e avaliado também sob o cenário com séries reais. A utilização destas séries permite avaliar o sistema sob o ponto de vista de situações não controladas e avaliar a extração das fontes independentes de forma cega. Neste ambiente, séries temporais com comportamentos distintos são estudadas.

Espera-se que o sistema de monitoração da QD seja capaz de adaptar-se às diversas situações apresentadas. Além disso, é possível analisar em quais situações e sob quais configurações a contribuição de ICA pode ser mais significativa. Espera-se que a introdução de ICA melhore o desempenho da monitoração da qualidade de dados. O sistema de monitoração da QD foi testado em dois contextos distintos: elétrico e financeiro.

As séries de carga elétrica foram divididas de duas formas distintas. A primeira configuração selecionou séries adjacentes ao horário de pico, reproduzindo uma monitoração de múltiplas séries. A segunda, utilizou apenas as séries de picos e de temperaturas, reproduzindo a monitoração de séries individuais com séries explicativas. Essa configuração também foi utilizada para efeitos de comparação com os modelos de uma competição.

As séries financeiras também foram configuradas de duas formas: com séries de um mesmo setor econômico e séries de uma empresa individual. As séries escolhidas se basearam em configurações de trabalhos anteriores [1] e, assim, foram utilizadas para comparação de resultados.

Capítulo 5

Séries Sintéticas

Neste capítulo, são apresentados os resultados obtidos com as séries sintéticas. Primeiramente, mostram-se os processos de geração das fontes e das séries-mistura. Em seguida, simula-se a separação cega de fontes. São mostrados estudos sobre a matriz de separação, a fonte no tempo e a distribuição de probabilidades das amostras. Ainda, a separação de fontes é avaliada em ambientes ruidosos e na presença de *outliers*. Os resultados obtidos com diversos algoritmos de ICA são comparados com os valores verdadeiros.

Após a separação de fontes, são mostrados os resultados do pré-processamento e da estimação dos modelos nos cenários com ICA, sem ICA e para o caso ideal de separação de fontes. Em seguida, o sistema é configurado para a detecção de *outliers* e avalia-se o impacto da introdução do método SCICA.

Por fim, são apresentados os resultados da monitoração e medição da QD. São mostrados os corredores de validação, os indicadores do modelo e os indicadores da qualidade de dados, num ambiente com e sem *outliers*. Os resultados também são comparados com e sem ICA.

5.1 Dados

As séries temporais podem ser definidas apenas como uma sequência de observações realizadas em intervalos de tempo [116]. No entanto, a sequência dos exemplo armazena informações que as tornam uma classe particular de dados. Além disso, algumas características são frequentemente observadas em diversos contextos. Assim, simularam-se séries com os principais comportamentos [22] presentes nas séries temporais, assim como tendências (estocástica e determinística), ciclos, sazonalidades e heterocedasticidades [1] [22] [116]. Esses componentes foram simulados como fontes independentes e, em cada uma delas, sinais senoidais foram adicionados. Optou-se por utilizar estes sinais por se observar frequentemente fenômenos na natureza com comportamentos similares a senoides.

Muitas séries temporais também apresentam uma componente estocástica marcante, tornando-se impossível prever exatamente o seu comportamento. Assim, em cada uma das fontes, adicionou-se um componente branco gaussiano ($N[0, \sigma]$), descorrelacionado espacial e temporalmente. Outra característica presente em situações reais são os ruídos, que aumentam a dificuldade na obtenção de padrões. Assim, uma das fontes foi simulada a partir de amostras retiradas de uma distribuição gaussiana aleatória. As fontes $\mathbf{s}[n]$ simuladas são descritas a seguir:

Tendência determinística:

$$\mathbf{s}_1[n] = an + b + A_0 sen(\omega_0 n) + N[\mu, \sigma^2] \tag{5.1}$$

Tendência estocástica:

$$\mathbf{s}_2[n] = a\mathbf{s}_2[n-1] + b + A_0 sen(\omega_0 n) + N[\mu, \sigma^2] \tag{5.2}$$

Sazonalidade:

$$\mathbf{s}_3[n] = a\mathbf{s}_3[n-ps] A_0 sen(\omega_0 n) + N[\mu, \sigma^2] \tag{5.3}$$

Ciclos:

$$\mathbf{s}_4[n] = A_1 sen(\omega_1 n) + A_2 sen(\omega_2 n) + A_3 sen(\omega_3 n) + N[\mu, \sigma^2] \tag{5.4}$$

Heterocedasticidade:

$$\mathbf{s}_5[n] = A_0 e^{n/T} sen(\omega_0 n) + N[\mu, \sigma^2] \tag{5.5}$$

Ruído:

$$\mathbf{s}_6[n] = N[\mu, \sigma^2] \tag{5.6}$$

onde a, b, c e d são os coeficientes das equações, A_i é a amplitude do sinal senoidal, $\omega_i = \frac{2\pi f_i}{Amostragem}$ são as frequências angulares (para $i = 0, 1, 2, 3$), ps é o período de sazonalidade e $N[\mu, \sigma^2]$ é o termo estocástico.

É importante que as energias das fontes não sejam demasiadamente discrepante umas das outras. Dessa forma, quando forem misturadas, garante-se que nenhuma fonte desbalanceie as misturas sintéticas. Assim, os parâmetros das séries foram ajustados empiricamente, através da observação das fontes resultantes, e normalizadas logo em seguida (veja Tabela 5.1). Ainda, dois conjuntos com a mesma quantidade de amostras (500 cada) foram simulados, um para desenvolvimento e outro para teste. A Figura 5.1 mostra as fontes simuladas.

A característica da tendência é apresentar um crescimento (ou decrescimento) ao

longo do tempo. Assim, a tendência determinística foi simulada por um polinômio de primeira ordem, simulando um crescimento linear do sinal. A tendência estocástica foi simulada através de inovações sobre a amostra imediatamente anterior.

A característica da sazonalidade é a repetição de padrões em intervalos regulares de tempo. Assim, a sazonalidade foi simulada através de senoides regulares (com frequência de 10 ciclos por amostragem) e inovações sobre as amostras passadas, para um período de sazonalidade (ps) igual a 5.

Os ciclos também têm a característica de repetição de padrões ao longo do tempo. No entanto, os períodos de tempo costumam ser maiores e mais irregulares que o padrão de sazonalidade. Dessa forma, os ciclos foram simulados através de senoides de baixa frequência, simulando longos ciclos irregulares. O ciclo foi simulado com frequências de 2, 2,2 e 2,4 ciclos por amostragem.

A heterocedasticidade é a alteração da variância em diferentes instantes de tempo. Em séries temporais, frequentemente observa-se o crescimento da variância ao longo do tempo, segundo algum fenômeno presente na formação da série. Este fenômeno foi simulado através de um crescimento exponencial da amplitude do sinal temporal, com o coeficiente de crescimento 1/T.

A característica do ruído é não ter correlação temporal nem espacial entre as amostras. Assim, amostras extraídas de uma distribuição gaussiana, gerada aleatoriamente, foram utilizadas para simular o ruído.

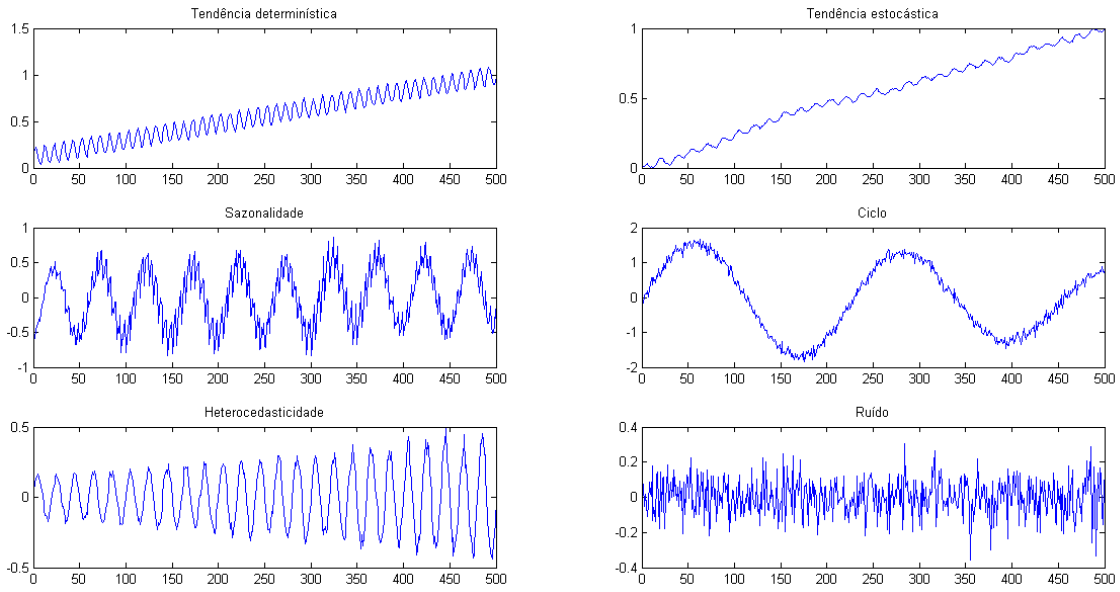

Figura 5.1: Fontes simuladas. Da esquerda para a direita e de cima para baixo, tem-se: tendência determinística, tendência estocástica, sazonalidade, ciclo, heterocedasticidade e ruído branco gaussiano.

Tabela 5.1: Parâmetros das séries sintéticas

Parâmetros	$\mathbf{s}_1[n]$	$\mathbf{s}_2[n]$	$\mathbf{s}_3[n]$	$\mathbf{s}_4[n]$	$\mathbf{s}_5[n]$	$\mathbf{s}_6[n]$
a	1	1				
b	70	0,3				
A_0	0,1	0,02	0,4		0,2	
A_1				1		
A_2				1		
A_3				1		
f_0	50	30	10		25	
f_1				2		
f_2				2,2		
f_3				2,4		
ps			5			
σ	0,1	0,5	0,1	0,1	0,1	[0,01, 4]
Normalização	min=0	min=0	μ=0	μ=0	μ=0	μ=0
	max=1	max=1	σ=1	σ=1	σ=1	
SNR (dB)	7,85	7,66	5,94	9,92	2,93	0
Exemplos	500	500	500	500	500	500

Sinais mistura

Para obter as séries temporais simuladas $\mathbf{x}(t)$, as fontes $\mathbf{s}(t)$ são misturadas através de uma transformação linear:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{5.7}$$

onde $\mathbf{A}$ (matriz-mistura) é a matriz quadrada cujos elementos são obtidos de forma aleatória, a partir de uma distribuição uniforme no intervalo [-1, 1].

$$\mathbf{A} = \begin{bmatrix} -0,56 & 0,28 & -0,18 & 0,46 & -0,31 & -0,81 \\ 0,45 & -0,99 & 0,63 & 0,91 & -0,88 & -0,10 \\ -0,86 & -0,32 & 0,74 & 0,31 & 0,43 & 0,73 \\ 0,93 & -0,45 & -0,95 & 0,48 & 0,92 & -0,22 \\ -0,58 & -0,91 & 0,45 & -0,31 & -0,68 & -0,49 \\ -0,68 & -0,81 & 0,69 & 0,77 & -0,17 & -0,29 \end{bmatrix}$$

A mistura aleatória permite que os componentes independentes sintetizados sejam ocultados nas séries resultantes. Na Figura 5.2, cada uma das séries sintéticas, também chamadas de sinais-mistura ($\mathbf{x}_1(t)$ a $\mathbf{x}_6(t)$), concentra uma fração das fontes. A fonte de ruído $\mathbf{s}_6(t)$ foi simulada com baixa intensidade ($\sigma = 0,01$) e pode ser utilizada para controlar o nível de ruído nas séries-mistura.

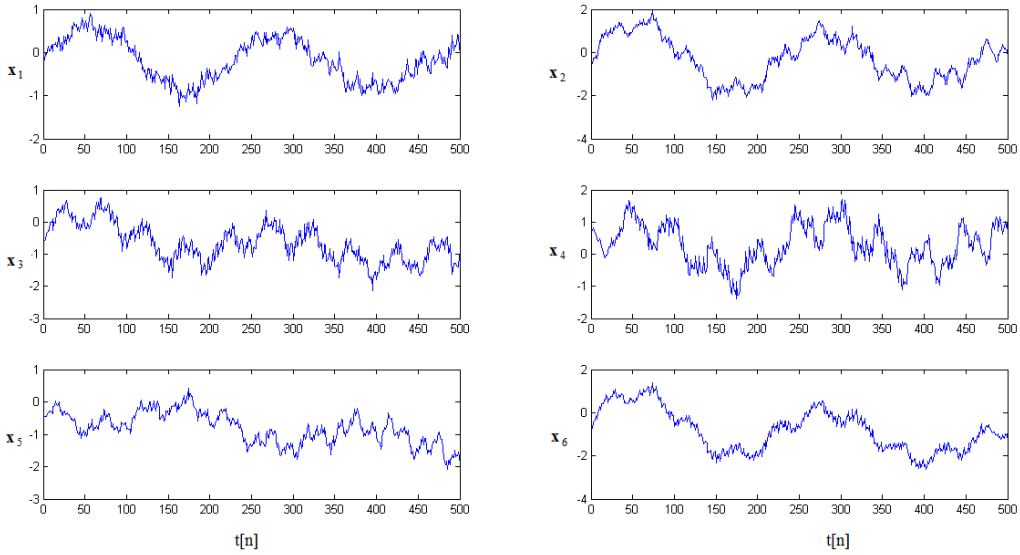

Figura 5.2: Sinais-mistura simulados. Da esquerda para a direita e de cima para baixo, tem-se: $\mathbf{x}_1(t), \mathbf{x}_2(t), \mathbf{x}_3(t), \mathbf{x}_4(t), \mathbf{x}_5(t)$ e $\mathbf{x}_6(t)$.

5.2 Separação de Fontes

As séries-mistura apresentadas na seção anterior são a base para a aplicação da metodologia ICA. Princípios de separação comumente utilizados na extração cega de fontes e princípios que exploram propriedades das séries temporais são aplicados na estimação da matriz separação $\mathbf{B}$. Dessa forma, estimam-se as fontes $\mathbf{y}(t)$ a partir das séries-mistura observadas $\mathbf{x}(t)$, a partir da seguinte expressão:

$$\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t) \tag{5.8}$$

onde, $\mathbf{B} = \hat{\mathbf{A}}^{-1}$ e $\mathbf{y} = \hat{\mathbf{s}}$.

As fontes foram separadas utilizando-se os algoritmos AMUSE [59], SOBI [114], SOBI-RO [114][70], FastICA [75][74][76] e JADE [55]. O primeiro algoritmo foi utilizado pela simplicidade na separação de fontes. O segundo foi escolhido por fazer uma extração de fontes clássica para séries temporais. Já o SOBI-RO é uma extensão de SOBI adaptada para ambientes ruidosos e, em situações práticas, espera-se encontrar situações hostis como esta. Já os dois últimos algoritmos, que utilizam estatística de ordem superior, são algoritmos populares de ICA padrão que têm mostrado bom desempenho em outros contextos. Com estes algoritmos, as fontes foram avaliadas sob três aspectos: matriz de separação, fontes no tempo e distribuição de probabilidade das fontes.

5.2.1 Avaliação da Matriz de Separação

Uma vez que a matriz $\mathbf{A}$ é conhecida, ao estimar $\mathbf{B}$, é possível avaliar o desempenho dos algoritmos, através do índice de separabilidade E_1 [30]:

$$E_1 = \sum_{i=1}^{m}(\sum_{j=1}^{m}\frac{|p_{ij}|}{\max_k |p_{ik}|} - 1) + \sum_{j=1}^{m}(\sum_{j=1}^{m}\frac{|p_{ij}|}{\max_k |p_{kj}|} - 1) \tag{5.9}$$

onde p_{ij} são os ij-ésimos elementos da matriz $\mathbf{P}_{m\times m} = \mathbf{BA}$.

Se as fontes foram separadas corretamente, $\mathbf{P}$ torna-se a matriz permutação (onde os elementos podem ter diferentes sinais, no entanto). Como a matriz permutação tem somente um dos elementos igual à unidade em cada uma de suas linhas e colunas, enquanto todos os demais elementos são zero, pode-se notar que o índice E_1 deve atingir o valor mínimo (zero) para o caso ideal. Quanto maior o valor de E_1, mais pobre é o desempenho estatístico do algoritmo.

A Tabela 5.2 mostra o índice de desempenho E_1 obtido para cada um dos algoritmos de ICA avaliados. No caso ideal, o indicador é zero, pois as fontes comparadas são idênticas e E_1 é calculado sob a matriz identidade. Para SOBI, SOBI-RO e AMUSE, é preciso definir o atraso τ (veja a Seção 3.4.1). Dessa forma, para evitar que a avaliação fosse tendenciosa, calcularam-se os indicadores para todos os atrasos possíveis (499 atrasos) e a média foi utilizada para comparação entre eles. Além disso, o desvio padrão (colocado entre parêntesis) de E_1 foi utilizado para avaliar o grau de dificuldade da escolha de τ. Observa-se que o menor valor de E_1 foi atingido com o algoritmo SOBI. O desvio-padrão menor indica que a escolha do atraso τ é menos crítica. O algoritmo SOBI-RO é um pouco pior quando é comparado com SOBI. O ganho de desempenho de SOBI é devido ao menor nível de ruído presente nas séries. Os algoritmos AMUSE, FastICA e JADE têm o pior desempenho, apresentando indicadores mais elevados. O pior desempenho de JADE e FastICA ocorreu pelo fato de desprezar a informação da sequência temporal das amostras. Já AMUSE, apesar de utilizar esta informação, é muito volátil (variância elevada em relação a SOBI e SOBI-RO) e, na média, tem o pior desempenho de todos. Este resultado dá um indicativo da maior separabilidade ocorrida para SOBI e SOBI-RO.

A Figura 5.3 mostra o índice E_1 para todos os atrasos possíveis. Observa-se que AMUSE é mais dependente do parâmetro τ, dado que a variância de E_1 é maior do que para SOBI e SOBI-RO. Essa variância elevada reflete uma dificuldade na estimação de τ.

Tabela 5.2: Índice de separabilidade E_1 para cada algoritmo testado. Para o caso de SOBI, SOBI-RO e AMUSE, calcula-se a média e o desvio padrão (entre parêntesis).

Algoritmo	E_1
ICA Ideal	0
SOBI	**0,11 (0,01)**
SOBI-RO	0,14 (0,02)
AMUSE	0,22 (0,05)
JADE	0,19
FastICA	0,24

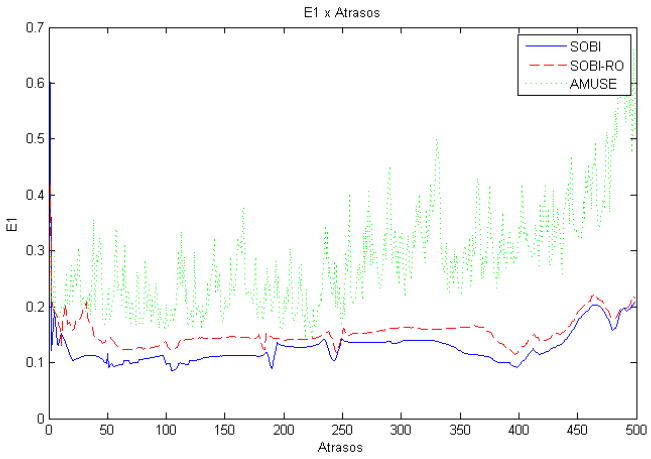

Figura 5.3: Índice de separabilidade E_1 para SOBI, SOBI-RO e AMUSE, para $\overline{SNR} = 11,5$ dB.

Presença de ruído

Os algoritmos também foram analisados sob o ponto de vista da presença de ruído (veja a Tabela 5.3). Observa-se que, com a elevação de σ da fonte simulada de ruídos (veja a Equação 5.6), a relação sinal-ruído (veja a Equação 5.10) é reduzida, mostrando a maior interferência nas séries mistura ($\mathbf{x}(t)$). Na coluna ($\overline{SNR}$), mostra-se a relação sinal-ruído média entre os sinais mistura para cada σ (em negrito). A partir de $\sigma = 1$, começam a surgir valores negativos, indicando que a energia do ruído é maior do que a energia dos sinais.

$$SNR_{\mathbf{x}_i(t)} = 10 \log_{10} \frac{RMS_{\mathbf{x}_i(t)_{semruido}}}{RMS_{ruido}} \tag{5.10}$$

onde, $RMS_{\mathbf{x}_i(t)_{semruido}}$ é o valor médio quadrático da série $\mathbf{x}_i(t)$, $i = 1, 2, ...6$, mapeada pela matriz-mistura sem a fonte de ruídos $\mathbf{s}_6(t)$ e RMS_{ruido} é o valor médio quadrático do ruído mapeado para a posição de $\mathbf{x}_i(t)$.

Tabela 5.3: Relação sinal-ruído (SNR) para os diversos níveis de σ da fonte de ruídos

SNR							
Ruído $N[0, \sigma]$	$SNR_{\mathbf{x}_1(t)}$ (dB)	$SNR_{\mathbf{x}_2(t)}$ (dB)	$SNR_{\mathbf{x}_3(t)}$ (dB)	$SNR_{\mathbf{x}_4(t)}$ (dB)	$SNR_{\mathbf{x}_5(t)}$ (dB)	$SNR_{\mathbf{x}_6(t)}$ (dB)	$\overline{SNR}$ (dB)
$\sigma = 0,01$	12,5	9,3	8,9	15,0	14,3	9,2	**11,5**
$\sigma = 0,04$	9,5	6,3	5,9	12,0	11,3	6,2	**8,5**
$\sigma = 0,09$	7,7	4,5	4,1	10,2	9,5	4,4	**6,7**
$\sigma = 0,16$	6,5	3,3	2,9	9,0	8,3	3,2	**5,5**
$\sigma = 0,25$	5,5	2,3	1,9	8,0	7,3	2,2	**4,5**
$\sigma = 1,00$	2,5	-0,7	-1,1	5,0	4,3	-0,8	**1,5**
$\sigma = 2,25$	0,7	-2,5	-2,9	3,2	2,5	-2,6	**-0,3**
$\sigma = 4,00$	-0,5	-3,7	-4,1	2,0	1,3	-3,8	**-1,5**

A Tabela 5.4 mostra o valor de E_1 em função de $\overline{SNR}$, para os diversos algoritmos estudados. Observa-se que os melhores resultados são também com SOBI e SOBI-RO (veja valores em negrito). No entanto, na média, SOBI-RO tem melhor desempenho quando o nível de ruído torna-se demasiadamente elevado. Quando a relação sinal-ruído começa a ficar negativa, SOBI-RO começa a se destacar. Apesar disso, o desvio padrão de E_1 ainda continua mais elevado do que o desvio para SOBI, o que continua prejudicando um pouco mais a escolha de τ. A despeito deste fato, se o atraso τ for bem definido, SOBI-RO terá o melhor desempenho, pois o valor mínimo (veja a Figura 5.3) de E_1 com o ruído elevado também ocorre para SOBI-RO.

Na Figura 5.4, é mostrado o comportamento de E_1 para $\overline{SNR} = -1,5dB$. Observa-se que índice E_1 melhora a partir do atraso 53 para SOBI-RO e, de ma-

Tabela 5.4: E_1 em função de $\overline{SNR}$

	E_1				
$\overline{SNR}$ (dB)	SOBI	SOBI-RO	AMUSE	JADE	FastICA
11,5	**0,11(0,015)**	0,14(0,025)	0,22(0,047)	0,18	0,23
8,5	**0,11(0,014)**	0,13(0,034)	0,22(0,049)	0,19	0,27
6,7	**0,11(0,014)**	0,13(0,032)	0,23(0,049)	0,20	0,28
5,5	**0,11(0,013)**	0,13(0,034)	0,23(0,049)	0,20	0,27
4,5	**0,12(0,013)**	0,14(0,028)	0,23(0,048)	0,22	0,27
1,5	**0,14(0,011)**	**0,14(0,036)**	0,24(0,044)	0,21	0,22
-0,3	0,15(0,010)	**0,14(0,040)**	0,24(0,042)	0,21	0,22
-1,5	0,15(0,013)	**0,14(0,035)**	0,24(0,012)	0,21	0,22

neira geral, permanece melhor do que SOBI até o último atraso possível. Os valores mínimos de E_1, tanto para SOBI quanto para SOBI-RO, foram obtidos com 103 atrasos. Observa-se também que à medida que os atrasos são maiores, o desempenho de E_1 pode se deteriorar. Isto pode ocorrer pela baixa correlação (não-linear) de amostras distantes e também pela redução do número de amostras para o cálculo da matriz de separação. Na Figura 5.3 e Figura 5.4, é possível observar este fenômeno a partir do atraso 400, para SOBI e SOBI-RO, e a partir do atraso 250, para AMUSE.

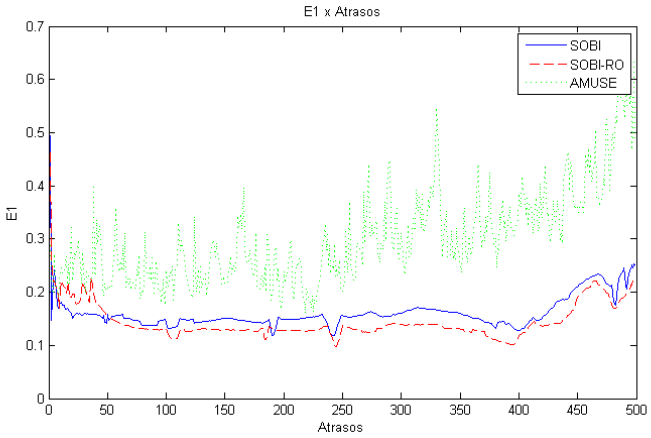

Figura 5.4: Índice de separabilidade E_1 para SOBI, SOBI-RO e AMUSE, para $\overline{SNR} = -1,5$ dB.

Estudos empíricos [117] mostram que, para dados com ruído, o parâmetro τ para SOBI e SOBI-RO deve ser configurado acima de 100 atrasos. Neste nosso experimento controlado, podemos notar que, de fato, a escolha do parâmetro se torna menos crítica após 100 atrasos. No entanto, a partir de certo número de atrasos, o desempenho de SOBI e SOBI-RO tende a se deteriorar. Assim, para estes dois algoritmos, podemos sugerir, como regra prática, a escolha de um valor τ entre 100 e a metade do número de amostras da série.

Presença de *outliers*

A presença de *outliers* nos dados, tanto nas séries de desenvolvimento quanto nas séries monitoradas, pode afetar todo o processo de modelagem do padrões de QD, culminando em uma perda de confiança nas medições da QD. Assim, avaliou-se o comportamento da separação de fontes com *outliers* introduzidos nas séries mistura.

Para simular estas anomalias nos dados, amostras fora da distribuição padrão das séries foram inseridas nos dados. A partir da avaliação da distribuição das séries integradas de primeira ordem, introduziram-se amostras afastadas da média em 2, 3, 4, 5 e 10 desvios padrão. Estas amostras espúrias foram distribuídas aleatoriamente na série, nas posições mostradas na Tabela 5.5, para evitar que algum tipo de padrão fosse formado. Na Figura 5.5, os *outliers* podem ser identificados sobrepostos a cada uma das séries.

Tabela 5.5: *Outliers* simulados

	Posições dos *outliers*				
Série sintética	2 DP	3 DP	4 DP	5 DP	10 DP
$\mathbf{x}_1(t)$	51	103	219	333	432
$\mathbf{x}_2(t)$	99	48	301	200	450
$\mathbf{x}_3(t)$	59	110	315	237	401
$\mathbf{x}_4(t)$	44	198	108	417	321
$\mathbf{x}_5(t)$	473	307	69	125	241
$\mathbf{x}_6(t)$	60	113	212	326	499

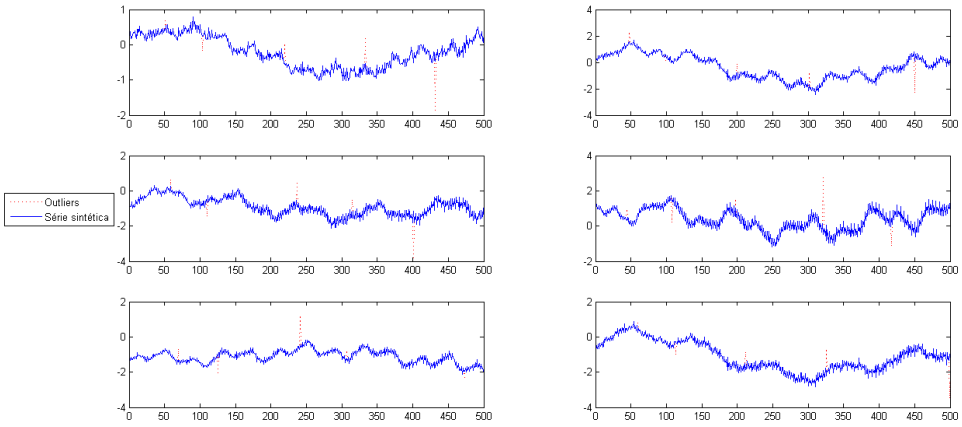

Figura 5.5: Séries sintéticas com *outliers*.

Quando há a presença de *outliers*, a separação das fontes é afetada. Na Tabela 5.6, o indicador E_1 quantifica a piora na separação das fontes. Apesar dos melhores resultados continuarem sendo com algoritmos que avaliam a estrutura temporal,

verifica-se que o valor do indicador se eleva excessivamente para todos os algoritmos. Dentre os algoritmos testados, o SOBI-RO obteve o melhor desempenho para as séries sintéticas com esse tipo de anomalia. Isso mostra a necessidade de um método de tratamento dos *outliers* presentes na série.

Tabela 5.6: E_1 para as séries sintéticas com *outliers*.

Algoritmo	E_1
SOBI	0,34 (0,02)
SOBI-RO	**0,28 (0,02)**
AMUSE	0,42 (0,08)
JADE	0,54
FastICA	0,52

5.2.2 Avaliação das Fontes no Tempo

Correlação não linear

Sob a perspectiva das fontes no tempo, avaliou-se a correlação não linear entre elas. Espera-se que as correlações entre as respectivas fontes estimadas e originais sejam próximas a 1. Para que a correlação seja não linear, basta que uma das fontes sofra uma transformação não linear $f(.)$ [30]. No entanto, é preciso definir a não-linearidade. Para nossos estudos, utilizou-se a função $f(\mathbf{y}(t)) = \tanh(\mathbf{y}(t))$, por permitir o mapeamento de todo o domínio da série para um conjunto imagem limitado ($-1 < \tanh(\mathbf{y}(t)) < 1$). Assim, a correlação não linear é expressa por:

$$\mathbf{R}_{sy}^{NL} = E\{\mathbf{s}(t).f(\mathbf{y}(t))\} \tag{5.11}$$

Indicadores extraídos das matrizes de correlação (não linear), para cada um dos algoritmos ICA testados, são mostradas na sequência. Ressalta-se que, devido à indefinição de ICA quanto a ordenação de fontes [30], as maiores correlações poderiam estar fora da diagonal principal. Assim, as linhas da matriz foram ordenadas segundo a coluna onde a correlação (não linear) com as fontes originais é mais expressiva. Dessa forma, através do valor quadrático médio (RMS) entre os $r_{i,j}$ fatores da matriz de correlação, extraem-se dois indicadores do desempenho da separação: RMS_{diag}, onde $i = j$ e $RMS_{crosstalk}$, onde $i \neq j$. Para uma boa separação, espera-se que RMS_{diag} seja próximo a 1 e $RMS_{crosstalk}$ próximo a zero.

Para efeitos de comparação, a matriz de correlação não linear entre as fontes originais (veja Tabela 5.7) também é analisada. Pode-se notar que a $RMS_{diag} = 1$. Já o $RMS_{crosstalk}$ não é zero ($RMS_{crosstalk} = 0,12$), pois, apesar de terem sido geradas de forma independente, há correlações estatística entre as séries. Observa-

se também que poderá haver uma dificuldade dos algoritmos em separar a primeira da segunda fonte. É possível verificar que a segunda fonte $\mathbf{s}_2(t)$ tem correlação significativa tanto com a primeira fonte $\mathbf{s}_1(t)$ quanto com a segunda fonte $\mathbf{s}_2(t)$.

Tabela 5.7: Matriz de Correlação Não-Linear R_{ss}^{NL}, indicadores RMS e percentual de ordenação correta das fontes para o caso de separação não cega (ICA ideal)

ICA Ideal	**s**(t)					
	1,00	0,99	-0,08	-0,27	-,0,02	0,00
	0,99	**1,00**	-0,08	-0,27	-,0,02	0,00
s(t)	-0,08	-0,08	**-1,00**	-0,02	-,0,01	0,00
	-0,27	-0,27	-0,02	**1,00**	-0,02	0,02
	-0,02	-0,02	0,02	-0,02	**1,00**	0,02
	0,00	0,00	-0,00	0,00	0,01	**1,00**
RMS_{diag}	1,00					
$RMS_{crosstalk}$	0,12					
Ordenação	100%					

Analisando-se os indicadores das matrizes de correlação não linear entre as fontes originais e as fontes estimadas, com os algoritmos ICA (veja a Tabela 5.8), observa-se um desempenho melhor para SOBI e, com resultados próximos, SOBI-RO vem logo a seguir (em negrito são mostrados os melhores resultados). O indicador RMS_{diag} atingiu o maior nível para SOBI e o menor para FastICA. Ainda, os valores do $RMS_{crosstalk}$ foram os menores para SOBI e SOBI-RO, indicando que estes algoritmos alcançaram uma maior independência estatística na separação das fontes. Para todos os casos, as fontes foram ordenadas corretamente pela correlação não linear, atingindo 100% de sucesso.

Tabela 5.8: Indicadores da matriz de correlação não linear obtidos com os algoritmos ICA (ICA estimada)

	SOBI	SOBI-RO	AMUSE	JADE	FastICA
RMS_{diag}	**0,85**	0,81	0,88	0,80	0,79
$RMS_{crosstalk}$	**0,06**	0,07	0,10	0,09	0,10
Ordenação	**100%**	**100%**	**100%**	**100%**	**100%**

Média e Desvio padrão do Erro

Apesar da correlação não-linear nos dar um indicador da separação de fontes, não necessariamente temos fontes estimadas próximas às fontes originais. Assim, avaliamos também a média e o desvio padrão do erro entre as fontes.

Para que seja possível a comparação entre as fontes estimadas e originais, ambas são normalizadas para terem a mesma energia. Essa normalização é necessária

devido a incerteza gerada pela transformação ICA na energia da fonte [30]. Na Tabela 5.9, observa-se que, para todas as fontes estimadas $\mathbf{y}(t)$, o erro é menor para algoritmos que utilizam a estrutura temporal para extrair as fontes. Tanto o erro médio quanto o desvio padrão foram melhores para SOBI e SOBI-RO nas fontes estimadas $\mathbf{y}_1(t)$, $\mathbf{y}_2(t)$, $\mathbf{y}_5(t)$ e $\mathbf{y}_6(t)$. AMUSE obteve melhor desempenho para as fontes $\mathbf{y}_3(t)$ e $\mathbf{y}_4(t)$.

Tabela 5.9: Média (EMQ) e Desvio Padrão (DP) da diferença entre as fontes originais e as fontes estimadas

Algoritmo	Indicador	$\mathbf{y}_1(t)$	$\mathbf{y}_2(t)$	$\mathbf{y}_3(t)$	$\mathbf{y}_4(t)$	$\mathbf{y}_5(t)$	$\mathbf{y}_6(t)$
SOBI	EMQ	1,10	**0,06**	0,05	0,17	**0,019**	0,04
	DP	0,8	**0,04**	0,03	0,09	**0,013**	0,03
SOBI-RO	EMQ	**0,08**	1,16	0,05	0,17	0,02	**0,04**
	DP	**0,05**	0,07	**0,02**	0,09	0,01	**0,03**
AMUSE	EMQ	0,90	0,32	**0,04**	**0,01**	0,10	0,12
	DP	0,67	0,20	0,03	**0,01**	0,07	0,08
JADE	EMQ	0,38	1,05	0,05	0,27	0,05	0,10
	DP	0,24	0,67	0,03	0,22	0,033	0,07
FastICA	EMQ	0,45	1,05	0,06	0,29	0,03	0,18
	DP	0,28	0,71	0,04	0,25	0,02	0,14

Na Figura 5.6, são mostradas as fontes estimadas com SOBI e as fontes originais normalizadas, para serem comparadas com a mesma energia. Observa-se que a fonte com tendência determinística ($\mathbf{s}_1(t)$) não foi separada corretamente. O algoritmo detectou apenas a senoide da fonte original. Isto pode ser explicado pela correlação não linear elevada ($R^{NL}_{s_1 s_2} = 0,99$) entre as tendências originais (veja Tabela 5.7) e a similaridade entre as funções de autocorrelação de cada uma delas. Assim, é de se esperar a dificuldade dos algoritmos ICA para separar estas fontes, principalmente aqueles cuja hipótese de separação é ter espectros distintos, assim como SOBI, SOBI-RO e AMUSE. É possível que grande parte da tendência determinística esteja misturada com a tendência estocástica, apesar da normalização poder ter mascarado este fato e ter levado a fonte $\mathbf{y}_2(t)$ a ficar quase sobreposta à fonte $\mathbf{s}_2(t)$. A estimativa de $\mathbf{y}_4(t)$ também mostra uma pequena distorção em relação à fonte original $\mathbf{s}_4(t)$, que pode ser explicada também pela presença de algum resquício das tendências. Para os outros casos, observam-se que as fontes originais e estimadas foram bem separadas. A sobreposição pode ser observada tanto para fonte com sazonalidade $\mathbf{s}_3(t)$ quanto para a fonte com heterocedasticidade $\mathbf{s}_5(t)$, em relação às fontes estimadas $\mathbf{y}_3(t)$ e $\mathbf{y}_5(t)$, respectivamente. Assim, as distorções encontradas não foram capazes de alterar a ordenação das fontes, que foi 100% correta para todos os casos avaliados. Por fim, o ruído foi corretamente separado, o que pode ser observado na sobreposição das fontes $\mathbf{y}_6(t)$ e $\mathbf{s}_6(t)$.

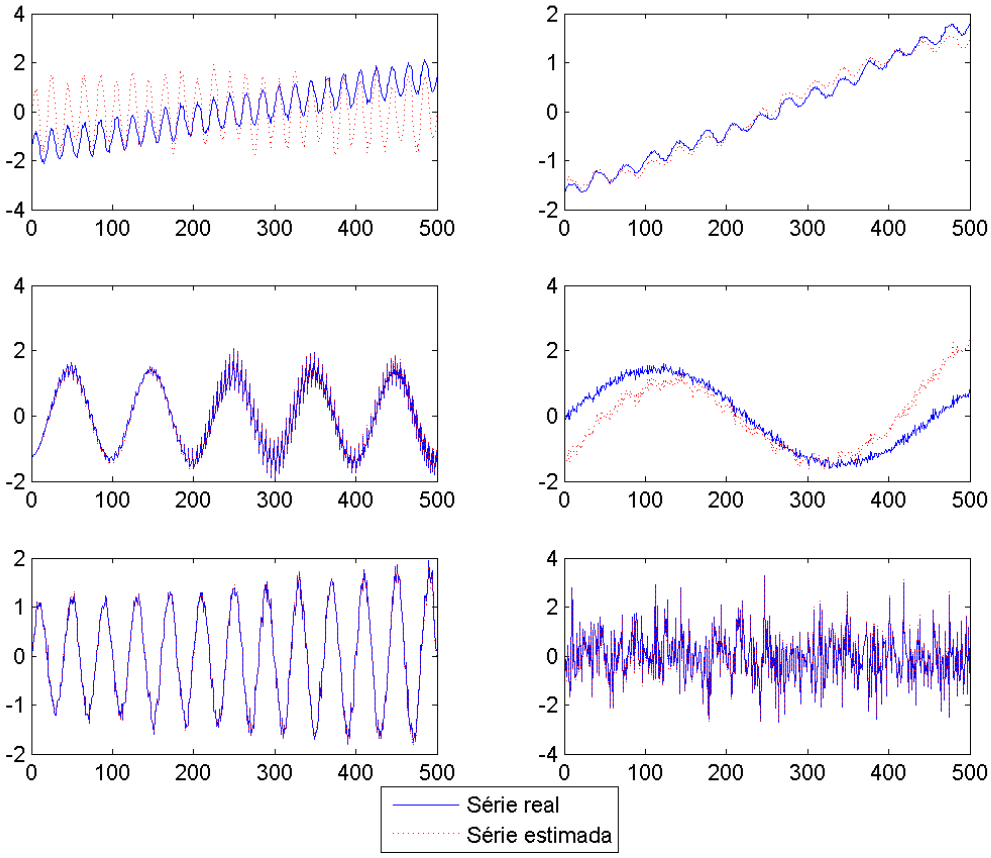

Figura 5.6: Fontes originais e estimadas com SOBI, normalizadas pela energia. Da esquerda para a direita e de cima para baixo, tem-se: tendência determinística, tendência estocástica, sazonalidade, ciclo, heterocedasticidade e ruído branco gaussiano.

5.2.3 Avaliação da Densidade de Probabilidade

As distribuições de probabilidade das fontes são utilizadas também para se avaliar a separação das fontes. Podemos estimar a informação mútua (IM) [49] entre as distribuições de probabilidade (FDP) ou utilizar alguma métrica para calcular a distância entre as distribuições, como a divergência de Kullback-Leibler (KL) [118]. Para ambos os casos, foram utilizados dois métodos para estimar a FDP: Janelas de Parzen [119] e histograma. No primeiro caso, é preciso se definir a função núcleo (*kernel*). No segundo caso, é preciso se definir a resolução do histograma. Nesta análise, o ponto crítico é a estimação da FDP, que pode sofrer variações de acordo com o método de estimação, a taxa de amostragem da função *kernel* e a resolução do histograma.

Informação mútua (IM)

Os resultados obtidos com a informação mútua foram calculados a partir da estimação da Função Densidade de Probabilidade com uma função kernel gaussiana e uma taxa de amostragem de 128. Estes parâmetros foram escolhidos após diversos testes, variando-se a função *kernel* e a taxa de amostragem da função gerada. A escolha destes parâmetros será discutida na Seção 5.2.4.

A mesma análise realizada dos indicadores RMS da Seção 5.2.2 pode ser aplicada para o caso da informação mútua. Espera-se que RMS_{diag} da matriz IM seja próxima a 1 e $RMS_{crosstalk}$ próximo a zero. No entanto, $RMS_{crosstalk}$=0,44 para o caso ideal (veja Tabela 5.10). Assim, pode ser difícil tirar conclusões por este último indicador.

Tabela 5.10: Matriz de Informação mútua IM_{ss}, indicadores RMS e percentual de ordenação correta das fontes para o caso de separação não cega (ICA ideal)

ICA Ideal	**s**(t)					
	1,00	0,99	0,36	0,97	0,34	0,06
	0,99	**1,00**	0,37	0,98	0,33	0,04
s(t)	0,36	037	**1,00**	0,49	0,57	0,32
	0,97	0,98	0,49	**1,00**	0,37	0,12
	0,34	0,33	0,58	0,37	**1,00**	0,29
	0,07	0,04	0,32	0,12	0,29	**1,00**
RMS_{diag}	1,00					
$RMS_{crosstalk}$	0,44					
Ordenação	100%					

Na Tabela 5.11, observa-se que RMS_{diag} é melhor para os casos de SOBI e SOBI-RO. O melhor caso foi obtido com SOBI (em negrito). Nota-se também que o desempenho de AMUSE é similar ao de JADE, contradizendo os indicadores utilizados até agora, que sempre apontaram uma melhor desempenho de JADE. Os

algoritmos JADE e FastICA também provocaram erros na ordenação das fontes, o que pode ser constatado no indicador abaixo de 100% para estes algoritmos.

Tabela 5.11: Indicadores RMS da matriz IM, para o caso de separação de fontes com algoritmos de ICA

	SOBI	SOBI-RO	AMUSE	JADE	FastICA
RMS_{diag}	**0,97**	0,96	0,95	0,95	0,94
$RMS_{crosstalk}$	0,47	0,40	0,39	0,42	0,43
Ordenação	**100%**	**100%**	**100%**	83,3%	83,3%

Kulback Lieber (KL)

A divergência KL mede a similaridade entre duas distribuições. Espera-se que a divergência entre as fontes estimadas e reais sejam mínimas e, dessa forma, a média da diagonal seja próxima de zero. Da mesma forma, espera-se que a KL cruzada seja mais elevada, indicando uma divergência maior entre distribuições de fontes distintas.

Novamente, para efeitos de comparação, avaliamos o comportamento de KL para o caso ideal (veja Tabela 5.12). Observa-se que, como esperado, os elementos da diagonal são todos zero e, dessa forma, $RMS^{KL}_{diag} = 0$. Observa-se também que o $RMS^{KL}_{crosstalk} \neq 0$, conforme esperado. No entanto, este valor não é tão afastado de zero e pode levar a análises inconclusivas. Uma explicação para estes valores próximos é a dificuldade de se estimar a FDP. Na Tabela 5.13, os algoritmos são comparados com o indicador RMS e o nível de ordenação. Observa-se que RMS_{diag} é menor para o algoritmo SOBI. Já o algoritmo AMUSE teve o pior desempenho, o que diverge das análises com a matriz de correlação e IM. Ainda, FastICA teve melhor desempenho em comparação com JADE e se igualou a SOBI-RO. Já o indicador $RMS_{crosstalk}$, seguindo a lógica de que os maiores são os melhores, aparentemente foi pior para SOBI e SOBI-RO. No entanto, verifica-se que os valores são próximos ao caso ideal. De fato, as análises de separação de fontes por este indicador são dificultadas pela proximidade entre os coeficientes da matriz KL. Assim, $RMS_{crosstalk}$ não foi conclusivo. Já a ordenação das fontes indicou que o melhor desempenho é obtido pelos algoritmos SOBI, SOBI-RO e AMUSE.

5.2.4 Discussão

Estimação das FDPs

A estimação da FDP é o fator mais crítico para as análises de IM e KL. Assim, para se definirem os melhores parâmetros da estimação, os indicadores $\overline{RMS}$ fo-

Tabela 5.12: Matriz KL_{ss}, indicadores RMS e percentual de ordenação correta das fontes para o caso de separação não cega (ICA ideal)

ICA Ideal	$\mathbf{s}$(t)					
	0,00	0,01	0,00	0,00	0,00	0,04
	0,01	0,00	0,02	0,03	0,02	0,04
$\mathbf{s}$(t)	0,00	0,02	0,00	0,00	0,00	0,06
	0,00	0,03	0,00	0,00	0,00	0,06
	0,00	0,02	0,00	0,00	0,00	0,06
	0,04	0,04	0,06	0,05	0,06	0,00
RMS_{diag}	0,00					
$RMS_{crosstalk}$	0,024					
Ordenação	100%					

Tabela 5.13: Indicadores RMS da matriz KL e nível de ordenação, para o caso de separação de fontes com algoritmos ICA (ICA estimada)

	SOBI	SOBI-RO	AMUSE	JADE	FastICA
RMS_{diag}	**0,004**	0,009	0,019	0,011	0,009
$RMS_{crosstalk}$	0,026	0,026	0,048	0,031	0,027
Ordenação	**100%**	**100%**	**100%**	83,3%	83,3%

ram analisados no cenário ideal para diferentes configurações. As Tabelas 5.14 (IM) e 5.15 (KL) mostram os indicadores $\overline{RMS}_{diag}$ e $\overline{RMS}_{crosstalk}$ para o conjunto total de séries sintéticas, utilizando-se quatro funções kernel (normal, Epanecknikov, triângulo e quadrada) e taxas de amostragem baixa, média e alta. Na Tabela 5.14, o $\overline{RMS}_{diag}$ permanece em 0,99 para os casos de amostragem média e alta. Para estas amostragens, $\overline{RMS}_{crosstalk}$ também não sofre grandes variações e permanece entre 0,42 e 0,44. Da mesma forma, na Tabela 5.15, não há grande sensibilidade para ambas as amostragens e $\overline{RMS}_{diag}$ permanece igual a zero e $\overline{RMS}_{crosstalk}$ permanece entre 0,02 e 0,03. Assim, deve-se optar por utilizar uma taxa de amostragem mais elevada quando se estima a FDP utilizando kernel. Os indicadores também não mostram grandes discrepâncias para o tipo de função kernel escolhida. Assim, pode-se optar pela função normal, que é comumente utilizada.

Os indicadores são também mostrados para o caso com estimação da PDF por histograma ao invés de função kernel, para 8, 32 e 128 *bins*. Tanto para IM quanto para KL, observa-se que os indicadores $\overline{RMS}_{diag}$ e $\overline{RMS}_{crosstalk}$ não se alteram com uma quantidade média ou alta de *bins*. Já o caso com 8 *bins* tende a ser mais crítico. Isso mostra que, quando se utiliza histograma, o $\overline{RMS}$ pode ser mais sensível em comparação com a estimação da FDP por kernel. Por exemplo, $\overline{RMS}_{diag}$ varia de 0,58 para 0,99 quando se utiliza 8 ou 128 *bins*, respectivamente. Assim, a opção de estimação de FDP foi por meio de funções kernel.

Tabela 5.14: Indicadores consolidados da matriz IM, para ICA ideal

Indicadores IM			
	Kernel		
Tx de amostragem	8	32	128
	Normal		
$\overline{RMS}^{IM}_{diag}$	0,99	0,99	0,99
$\overline{RMS}^{IM}_{crosstalk}$	0,30	0,44	0,44
	Epanechnikov		
$\overline{RMS}^{IM}_{diag}$	0,97	0,99	0,99
$\overline{RMS}^{IM}_{crosstalk}$	0,30	0,43	0,44
	Triângulo		
$\overline{RMS}^{IM}_{diag}$	0,98	0,99	0,99
$\overline{RMS}^{IM}_{crosstalk}$	0,30	0,44	0,44
	Quadrada		
$\overline{RMS}^{IM}_{diag}$	0,99	0,99	0,99
$\overline{RMS}^{IM}_{crosstalk}$	0,34	0,42	0,42
	Histograma		
Bins	8	32	128
$\overline{RMS}^{IM}_{diag}$	0,96	0,99	0,99
$\overline{RMS}^{IM}_{crosstalk}$	0,58	0,82	0,99

Tabela 5.15: Indicadores consolidados da matriz KL, para ICA ideal

Indicadores KL			
	Kernel		
Tx de amostragem	8	32	128
Kernel	**Normal**		
$\overline{RMS}^{KL}_{diag}$	0,00	0,00	0,00
$\overline{RMS}^{KL}_{crosstalk}$	0,03	0,02	0,02
	Epanechnikov		
$\overline{RMS}^{KL}_{diag}$	0,00	0,00	0,00
$\overline{RMS}^{KL}_{crosstalk}$	0,03	0,03	0,03
	Triângulo		
$\overline{RMS}^{KL}_{diag}$	0,00	0,00	0,00
$\overline{RMS}^{KL}_{crosstalk}$	0,11	0,03	0,03
	Quadrada		
$\overline{RMS}^{KL}_{diag}$	0,00	0,00	0,00
$\overline{RMS}^{KL}_{crosstalk}$	0,04	0,03	0,03
	Histograma		
Bins	8	32	128
$\overline{RMS}^{KL}_{diag}$	0,00	0,00	0,00
$\overline{RMS}^{KL}_{crosstalk}$	0,05	0,03	0,03

Utilizando as melhores estimativas da FDP, as análises para a estimação das fontes independentes estão resumidas na Tabela 5.16. Em negrito estão os melhores resultados obtidos. Observa-se que para todos os casos, os métodos que utilizam a informação temporal para extrair as fontes independentes obtiveram melhor desempenho. Entre os três algoritmos que utilizam este princípio, SOBI se destaca como o que obteve melhor desempenho e SOBI-RO chega a ter melhor desempenho, na média, em casos onde o ruído é elevado e na presença de *outliers*. Sob o ponto de vista de $RMS^{NL}_{diag_{sy}}$, o melhor caso obtido com AMUSE chega a ter um melhor desempenho que SOBI e SOBI-RO, mas há que se ponderar a dificuldade de se estimar o parâmetro de atraso para este algoritmo. Por fim, pode-se concluir que, dentre os algoritmos avaliados, SOBI ou SOBI-RO são os mais indicados para estimação cega de fontes independentes em séries temporais. Quando o ruído não é elevado ou não há *outliers* nas amostras, SOBI deve ser considerado preferencialmente. Já em situações adversas, SOBI-RO tem melhor desempenho. Assim, em situações nas quais não se conhece *a priori* as condições dos dados, a ortogonalização robusta pode ser necessária.

Tabela 5.16: Resumo da avaliação da separação de fontes para diferentes algoritmos ICA.

	ICA ideal	SOBI	SOBI-RO	AMUSE	JADE	FastICA
E_1 (11,6dB)	0	**0,11(0,01)**	0,15(0,02)	0,22(0,05)	0,23	0,18
$E_1^{\text{ruído}}$ (-0,1dB)	0	0,15(0,01)	**0,14(0,03)**	0,24	0,20	0,22
E_1^{outliers}	0	0,34(0,02)	**0,28(0,02)**	0,42	0,54	0,52
Erro (média)	0	**0,23**	0,25	0,25	0,31	0,34
$RMS^{Corr_{NL}}_{diag_{sy}}$	1,0	0,85	0,81	**0,88**	0,78	0,75
$RMS^{Corr_{NL}}_{cross_{sy}}$	0,12	**0,06**	0,07	0,10	0,10	0,10
$RMS^{Corr_{NL}}_{cross_{yy}}$	0,12	**0,0004**	0,04	0,04	0,03	0,03
$RMS^{IM}_{diag_{yy}}$	0,99	**0,96**	**0,96**	0,95	0,94	0,95
$RMS^{KL}_{diag_{yy}}$	0	**0,004**	0,008	0,019	0,011	0,009

5.3 Pré-processamento

Esta tese propõe a introdução de ICA antes do pré-processamento proposto em [1] (veja a Seção 4.3) Assim, espera-se que o próprio pré-processamento seja facilitado devido aos sinais mais estruturados após a aplicação de ICA. Nesta seção, o pré-processamento é avaliado com as fontes originais (ICA ideal), no cenário com fontes estimadas pelos melhores métodos encontrados na seção anterior (ICA estimada) e

sem a aplicação de ICA.

5.3.1 ICA ideal

No cenário ideal, as fontes que geraram as misturas são conhecidas. Assim, possibilita-se uma melhor avaliação do que ocorre durante os testes e análises de pré-processamento. Sabem-se exatamente quais os componentes estão presentes nas séries. Os resultados dos testes e análises de pré-processamento são mostrados na Tabela 5.17, onde são mostrados o pré-processamento da heterocedasticidade (H), tendência estocástica (TE), sazonalidade (S), tendência linear (TL), ciclos (C) e a normalização (N). Os valores nulos indicam que o respectivo componente não foi detectado e valores não nulos informam os parâmetros relevantes do pré-processamento.

A presença de raízes unitárias foi detectada na fonte $\mathbf{s}_2(t)$ e $\mathbf{s}_4(t)$, através do teste combinado de Dickey-Fuller Aumentado [108] e Phillips-Perron (ADF-PP) [109]. No entanto, este teste pode falhar quando o fenômeno é um ciclo de baixa frequência ao invés de uma tendência, o que é o caso da série $\mathbf{s}_4(t)$. Assim, o teste combinado ADF-PP deve ser aplicado antes e depois da avaliação da presença de ciclos, para evitar uma conclusão indevida. De fato, quando se retira o ciclo mais relevante de baixa frequência de $\mathbf{s}_4(t)$, o teste não detecta a presença de raízes unitárias na série resultante. Assim, a princípio, o operador de primeira diferença deveria ser aplicado apenas na fonte $\mathbf{s}_2(t)$. Por outro lado, para a fonte $\mathbf{s}_4(t)$, os ciclos não inteiros (2,2 e 2,4) não foram mapeados com precisão suficiente para o espectro de Fourier, ficando misturados nas adjacências da segunda componente de frequência. Isso pode causar distorções ao se retirar os ciclos do espectro. Por exemplo, estes componentes estão misturados na terceira frequência, na qual não se espera nenhum ciclo. Assim, ao se retirar por completo este componente, o pré-processamento leva a uma modelagem indevida. Dessa forma, com o intuito de estacionarizar esta série, o operador de primeira diferença também foi aplicado. Esta estacionarização é útil, pois muitos métodos de modelagem assumem séries estacionárias na entrada.

A sazonalidade presente na fonte $\mathbf{s}_3(t)$ pode ser detectada através da análise da função de autocorrelação. Correlações significativas em um período constante ($ps = 5$) mostram a presença deste componente. No entanto, a presença do ciclo dificulta a análise, pois induz uma senoide na função de autocorrelação. Para uma melhor análise, o ciclo pode ser retirado da série através da retirada da componente cíclica do espectro de Fourier ou mesmo pela diferenciação de primeira ordem. Após a retirada do ciclo sazonal, as inovações sazonais puderam ser visualizadas com mais clareza na função de autocorrelação.

Para a série $\mathbf{s}_1(t)$, dado o conhecimento a *a priori* dos dados e a avaliação visual

da série no tempo, verifica-se uma tendência linear, que foi extraída por um polinômio de primeira ordem. No entanto, inversões de tendência poderiam afetar esta abordagem do problema. Na prática, seria complicado modelar esta componente a cada inversão de tendência. Além disso, nem sempre tem-se conhecimento *a priori* sobre os dados e a análise da série no tempo poderia ser distorcida por avaliações subjetivas. Nestes casos, seria preferível não se fazer este pré-processamento e deixar a modelagem deste fenômeno para os métodos de estimação, como redes neurais.

A heterocedasticidade, presente na fonte $\mathbf{s}_5(t)$, foi retirada aplicando-se o fator e^{-ct}, para c=1/2 (coluna H). Avaliando-se visualmente o comportamento da série no tempo, assumiu-se que a variância cresce exponencialmente (veja Figura 5.1) e o parâmetro c foi estimado empiricamente. Novamente, em situações práticas, nem sempre a heterocedasticidade se comporta de forma que possa ser extraída assumindo-se hipóteses deste tipo. Nestes casos, a modelagem do fenômeno deve ser deixada para os métodos de estimação.

Por fim, as séries foram normalizadas com média zero e desvio padrão 1, para permanecerem abaixo dos limites de saturação dos métodos de estimação. Nota-se que, para esta abordagem, *outliers* devem ser tratados para não ocultar os padrões presentes nas séries.

Tabela 5.17: Pré-processamento após ICA (ideal).

Fontes	H	Te	S	Tl	C	N
$\mathbf{s}_1(t)$	0	0	0	1	0	1
$\mathbf{s}_2(t)$	0	1	0	0	0	1
$\mathbf{s}_3(t)$	0	0	5	0	0	1
$\mathbf{s}_4(t)$	0	0	0	0	1	1
$\mathbf{s}_5(t)$	1/2	0	0	0	0	1
$\mathbf{s}_6(t)$	0	0	0	0	0	1

5.3.2 ICA estimada

A Tabela 5.18 mostra os resultados dos testes e análises de pré-processamento para as fontes estimadas com SOBI ($\tau = 103$), ordenadas de acordo com análise da correlação não linear com as fontes originais. As mesmas particularidades encontradas no pré-processamento no cenário ideal também foram encontradas aqui. Outras questões também surgiram como, por exemplo, a não detecção da tendência linear na fonte $\mathbf{y}_1(t)$. A separação de fontes não foi capaz de distinguir os dois tipos de tendência e concentrou ambos componentes na fonte $\mathbf{y}_2(t)$, conforme explicado anteriormente. Nesta fonte, detectou-se apenas a presença de raiz unitária, que foi retirada através do operador de primeira diferença. Na fonte $\mathbf{y}_4(t)$, também se de-

tectou a presença de raiz unitária. No entanto, após a retirada do ciclo, a raiz não foi mais detectada, mostrando que não se tratava de uma tendência. O mesmo ocorreu na fonte $\mathbf{y}_3(t)$, que teve uma raiz unitária detectada. Apesar disso, após a retirada do ciclo sazonal no espectro de Fourier, o teste não se confirmou. Assim, somente a aplicação do operador diferença de quinta ordem foi necessário no pré-processamento desta fonte.

Por fim, da mesma forma que ICA ideal, a heterocedasticidade foi detectada visualmente (veja a Figura 5.6) e retirada aplicando-se o fator e^{-ct}, onde $c = 1/2$. Em seguida, todas as fontes foram normalizadas.

Tabela 5.18: Pré-processamento após ICA (SOBI).

Fontes estimadas	H	Te	S	Tl	C	N
$\mathbf{y}_1(t)$	0	0	0	0	0	1
$\mathbf{y}_2(t)$	0	1	0	0	0	1
$\mathbf{y}_3(t)$	0	0	5	0	0	1
$\mathbf{y}_4(t)$	0	0	0	0	1	1
$\mathbf{y}_5(t)$	1/2	0	0	0	0	1
$\mathbf{y}_6(t)$	0	0	0	0	0	1

5.3.3 Sem ICA

Para o caso sem ICA, os componentes estão misturados nas séries. Assim, os testes e análises para o pré-processamento são dificultadas (veja a Tabela 5.19). Por exemplo, em uma das séries não foi detectada a presença de raiz unitária (sinal-mistura $\mathbf{x}_4(t)$). Isso mostra que o teste não foi totalmente eficiente, pois se sabe, a priori, que todas as séries tem tendência misturada. A heterocedasticidade e os ciclos também estão misturados nas séries. Na análise do espectro de Fourier, nota-se uma dificuldade maior em se distinguir os ciclos relevantes dos outros ciclos. Esta dificuldade também é percebida na avaliação da sazonalidade, que fica mascarada na função de autocorrelação das séries. Somente após a diferenciação de primeira ordem é possível identificar com clareza este componente. O mesmo ocorre na análise da heterocedasticidade, através das séries no domínio do tempo. Esta componente não foi percebida em nenhuma das séries. Por fim, além da normalização, todas as séries devem ser pré-processadas com a retirada da sazonalidade e da tendência.

Tabela 5.19: Pré-processamento sem ICA.

Séries sintéticas	H	Te	S	Tl	C	N
$\mathbf{x}_1(t)$	0	1	5	0	0	1
$\mathbf{x}_2(t)$	0	1	5	0	0	1
$\mathbf{x}_3(t)$	0	1	5	0	0	1
$\mathbf{x}_4(t)$	0	0	5	0	0	1
$\mathbf{x}_5(t)$	0	1	5	0	0	1
$\mathbf{x}_6(t)$	0	1	5	0	0	1

5.4 Modelagem

5.4.1 ICA Ideal

Para a modelagem das séries, foram utilizados modelos simplificados, lineares e não lineares. Para o caso de modelos não lineares utilizando-se redes neurais, avaliaram-se as redes não recorrentes (MLP) e redes recorrentes (Elman). Analisou-se o erro médio quadrático entre o sinal desejado e o estimado, normalizado pela energia da série alvo (NMSE). Na Tabela 5.20, são mostrados os resultados para ambas as redes (em negrito estão os melhores resultados). Observa-se que o desempenho foi similar entre redes recorrentes e não-recorrentes. Isso indica que redes mais complexas e que utilizam variáveis recorrentes não necessariamente melhoram o processo de modelagem. Assim, dado que o método (testes de seleção de parâmetros) busca por modelos parcimoniosos, estes resultados indicam que a utilização de redes neurais MLP pode ser o caminho mais adequado para se seguir nas análises com modelos não lineares.

Tabela 5.20: Redes neurais MLP e Elman no cenário de ICA ideal.

	NMSE	
Fontes	MLP	ELMAN
$\mathbf{s}_1(t)$	**0,15**	**0,15**
$\mathbf{s}_2(t)$	0,82	**0,81**
$\mathbf{s}_3(t)$	**0,15**	0,16
$\mathbf{s}_4(t)$	**0,88**	**0,88**
$\mathbf{s}_5(t)$	**0,16**	0,17
$\mathbf{s}_6(t)$	**1,00**	**1,00**

Uma vez que o pré-processamento no cenário ideal é mas fidedigno com os verdadeiros componentes presentes na série, é de se esperar que a modelagem seja facilitada. Na Tabela 5.21, são mostrados os resultados da modelagem para o número máximo de atrasos (Lag Máximo), a presença de séries explicativas (Explicativa) e

o número de neurônios na camada escondida (NNE) da rede neural MLP. O Lag Máximo é definido pelo maior atraso relevante na função de correlação linear ou não linear. Para padronizar esta saída e facilitar as análises, optou-se por mostrar apenas o atraso máximo ao invés de todos os atrasos relevantes possíveis. Outra particularidade da saída está na coluna NNE. Se os testes de hipótese indicam que modelos não-lineares não são os mais apropriados, então NNE=0. Isso significa que, dependendo da existência de atrasos relevantes/séries explicativas, o modelo pode ser ou linear ou um modelo simplificado. Por exemplo, a fonte $\mathbf{s}_6(t)$ não tem nenhum atraso relevante ou série explicativa. Assim, não é necessário se buscar modelos complexos e o método sugere simplesmente o melhor estimador de caminho aleatório para esta fonte. Este fato era de se esperar, já que se trata da fonte de ruído.

ICA tem a vantagem de apresentar fontes estatisticamente independentes na entrada do modelo. Assim, é esperado que não haja necessidade de séries explicativas, o que facilita o processo de modelagem. Analisando as séries explicativas, nota-se que não foi necessária nenhuma fonte adicional para a modelagem da respectiva fonte-alvo. Apenas os atrasos da própria fonte alvo foram utilizados na modelagem.

Por fim, notam-se que, com exceção da fonte $\mathbf{s}_6(t)$ (ruído), todas as outras séries foram modeladas com a rede MLP. De fato, esperam-se modelos não lineares, uma vez que o resíduo é composto das senoides que foram simuladas. Além disso, observa-se que foram encontrados modelos parcimoniosos, com poucos neurônios (entre 1 e 4 neurônios) na camada escondida.

Tabela 5.21: Modelagem para o cenário com ICA (ideal).

Fontes	Lag Máximo	Explicativa	NNE
$\mathbf{s}_1(t)$	4	0	4
$\mathbf{s}_2(t)$	5	0	1
$\mathbf{s}_3(t)$	5	0	2
$\mathbf{s}_4(t)$	1	0	1
$\mathbf{s}_5(t)$	5	0	4
$\mathbf{s}_6(t)$	0	0	0

5.4.2 ICA Estimada

Assim como com ICA ideal, avaliou-se o comportamento dos modelos não lineares com redes MLP e Elman (veja a Tabela 5.22). Observa-se que o desempenho com as redes MLP foram melhores. Assim, estas redes foram utilizadas nas análises subsequentes. Observa-se ainda que NMSE para ICA estimada é similar a NMSE para ICA ideal, indicando mais uma vez que houve uma boa separação de fontes.

Na Tabela 5.23, são mostrados os resultados da modelagem nos resíduos. Observa-se que, assim como no caso ideal, a fonte ($\mathbf{y}_6(t)$) foi adequadamente tratada como um ruído. Já para a fontes $\mathbf{y}_1(t)$ e $\mathbf{y}_4(t)$, o número de atrasos significativos foi diferente em relação à ICA ideal, devido às diferenças no pré-processamento. Outro motivo da diferença em relação a ICA ideal é o fato da fonte $\mathbf{y}_2(t)$ concentrar tanto a tendência determinística (fonte $\mathbf{s}_1(t)$) quanto a tendência estocástica (fonte $\mathbf{s}_2(t)$). Já o número de lags para as fontes $\mathbf{y}_3(t)$, $\mathbf{y}_4(t)$ e $\mathbf{y}_6(t)$ foram idênticos ao caso com ICA ideal. Assim como ICA ideal, as fontes estimadas foram também modeladas com modelos não lineares (MLP) para as $\mathbf{y}_1(t)$, $\mathbf{y}_2(t)$, $\mathbf{y}_3(t)$, $\mathbf{y}_4(t)$ e $\mathbf{y}_5(t)$ com 1 até 4 neurônios na camada escondida. Os modelos para as séries $\mathbf{y}_3(t)$, $\mathbf{y}_5(t)$ e $\mathbf{y}_6(t)$ são praticamente iguais ao cenário de ICA ideal, indicando a separação dessas fontes. O modelo para $\mathbf{y}_2$ também foi igual, indicando que o pré-processamento pela diferenciação extrai tanto a tendência estocástica quanto a tendência linear, assim o resíduo foi modelado com ICA e ICA ideal. Já a série $\mathbf{y}_4(t)$ foi modelada de forma diferente, dado que a tendência linear também não foi totalmente separada desta fonte.

Tabela 5.22: Redes neurais MLP e Elman no cenário de ICA estimada.

	NMSE	
Fontes	MLP	ELMAN
$\mathbf{y}_1(t)$	**0,16**	0,17
$\mathbf{y}_2(t)$	**0,90**	**0,90**
$\mathbf{y}_3(t)$	**0,15**	0,17
$\mathbf{y}_4(t)$	**0,90**	0,91
$\mathbf{y}_5(t)$	**0,17**	0,19
$\mathbf{y}_6(t)$	**1,00**	**1,00**

Tabela 5.23: Modelagem para o cenário com ICA (SOBI).

Fontes estimadas	Lag máximo	Explicativa	NNE
$\mathbf{y}_1(t)$	1	1	1
$\mathbf{y}_2(t)$	1	1	1
$\mathbf{y}_3(t)$	5	0	2
$\mathbf{y}_4(t)$	4	0	2
$\mathbf{y}_5(t)$	5	0	4
$\mathbf{y}_6(t)$	0	0	0

5.4.3 Sem ICA

Sem ICA, novamente, as redes MLP tiveram um melhor desempenho (veja a Tabela 5.24). Com exceção dos desempenhos idênticos para as séries $\mathbf{x}_1(t)$, $\mathbf{x}_5(t)$ e $\mathbf{x}_6(t)$, para todas as outras séries, a rede MLP teve um melhor desempenho. Na Tabela 5.25, observam-se que são sugeridas séries explicativas para todas as séries. A necessidade de séries explicativas pode dificultar a busca por modelos mais adequados, uma vez que o espaço a ser modelado é maior. Ainda, sem o tratamento do ruído que se encontra misturado, pode haver dificuldade ainda maior de se encontrar padrões nos dados. Por outro lado, padrões complexos podem também exigir modelos mais complexos, caso sejam detectados pelo método de modelagem. Assim, seja pela dificuldade de encontrar padrões, seja pela dificuldade de modelar padrões complexos, o que se nota é que a modelagem pode ser prejudicada ao se utilizar o modelo não adequado. Por exemplo, na série $\mathbf{x}_1(t)$ sugeriu-se um modelo linear. Isso mostra que o método não foi tão eficaz na sugestão do modelo, uma vez que sabemos que há não linearidades nas séries. O modelo para a série $\mathbf{x}_5$ tem mais neurônios do que qualquer modelo para ICA. Isso mostra que foi preciso modelos mais complexos para se encontrar padrões que poderiam ser modelados de forma mais simples. Por fim, verifica que sem ICA foram necessários mais neurônios do que com ICA estimada ou ideal, o que reflete a maior complexidade encontrada.

Tabela 5.24: Modelo para o cenário sem ICA.

	MSE	
Fontes	MLP	ELMAN
$\mathbf{x}_1(t)$	**0,87**	**0,87**
$\mathbf{x}_2(t)$	**0,74**	0,75
$\mathbf{x}_3(t)$	**0,80**	0,81
$\mathbf{x}_4(t)$	**0,43**	0,56
$\mathbf{x}_5(t)$	**0,76**	**0,76**
$\mathbf{x}_6(t)$	**0,71**	**0,71**

Tabela 5.25: Modelagem para o cenário sem ICA.

Séries sintéticas	Lag máximo	Explicativa	NNE
$\mathbf{x}_1(t)$	1	1	0
$\mathbf{x}_2(t)$	5	1	2
$\mathbf{x}_3(t)$	5	1	2
$\mathbf{x}_4(t)$	5	1	5
$\mathbf{x}_5(t)$	5	1	2
$\mathbf{x}_6(t)$	5	1	3

5.5 Detecção de Anomalias/Outliers

Os *Outliers* em séries temporais são classificados como amostras fora do padrão exibido pelas amostras vizinhas. Já as anomalias são amostras que podem estar dentro dos padrões mas contém algum tipo de problema como, por exemplo, uma diferença em relação ao valor real. No entanto, uma amostra inicialmente identificada como *outlier* pode vir a ser considerada como uma anomalia. Estes dados podem comprometer o desempenho da modelagem dinâmica. Por uma questão de simplificação, em alguns casos ambas as palavras serão utilizadas como sinônimas. Assim, ao se deparar com um *outlier/anomalia*, o sistema deve identificá-la e corrigi-la para o próprio benefício da monitoração. Por outro lado, a detecção ocorre de forma estatística, não sendo possível identificá-las de forma exata e podendo ocasionar falsos alarmes. Assim, ao detectar um provável *outlier*/anomalia, a decisão da reposição da amostra entrante deve ser levada ao usuário.

5.5.1 SCICA

Nesta seção, propõe-se utilizar o método SCICA para a detecção de *outliers*. O método é aplicado nas séries individualmente, podendo detectar *outliers* unidimensionais. Sob cada uma das séries sintéticas, aplicou-se o método SCICA com o algoritmo SOBI-RO, que se mostrou mais robusto à presença de *outliers*. Para os estudos com SCICA, avaliou-se apenas amostras que pudessem ser um problema para a própria modelagem. Neste caso, o próprio sistema de monitoração é o "usuário"dos dados.

Ao se aplicar diversos atrasos nas séries, notou-se uma grande dificuldade na separação das fontes. Somente entre o quinto e o décimo quinto atraso começam a surgir estruturas além da estrutura da própria série. Apesar disso, as fontes continuam bastante misturadas. A partir do trigésimo atraso, as fontes começam a se separar e somente a partir do quinquagésimo atraso as fontes começam a ficar mais nítidas. Ainda assim, algumas não foram bem identificadas.

A aplicação de muitos atrasos na série de entrada dificulta a utilização das fontes extraídas, dado que a sequência do tempo nas fontes fica distorcida. Por outro lado, quando se aplica apenas um atraso, o método é útil para isolar as amostras fora do padrão, assim como ruídos, *outliers* e outras anomalias. A partir da série individual, o método produz duas fontes: uma preservando a estrutura original da série e a outra contendo amostras fora da estrutura (veja Figura 5.7). Os mesmos procedimentos utilizados para ICA para a escolha dos parâmetros foram utilizados com SCICA. O atraso que obteve melhor desempenho foi $\tau = 141$.

Inicialmente, a tentativa foi utilizar a fonte estruturada como a fonte livre de anomalias. No entanto, este método não se mostrou eficaz, pois os *outliers* não são

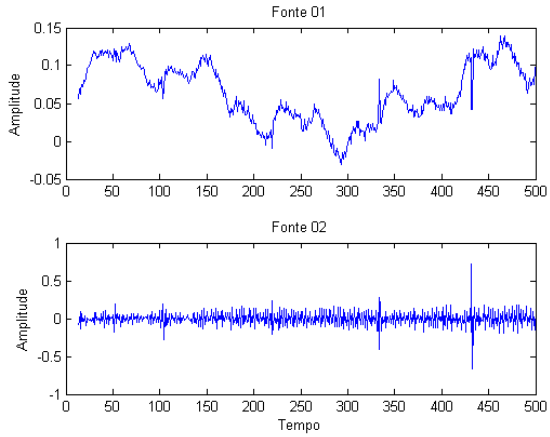

Figura 5.7: Fontes obtidas com o método SCICA.

totalmente filtrados pelo método. A outra opção, que se mostrou mais eficiente, utilizou a fonte desestruturada (fonte 02) como detector de anomalias. Nesta fonte, construiu-se corredores de validação utilizando a mesma metodologia que tem sido utilizada em séries múltiplas. A outra fonte (Fonte 01) é desconsiderada e, portanto, não se faz a transformação inversa de ICA.

Não se espera que os componentes frequentemente presentes nas séries temporais sejam mapeados para a fonte desestruturada. É esperado que estes componentes permaneçam misturados na fonte estruturada. No entanto, é útil aplicar o pré-processamento, principalmente devido a incerteza na amplitude da fonte. Assim, a normalização deve sempre ser realizada. Caso algum resquício dos componentes ainda estejam presentes na fonte, é útil também aplicar os testes de todo o pré-processamento aqui proposto. Para as séries estudadas, os testes não detectaram nenhuma estrutura particular na fonte. Dessa forma, somente a normalização (média zero e desvio padrão 1) foi realizada.

Os testes para especificação dos modelos indicaram a necessidade de modelos não lineares. Isto mostra que podem haver padrões nas séries e não apenas ruído. Estes componentes devem ser retirados para que somente os componentes sem padrão definido sejam detectados, assim como os *outliers*. Para todas as fontes avaliadas, foram necessários dois neurônios na camada escondida da rede neural.

Detecção de *Outliers*

As séries sintéticas com anomalias simuladas (veja a Seção 5.2.1) foram utilizadas para avaliar os método SCICA, ICA e sem ICA. Para os três métodos, configurou-se os corredores de validação para detectar valores extremos nas séries. Utilizou-se $k = 3$, para se buscar amostras afastadas 3σ do padrão da série (veja a Equação 4.1). Na Tabela 5.26, nota-se que o método SCICA detecta os *outliers* com a

mesma precisão que os métodos ICA e sem ICA. No entanto, SCICA emite menos falsos alarmes. Com exceção de $\mathbf{x}_4(t)$, para todas as outras séries o percentual de falsos alarmes foi menor com SCICA. Isso reforça a utilização do método na detecção de *outliers* integrado ao método ICA. O bom desempenho de SCICA se dá pelo foco na série desestruturada. Os fatores da matriz de separação de SCICA funcionam como coeficientes de filtros passa-baixa. Assim, o resíduo de filtragem contém componentes de alta-frequência, assim como ruídos e *outliers*.

Tabela 5.26: Detecção de *outliers* e falsos alarmes

Séries sintéticas	Indicadores	SCICA	ICA	Sem ICA
$\mathbf{x}_1(t)$	Falso alarme (%)	**1,2**	3,27	2,04
	Detecção (%)	80	**100**	80
$\mathbf{x}_2(t)$	Falso alarme (%)	**2,25**	5,95	7,6
	Detecção (%)	80	80	**100**
$\mathbf{x}_3(t)$	Falso alarme (%)	**1,82**	7,6	11,7
	Detecção (%)	**100**	**100**	**100**
$\mathbf{x}_4(t)$	Falso alarme (%)	2,45	8,02	**2,24**
	Detecção (%)	**100**	80	**100**
$\mathbf{x}_5(t)$	Falso alarme (%)	**2,64**	7,61	3,07
	Detecção (%)	**80**	60	**80**
$\mathbf{x}_6(t)$	Falso alarme (%)	**1,41**	7,39	7,6
	Detecção (%)	**100**	**100**	80

5.6 Qualidade de Dados

No modelo proposto para séries temporais, os corredores são associados a dimensões da QD. Os corredores são divididos em diversos níveis para realizar a qualificação da série ou do conjunto de séries. Além disso, verificam-se dados faltantes para compor a métrica de qualidade de dados. Da mesma forma, mede-se a qualidade dos modelos que produziram os padrões de QD, para formarem outra métrica. Estes indicadores são chamados de Indicadores da Qualidade de Dados (IQD) e Indicadores da Qualidade do Modelo (IQM) e os resultados são mostrados a seguir.

5.6.1 Indicadores da Qualidade dos Dados (IQD)

Para se medir a acurácia dos dados, a Equação 2.1 foi adaptada para a utilização em séries temporais (veja a Equação 5.12). O exemplo incorreto foi definido com base na precisão estabelecida pelo corredor. Assim, exemplos fora do corredor são classificados como suspeitos de serem incorretos. Para efeitos de análise, a perspectiva

do usuário em relação à acurácia foi estabelecida em 3 níveis: baixo (B), médio (M) e alto (A). Já a quantidade total de dados foi estabelecida com todas as amostras da série temporal. Ainda, ajustou-se esta equação para que os níveis de acurácia oscilassem entre 0 (nível mínimo) e 1 (nível máximo).

$$\text{Acurácia} = 1 - \frac{\sum \text{Fora do corredor}}{\text{Quantidade Total de Dados}}. \tag{5.12}$$

O nível mais alto da acurácia exige corredores mais estreitos, enquanto o nível mais baixo tem corredores mais largos. Com base na discussão da Seção 4.2.2, definem-se os três níveis alto (A), médio (M) e baixo (B) de acurácia com corredores simétricos $k_1 = k_2 = 1$, $k_1 = k_2 = 2$ e $k_1 = k_2 = 3$, respectivamente.

A dimensão da completude foi quantificada adaptando-se a Equação 2.3 para que o nível mínimo da completude fosse 0 e o máximo 1. Os exemplos faltantes são identificados com base nos instantes de tempo de referência (meta-informação). A presença ou não destes exemplos nos instante pré-definido define se o exemplo é faltante ou não. Já a quantidade total de dados é definida como a quantidade total de instantes de tempo esperados.

$$\text{Completude} = 1 - \frac{\sum \text{Quantidade de exemplos faltantes}}{\text{Quantidade Total de Dados}}. \tag{5.13}$$

Estes indicadores podem ser utilizados para compor a métrica da QD e são ponderados para se chegar a um indicador geral da qualidade de dados (IGQD). Esta ponderação deve ser realizada pelos usuários dos dados. No entanto, para se realizarem as análises, assume-se este papel de usuário e pondera-se igualmente todas as dimensões.

5.6.2 Indicadores da Qualidade do Modelo (IQM)

Para se medir a qualidade do modelo, utilizou-se o Erro Médio Quadrático Normalizado (NMSE) para se medir o erro do corredor, o $NMSE_1$ para se comparar com o estimador médio ($\hat{\mathbf{x}}(t) = \overline{\mathbf{x}}(t)$) e o $NMSE_2$ para se comparar com o melhor estimador de caminho aleatório ($\hat{\mathbf{x}}(t) = \mathbf{x}(t-1)$). Ainda, utilizou-se o indicador R para medir o efeito do atraso nas estimativas e $Lag_{n=0}$ para medir a correlação não linear entre o modelo estimado e o modelo real. Além disso, estes indicadores foram normalizados e ponderados igualmente, de modo a estabelecer um indicador geral da qualidade do modelo (IGQM).

$$NMSE = E\left\{\frac{\sqrt{[\hat{\mathbf{x}}(t) - \mathbf{x}(t)]^2}}{\sqrt{(\mathbf{x}^2(t))}}\right\} \tag{5.14}$$

$$NMSE_1 = E\left\{\frac{\sqrt{[\hat{\mathbf{x}}(t)-\mathbf{x}(t)]^2}}{\sqrt{[\overline{\mathbf{x}}(t)-\mathbf{x}(t)]^2}}\right\} \tag{5.15}$$

$$NMSE_2 = E\left\{\frac{\sqrt{[\hat{\mathbf{x}}(t)-\mathbf{x}(t)]^2}}{\sqrt{[\mathbf{x}(t-1)-\mathbf{x}(t)]^2}}\right\} \tag{5.16}$$

$$R = \frac{Lag_{n=0}}{Lag_{n=1}} \tag{5.17}$$

$$Lag_{n=t} = corr\left\{\mathbf{x}(n), \hat{\mathbf{x}}(n-t)\right\} \tag{5.18}$$

$$IGQM = \frac{1}{5}\left[(1/e^{NMSE}) + (1/e^{NMSE_1}) + (1/e^{NMSE_2}) + (1-1/e^{|R|}) + (|Lag_{n=0}|)\right] \tag{5.19}$$

5.6.3 Resultados

Dados sem *outliers*

Na Tabela 5.27, são mostrados os resultados para os indicadores de qualidade do modelo (os melhores resultados estão em negrito). Os resultados com ICA ideal são utilizados apenas para comparação. Observa-se que, para todos os casos, o desempenho foi melhor quando se introduz ICA. Nota-se também que o algoritmo SOBI atinge um desempenho similar ao caso ideal, o que mostra que as fontes extraídas foram próximas das fontes verdadeiras. Por fim, verifica-se que o IGQM foi capaz de refletir, de forma consolidada, a melhora da qualidade dos modelos ICA.

Na Tabela 5.28, são mostrados os resultados para a medição da qualidade de dados com e sem ICA (os maiores indicadores estão em negrito). Dado que as séries foram simuladas sem a presença de dados faltantes, a completude atingiu o valor máximo, para todas as séries. Ainda, uma vez que as séries não foram simuladas com outras anomalias, é de se esperar que os níveis de qualidade sejam elevados.

Em geral, observa-se que ICA atinge um desempenho similar ao ICA ideal. Ainda, para o nível de acurácia baixo, a QD média sem ICA é idêntica às medições com ICA. No entanto, para os níveis médio e alto de acurácia, o desempenho com ICA é superior, medindo a QD de forma mais fidedigna. Por exemplo, acurácia (M) da série $\mathbf{x}_4(t)$ foi 0,75, enquanto com ICA foi de 0,99.

Analisando-se o conjunto de séries (BD), com ICA, o indicador de acurácia média foi acima de 0,99, enquanto que, para o caso sem ICA, o nível foi de 0,94. Para o nível mais alto da acurácia (A), o indicador de qualidade com ICA foi acima de 0,75 enquanto, sem ICA, este indicador foi cerca de 0,58. Por fim, analisando-se o indicador consolidado (IGQD), observa-se o melhor desempenho da medição da QD com ICA, atingindo 0,93, contra 0,88 para o caso sem ICA.

O fato dos indicadores de QD mais elevados mostrarem melhor desempenho só é

Tabela 5.27: Indicadores de Qualidade do modelo (IQM)

ICA ideal						
Séries sintéticas	$NMSE$	$NMSE_1$	$NMSE_2$	R	Lag_0	IGQM
$\mathbf{x}_1(t)$	0,2369	0,0792	0,7000	0,9776	0,9672	0,8127
$\mathbf{x}_2(t)$	0,1103	0,0164	0,1545	1,0388	0,9919	0,8743
$\mathbf{x}_3(t)$	0,1017	0,0482	0,1149	1,1986	0,9767	0,8821
$\mathbf{x}_4(t)$	0,1112	0,0185	0,0384	1,2956	0,9907	0,8904
$\mathbf{x}_5(t)$	0,0619	0,0555	0,1670	1,1406	0,9724	0,8768
$\mathbf{x}_6(t)$	0,0685	0,0141	0,1031	1,0571	0,9929	0,8762
ICA						
$\mathbf{x}_1(t)$	**0,2407**	**0,0813**	**0,7181**	**0,9781**	**0,9609**	**0,8113**
$\mathbf{x}_2(t)$	**0,1137**	**0,0178**	**0,1676**	**1,0383**	**0,9920**	**0,8742**
$\mathbf{x}_3(t)$	**0,1012**	**0,0483**	**0,1151**	**1,1986**	**0,9772**	**0,8822**
$\mathbf{x}_4(t)$	**0,1101**	**0,0181**	**0,0374**	**1,3018**	**0,9911**	**0,8910**
$\mathbf{x}_5(t)$	**0,0606**	**0,0534**	**0,1604**	**1,1543**	**0,9738**	**0,8785**
$\mathbf{x}_6(t)$	**0,0713**	**0,0147**	**0,1077**	**1,0578**	**0,9929**	**0,8763**
Sem ICA						
$\mathbf{x}_1(t)$	0,2642	0,0939	0,8302	0,9573	0,9526	0,7998
$\mathbf{x}_2(t)$	0,1627	0,0354	0,3336	0,9970	0,9825	0,8468
$\mathbf{x}_3(t)$	0,1786	0,1453	0,3462	0,9743	0,9253	0,8485
$\mathbf{x}_4(t)$	0,4876	0,3447	0,7172	0,9992	0,9215	0,7997
$\mathbf{x}_5(t)$	00823	0,1041	0,3107	0,9997	0,9467	0,8585
$\mathbf{x}_6(t)$	0,1207	0,0450	0,3294	0,99871	0,9776	0,8630

válido para este caso simulado em que a QD é controlada. Indicadores mais elevados para um mesmo banco de dados não significam necessariamente melhor desempenho. Assim, é preciso olhar paralelamente tanto o indicador de QD quanto os indicadores consolidados de qualidade do modelo (IGQM). O IGQM para o conjunto de séries (BD), quando se utiliza ICA, é próximo a 0,88 enquanto que, sem ICA, é cerca de 0,86. Assim, podemos confiar mais na medição da QD quando ICA é aplicada. A maior QD obtida com ICA corrobora esta análise.

Tabela 5.28: Indicadores de Qualidade de Dados (IQD)

ICA ideal					
Séries sintéticas	Compl.	Acurácia (B)	Acurácia (M)	Acurácia (A)	IGQD
$\mathbf{x}_1(t)$	1,00	0,99	0,99	0,75	0,94
$\mathbf{x}_2(t)$	1,00	0,99	0,98	0,76	0,94
$\mathbf{x}_3(t)$	1,00	0,99	0,98	0,72	0,93
$\mathbf{x}_4(t)$	1,00	0,99	0,99	0,75	0,94
$\mathbf{x}_5(t)$	1,00	0,99	0,99	0,76	0,94
$\mathbf{x}_6(t)$	1,00	0,99	0,99	0,77	0,94
BD	1,00	0,99	0,99	0,75	0,94
ICA					
$\mathbf{x}_1(t)$	**1,00**	**0,99**	**0,99**	**0,72**	**0,93**
$\mathbf{x}_2(t)$	**1,00**	**0,99**	**0,98**	**0,76**	**0,94**
$\mathbf{x}_3(t)$	**1,00**	**0,99**	**0,98**	**0,72**	**0,93**
$\mathbf{x}_4(t)$	**1,00**	**0,98**	**0,99**	**0,76**	**0,94**
$\mathbf{x}_5(t)$	**1,00**	**0,99**	**0,99**	**0,76**	**0,94**
$\mathbf{x}_6(t)$	**1,00**	**0,99**	**0,99**	**0,76**	**0,94**
BD	**1,00**	**0,99**	**0,99**	**0,75**	**0,94**
Sem ICA					
$\mathbf{x}_1(t)$	**1,00**	**0,99**	**0,99**	0,68	0,92
$\mathbf{x}_2(t)$	**1,00**	**0,99**	**0,98**	0,64	0,91
$\mathbf{x}_3(t)$	**1,00**	**0,99**	**0,98**	0,50	0,87
$\mathbf{x}_4(t)$	**1,00**	**0,98**	0,75	0,43	0,79
$\mathbf{x}_5(t)$	**1,00**	**0,99**	0,98	0,64	0,91
$\mathbf{x}_6(t)$	**1,00**	**0,99**	0,97	0,56	0,88
BD	**1,00**	**0,99**	0,94	0,58	0,88

Os corredores de validação são mostrados na Figura 5.8, obtidos com a aplicação de ICA e mostrado para as 30 primeiras amostras das séries sintéticas. Observa-se que os corredores acompanham dinamicamente o comportamento das séries e os 3 níveis de acurácia definem os patamares para se definir a QD. Pelo ponto de vista da monitoração, as amostras poderiam ser classificadas como aceitáveis (amostras dentro do nível mais alto de acurácia), prováveis anomalias (amostras entre os limites alto e médio) e anômalas (acima do limite mais baixo de acurácia).

Para estas amostras, avisos devem ser emitidos para os usuários, que avaliam a amostra suspeita de baixa QD.

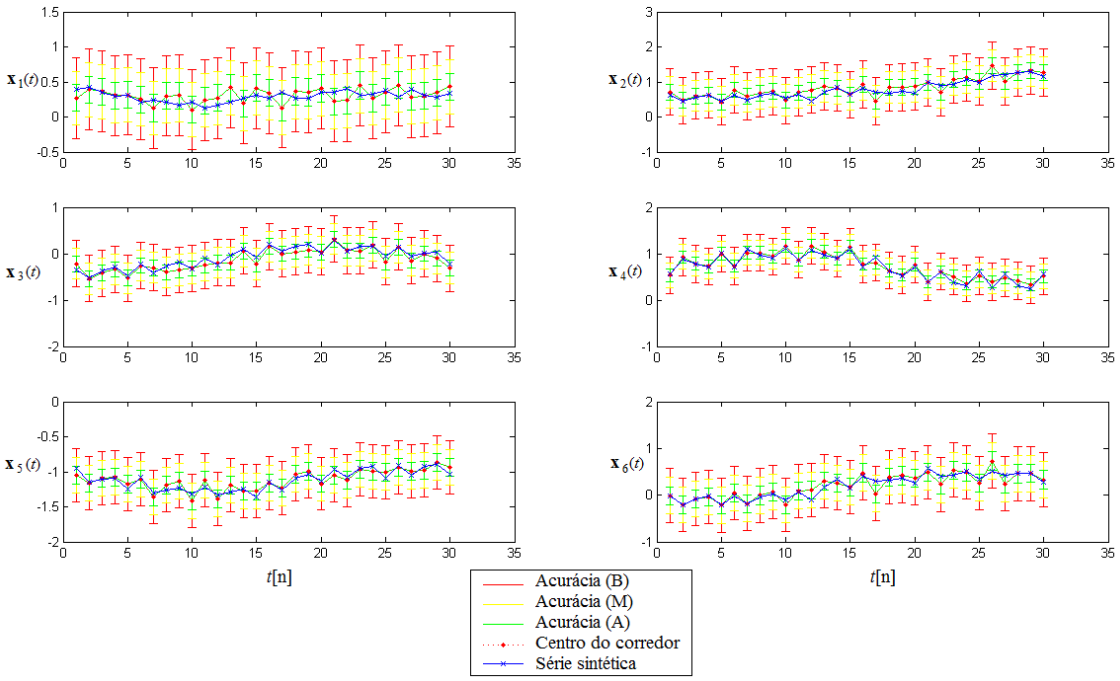

Figura 5.8: Corredores de validação para as séries sintéticas

Dados com *outliers*

Na Tabela 5.29, são mostrados os resultados da medição da qualidade do modelo (IQM) para as séries simuladas com anomalias. Novamente, observa-se que ICA teve um melhor desempenho. Com exceção da série $\mathbf{x}_5(t)$, o IGQM obtido com ICA foi melhor para todas as outras séries. No entanto, na comparação com as amostras sem *outliers*, observa-se que a maioria dos modelos foram um pouco piores. Somente o modelo para $\mathbf{x}_1(t)$ não foi pior. A piora dos modelos pode ser explicada pela influência dos *outliers* na série. Apesar disso, a diferença não foi demasiadamente discrepante, devido ao tratamento dado por SCICA. Isso evita que os corredores sejam demasiadamente distorcidos.

Já a QD (veja a Tabela 5.30), era de se esperar que fosse menor, devido às anomalias introduzidas nas séries. Era esperado que a QD reduzisse em aproximadamente 1%, uma vez que essa foi a quantidade de anomalias simuladas por série. No entanto, a queda foi de aproximadamente 4%, devido à medição da QD com modelos menos confiáveis. Apesar disso, a QD medida com ICA continua sendo mais confiável do que a QD medida sem ICA. Isso é refletido nos indicadores IQD, nos quais se observam valores mais fidedignos à QD esperada.

Tabela 5.29: Indicadores de Qualidade do modelo (IQM) para séries com anomalias

Séries sintéticas	$NMSE$	$NMSE_1$	$NMSE_2$	R	Lag_0	IGQM
ICA						
$\mathbf{x}_1(t)$	**0,2018**	**0,0603**	**0,5690**	**1,0086**	**0,9709**	**0,8325**
$\mathbf{x}_2(t)$	**0,1620**	**0,0357**	**0,3425**	**1,0369**	**0,9828**	**0,8616**
$\mathbf{x}_3(t)$	**0,1220**	**0,0799**	**0,1926**	**1,2376**	**0,9608**	**0,8801**
$\mathbf{x}_4(t)$	0,2312	0,0932	0,1933	**1,3374**	0,9525	**0,8814**
$\mathbf{x}_5(t)$	0,0871	0,1188	0,3641	**1,1911**	0,9388	0,8613
$\mathbf{x}_6(t)$	**0,1012**	**0,0325**	**0,2347**	**1,0609**	**0,9840**	**0,8719**
Sem ICA						
$\mathbf{x}_1(t)$	0,2404	0,0798	0,7338	0,9667	0,9597	0,8087
$\mathbf{x}_2(t)$	0,1717	0,0393	0,4132	0,9950	0,9808	0,8510
$\mathbf{x}_3(t)$	0,1969	0,1685	0,4573	0,9726	0,9146	0,8302
$\mathbf{x}_4(t)$	**0,1417**	**0,0287**	**0,0593**	1,2732	**0,9855**	0,8801
$\mathbf{x}_5(t)$	**0,0678**	**0,0667**	**0,2166**	1,0740	**0,9668**	**0,8703**
$\mathbf{x}_6(t)$	0,1237	0,0483	0,3588	1,0183	0,9770	0,8579

Tabela 5.30: Indicadores de Qualidade de Dados (IQD)

Séries sintéticas	Compl.	Acurácia (B)	Acurácia (M)	Acurácia (A)	IGQD
ICA					
$\mathbf{x}_1(t)$	**1,00**	**0,99**	**0,92**	**0,67**	**0,89**
$\mathbf{x}_2(t)$	**1,00**	**0,98**	**0,94**	**0,69**	**0,90**
$\mathbf{x}_3(t)$	**1,00**	**0,98**	**0,93**	**0,66**	**0,89**
$\mathbf{x}_4(t)$	**1,00**	**0,98**	**0,90**	**0,65**	**0,88**
$\mathbf{x}_5(t)$	**1,00**	**0,98**	**0,92**	**0,65**	**0,89**
$\mathbf{x}_6(t)$	**1,00**	**0,99**	**0,94**	**0,68**	**0,90**
BD	**1,00**	**0,98**	**0,92**	**0,67**	**0,89**
Sem ICA					
$\mathbf{x}_1(t)$	**1,00**	0,98	0,90	0,62	0,88
$\mathbf{x}_2(t)$	**1,00**	0,92	0,81	0,51	0,81
$\mathbf{x}_3(t)$	**1,00**	0,88	0,72	0,42	0,75
$\mathbf{x}_4(t)$	**1,00**	0,98	0,89	0,61	0,89
$\mathbf{x}_5(t)$	**1,00**	0,97	0,87	0,61	0,86
$\mathbf{x}_6(t)$	**1,00**	0,92	0,82	0,52	0,82
BD	**1,00**	0,94	0,84	0,55	0,83

Capítulo 6

Séries de Carga Elétrica

Neste capítulo, são analisadas conjuntos de séries de carga elétrica. Estas séries são dotadas de estruturas definidas frequentemente presentes em séries temporais. A estas séries, associa-se uma série explicativa de temperatura diária, que também foi incluída nas análises. As séries de carga foram configuradas de duas formas: série de picos diários e séries adjacentes ao horário de picos. Assim, os conjuntos foram avaliados aplicando-se a mesma metodologia utilizadas nas séries sintéticas para a monitoração da QD. O estudo da introdução de ICA foi concentrado na avaliação de SOBI e SOBI-RO. O pré-processamento e a modelagem das séries é avaliado através dos testes objetivos e análises subjetivas, nos contextos com e sem ICA. A detecção de *outliers* é avaliada com os métodos SCICA, ICA e sem ICA. Por fim, avaliam-se os resultados das medidas da qualidade dos modelos e da qualidade dos dados.

6.1 Dados

As séries de carga elétrica e as séries de temperatura são dados cedidos por uma concessionária de energia europeia (*East-Slovakia Power Distribution Company*), e foram utilizados em uma competição promovida em 2001 pela *European Network on Intelligent Technologies for Smart Adaptive Systems – EUNITE* [120]. A base de dados de cargas elétricas representa a demanda/oferta de energia, medida em mega-watts (MW). Os dados são coletados a cada trinta minutos durante 24 horas, gerando 48 variáveis com amostragem diária. Os dados utilizados para o desenvolvimento dos modelos correspondem ao período de 1° de janeiro de 1997 a 31 de dezembro de 1998 (730 amostras). Além dessas séries, tem-se a série correspondente à temperatura média diária em °C, para o mesmo período das séries de carga elétrica. Assim, o conjunto total destas séries corresponde a uma matriz 49x730, onde o número de linhas representa as variáveis e o número de colunas representa as amostras no tempo.

Na competição realizada em 2001, a tarefa dos competidores foi desenvolver

modelos para previsão do pico diário de carga. As amostras do mês de janeiro de 1999 foram utilizadas para comparação entre os modelos dos competidores. Dessa forma, uma das configurações dos dados aqui avaliados utiliza as amostras de janeiro de 1999 para o teste do SMQD-ST. Apesar da competição não focar na qualidade das séries, os modelos de predição seguem a mesma lógica do modelo aqui proposto para monitorar a qualidade de dados. A outra configuração utiliza as amostras do ano de 1997 para desenvolvimento e as amostras subsequentes para teste, para equilibrar ambos os conjuntos.

6.2 Separação de fontes

Para o estudo da metodologia proposta, agruparam-se as séries de carga elétrica de duas formas: séries adjacentes ao horário de pico e séries de picos diários de energia.

6.2.1 Séries adjacentes

Para o estudo proposto com ICA, é preciso selecionar as séries que irão compor o modelo. Assume-se que as séries de carga elétrica (séries-mistura) são compostas por fontes independentes misturadas e desconhecidas [30]. Assim, dado que variáveis com correlação diferente de zero não são independentes, a correlação é avaliada para se montar o conjunto de séries de entrada. Basta existir correlação linear para que as séries não sejam estatisticamente independentes.

Além da avaliação da correlação, foram selecionadas apenas as séries de carga relevantes para o contexto. Estas variáveis correspondem às séries relacionadas aos picos de energia. Na Figura 6.1, observa-se a variação média da carga ao longo do dia e o pico de oferta/demanda as 20:00h.

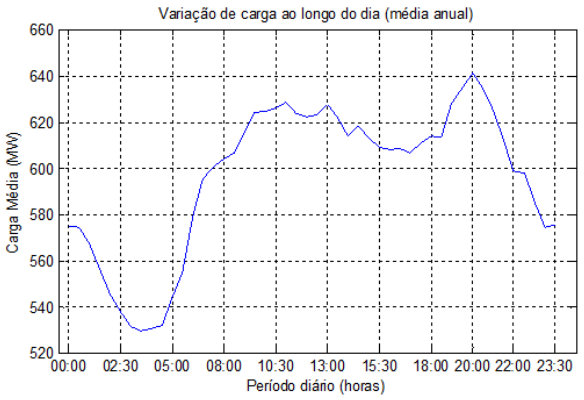

Figura 6.1: Média da variação de carga ao longo do dia.

As duas metodologias foram utilizadas para montar o conjunto de séries-mistura. Com a série do horário de pico (20:00h), avaliou-se as correlações entre as séries

adjacentes, e selecionou-se aquelas séries cuja correlação fosse acima de 0,95 (veja Figura 6.2). Assim, as séries entre 18:30h e 21:30h foram selecionadas. Além disso, utilizou-se a série de temperaturas como série explicativa. Dessa forma, o conjunto de séries-mistura é composto por uma matriz 8 x 730 (veja Figura 6.3).

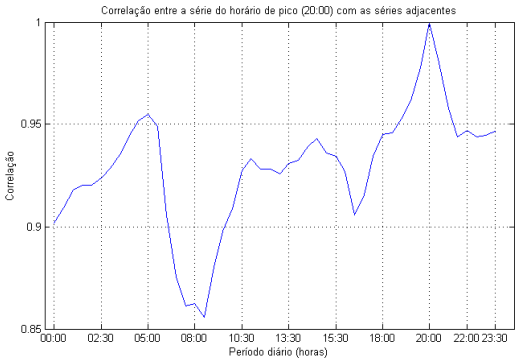

Figura 6.2: Correlação entre a série do horário de pico (20:00h) com as séries adjacentes.

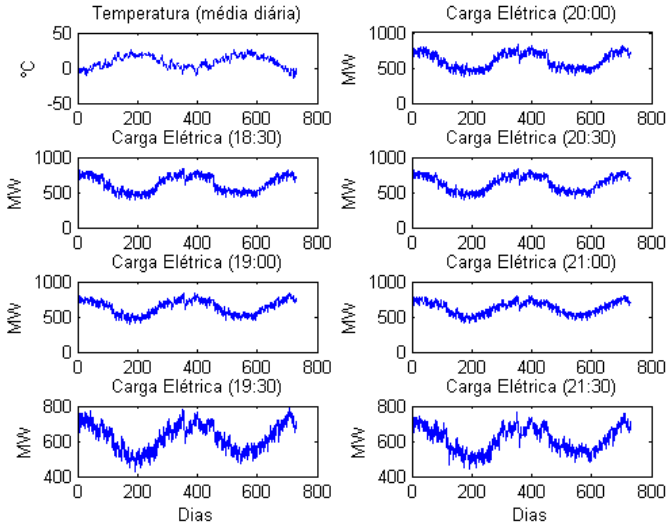

Figura 6.3: Série de temperatura e séries de demanda/oferta de carga elétrica.

Dado este conjunto de séries-mistura, a Análise de Componentes Independentes faz a extração cega das fontes, utilizando-se algoritmos de aplicação específica para séries temporais (SOBI e SOBI-RO). No entanto, uma vez que a matriz mistura não é conhecida, não é possível avaliar o índice de separabilidade E_1 [30] nem os outros indicadores utilizados na Seção 5.2. Assim, a definição dos parâmetros do modelo ICA foi realizada através da própria figura de mérito dos algoritmos utilizados para extração de fontes. Para SOBI e SOBI-RO, avaliou-se o indicador mostrado na Equação 3.11.

O indicador mede o nível de diagonalização das matrizes de covariância atrasadas, para atrasos de 1 até $n-1$ (onde n é o número de amostras da série). Espera-se que uma separação de fontes bem sucedida minimize $off(\mathbf{M})$ em direção a zero.

Apenas como parâmetro de comparação, $off(\mathbf{M})$ foi também calculado para as séries mistura, sem a introdução do bloco ICA (veja Figura 6.4). Observa-se que, dependendo do atraso, o indicador pode variar. Espera-se que os valores com SOBI e SOBI-RO estejam abaixo da linha produzida pelo indicador calculado sem ICA.

De fato, quando os algoritmos ICA são aplicados, a linha do indicador move-se em direção a zero. Em geral, SOBI-RO aparenta ter um melhor desempenho em comparação com SOBI. Para SOBI-RO, o valor mínimo de $off(\mathbf{M})$ ocorre no atraso 82, atingindo 0,27. Se utilizarmos a regra prática proposta na Seção 5.2.1, o valor mínimo ocorre para o atraso 141, atingindo 0,28. Já com SOBI, o valor mínimo ocorre no atraso 315, atingindo 0,31. Considerando-se a regra prática, o valor mínimo ocorre no atraso 170, cujo valor é 0,38. Estes valores são abaixo da referência sem ICA. Por fim, optou-se por SOBI-RO com 141 atrasos para a continuação do estudo.

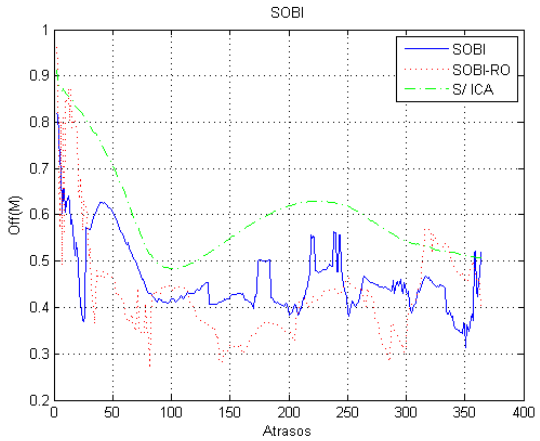

Figura 6.4: Variação do indicador $off(\mathbf{M})$ em função do numero de atrasos, para a configuração de séries de carga adjacentes.

A Figura 6.5 mostra as fontes independentes obtidas com o algoritmo SOBI-RO ($\tau = 141$). Visualmente, é possível verificar que a dinâmica da temperatura está concentrado na Fonte $\mathbf{y}_1(t)$. Na fonte $\mathbf{y}_2(t)$, é possível identificar uma componente parecida com uma tendência. Ressalta-se apenas que ela pode estar invertida, dada a incerteza na fase ocasionada por ICA. Na fonte $_3$ é possível observar ciclos quadrimestrais. Já nas outras fontes não foi possível observar nenhum comportamento especial.

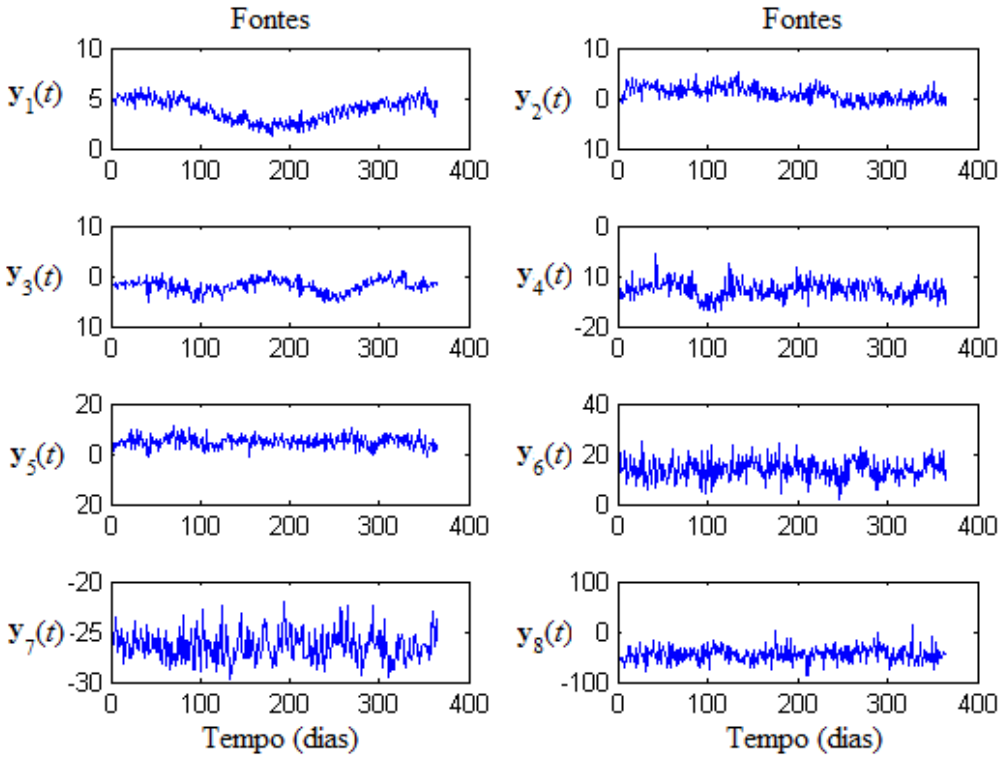

Figura 6.5: Fontes independentes obtidas a partir da configuração de séries adjacentes.

6.2.2 Série de picos

A outra configuração estudada foi obtida através da amostragem do pico diário. Esta série foi utilizada na competição da EUNITE [120]. Além disso, baseado no conhecimento especialista sobre estes dados [121], utilizou-se a temperatura como série explicativa para todas as configurações. A partir destas séries, as fontes foram extraídas (veja Figura 6.6). Observa-se que, da mesma forma que a configuração de séries adjacentes, a fração do consumo relacionada a variação anual da temperatura é mapeada para a primeira fonte. Na segunda fonte, visualmente, não se identificam estruturas.

Para a escolha do parâmetro, avaliou-se também o comportamento do índice de separabilidade $off(\mathbf{M})$ (veja Figura 6.7). Nota-se que SOBI-RO teve melhor desempenho. O menor valor de $off(\mathbf{M})$ para SOBI-RO foi com 13 atrasos, chegando a 0,0719. Para SOBI, $off(\mathbf{M})$ chegou ao mínimo de 0,1420, para 685 atrasos. Utilizando-se a regra prática, os menores valores foram 0,4221 ($\tau = 121$) e 0,108 ($\tau = 358$) para SOBI e SOBI-RO, respectivamente. Assim, para o estudo da série de picos, optou-se por utilizar o algoritmo SOBI-RO com 358 atrasos. Observa-se ainda que, nesta configuração, o desempenho de SOBI-RO foi bem melhor do que SOBI. Isso ocorre pela tendência do algoritmo em separar as fontes que têm menos correlações atrasadas umas com as outras. Nesse sentido, o ruído tende a ser uma

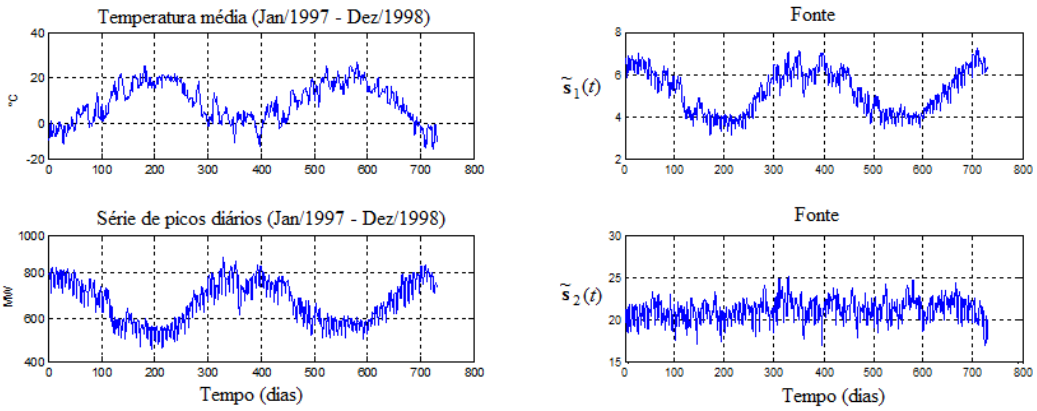

Figura 6.6: Na esquerda, série de temperaturas (acima) e série de picos (abaixo). Na direita, as duas fontes independentes obtidas com SOBI(141).

das fontes. Além disso, dada a configuração com duas séries, assumiu-se a existência de apenas duas fontes. Assim, uma fonte desestruturada e mais ruidosa é extraída. Isso leva o método SOBI-RO, que é menos sensível ao ruído, a ter um resultado bem melhor do que SOBI.

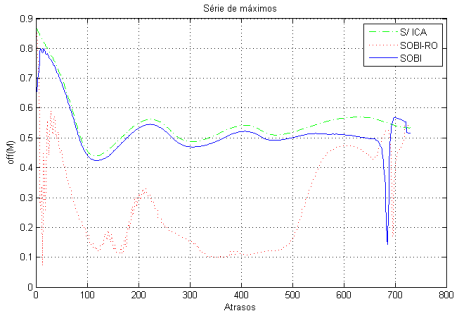

Figura 6.7: Variação do indicador $off(\mathbf{M})$ em função do número de atrasos, para a série de picos.

6.3 Pré-processamento

A mesma metodologia aplicada para pré-processar as séries sintéticas foi aplicada para as séries de carga elétrica. Antes de se aplicar o pré-processamento, avaliam-se os testes objetivos e as análises subjetivas para ambas as configurações estudadas.

6.3.1 Séries adjacentes

ICA estimada

Após extrair as fontes com SOBI-RO ($\tau = 141$), aplicaram-se os testes e análises de pré-processamento das fontes obtidas (veja a Tabela 6.1). Observa-se que a

heterocedasticidade não é identificada em nenhuma das fontes (coluna H com valores iguais a zero). De fato, observando-se o comportamento das fontes no tempo (veja a Figura 6.5), não se nota a alteração da variância. Assim, este componente não foi extraído para nenhuma das fontes.

Tabela 6.1: Pré-processamento após ICA (SOBI).

Fontes	H	TE	S	TL	C	N
$\mathbf{y}_1(t)$	0	0	7	0	0	1
$\mathbf{y}_2(t)$	0	1	0	0	0	1
$\mathbf{y}_3(t)$	0	1	0	0	0	1
$\mathbf{y}_4(t)$	0	0	0	0	0	1
$\mathbf{y}_5(t)$	0	0	0	0	0	1
$\mathbf{y}_6(t)$	0	0	0	0	0	1
$\mathbf{y}_7(t)$	0	0	0	0	0	1
$\mathbf{y}_8(t)$	0	0	0	0	0	1

A sazonalidade foi detectada apenas na fonte $\mathbf{y}_1(t)$, com período de 7 dias. Conforme pode ser observado na função de autocorrelação (veja a Figura 6.8), a fonte $\mathbf{y}_1(t)$ apresenta correlações significantes com frequência semanal. No entanto, antes de estimar a função de autocorrelação, aplicou-se o operador de primeira diferença, pois uma expressiva correlação cíclica dificultava a visualização da sazonalidade. Após a retirada da sazonalidade, com a aplicação do operador diferença sazonal, a função de autocorrelação não tem mais correlações periódicas significativas (veja Figura 6.9) e o comportamento periódico não é mais identificado. Apenas uma correlação significativa permaneceu no sétimo atraso. Observa-se também que, antes da retirada da sazonalidade, o teste de raiz unitária indicou a presença de tendência estocástica. No entanto, após a retirada da sazonalidade, o teste não se confirmou. Avaliou-se também a retirada do ciclo, extraindo-se a primeira componente do espectro de frequências (veja a Figura 6.10). No entanto, quando se recupera o ciclo no ano seguinte (conjunto de testes), percebe-se uma divergência elevada entre a série estimada e a série real. Pode-se explicar esta piora no desempenho devido ao ciclo detectado ser estocástico. Assim, a frequência e a amplitude do ciclo tem uma variação estocástica ano a ano. Isso fez com que a modelagem determinística através da componente senoidal de frequência não fosse bem sucedida.

Já a tendência estocástica foi detectada apenas na fonte $\mathbf{y}_2(t)$, através dos testes combinados de Dickey-Fuller e Phillips-Perron a um nível de significância de 5%. Esta tendência também pode ser notada na função de autocorrelação, que tem os componentes defasados decaindo lentamente (veja a Figura 6.11). Assim, aplicou-se o operador de primeira diferença nesta fonte. Após o pré-processamento, a função de autocorrelação segue um padrão de uma série estacionária (veja a Figura 6.12).

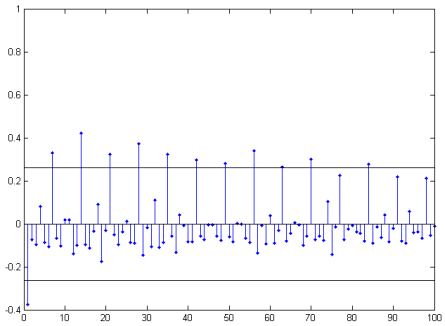

Figura 6.8: Função de autocorrelação da fonte $\mathbf{y}_1(t)$

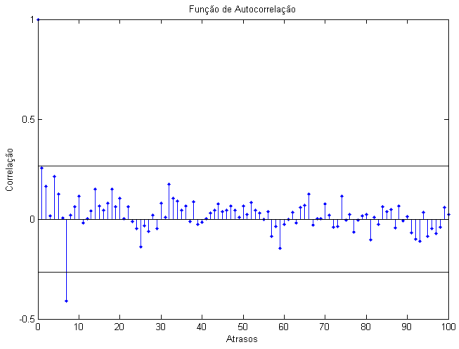

Figura 6.9: Função de autocorrelação da fonte $\mathbf{y}_1(t)$ após a retirada da sazonalidade

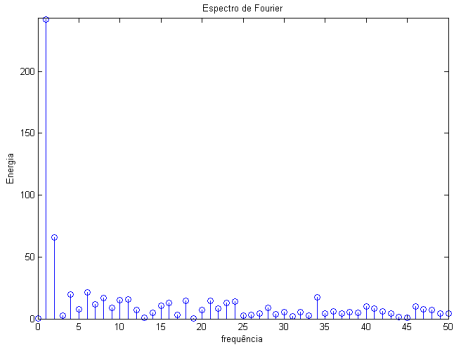

Figura 6.10: Espectro de frequências da fonte $\mathbf{y}_1(t)$, antes do pré-processamento

Observa-se que apenas o primeiro atraso é relevante. Nenhuma outra tendência (estocástica ou linear) foi identificada.

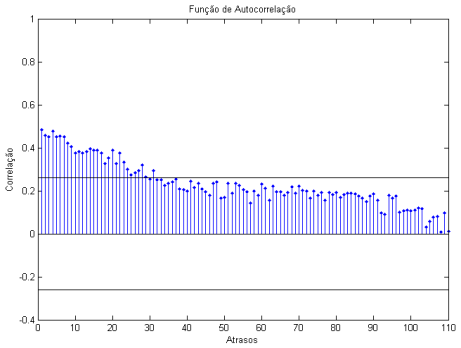

Figura 6.11: Função de autocorrelação da fonte $\mathbf{y}_2(t)$

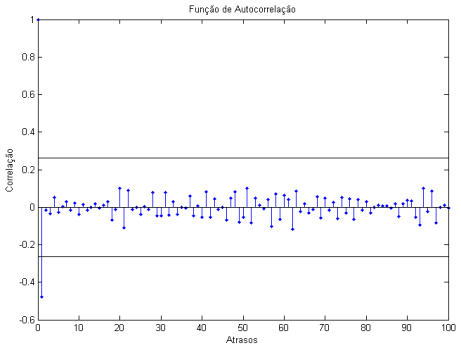

Figura 6.12: Função de autocorrelação da fonte $\mathbf{y}_2(t)$ após a retirada da tendência

Os ciclos foram detectados na fonte $\mathbf{y}_3(t)$, através da análise do espectro de frequências de Fourier (veja a Figura 6.13). Ciclos semestrais (segundo componente de frequência) e quadrimestrais (terceiro componente frequência) foram detectados. Visualmente, também é possível perceber a presença dos ciclos na própria fonte (veja a Figura 6.5). No entanto, quando estas frequências foram isoladas do espectro, ocorreu o mesmo problema ocorrido na fonte $\mathbf{y}_1(t)$ com a retirada do ciclo. Assim, fez-se apenas a diferenciação de primeira ordem para torná-la estacionária.

Para se certificar que não há nenhuma tendência, sazonalidade ou ciclo, avaliaram-se também as funções de autocorrelação e os espectros de frequências das outras fontes. Nas fontes $\mathbf{y}_4(t)$ e $\mathbf{y}_7(t)$, não se detectou nenhuma periodicidade sazonal e os testes de raiz unitária não identificaram tendência estocástica. Assim, a série já é estacionária e os possíveis padrões identificados na função de autocorrelação podem ser modelados na sequência. As fontes $\mathbf{y}_5(t)$, $\mathbf{y}_6(t)$ e $\mathbf{y}_8(t)$ não apresentaram correlações atrasadas significativas, já sendo estacionárias após a transformação ICA, o que reduz a necessidade de análises e facilita o pré-processamento. Por fim,

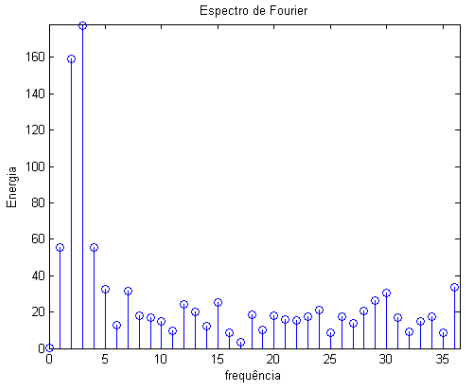

Figura 6.13: Espectro de frequências da fonte $\mathbf{y}_3(t)$

todas as séries foram normalizadas para média zero e desvio-padrão unitário e apresentadas ao bloco de modelagem.

Sem ICA

O pré-processamento também foi avaliado sem a aplicação de ICA e resultados dos testes e análises são mostrados na Tabela 6.2. Da mesma forma que no pré-processamento com ICA, não se detectou heterocedasticidades nas séries de carga elétrica e nem na série de temperatura (veja a Figura 6.3). Nota-se que, de fato, as variâncias não se alteram com o decorrer do tempo.

Tabela 6.2: Pré-processamento sem ICA.

Séries	H	TE	S	TL	C	N
$\mathbf{x}_1(t)(temperatura)$	1	0	0	0	0	1
$\mathbf{x}_2(t)(18:30)$	0	0	7	0	0	1
$\mathbf{x}_3(t)(19:00)$	0	0	7	0	0	1
$\mathbf{x}_4(t)(19:30)$	0	0	7	0	0	1
$\mathbf{x}_5(t)(20:00)$	0	0	7	0	0	1
$\mathbf{x}_6(t)(20:30)$	0	0	7	0	0	1
$\mathbf{x}_7(t)(21:00)$	0	0	7	0	0	1
$\mathbf{x}_8(t)(21:30)$	0	0	7	0	0	1

Analisando-se a função de autocorrelação, observa-se que as séries não são estacionárias (veja a Figura 6.14). Ressalta-se que este padrão é seguido por todas as outras séries de carga, como era de se esperar, já que são séries muito similares.

Inicialmente, o teste de raiz unitária identificou a presença de tendência estocástica nas séries. Ao se retirar esta tendência, aplicando-se o operador de primeira diferença, a função de autocorrelação passa a revelar a presença de sazonalidade semanal (veja a Figura 6.15). No entanto, quando a sazonalidade foi retirada, o teste de raiz unitária não se confirmou. Assim, somente a sazonalidade foi retirada para

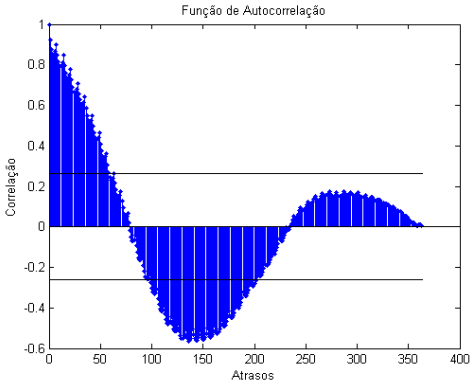

Figura 6.14: Função de autocorrelação para a série de carga (20:00), antes do pré-processamento

tornar estacionária a série (veja a Figura 6.16) e 4 atrasos significativos podem ser utilizados para identificar os padrões presentes nesta série.

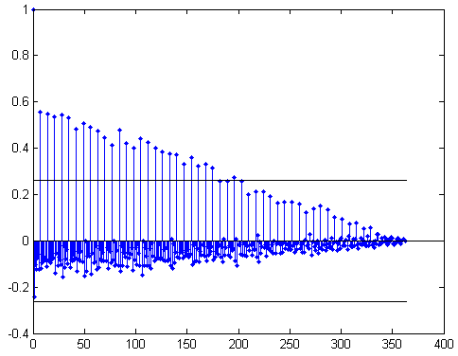

Figura 6.15: Função de autocorrelação para a série de carga (20:00), após a aplicação do operador de primeira diferença

Na série de temperatura, identificou-se a presença de um ciclo anual. No entanto, a retirada do ciclo apresentou o mesmo problema ocorrido com a fonte estimada $\mathbf{y}_1(t)$, na Seção anterior. Assim, optou-se apenas por tornar estacionária a série aplicando-se o operador de primeira diferença. Por fim, todas as séries foram normalizadas e passadas ao bloco de modelagem.

6.3.2 Série de picos

ICA estimada

A segunda configuração estudada é através da série de picos. As séries e as fontes extraídas são mostradas na Figura 6.6. Na primeira fonte, o teste de raiz unitária detectou a presença de tendência estocástica. Ao se retirá-la, assim como no caso com séries adjacentes, identifica-se mais facilmente a presença de sazonalidade se-

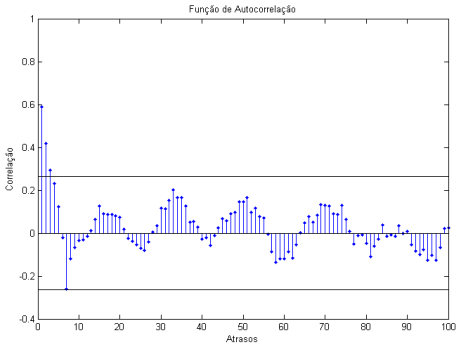

Figura 6.16: Função de autocorrelação para a série de carga (20:00), após o pré-processamento com a retirada da sazonalidade

manal na função de autocorrelação. Se a sazonalidade for retirada antes do teste de raiz unitária, o teste não se confirma. Além disso, a retirada da periodicidade anual através da extração de componentes de frequência também apresenta o mesmo problema reportado nas seções anteriores com $\mathbf{y}_1(t)$. Assim, a série é tornada estacionária retirando-se apenas a sazonalidade.

Já na segunda fonte, apesar de não ter sido possível identificar visualmente nenhuma estrutura no tempo, identifica-se um padrão de sazonalidade semanal (veja Figura 6.17).

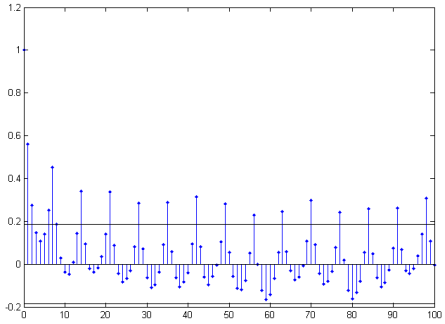

Figura 6.17: Função de autocorrelação para a série de picos, antes do pré-processamento

Os resultados dos testes e análises são mostrados na Tabela 6.3. Nota-se que nenhuma heterocedasticidade, tendência ou ciclos foram extraídos e que a normalização foi realizadas para ambas as fontes.

Tabela 6.3: Pré-processamento nas fontes para configuração de serie de picos.

Fontes	H	TE	S	TL	C	N
$\tilde{\mathbf{s}}_1(t)$	0	0	7	0	0	1
$\tilde{\mathbf{s}}_2(t)$	0	0	7	0	0	1

Sem ICA

Para a configuração de série de picos, os testes e análises indicaram o mesmo pré-processamento para a configuração de séries adjacentes (veja a Tabela 6.4). Este fato era de se esperar, dado que muitas amostras da série de picos foram retiradas das séries adjacentes. Assim, espera-se que mantenham padrão similar. Os testes e análises para a série de temperatura também foram similares aos realizados na configuração de séries adjacentes. Isto ocorreu pelo fato de que, para o primeiro ano de amostragem, a série é idêntica à série utilizada na configuração adjacente. Além disso, o comportamento da temperatura no segundo ano seguiu uma padrão similar ao primeiro ano.

Tabela 6.4: Pré-processamento para a configuração de série de picos sem a aplicação de ICA.

Séries	H	TE	S	TL	C	N
Temperatura	0	1	0	0	0	1
Série de picos	0	0	7	0	0	1

6.4 Modelagem

6.4.1 Séries adjacentes

ICA estimada

Os parâmetros dos modelos obtidos são mostrados na Tabela 6.5. Espera-se que ao aplicar ICA, as estrutura dos dados sejam mapeadas nas fontes independentes e facilite a busca por padrões nos dados (padrões de QD). Da mesma forma, espera-se que os sinais sem estruturas e padrões, assim como os ruídos, sejam isolados e mais facilmente identificados pelo método de modelagem.

Observa-se que o método não detecta atrasos relevantes para se modelar o resíduo das fontes $\mathbf{y}_6(t)$ e $\mathbf{y}_7(t)$. Tampouco, o método detectou séries explicativas para ambos. Assim, nenhum modelo neural ou linear foi gerado. Estes resíduos foram tratados como ruído e apenas substituídos pela média das amostras anteriores. Já para o resíduo da fonte $\mathbf{y}_1(t)$, foi necessário apenas um modelo linear, com 4 atrasos e uma série explicativa. Da mesma forma, o resíduo de $\mathbf{y}_3(t)$ foi modelado linearmente com 4 atrasos, sem série explicativa. O último modelo linear foi para $\mathbf{y}_5(t)$, que utilizou apenas um atraso na entrada. Já os resíduos das fontes $\mathbf{y}_2(t)$, $\mathbf{y}_4(t)$ e $\mathbf{y}_8(t)$ foram modelados não linearmente. Os modelos não lineares foram gerados com redes neurais MLP, utilizando no máximo 2 neurônios na camada escondida.

Nota-se que, uma vez que os resíduos das fontes são independentes, o método não detecta a necessidade de séries explicativas para a maioria dos modelos ($\mathbf{y}_3(t)$,$\mathbf{y}_4(t)$, $\mathbf{y}_5(t)$, $\mathbf{y}_6(t)$ e $\mathbf{y}_7(t)$). Além disso, alguns resíduos são classificadas como ruídos, o que simplifica a modelagem. No entanto, em alguns casos, o método detecta a necessidade de séries explicativas. Este fato pode ter ocorrido devido à extração de mais fontes do que na realidade existem. Assim, algumas fontes que deveriam ser independentes podem conter informação uma da outra, como é o caso das fontes $\mathbf{y}_1(t)$,$\mathbf{y}_2(t)$ e $\mathbf{y}_7(t)$. Por fim, observa-se que ICA permitiu a construção de modelos ajustados a cada um dos resíduos das fontes, desde modelos simplificados até modelos não lineares.

Tabela 6.5: Modelagem para o cenário com ICA (SOBI)

Fontes	Max Lag	Explicativa	NNE
$\mathbf{y}_1(t)$	7	1	0
$\mathbf{y}_2(t)$	1	1	2
$\mathbf{y}_3(t)$	4	0	0
$\mathbf{y}_4(t)$	5	0	2
$\mathbf{y}_5(t)$	1	0	0
$\mathbf{y}_6(t)$	0	0	0
$\mathbf{y}_7(t)$	0	0	0
$\mathbf{y}_8(t)$	2	1	1

Sem ICA

A metodologia de modelagem anterior foi também avaliada sem o bloco ICA. Os parâmetros obtidos para os modelos são mostrados na Tabela 6.6. Nota-se que o resíduo da série de temperatura não tem atrasos significativos e nem precisa de séries explicativas. Assim, nenhum modelo mais complexo foi utilizado, apenas a média das amostras passadas. O fato da temperatura não ter séries explicativas já era esperado, uma vez que as séries de carga elétrica não influenciam a temperatura.

Já as séries de carga necessitaram da série de temperatura para serem explicadas e 4 atrasos. Ainda, todos os modelos são não lineares, variando-se apenas o número de neurônios da camada escondida. Foram utilizados entre 1 e 4 neurônios para modelar os resíduos das séries de carga. Esta semelhança entre os modelos era de se esperar, uma vez que as séries de carga são similares e tiveram o mesmo pré-processamento. Por fim, nota-se que, de maneira geral, o método sem ICA propôs modelos mais complexos (sem modelos simplificados ou lineares e com mais neurônios na camada escondida da rede neural) do que no caso com ICA.

Tabela 6.6: Modelagem para o cenário com sem ICA

Séries	Max Lag	Explicativa	NN
$\mathbf{x}_1(t)(Temperatura)$	0	0	0
$\mathbf{x}_2(t)(18:30)$	7	1	2
$\mathbf{x}_3(t)(19:00)$	7	1	4
$\mathbf{x}_4(t)(19:30)$	7	1	2
$\mathbf{x}_5(t)(20:00)$	7	1	1
$\mathbf{x}_6(t)(20:30)$	7	1	2
$\mathbf{x}_7(t)(21:00)$	7	1	2
$\mathbf{x}_8(t)(21:30)$	7	1	1

6.4.2 Série de picos

Os modelos para as séries de picos podem ser comparados com os melhores modelos desenvolvidos na competição da EUNITE [120] (veja a Figura 6.18). Os modelos foram comparados com base no MAPE [1] [121] e no valor do erro máximo (MAX) [121]. Quanto menor for o MAPE e o MAX, melhor classificado é o modelo. O MAPE para os três primeiros colocados foi de 1,98%, 2,14% e 2,49%.

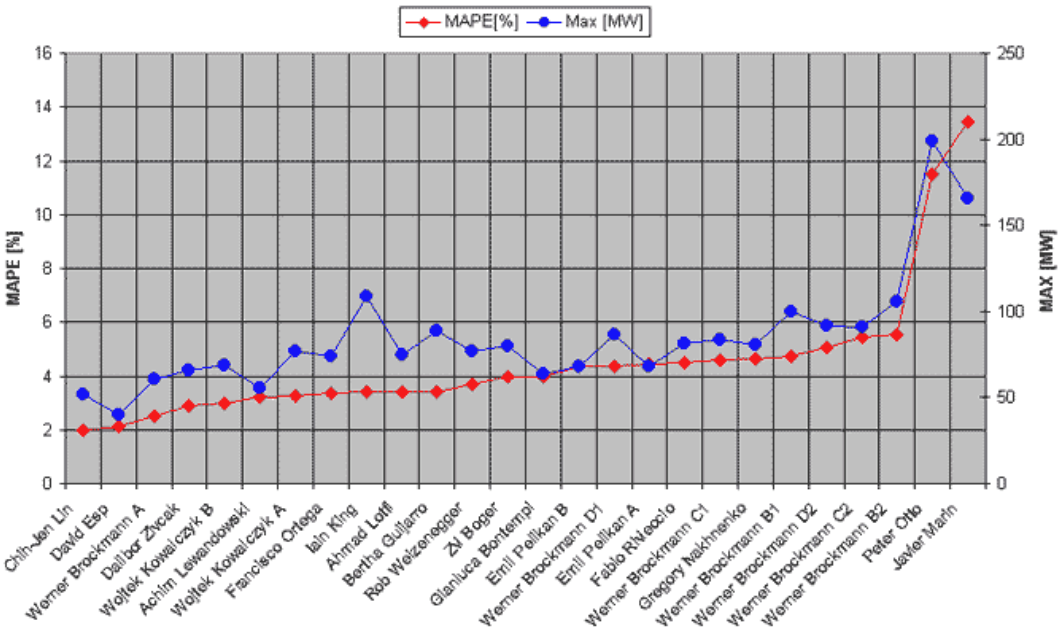

Figura 6.18: Resultado da competição para modelagem de séries de carga elétrica. Extraído de [120].

ICA estimada

Os parâmetros das fontes extraídas com ICA são mostrados na Tabela 6.7. Para o resíduo da fonte y_1 foram necessários 9 atrasos para construir um modelo não linear com dois neurônios na camada escondida. O resíduo da fonte y_2 foi modelado com 10

atrasos, série explicativa e um neurônio na camada escondida. O MAPE alcançado com este modelo foi de 2,37%, o que colocaria o modelo entre os melhores colocados na competição da EUNITE. Este resultado mostra que o método de modelagem proposto é bastante eficiente pois, a partir de modelos parcimoniosos (no máximo 2 neurônios na camada escondida), foi capaz de extrair bons padrões de qualidade para os dados. Além disso, o método não é projetado com foco de especialista em carga elétrica e, assim, o fato estar entre os melhores modelos reforça a aplicação de ICA e de toda a metodologia aplicada como uma ferramenta para se extrair bons padrões de QD.

Tabela 6.7: Modelo da série de picos, para o cenário com ICA

Fontes	Max Lag	Explicativa	NN
$\tilde{\mathbf{s}}_1(t)$	9	1	2
$\tilde{\mathbf{s}}_2(t)$	10	1	1

Sem ICA

Os parâmetros do modelo sem ICA são mostrados na Tabela 6.8. Um modelo linear com dois atrasos na entrada e nenhuma série explicativa foi utilizado para modelagem do resíduo da temperatura. Nota-se que o modelo difere ligeiramente do modelo proposto na configuração de séries adjacentes. Este fato ocorreu devido a diferença entre os dois tipos de configuração. Para séries adjacentes, a amostragem foi de um ano. Já na configuração de séries de picos, foram utilizados dois anos de amostras pra o desenvolvimento. Neste universo mais rico de informações, foi possível detectar correlação significativa com as amostras passadas da série de temperatura.

O resíduo da série de picos diários foi modelado de forma semelhante as séries adjacentes. Foi necessária a série de temperatura como série explicativa e também 7 atrasos para construir o modelo. Este número de atraso refere-se a padrões de consumo semanais. O modelo sugerido foi não-linear, com 2 neurônios na camada escondida da rede neural MLP, o que é semelhante aos encontrados em séries adjacentes. Por fim, o valor encontrado do MAPE para a série de picos foi de 3,42%, o que classifica este modelo apenas entre os 10 melhores da competição. Observa-se também que o próprio método ICA emitiu menos falsos alarmes do que sem ICA.

Tabela 6.8: Modelo para a série de picos, para o cenário sem ICA

Séries	Max Lag	Explicativa	NN
Temperatura	2	0	0
Séries de picos	7	1	2

6.5 Detecção de *Outliers*

Para se avaliar a detecção de *outliers* nas series de carga, selecionou-se as amostras afastadas da média da distribuição das séries diferenciadas, com três desvios padrão. Essas amostras extremas foram marcadas como *outliers* e é a figura de mérito para a detecção. Assim, espera-se que o melhor método seja capaz de detectar estas amostras, emitindo o mínimo de falso alarme. Os resultados são mostrados na Tabela 6.9. Em negrito estão os melhores resultados e, na fonte $\mathbf{x}_4(t)$, não há *outliers* para se detectar. Para as demais séries, nota-se que não houve diferença na detecção com SCICA, ICA e Sem ICA. No entanto, com exceção das séries $\mathbf{x}_1(t)$, $\mathbf{x}_5(t)$ e $\mathbf{x}_6(t)$, SCICA teve o menor nível de falso alarme. Assim como apontado na situação simulada da seção anterior, estes indicadores indicam que este método pode ser utilizado como detector de *outliers*. Ainda, observa-se que o método com ICA também é superior ao método sem ICA, emitindo, em geral, menos falsos alarmes enquanto detecta o mesmo número de *outliers*.

Tabela 6.9: Nível de alarmes, com substituição

Séries	Indicadores	SCICA	ICA	Sem ICA
$\mathbf{x}_1(t)$	Falso alarme (%)	0,86	**0,57**	1,15
	Detecção (%)	**100%**	**100%**	**100%**
$\mathbf{x}_2(t)$	Falso alarme (%)	**0,58**	0,86	2,01
	Detecção (%)	**100%**	**100%**	**100%**
$\mathbf{x}_3(t)$	Falso alarme (%)	**0,29**	0,86	0,86
	Detecção (%)	**100%**	**100%**	**100%**
$\mathbf{x}_4(t)$	Falso alarme (%)	**0,29**	0,57	1,44
	Detecção (%)	**-**	**-**	**-**
$\mathbf{x}_5(t)$	Falso alarme (%)	0,58	**0,29**	1,15
	Detecção (%)	**100%**	**100%**	**100%**
$\mathbf{x}_6(t)$	Falso alarme (%)	0,86	**0,29**	**0,29**
	Detecção (%)	**100%**	**100%**	**100%**
$\mathbf{x}_7(t)$	Falso alarme (%)	**0,58**	1,15	1,72
	Detecção (%)	**100%**	**100%**	**100%**
$\mathbf{x}_8(t)$	Falso alarme (%)	**0,29**	0,86	0,57
	Detecção (%)	**100%**	**100%**	**100%**

6.6 Qualidade de Dados

Para a medição da QD, mediu-se tanto a qualidade das séries de carga elétrica quanto a qualidade dos modelos utilizados.

6.6.1 Qualidade dos Modelos

Os indicadores de qualidade dos modelos (IQM) são mostrados na Tabela 6.10. Os melhores resultados estão em negrito. Analisando-se as séries de carga para a configuração de séries adjacentes (séries $\mathbf{x}_2(t)$ a $\mathbf{x}_8(t)$), observa-se que, com exceção das séries $\mathbf{x}_2(t)$, $\mathbf{x}_4(t)$ e $\mathbf{x}_8(t)$, $NMSE$, $NMSE_1$ e $NMSE_2$ são menores para as séries modeladas com ICA. Além disso, todos os valores de $NMSE_1$ e $NMSE_2$ estão abaixo de 1, indicando que a metodologia proporciona melhores resultados do que modelos simplificados (média ou o melhor estimador de caminho aleatório). Essa melhora vale tanto para o caso ICA quanto para o caso sem ICA. Analisando-se o indicador do atraso R, verifica-se que, com exceção dos modelos das séries $\mathbf{x}_2(t)$ e $\mathbf{x}_6(t)$, todos os outros modelos que utilizam ICA são menos sensíveis ao problema do atraso. A maioria dos resultados para Lag_0 também indicam essa melhora. Ainda, observa-se que, tanto para o caso com ICA quanto para o caso sem ICA, todos os indicadores de atraso R estão acima de 1. Novamente, isso mostra que a metodologia produz bons modelos mesmo sem ICA. Por fim, a maioria dos indicadores consolidados ($IGQM$) apontam para melhora com a introdução de ICA.

Avaliando-se a configuração com a série de picos, observa-se que todos os indicadores também mostram esta melhora. Além disso, $NMSE_1$ e $NMSE_2$ são abaixo de 1, mostrando também a melhora em relação aos modelos simplificados. Já valor de R acima de 1 mostra que o modelo também não sofre do problema de atraso. Além disso, observa-se que Lag_0 é maior do que o caso sem ICA, indicando a maior correlação dos padrões de QD com o valor esperado.

Por outro lado, para a série de temperatura, nota-se que o desempenho não foi satisfatório para nenhum dos dois cenários (com e sem ICA). Em ambos os casos, o indicador R ficou abaixo de 1. Isso ocorre pela dificuldade natural de se identificar padrões nos resíduos deste tipo de série utilizando-se apenas amostras passadas. Assim, os padrões de QD obtidos estão sujeitos ao problema do atraso na estimação. Este fato também pode ser notado no indicador $NMSE_2$ que é muito próximo a 1 (para o caso sem ICA). Isso indica que o modelo encontrado é próximo ao melhor estimador de caminho aleatório. Ainda, os indicadores apontam para uma piora no desempenho com ICA nesta série. Esse fato pode ser explicado pela forma de configuração utilizada. Uma vez que o objetivo é monitorar as séries de carga e não a série de temperatura, a configuração incluiu esta série para explicar as séries de carga. O contrário não é verdadeiro, pois não se espera que a série de carga explique a temperatura. Assim, a série de temperatura, que já é independente das outras, foi transformada por aproximações estatísticas que a distorceram e produziram modelos piores. Assim, caso se queira monitorar a temperatura, outras configurações devem ser propostas.

Tabela 6.10: Indicadores de Qualidade do modelo (IQM) nos cenários com e sem a introdução de ICA, para as configurações de séries adjacentes e a série de picos

Séries	$NMSE$	$NMSE_1$	$NMSE_2$	R	Lag_0	IGQM
ICA						
$\mathbf{x}_1(t)(Temperatura)$	0,2857	0,1893	2,85	**0,9673**	0,9109	0,7054
$\mathbf{x}_2(t)(18:30)$	0,0462	0,1058	0,6890	1,0335	0,9485	0,8188
$\mathbf{x}_3(t)(19:00)$	**0,0456**	**0,1036**	**0,7457**	**1,0450**	**0,9501**	**0,8145**
$\mathbf{x}_4(t)(19:30)$	0,0424	0,1057	0,8049	**1,0382**	0,9489	0,8084
$\mathbf{x}_5(t)(20:00)$	**0,0412**	**0,1082**	**0,7461**	**1,0339**	**0,9473**	**0,8131**
$\mathbf{x}_6(t)(20:30)$	**0,0380**	**0,1251**	**0,6860**	1,0331	**0,9387**	**0,8171**
$\mathbf{x}_7(t)(21:00)$	**0,0331**	**0,1264**	**0,7110**	**1,0436**	**0,9387**	**0,8154**
$\mathbf{x}_8(t)(21:30)$	0,0368	0,1741	0,9259	**1,0337**	0,9158	0,7906
Série de picos	**0,0235**	**0,4429**	**0,4434**	**1,2469**	**0,7635**	**0,8005**
Sem ICA						
$\mathbf{x}_1(t)(Temperatura)$	**0,1687**	**0,0671**	**1,0149**	**0,9666**	**0,9665**	**0,7892**
$\mathbf{x}_2(t)(18:30)$	**0,0425**	**0,0941**	**0,6133**	**1,0372**	**0,9532**	**0,8277**
$\mathbf{x}_3(t)(19:00)$	0,0460	0,1453	0,7816	1,0398	0,9472	0,8102
$\mathbf{x}_4(t)(19:30)$	**0,0374**	**0,0883**	**0,6731**	1,0289	**0,9566**	**0,8217**
$\mathbf{x}_5(t)(20:00)$	0,0427	0,1041	0,7948	1,0255	0,9426	0,8071
$\mathbf{x}_6(t)(20:30)$	0,0386	0,1263	0,6954	**1,0469**	0,9373	0,8168
$\mathbf{x}_7(t)(21:00)$	0,0392	0,1778	1,0021	1,0430	0,9155	0,7853
$\mathbf{x}_8(t)(21:30)$	**0,0321**	**0,1410**	**0,7504**	1,0275	**0,9292**	**0,8085**
Série de picos	0,0338	0,9466	0,9349	1,1971	0,5608	0,6607

6.6.2 Qualidade das Séries de Carga Elétrica

A Tabela 6.11 mostra os indicadores da qualidade das séries de carga elétrica e temperatura. Em negrito, são mostrados os valores mais elevados. Nota-se que os dados avaliados não contêm nenhum valor faltante. Assim, a completude atingiu o seu valor máximo de 1 para todos os casos. Analisando-se os indicadores da acurácia para o banco de séries completo (BD), nota-se que as medidas foram de 0,9974, 0,9585 e 0,7273, para os níveis baixo (B), médio (M) e alto (A) de acurácia, respectivamente. O nível de qualidade está dentro do esperado para os valores definidos para k, 3, 2 e 1, respectivamente. Pois se espera que, se os padrões foram bem modelados, o erro produzido tenda a uma distribuição normal.

Observa-se que, em geral, os indicadores de QD com ICA são mais elevados. Assumindo-se que os dados estudados são de boa qualidade, poderia se dizer que a medição com ICA tem um melhor desempenho. De fato, podemos afirmar apenas que as medidas com ICA são mais confiáveis, pois os modelos utilizados na medição têm melhor desempenho. Quando os modelos utilizados são menos confiáveis como, por exemplo, o modelo para série de temperatura, o resultado da medição da QD é menos confiável. Nota-se, inclusive, que o IGQD desta série é menor quando ICA é introduzida. O mesmo ocorre para a série $\mathbf{x}_8(t)$, que é melhor modelada sem ICA. Assim, o indicador de IGQD para esta série foi maior sem ICA. Já a série $\mathbf{x}_6(t)$, apesar de parecer contrariar esta lógica, tem a qualidade dos modelos muito parecidas (veja o IGQM para a série $\mathbf{x}_6(t)$ na Tabela 6.10). Assim, o que se observa é que o IGQD é também bastante similar (IGQD=0,92) .

As medidas de QD permitem a gestão da qualidade desde indicadores consolidados (IGQD do BD) até indicadores pontuais para cada uma das séries e para cada uma das dimensões de QD. Por exemplo, analisando-se o indicador consolidado IGQD, observa-se que todas as séries de carga elétrica têm valores similares, entre 0,91 e 0,93. Estes valores podem fazer sentido para a gestão da QD, pois, de fato, as séries tem a mesma origem e espera-se que tenham a mesma qualidade. Ainda, uma vez que os dados foram fornecidos para uma competição, espera-se que os dados tenham sido pré-certificados antes de serem distribuídos. Assim, um nível de qualidade elevado, acima de 0,9, poderia ser esperado. Outra característica que é mostrada pelos indicadores é o nível de regularidade das séries. Por exemplo, o indicador de acurácia (A) mostra que o nível de amostras fora do corredor é próximo a uma distribuição normal de erros, ou seja, próximo a 69% para 1 desvio-padrão. Outra visão da qualidade poderia ser através da associação do indicador de acurácia (B) com o nível de *outliers* da série. Observa-se que, para o caso com ICA, este nível mantém-se próximo de 0,9974. Este valor está dentro do esperado para *outliers* suaves em uma série estocástica (1 a cada 150 amostras [1]). Já o indicador de

acurácia (M), poderia indicar o nível de credibilidade das amostras. Este indicador muito inferior ao esperado poderia indicar que os dados não tem credibilidade.

Tabela 6.11: Indicadores de Qualidade de Dados (IQD)

ICA					
Séries	Completude	Acurácia (B)	Acurácia (M)	Acurácia (A)	IGQD
$\mathbf{x}_1(t)(Temperatura)$	**1,00**	**0,9914**	0,9255	0,6476	0,8911
$\mathbf{x}_2(t)(18:30)$	**1,00**	**0,9971**	**0,9542**	**0,7736**	**0,9312**
$\mathbf{x}_3(t)(19:00)$	**1,00**	**0,9971**	**0,9685**	**0,7622**	**0,9319**
$\mathbf{x}_4(t)(19:30)$	**1,00**	**0,9971**	**0,9628**	**0,7479**	**0,9269**
$\mathbf{x}_5(t)(20:00)$	**1,00**	**1,0000**	**0,9599**	**0,7249**	**0,9212**
$\mathbf{x}_6(t)(20:30)$	**1,00**	**0,9971**	0,9599	0,7364	0,9234
$\mathbf{x}_7(t)(21:00)$	**1,00**	**1,0000**	**0,9742**	**0,7622**	**0,9341**
$\mathbf{x}_8(t)(21:30)$	**1,00**	**0,9971**	**0,9542**	0,7135	0,9162
Série de picos	**1,00**	**1,0000**	**0,9677**	**0,6774**	**0,9113**
BD	**1,00**	**0,9974**	**0,9585**	**0,7273**	**0,9208**
Sem ICA					
$\mathbf{x}_1(t)(Temperatura)$	**1,00**	**0,9914**	**0,9483**	**0,6868**	**0,9066**
$\mathbf{x}_2(t)(18:30)$	**1,00**	0,9770	0,9483	0,7098	0,9088
$\mathbf{x}_3(t)(19:00)$	**1,00**	0,9856	0,9598	0,7213	0,9167
$\mathbf{x}_4(t)(19:30)$	**1,00**	0,9799	0,9339	0,7184	0,9080
$\mathbf{x}_5(t)(20:00)$	**1,00**	0,9828	0,8994	0,6379	0,8800
$\mathbf{x}_6(t)(20:30)$	**1,00**	0,9943	**0,9626**	**0,7443**	**0,9253**
$\mathbf{x}_7(t)(21:00)$	**1,00**	0,9770	0,9109	0,6293	0,8793
$\mathbf{x}_8(t)(21:30)$	**1,00**	0,9943	0,9511	**0,7443**	**0,9224**
Série de picos	**1,00**	**1,0000**	0,8824	0,5882	0,8676
BD	**1,00**	0,9864	0,9310	0,7140	0,9010

Capítulo 7

Séries Financeiras

Neste capítulo, utiliza-se o sistema de monitoração da QD no contexto de séries financeiras. Assim como nos estudos anteriores, cada uma das etapas da metodologia é avaliada. O sistema é avaliado paralelamente com dois conjuntos de séries: séries de preços de ações de empresas de um mesmo setor e séries de níveis de preços diários (abertura, máximo, mínimo e fechamento) para uma ação individual. A separação de fontes é avaliada para os melhores algoritmos de séries temporais (SOBI e SOBI-RO). Na sequência do processamento, apenas o melhor método é utilizado para comparações com e sem ICA. No processo de detecção de *outliers*, as amostras anômalas são avaliadas em detalhes. Por fim, a qualidade dos modelos e a qualidade dos dados é medida, avaliada e comparada entre os métodos.

7.1 Dados

As séries temporais financeiras referem-se aos preços de ações negociadas em bolsa de valores, obtidos de fontes "pseudo-certificadas", a partir do banco de dados de um software pago [122], e fontes não certificadas, obtidas gratuitamente na Internet [123]. O provedor gratuito não garante a certificação das séries. Já o fabricante do software, afirma que as séries são certificadas. Isso permite que se faça uma comparação para a busca de problemas de qualidade.

Avaliaram-se as séries financeiras em duas configurações distintas. A primeira configuração utilizada incluiu dois anos de amostras diárias dos preços de fechamento das ações da SUN, IBM e Microsoft (504 amostras cada), para o período entre abril de 2004 e março de 2006 (veja a Figura 7.1). Esta configuração foi baseada na hipótese de que as séries das empresas que disputam o mesmo mercado são relacionadas entre si. Fazendo uma análise visual, já é possível identificar que a estrutura da série IBM é similar à serie MSFT, o que mostra que estas séries podem conter informação uma da outra.

A outra configuração estudada avalia as séries de preços de abertura, máximo,

mínimo e fechamento da ação da AMD, entre janeiro de 1999 e dezembro de 2000 (veja a Figura 7.1). Esta configuração é conhecida como $OHLC$, em referência aos termos em inglês *Open*, *High*, *Low* e *Close*. Observa-se que há uma grande similaridade entre elas, o que indica a possibilidade de haver fontes ocultas nestas séries.

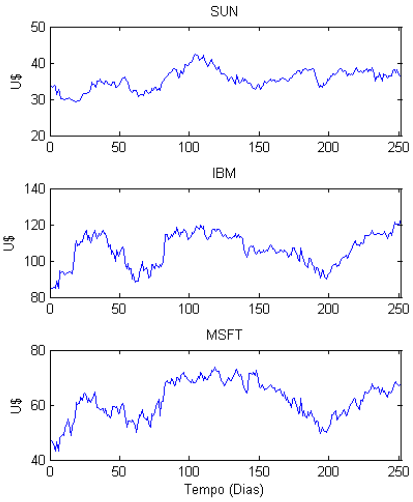

Figura 7.1: Series de preços de fechamento certificadas para o período de 04/2004 até 03/2005 (séries de treino), utilizadas no desenvolvimento do sistema de monitoração da qualidade de dados. De cima para baixo, SUN, IBM e Microsoft (MSFT).

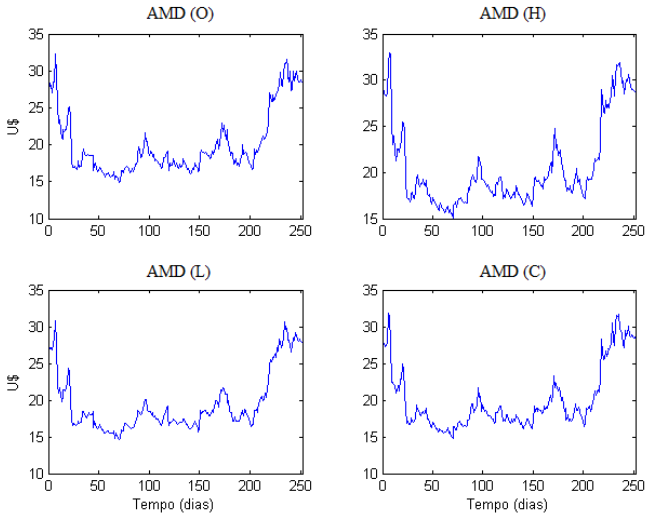

Figura 7.2: Série de preços das ações no instante da abertura (AMD (O)), máximo do dia (AMD (H)), mínimo do dia (AMD (L)) e fechamento (AMD (C)).

7.2 Separação de fontes

Dado os bons resultados com SOBI e SOBI-RO, ambos os algoritmos foram testados também no contexto de séries financeiras.

7.2.1 SUN-IBM-MSFT

Para se determinar o atraso para os algoritmos SOBI e SOBI-RO, utilizou-se o indicador $off(\mathbf{M})$. Na Figura 7.3, o indicador é mostrado para todos os atrasos possíveis (251 atrasos), dado o número de amostras disponíveis. Nota-se que o indicador também é calculado para o conjunto de séries sem a aplicação de ICA. Esse valor serve como base de referência em cada atraso. Em geral, podemos notar que o algoritmo SOBI-RO aproxima-se mais de zero e, portanto, produz uma melhor separação. O valor mínimo para SOBI-RO chegou a 0,2340 no atraso 122. Utilizando-se a regra prática sugerida na Seção 5.2.1, observa que o valor do mínimo permanece o mesmo. Já o valor mínimo de SOBI foi de 0,2926 para o atraso 69. Utilizando-se a regra prática, o valor mínimo foi de 0,2926 para o atraso 126. Este melhor desempenho de SOBI-RO era de supor, uma vez que este tipo de série é bastante complexa e, por vezes, comparada a um ruído imprevisível. Assim, opta-se por utilizar SOBI-RO ($\tau = 122$) nos estudos da configuração SUN-IBM-MSFT.

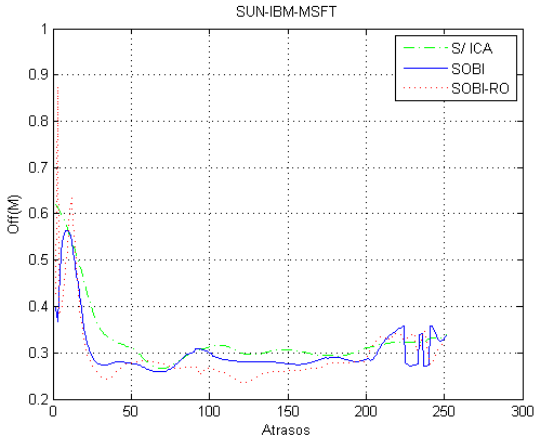

Figura 7.3: Indicador de separabilidade $off(\mathbf{M})$ em função do número de atrasos utilizados na separação de fontes com SOBI, SOBI-RO e valores de referência quando ICA não é aplicada.

Na Figura 7.4, são mostradas as fontes extraídas com SOBI-RO. Observa-se que a estrutura das séries IBM e MSFT têm uma forma similar à terceira fonte, exceto pela fase que está invertida. Isto mostra que ICA tenta agrupar as estruturas similares e concentrar a redundância de informação. Outra característica que se observa é uma forma similar a uma tendência na segunda fonte.

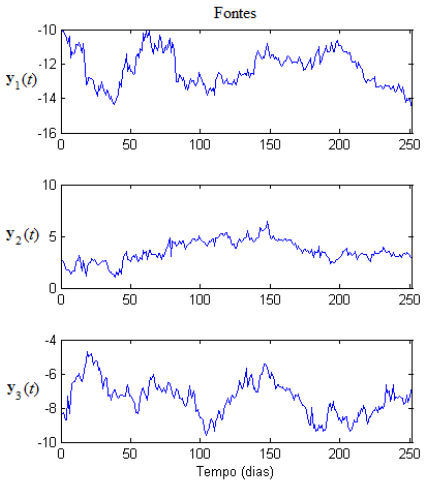

Figura 7.4: Fontes independentes extraída a partir das séries SUN, IBM e MSFT.

7.2.2 AMD OHLC

Na Figura 7.5, o indicador de separabilidade é mostrado para a configuração OHLC da série AMD. Os valores para as curvas produzidas são bastante próximos. Assim, para que seja notada a diferença entre cada um dos algoritmos, reduziu-se a visualização para os intervalos entre as amostras 95 e 130, que abrange o intervalo sugerido na Seção 5.2.1 para a escolha do atraso. De fato, os valores mínimos para SOBI e SOBI-RO ocorreram próximo a este intervalo (atraso 98). O algoritmo SOBI atinge 0,2319 e SOBI-RO chega a 0,2316. Estes valores estão abaixo do valor de referencia ($off(\mathbf{M}) = 0,2334$ para o atraso 98). Apesar disso, nota-se que estes valores são bastante próximos uns dos outros. Em certos atrasos do intervalo, $off(\mathbf{M})$ para SOBI e SOBI-RO fica acima da curva de referência. Isso poderia refletir certa dificuldade em separar as fontes. No entanto, observando-se as fontes separadas com SOBI-RO ($\tau = 98$) (veja a Figura 7.6), observa-se uma clara distinção das fontes extraídas. Assim, optou-se por utilizar esta configuração nas análises seguintes.

Na Figura 7.6, são mostradas as fontes independentes extraídas para o conjunto de série de preços da AMD. Observa-se que a estrutura redundante das séries foi mapeada para a primeira fonte (fonte $\tilde{\mathbf{s}}_1(t)$). Nas outras fontes, não é possível identificar estruturas visualmente. O que se nota são fontes desestruturadas e com amostras suspeitas de serem *outliers* ou alguma outra característica fora dos padrões, assim como ruídos. Por exemplo, na posição 218 da segunda fonte, há uma amostra discrepante das demais.

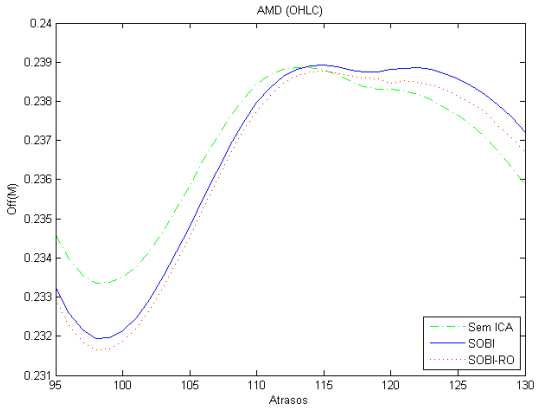

Figura 7.5: Indicador de separabilidade $off(\mathbf{M})$ em função do número de atrasos utilizados na separação de fontes com SOBI, SOBI-RO e valores de referência quando ICA não é aplicada, para a configuração AMD OHLC.

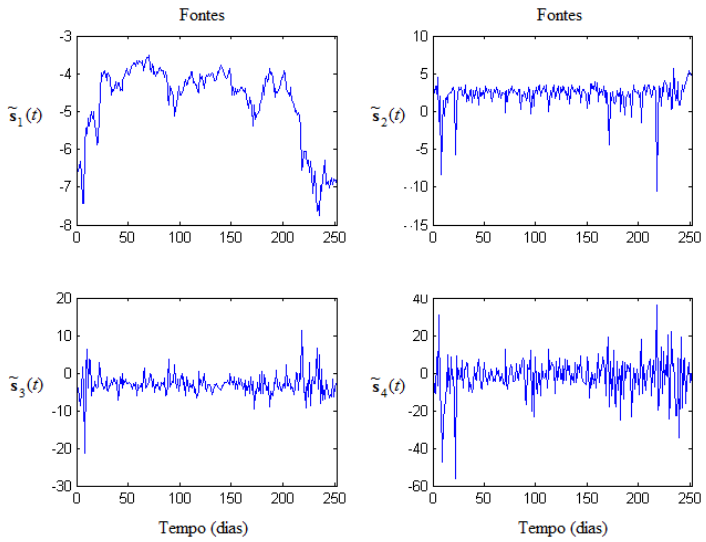

Figura 7.6: Fontes estimadas para a configuração AMD OHLC.

7.3 Pré-processamento

7.3.1 SUN-IBM- MSFT

ICA estimada

Os resultados dos testes e análises de pré-processamento no cenário com ICA são mostrados na Tabela 7.1. Através dos testes de raiz unitária, observa-se que somente a segunda fonte concentra algum tipo de estrutura de tendência estocástica. De fato, visualmente, já havíamos constatado uma possível existência de tendência. Pode-se constatar na Figura 7.7 que a correlação decai lentamente e tem um padrão típico de uma tendência ou mesmo um ciclo. De fato, no espectro de Fourier, é possível encontrar frequências significantes. Apesar disso, é pouco provável que estes ciclos representem algum padrão que se repita nas amostras subsequentes. Após aplicar o operador de primeira diferença, notam-se apenas duas correlações significativas (veja a Figura 7.8). Uma correlação negativa no primeiro atraso e outra no trigésimo primeiro atraso. O primeiro caso pode indicar a tendência de inversão do preço do dia anterior. Essa característica já havia sido identificada em [1]. A outra correlação pode indicar algum tipo de padrão particular ocorrendo no intervalod de um mês e meio ou alguma correlação espúria.

Apesar de não terem sido identificadas tendências estocásticas nas outras duas fontes (TE=0), ainda assim cabe ao usuário uma avaliação dos resultados do teste antes de pré-processar as séries. Isso ocorre pela possibilidade do teste dar resultados divergentes a níveis de significância diferentes. De fato, aumentando-se o nível de significância do teste para 10%, é possível detectar a presença de tendências. Assim, ambos os pré-processamentos com e sem diferenciação foram avaliados. No caso de diferenciação as séries, não se detecta nenhum atraso significativo nas duas fontes, padrão típico de um ruído (veja a Figura 7.9). Por fim, todas as fontes foram normalizadas com média zero e desvio padrão 1 para serem modeladas.

Tabela 7.1: Testes e análises de pré-processamento das fontes extraídas de SUN-IBM-MSFT após ICA.

Fontes	H	TE	S	TL	C	N
$\mathbf{y}_1(t)$	0	0	0	0	0	1
$\mathbf{y}_1(t)$	0	1	0	0	0	1
$\mathbf{y}_3(t)$	0	0	0	0	0	1

Sem ICA

Sem aplicar ICA, o teste de raiz unitária detecta tendência estocástica em todas as séries (veja Tabela a 7.2). Isto reforça a tese de ICA ter mapeado esta tendência

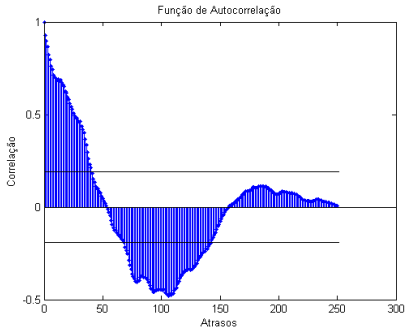

Figura 7.7: Função de autocorrelação da segunda fonte de SUN-IBM-MSFT, antes da retirada da tendência.

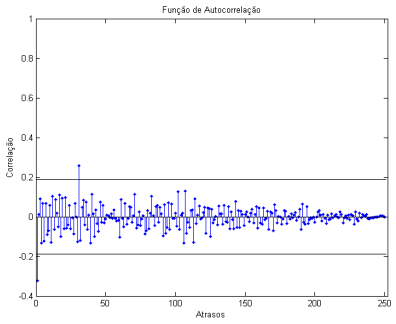

Figura 7.8: Função de autocorrelação da segunda fonte de SUN-IBM-MSFT, após a retirada da tendência.

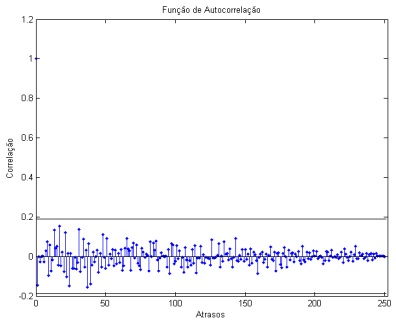

Figura 7.9: Função de autocorrelação da primeira fonte de SUN-IBM-MSFT, após ser tornada estacionária.

para uma das fontes, uma vez que esta componente estaria misturada em todas as três séries. Para extrai-la, aplicou-se o operador de primeira diferença. Após este pré-processamento, não se detectou nenhum atraso com correlação significativa nas séries. Ao contrário do que se observou nas fontes independentes, nem mesmo a tendência de inversão foi detectada. Isso mostra que o fato de haver componentes misturados pode mascarar a presença de possíveis padrões nas séries. Por fim, todas as séries foram normalizadas e apresentadas à entrada do método de modelagem.

Tabela 7.2: Testes e análises de pré-processamento sem ICA.

Séries	H	TE	S	TL	C	N
SUN	0	1	0	0	0	1
IBM	0	1	0	0	0	1
MSFT	0	1	0	0	0	1

7.3.2 AMD OHLC

ICA estimada

Na configuração OHLC para a ação da AMD, a estrutura das séries foi novamente mapeada para a primeira fonte. Ao nível de significância de 5%, o teste da raiz unitária identificou a presença de tendência estocástica apenas nesta fonte, que foi pré-processada pelo operador de primeira diferença (veja a Tabela 7.3). Alterando-se a significância do teste para 10%, verificou-se a presença da tendência estocástica na segunda fonte, que foi tornada estacionária pela diferenciação de primeira ordem. Após o pré-processamento, foi possível identificar na fonte resultante uma tendência de inversão, por meio da correlação negativa no primeiro atraso (veja a Figura 7.10). Uma correlação significativa também foi detectada no atraso 210. Essa correlação pode ser algum valor espúrio, devido a proximidade com o limiar de significância. Já as outras fontes não apresentaram nenhuma correlação significativa. Nesse sentido, estas fontes não precisariam ser estacionarizadas e poderiam ser classificadas como ruídos. Apenas a normalização deve ser realizada para todas as séries.

Tabela 7.3: Testes e análises de pré-processamento após ICA.

Fontes	H	TE	S	TL	C	N
$\tilde{\mathbf{s}}_1(t)$	0	1	0	0	0	1
$\tilde{\mathbf{s}}_1(t)$	0	0	0	0	0	1
$\tilde{\mathbf{s}}_3(t)$	0	0	0	0	0	1
$\tilde{\mathbf{s}}_4(t)$	0	0	0	0	0	1

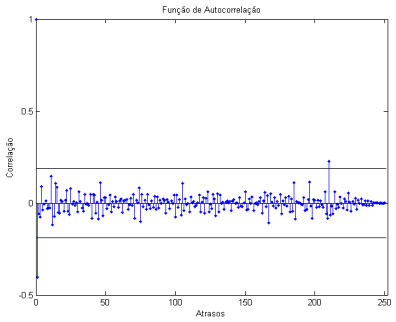

Figura 7.10: Função de autocorrelação da segunda fonte de AMD-OHLC, após tornar-se estacionária.

Sem ICA

Assim como ocorreu com as séries SUN-IBM-MSFT, a configuração AMD-OHLC sem ICA também apresenta tendência estocástica para todas as séries. Os resultados do pré-processamento são mostrados na Tabela 7.4. O teste de raiz unitária identificou esta característica para todas elas. As fontes foram então pré-processadas com o operador de primeira diferença. No entanto, não se verificou nenhuma correlação significativa na sequência. Dessa forma, o método pode tratar estas séries como ruído. No entanto, foi observado nas fontes obtidas com ICA a presença de correlações significativas, principalmente a tendência de inversão das séries. Isso reforça a tese da existência desta tendência e que não foi possível detectá-la sem ICA.

Tabela 7.4: Testes e análises de pré-processamento sem ICA.

Séries	H	TE	S	TL	C	N
AMD(O)	0	1	0	0	0	1
AMD(H)	0	1	0	0	0	1
AMD(L)	0	1	0	0	0	1
AMD(C)	0	1	0	0	0	1

7.4 Modelagem

7.4.1 SUN-IBM- MSFT

ICA estimada

Aplicando-se ICA na configuração SUN-IBM-MSFT, a metodologia de modelagem detecta atrasos relevantes apenas para o resíduo da segunda fonte. De fato, nota-se que há uma correlação negativa no atraso imediatamente anterior e no atraso de 31 dias úteis (veja a Figura 7.8).

Na Tabela 7.5, observa-se que o método de modelagem propôs a utilização de um modelo não linear (redes neurais MLP), com um neurônio na camada escondida para o resíduo da segunda fonte. O fato do sistema detectar a necessidade de uma série explicativa pode ter ocorrido devido ao pré-processamento tornar evidente alguma relação entre as fontes. Por exemplo, a correlação no trigésimo primeiro atraso tornou-se evidente após o pré-processamento. Por fim, nota-se que as outras fontes não têm atrasos relevantes e nem séries explicativas e, assim, o método sugere tratá-lo como um ruído. A opção pela média ao invés de usar o melhor estimador de caminho aleatório pode ser o melhor caminho, devido a média ser menos sensível a presença dos ruídos e *outliers*.

Tabela 7.5: Modelagem para o cenário com ICA (SOBI)

Fontes	Max Lag	Explicativa	NNE
$\mathbf{y}_1(t)$	0	0	0
$\mathbf{y}_2(t)$	31	1	1
$\mathbf{y}_3(t)$	0	0	0

Sem ICA

Sem aplicação de ICA, os testes de hipótese não detectaram nenhuma estrutura nos resíduos (veja a Tabela 7.6). Em [1], foi reportado uma pequena correlação negativa com a primeira amostra passada da série, indicando uma tendência oposta ao valor do dia anterior. No entanto, esta correlação está tão próxima dos patamares de relevância que, pequenas alterações na amostragem da série podem afetar a detecção desta componente. Assim, o método não identificou como relevante a correlação com as amostras passadas e, dessa forma, o resíduo tratado como ruído.

Tabela 7.6: Modelagem para o cenário sem ICA

Séries	Max Lag	Explicativa	NNE
SUN	0	0	0
IBM	0	0	0
MSFT	0	0	0

7.4.2 AMD OHLC

ICA estimada

Para esta configuração, observa-se que o modelo mais elaborado (não linear) foi sugerido para a primeira fonte (veja Tabela 7.7). De fato, visualmente, é possível

notar (veja a Figura 7.6) que a maior parte da estrutura da série está concentrada nesta fonte. Na segunda fonte, os atrasos significantes foram modelados linearmente. As outras fontes não têm atrasos significantes e nem séries explicativas e, assim, foram tratadas como ruído. Isso já era esperado, uma vez que ICA tem a capacidade de isolar fontes deste tipo.

Tabela 7.7: Modelagem para o cenário com ICA

Fontes	Max Lag	Explicativa	NNE
$\tilde{\mathbf{s}}_1(t)$	2	1	7
$\tilde{\mathbf{s}}_2(t)$	210	0	0
$\tilde{\mathbf{s}}_3(t)$	0	0	0
$\tilde{\mathbf{s}}_4(t)$	0	0	0

Sem ICA

Para o caso sem ICA, observa-se que, com exceção da série de mínimos (AMD(L)), nenhum atraso significante foi detectado nas séries alvo (veja a Tabela 7.8). A série de mínimos tem um atraso significante, que é a tendência de inversão. É provável que esta tendência não tenha sido detectada nas outras séries devido à respectiva componente de correlação ser próxima aos patamares de significância.

As séries de abertura (AMD(O)), máximo (AMD(H)) e mínimo (AMD(L)) são modeladas apenas com séries explicativas e estimativas lineares (NNE=0). Para estas séries, padrões de QD lineares foram sugeridos. Analisando-se necessidade de séries explicativas, observa-se que a série de abertura é explicadas pela série de fechamento (AMD(C)). De fato, apesar do valor do preço de abertura não ser exatamente o preço de fechamento do dia anterior, estes valores são bastante similares. Já os preços de máximo e mínimo são influenciados pela série de abertura. De fato, muitas vezes o preço de abertura é o valor máximo ou o mínimo do dia. Nota-se também que, diferentemente do caso com ICA, nenhum modelo não linear foi detectado. Isso mostra que os padrões não lineares presentes nas séries não foram detectados sem ICA.

Tabela 7.8: Modelagem para o cenário sem ICA

Séries	Max Lag	Explicativa	NNE
AMD(O)	0	1	0
AMD(H)	0	1	0
AMD(L)	1	1	0
AMD(C)	0	0	0

7.5 Detecção de Anomalias/Outliers

7.5.1 SUN-IBM-MSFT

As anomalias foram avaliadas com as metodologias SCICA, ICA e sem ICA. Para a marcação prévia dos *outliers*, selecionou-se amostras afastadas três desvios da média das séries diferenciadas, a partir do conhecimento *a priori* de todo o conjunto de amostras. Na Figura 7.11, mostra-se a série IBM diferenciada e os respectivos *outliers* marcados. Na Tabela 7.10, são mostradas as posições dos *outliers* encontrados para as três séries e, em negrito, os *outliers* detectados. Nota-se que SCICA detectou todos os *outliers*. Já os métodos ICA padrão e sem ICA deixaram de detectar um outlier cada, nas posições 49 e 189, respectivamente. Na tabela 7.10, resume-se o comportamento dos métodos através dos indicadores de falso alarme e detecção (os melhores resultados estão em negrito). Observa-se que, para as séries SUN e IBM, o método SCICA teve um melhor desempenho, detectando mais *outliers* ao passo que emite a mesma quantidade de falsos alarmes. Na série MSFT, SCICA emitiu apenas um falso alarme a mais que os outros dois métodos. Considerando-se que SCICA detectou bem melhor os *outliers*, essa pequena perda torna-se irrelevante. Assim, esses resultados apontam para a utilização de SCICA como detector de *outliers*.

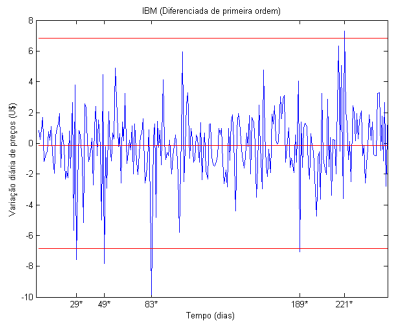

Figura 7.11: Série IBM diferenciada de primeira ordem e *outliers* detectados a priori.

7.5.2 AMD-OHLC

Para as séries AMD-OHLC, fez-se um estudo qualitativo das anomalias detectadas em maiores detalhes. Em [1], são discutidas as diferenças entre série certificadas [122] e não certificadas [123] para as séries da AMD. Na Figura 7.12, essa diferença é mostrada para as séries (sobrepostas) no período de 1996 até 2005. Observa-se que, a partir de 2002, houve uma mudança que tornou as séries não certificadas iguais às séries certificadas. Ainda assim, o provedor [123] não garante a certificação das séries. Dessa forma, selecionou-se as amostras entre 1999 e 2000 para se avaliar as diferenças ocorridas. Este conjunto de amostras tem um período com maior

Tabela 7.9: Posição dos *outliers* na série de teste

Séries	*Outliers*	SCICA	ICA	Sem ICA
SUN	158, 189	**158, 189**	**189**	**189**
IBM	29, 49, 83, 189, 221	**29, 49, 83, 189, 221**	**29, 83, 189, 221**	**29, 49, 83, 221**
MSFT	105	**105**, 107, 159	**105**, 189	**105**, 189

Tabela 7.10: Níveis de alarme e detecção

Séries	Indicadores	SCICA	ICA	Sem ICA
SUN	Falso alarme	**0%**	**0%**	**0%**
	Detecção	**100%**	50%	50%
IBM	Falso alarme	**0%**	**0%**	**0%**
	Detecção	**100%**	80%	80%
MSFT	Falso alarmeAlarmes	0,4%	**0,2%**	**0,2%**
	Detecção	**100%**	**100%**	**100%**

qualidade (1999), que serão utilizados para desenvolvimento do modelo, e o período subsequente (utilizadas como conjunto de teste) com diferenças relevantes. Para se avaliar estas amostras, apenas as diferenças afastadas 3 desvios padrão da média foram avaliadas. Ainda, selecionou-se as amostras mais afastadas da média das séries diferenciada (com 2 e 3 desvios padrão), classificadas como *outliers*. Na Tabela 7.11, são mostradas as posições de cada uma das anomalias para a série de testes.

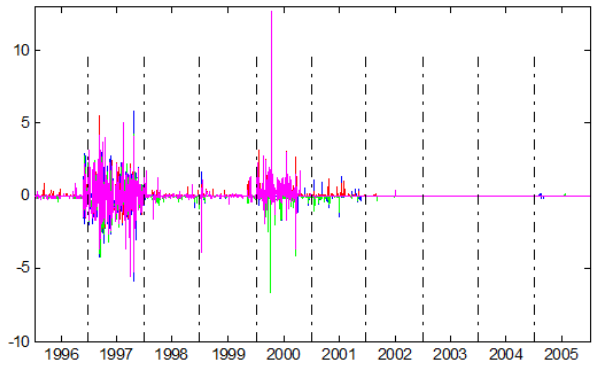

Figura 7.12: Diferenças entre [122] e [123], para as séries AMD-OHLC, entre 1996 e 2005

Na Tabela 7.12, é mostrado o nível de detecção e falso alarme com ICA (SCICA) e sem ICA. Nota-se que, para ambos os casos, todas as amostras classificadas como

Tabela 7.11: Qualificação das anomalias encontradas nas respectivas posições da série de teste

Séries	Diferenças entre [122] e [123]	*Outliers* (2σ)	*Outliers* (3σ)
AMD(O)	13, 44, 72, 117, 150	67, 72, 75, 103, 104, 118, 131	162
AMD(H)	12, 16, 44, 47, 71, 74, 79, 103, 106, 135, 138, 149	45, 70, 91, 117, 128	103, 162
AMD(L)	46, 57, 67, 90, 108, 117, 162	65, 68, 72, 89, 100, 103, 118	66, 162
AMD(C)	9, 16, 34, 38, 45, 50, 54, 61, 63, 71, 78, 90, 137, 138, 140, 145, 151	45, 66, 91, 117, 128	103, 162

outliers foram detectadas corretamente. As diferenças entre [122] e [123] também foram igualmente detectadas entre os métodos. No entanto, o nível de acertos foi bem menor. De fato, estas diferenças são bem mais difíceis de serem detectadas, uma vez que são menores do que valores extremos da série. Apesar da dificuldade na detecção das diferenças, para o caso da série AMD (O), o índice de detecção chega a 80%. Isto pode ser explicado pela maior previsibilidade desta série. De fato, a série de preços de fechamento do dia anterior é uma boa série explicativa para a série de abertura do dia seguinte. Por outro lado, avaliando-se o nível de falsos alarmes, observa-se que SCICA teve um melhor desempenho em todas as séries monitoradas. Este resultado reforça a utilização deste método para a detecção de *outliers*.

Tabela 7.12: Percentual de detecção e falsos alarmes na série de testes

Séries		Diferenças entre [122] e [123]	*Outliers* (2σ)	*Outliers* (3σ)	Falsos alarmes
AMD(O)	SCICA	**80%**	**100%**	**100%**	**5.5%**
	S/ ICA	**80%**	**100%**	**100%**	9,9%
AMD(H)	SCICA	**33,3%**	**100%**	**100%**	**2,8%**
	S/ ICA	**33,3%**	**100%**	**100%**	4,3%
AMD(L)	SCICA	**28%**	**100%**	**100%**	**5,91%**
	S/ ICA	**28%**	**100%**	**100%**	8,7%
AMD(C)	SCICA	**22%**	**100%**	**100%**	**5,52%**
	S/ ICA	**22%**	**100%**	**100%**	9,1%

7.6 Qualidade de Dados

7.6.1 Qualidade do Modelo

Os indicadores de qualidade do modelo (IQM) são mostrados na Tabela 7.13. Os melhores resultados estão em negrito. Analisando-se o primeiro grupo de séries (SUN-IBM-MSFT), observa-se que os indicadores são similares para os modelos com em sem ICA, com exceção de $NMSE_2$ para as séries SUN e MSFT que foi melhor para o caso com ICA. Ainda assim, o indicador permaneceu um pouco acima de 1, mostrando a dificuldade de se encontrar padrões nestas séries. O indicador R abaixo de 1 também mostra esta dificuldade, indicando que há atraso na modelagem. De fato, este grupo de séries é composto por preços de fechamento diário, que, geralmente, são de difícil modelagem. Este resultado também foi constatado em [1], que encontrou valores similares.

Analisando o grupo de séries AMD OHLC, nota-se que ICA introduziu uma melhora nos indicadores dos modelos, principalmente na série AMD(O). Todos os indicadores expressam a melhora para esta série, com exceção de Lag_0 que é igual para ambos os casos com e sem ICA. Os indicadores para a série da série de mínimos também apontam para uma pequena melhora com ICA. Ainda, nota-se que o modelo sem ICA para esta série AMD(L) sofre de atraso. Observa-se também que ICA teve um desempenho um pouco melhor, mas ainda sim não resolveu o problema do atraso para as estimativas do padrão de QD. Já os melhores indicadores da série AMD(H) tiveram uma pequena divergência para os casos com e sem ICA. Enquanto Lag_0 aponta para uma melhora com ICA, $NMSE_2$ aponta para uma piora. Avaliando-se o IGQM desta série, nota-se que não houve melhora. A série AMD(C) obteve o pior desempenho. Nota-se que,tanto para o caso ICA quanto sem ICA, $NMSE_2$ é maior que 1 e R é menor que 1. Isso mostra que nenhum padrão foi detectado e que a série poderia ser tratada como ruído. De fato, esta é a série de preços de fechamento e, conforme constatado nas séries de fechamento para SUN-IBM-MSFT, há uma grande dificuldade na modelagem.

7.6.2 Qualidade das séries financeiras

Na Tabela 7.14, a qualidade das séries financeiras é medida. Nota-se que os dados avaliados não contêm nenhum valor faltante. Assim, a completude atingiu o seu valor máximo de 1 para todos os casos. De fato, esse não foi um problema encontrado em nenhuma das séries aqui estudadas.

Analisando-se o indicador consolidado do conjunto total de séries (IGQD/BD), nota-se que à medida da QD é igual para os casos com e sem ICA (IGQD=0,87). Este resultado era de se esperar, uma vez que não se notou grandes diferenças nos

Tabela 7.13: Indicadores de Qualidade do modelo (IQM)

Séries financeiras	$NMSE$	$NMSE_1$	$NMSE_2$	R	Lag_0	IGQM
ICA						
SUN	**0,01**	**0,03**	**1,04**	**0,98**	**0,98**	**0,79**
IBM	**0,02**	**0,02**	**1,02**	**0,99**	**0,99**	**0,80**
MSFT	**0,02**	**0,05**	**1,01**	**0,98**	**0,98**	**0,79**
AMD (O)	**0,03**	**0,01**	**0,61**	**1,00**	**0,99**	**0,83**
AMD (H)	**0,04**	**0,02**	0,98	**1,00**	**0,99**	**0,80**
AMD (L)	**0,04**	**0,02**	**0,90**	**1,00**	**0,99**	**0,81**
AMD (C)	**0,04**	**0,02**	1,06	**0,99**	**0,99**	**0,80**
Sem ICA						
SUN	**0,01**	0,04	1,05	**0,98**	**0,98**	**0,79**
IBM	**0,02**	**0,02**	**1,02**	**0,99**	**0,99**	**0,80**
MSFT	**0,02**	**0,05**	1,02	**0,98**	**0,98**	**0,79**
AMD (O)	0,04	0,02	0,75	0,99	**0,99**	0,82
AMD (H)	**0,04**	**0,02**	**0,95**	0,99	**0,99**	**0,80**
AMD (L)	**0,04**	**0,02**	**0,90**	0,99	**0,99**	**0,81**
AMD (C)	**0,04**	**0,02**	**1,01**	**0,99**	**0,99**	**0,80**

modelos produzidos em cada um dos cenários.

Avaliando-se o indicadores de acurácia (B) do BD, nota-se que a QD está abaixo da esperada. Este fato pode ser explicado pela baixa qualidade das séries AMD, que foram obtidas de fonte não certificada [123]. Na Tabela 7.11, é possível se verificar a grande quantidade de problemas e anomalias encontradas.

Comparando-se as medidas de qualidade entre os dois conjuntos estudados, nota-se que, para a configuração SUN-IBM-MSFT, o valor é próximo de 0,93, enquanto, para a configuração AMD OHLC, o valor é próximo de 0,83. A QD inferior pode ser devido a obtenção das amostras da configuração AMD-OHLC de fonte não certificada.

A metodologia de montagem dos conjuntos também pode afetar os modelos e, assim, acabar afetando a medição da QD. Dessa forma, é fundamental manter constante a forma de se medir a qualidade, de acordo com as necessidades do usuário e o contexto, de forma que se tenham parâmetros para comparação temporal ou espacial. Por exemplo, nota-se que entre as séries individuais SUN, IBM e MSFT, o IGQD é próximo entre as séries. Isso poderia ser interpretado como um nível de qualidade semelhante entre estas séries. A experiência dos usuários na medição da QD deve dizer se os dados são de alta ou baixa qualidade.

Por fim, na Figura 7.13, é mostrado um exemplo da monitoração da QD com

Tabela 7.14: Indicadores de Qualidade de Dados (IQD)

Séries financeiras	Completude	Acurácia (B)	Acurácia (M)	Acurácia (A)	IGQD
ICA SOBI					
SUN	**1,00**	**0,99**	**0,96**	0,76	**0,93**
IBM	**1,00**	**0,98**	**0,94**	0,78	0,93
MSFT	**1,00**	**0,99**	**0,98**	0,80	0,94
AMD (O)	**1,00**	**0,90**	**0,79**	0,58	**0,82**
AMD (H)	**1,00**	0,90	**0,81**	**0,65**	**0,84**
AMD (L)	**1,00**	**0,87**	0,75	**0,59**	0,80
AMD (C)	**1,00**	**0,90**	**0,80**	**0,60**	**0,83**
BD	**1,00**	**0,93**	**0,86**	0,68	**0,87**
Sem ICA					
SUN	**1,00**	**0,99**	**0,96**	**0,77**	**0,93**
IBM	**1,00**	**0,98**	**0,96**	**0,83**	**0,94**
MSFT	**1,00**	**0,99**	**0,98**	**0,81**	**0,95**
AMD (O)	**1,00**	0,85	0,76	**0,62**	0,81
AMD (H)	**1,00**	**0,91**	**0,81**	0,64	**0,84**
AMD (L)	**1,00**	**0,87**	**0,77**	0,58	**0,81**
AMD (C)	**1,00**	0,88	0,78	0,59	0,81
BD	**1,00**	0,92	**0,86**	**0,69**	**0,87**

os corredores de validação produzidos para 50 amostras de teste da série SUN. Amostras anômalas foram simuladas nas posições 5, 10 e 50. Valores faltantes foram simulados nas posições 35, 40 e 45. Observa-se que os corredores adaptam-se às variações estatísticas da série e sugerem correções (centro do corredor) para as amostras detectadas com problemas.

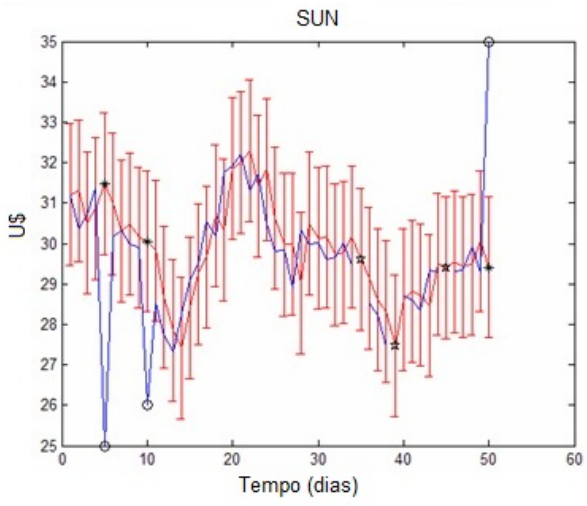

Figura 7.13: Corredores de validação para a monitoração da QD na série de preços da SUN

Capítulo 8

Análises e Conclusões

Nesta seção, os resultados obtidos nos capítulos anteriores são analisados para os dois tipos de séries estudadas: sintéticas e reais. A metodologia de estudo através de séries sintéticas (em condições controladas e fontes conhecidas) permitiu que diversas técnicas fossem avaliadas antes de serem aplicadas em um ambiente real e desconhecido (separação cega de fontes).

8.1 Séries Sintéticas

No ambiente simulado, dado o conhecimento a priori da matriz mistura, das fontes originais e suas respectivas distribuições de probabilidade, foi possível fazer uma análise dos métodos de separação utilizados. O que se nota é que a informação contida na sequência temporal não deve ser desprezada para a extração das fontes independentes. Esse foi o principal motivo dos métodos de ICA para séries temporais terem o melhor desempenho em relação aos algoritmos de ICA básica. Os indicadores de separação de fontes mostraram a tendência de melhor desempenho destes algoritmos que utilizam apenas estatística de segunda ordem.

O algoritmo SOBI obteve o melhor desempenho quando são utilizados os sinais sintéticos com baixo nível de ruído. Quando se eleva o nível de ruído ou se introduzem *outliers* nas séries, SOBI-RO tende a ter um melhor desempenho. A explicação para SOBI-RO se sobressair em condições adversas é a forma de branqueamento das séries. Enquanto o branqueamento realizado em SOBI diagonaliza apenas uma matriz de covariância, SOBI-RO utiliza uma combinação de matrizes de covariâncias para branquear os dados, estimadas em períodos de tempo distintos. Esta combinação distribui o efeito do ruído e das anomalias no branqueamento dos dados e permite uma melhor separação das fontes.

No pré-processamento, observa-se uma similaridade nos testes e análises com ICA estimada e ICA ideal. Isto mostra mais uma vez que o método de separação cega de fontes fez uma boa aproximação. Apenas a tendência linear não foi bem separada da

tendência estocástica, o que refletiu na diferença de pré-processamento dessas fontes. Este fato pode ser explicado pela elevada correlação não linear entre as próprias fontes originais e pela similaridade entre a função de autocorrelação de ambas. Isso dificulta a separação pelos algoritmos, uma vez que ICA assume a hipótese das fontes terem espectros disjuntos. Apesar desta falha na separação, ainda assim o pré-processamento no espaço ICA foi facilitado em relação ao pré-processamento sem ICA. Todos os outros componentes presentes nas séries (heterocedasticidade, tendência estocástica, ciclos e sazonalidades) puderam ser identificados pelos testes e métodos de análise de pré-processamento (teste da raiz unitária, análise da função de autocorrelação e do espectro de frequências e observação da própria série). Sem ICA, alguns componentes ficaram mascarados nas séries, assim como a tendência linear, a heterocedasticidade e os ciclos. A dificuldade sem ICA ocorre pela sobreposição dos componentes no processo de mistura das fontes, o que dificulta a análise. Mesmo quando testes objetivos são aplicados, a análise do usuário ainda é fundamental. Por exemplo, longos ciclos presentes nas séries podem induzir o teste de raiz unitária a indicar a presença de tendência. Ou, caso não se tenha uma visão completa das séries, fenômenos que são na realidade ciclos podem ser identificados como tendência. Apesar dessa dificuldade intrínseca ao pré-processamento, ICA permite a melhor prospecção dos dados e facilita tanto as análises subjetivas do usuário quanto os testes objetivos.

Na modelagem, ICA facilita a busca por modelos mais adequados a cada tipo de série. Com ICA introduzido na cadeia de pré-processamento, a metodologia é capaz de ajustar desde modelos simplificados (média e o melhor estimador de caminho aleatório), passando por modelos lineares (AR), até não lineares (MLP). A definição do modelo ficou a cargo de testes de hipótese, que partem do princípio de modelos parcimoniosos e, gradualmente, aumentam a complexidade até encontrar o modelo mais adequado aos dados. Os testes tiveram um melhor desempenho quando aplicados nas fontes independentes. O ganho introduzido por ICA se reflete principalmente na não necessidade de séries explicativas. Isto ocorre porque as fontes independentes não armazenam informação umas das outras. Isso evita que informação redundante seja utilizada na construção do modelo. Assim, o processo de modelagem faz o mapeamento em um espaço de dimensões reduzidas e utiliza apenas a informação relevante. Outro aspecto que se observa com a introdução de ICA é a extração de ruídos. Assim, este componente deixa de dificultar a modelagem de outras séries. Observa-se também que os modelos das fontes obtidas com ICA foram semelhantes aos modelos das fontes ideais. Novamente, este aspecto denota que a separação das fontes foi bem sucedida.

O ganho proporcionado por ICA no pré-processamento e na modelagem se reflete na construção de melhores padrões de QD. Padrões mais acurados permitem que a

qualidade seja monitorada de forma mais eficiente. Isto pode ser notado no estudo das séries com amostras anômalas. Com ICA e SCICA, o sistema foi capaz de detectar melhor os problemas e diminuiu os falsos alarmes. SCICA aumenta ainda mais a eficiência, pois muda o foco da detecção. Na metodologia ICA padrão, a detecção busca o padrão de QD. O foco é na solução dos problemas. Já SCICA busca diretamente amostras anômalas. Assim, o foco passa a ser nos problemas. Observa-se que, para todas as séries, os indicadores de qualidade dos modelos (IQM) apontam para uma melhora quando ICA é introduzida. Isso mostra que estes modelos geram corredores mais estreitos e acurados. O IGQM para o banco completo de séries sintéticas resume a melhora no processo de monitoração da QD com ICA. O método ICA tem o maior IGQM tanto para séries simuladas sem anomalias, quanto para séries simuladas com anomalias. Uma vez que os indicadores de qualidade do modelo (IQM) são melhores, os indicadores de qualidade de dados (IQD) também se tornam mais confiáveis. Assim, o método se mostrou promissor para ser testado na medição e monitoração da qualidade em um ambiente real.

8.2 Séries Reais

Após a avaliação da metodologia em condições controladas, os melhores procedimentos foram aplicados em séries reais. O sistema de monitoração da QD foi avaliado em dois ambientes distintos: com séries de carga elétrica e com séries financeiras.

8.2.1 Séries de Carga Elétrica

Para as séries de carga elétrica, espera-se que haja estruturas de componentes frequentemente presentes em séries temporais, assim como tendências, sazonalidades, ciclos e heterocedasticidades. Estes padrões são esperados devido ao comportamento que se observa para o consumo de energia. O que se nota no dia-a-dia é a repetição de rotinas de consumo, assim como padrões de comportamento semanais, mensais e anuais. Para a análise neste contexto, optou-se por dados de uma competição, pelo fato de terem modelos que pudessem ser comparados com o modelo aqui proposto para extração de padrões de QD. Por esta razão, uma das configurações utiliza o conjunto de séries no mesmo formato da competição (configuração com a série de picos). Além dessa configuração, as séries foram também agrupadas de uma forma que se julgou mais adequada para a monitoração da QD (configuração de séries adjacentes). Devido a proximidade temporal das séries (amostradas de 30 em 30 minutos), esperam-se fontes ocultas misturadas nas séries.

Na separação de fontes, SOBI-RO apresentou melhor desempenho, devido a presença de ruídos (componentes não previsíveis) intrínseco a este tipo de série. Con-

forme explicado na Seção 8.1, SOBI-RO é menos sensível a este tipo de problema. Já nas fontes extraídas, é possível identificar visualmente alguns tipos de estruturas e padrões. Tanto na configuração de picos quanto na configuração de séries adjacentes, verifica-se que a influência da temperatura é mapeada para uma das fontes, de forma cega. A influência dessa fonte é relatada por especialistas, que identificam um forte relacionamento do consumo de energia com a variação de temperatura. Nas fontes obtidas com a configuração de séries adjacentes, observam-se também padrões cíclicos. Ciclos quadrimestrais podem ser visualmente identificados e são constatados no espectro de frequências (amplitude significante na terceira frequência). Este comportamento pode estar associado a padrões de consumo relacionados às estações do ano. No espectro de frequências, a segunda componente de frequência também é destacada, o que pode corresponder a períodos do ano nos quais a diferença de comportamento do consumo é mais evidente (inverno e verão). Uma outra fonte identificada, tanto para a configuração de picos quanto para séries adjacentes, é a sazonalidade semanal. Esta característica pode ser percebida na função de autocorrelação da fonte. A correlação significativa no sétimo atraso indica que esta fonte foi bem mapeada. Finalmente, nas outras fontes não foi detectado nenhum padrão pelos testes de pré-processamento. Isso indica que algumas dessas fontes devem ser tratadas como ruídos, o que facilita o processo de monitoração da QD.

A concentração das estruturas existentes em poucas fontes proporcionou formas diferentes de pré-processamento em relação ao caso sem ICA. Apenas uma das fontes foi pré-processada com a retirada da sazonalidade. Todas as outras são apenas tornadas estacionárias com o operador de primeira diferença. Mesmo a fonte contendo ciclos não foi pré-processada com a retirada das frequências do espectro. Isso porque os ciclos são estocásticos e a frequência e a amplitude podem variar de ano para ano. Assim, a melhor opção foi apenas tornar estacionária a série e deixar que os modelos (lineares ou não lineares) mapeiem este comportamento. Sem ICA, cada uma das séries de carga é pré-processada com a retirada da sazonalidade. Esse pré-processamento similar para as séries de carga era esperado, uma vez que são similares. Apenas a fonte de temperatura é tornada estacionária com o operador de primeira diferença. Isso porque o ciclo é estocástico e foi deixado para o modelo neural estimá-lo.

Na modelagem, a introdução de ICA também reduziu a redundância na entrada dos modelos. Isso facilita a estimação dos resíduos que precisam ser mapeados com menos variáveis para se obter os padrões de QD. O ruído foi tratado de maneira adequada, com modelos simplificados (média). A variedade de modelos utilizados (simplificados, lineares e não lineares) reflete a busca por modelos mais adequados a cada tipo de série. Os poucos neurônios na camada escondida da rede para mapear as não linearidades reflete a busca por modelos parcimoniosos. Assim, características

lineares ou não lineares nos dados foram tratadas com o respectivo modelo. Este tratamento mais adequado é refletido na redução do Erro Percentual Absoluto Médio (MAPE) e no erro máximo (MAX) das estimativas. Na comparação com modelos externos a esta tese, obtidos de uma competição, os modelos com ICA ficaram entre os primeiros colocados, na frente do método sem ICA. Mesmo sem ICA, ainda assim a metodologia atingiu uma boa colocação na competição. Isso mostra que as outras etapas da metodologia também contribuem para estes bons resultados.

Na detecção de outliers, ICA gerou menos falsos alarmes do que o método sem ICA. Além disso, a nova proposta para detecção de outliers com SCICA refinou ainda mais o método com ICA. Para um mesmo nível de detecção, SCICA produziu os menores índices de falsos alarmes. Esta melhora foi relatada na 8.1 e pode ser explicada pelo foco dado por cada uma das metodologias: nos problemas ou nas soluções.

Por fim, os indicadores permitiram medir a qualidade dos dados e produziram desde visões consolidadas até visões mais detalhadas da QD. Com ICA, os modelos são melhores, o que pode ser constatado pela maioria dos indicadores consolidados IGQM. Assim, tem-se uma medição mais fidedigna da QD nestas séries. Isso pode ser constatado nos maiores IGQD's para a maioria dos casos analisados com ICA. Uma vez que os dados não foram coletados "in natura"e foram pré-selecionados antes de serem disponibilizados para a competição, de fato, espera-se que a qualidade seja mais elevada.

8.2.2 Séries Financeiras

Para as séries financeiras, notou-se novamente que SOBI-RO tem um melhor desempenho na separação das fontes. Mais uma vez, a explicação está no tipo de série monitorada, com padrões pouco previsíveis e indefinidos, o que aproxima o comportamento dessas séries ao comportamento de um ruído. Analisando-se visualmente as fontes extraídas, observa-se que SOBI-RO foi capaz de extrair algumas fontes que estavam ocultas nas séries. Para a configuração SUN-IBM-MSFT, pode-se perceber a presença de tendência estocástica. Os testes de raiz unitária corroboram esta análise. Esta componente pode ser explicada pela tendência de crescimento do setor de tecnologia da informação (TI) na última década. Na configuração AMD-OHLC, a ICA foi capaz de extrair fontes de ruídos, que foram tratadas adequadamente. A extração de ruídos permite a criação de modelos parcimoniosos nas outras fontes, pois a informação imprevisível não é processada. Algumas fontes não precisaram sequer do pré-processamento (apenas normalização), pois já eram estacionárias.

Na modelagem, nota-se que a introdução de ICA evidenciou padrões que se encontravam ocultos nas séries. Por exemplo, a tendência de reversão do dia anterior

fica clara na função de autocorrelação quando se introduz ICA. De fato, observando-se diretamente a autocorrelação das séries, nem sempre isso é evidente. Assim, modelos adequados a cada tipo de fonte foram propostos, desde simplificações para o tratamento do ruído, passando por modelos lineares, até modelos não lineares com redes neurais MLP. Sem ICA, para todos os resíduos do preços de fechamento, o tratamento foi o mesmo que seria dado a um ruído. Isso porque não foram encontradas nenhuma estrutura nestes resíduos. A diferença na modelagem também é notada da série preços de abertura (AMD(O)). Sem ICA, utiliza-se um modelo linear, enquanto ICA encontra não linearidades nas fontes mais estruturadas.

Na detecção de outliers, nota-se que o método SCICA obtém os melhores resultados, identificando mais anomalias e emitindo menos falsos alarmes. Na avaliação qualitativa das amostras identificadas, nota-se que o método é capaz de identificar até mesmo diferenças entre fontes certificadas e não certificadas. Apesar de serem difíceis de identificar, o nível de detecção foi elevado para a séries com padrões mais definidos, assim como a série de preços de abertura - AMD(O). Os valores entrantes desta série são explicadas pelo preço de fechamento do dia anterior.

Na medição e monitoração da QD, o desempenho também foi melhor quando há padrões definidos nas séries, assim como AMD(O). Para as séries de preços de fechamento, os indicadores não mostraram ganhos expressivos. Isso mostra que a seleção do conjunto de séries é fundamental para o bom desempenho do método de monitoração da QD. Do contrário, a introdução de ICA pode significar apenas uma elevação na complexidade dos cálculos.

8.3 Conclusão

Este trabalho propôs um sistema de monitoração da qualidade de dados em séries temporais, utilizando ICA para extrair fontes estruturadas de forma cega. A ICA é inserida na cadeia de pré-processamento para se extrair padrões de qualidade mais acurados. A partir da incerteza gerada pelos modelos de qualidade, constroem-se corredores de validação para medir e monitorar a qualidade das séries. O corredor se adapta às variações estatísticas dos dados e o sistema alerta o usuário quando há amostras afastadas dos padrões de qualidade (fora dos limites dos corredores). Caso seja necessário ou solicitado, o centro do corredor é sugerido para a correção dos problemas. As amostras marcadas como suspeitas são contabilizadas para se produzir indicadores da qualidade. Foram propostos indicadores para a qualidade do modelo e para a qualidade dos dados. O método proporciona indicadores mais acurados e confiáveis.

Na metodologia proposta, redes neurais ou modelos lineares estimam o padrão de qualidade no espaço de fontes independentes, pré-processadas de forma clássica.

As fontes são obtidas com algoritmos que utilizam estatística de ordem superior ou estatística de segunda ordem associada à informação temporal. O método facilita o pré-processamento e a modelagem dos padrões de QD. O impacto do sistema na medição e monitoração da qualidade dos dados foi analisado em séries sintéticas e reais. O sistema foi testado em condições ruidosas, com *outliers* e sob configurações distintas das séries. De maneira geral, ICA se adaptou às diversas situações apresentadas e produziu corredores mais estreitos e acurados, o que tem impacto positivo na monitoração e medição da QD.

8.4 Trabalhos Futuros

Como extensão deste trabalho, propõe-se avaliar o SMQD-ST com outras configurações dos dados. Nas séries financeiras, sugere-se testar o sistema com um conjunto maior de séries de ações. Sugere-se o estudo de todas as séries negociadas em uma mesma bolsa de valores, como a Bovespa. É recomendável também que se faça uma pesquisa com especialistas do mercado de ações, para a seleção de variáveis explicativas. Além disso, sugere-se aplicar uma pesquisa para mapear os diversos problemas encontrados por diferentes tipos de usuários e contextos de aplicação das séries temporais.

Para a extração de fontes, devem ser incluídos outros algoritmos de ICA que combinem ambos os princípios de separação de ICA básica e de séries temporais, assim como algoritmos que diagonalizem matrizes de cumulantes deslocadas no tempo. Na Seção 3.4, são apresentados alguns algoritmos que podem ser aplicados. Recomenda-se também a utilização de outros métodos de separação cega de fontes, assim como a Fatoração de Matrizes Não Negativas (NMF). Para o pré-processamento, sugere-se a introdução de mais testes objetivos e a criação de métodos para a redução da subjetividade nesta etapa. Em [107], são apresentados testes de heterocedasticidade que podem ser aplicados. Para a modelagem, sugere-se incluir os métodos não lineares recorrentes (redes neurais Elman) no teste de hipótese de seleção de modelos. Por fim, sugere-se aprofundar no estudo de outras formas de construção dos corredores de validação, assim como corredores não simétricos e métodos não paramétricos.

Para a detecção de *outliers*, os estudos indicaram que o método SCICA é promissor que deve ser investigado em maiores detalhes. A avaliação tanto da fonte de ruídos quanto da fonte estruturada deve ser aprofundada. Sugere-se avaliar a fonte estruturada como sendo a série de *outliers* filtrados.

Para a monitoração da QD, sugere-se aprofundar as análises com e sem a reposição das amostras identificadas com problemas. Outra sugestão é utilizar SCICA para monitorar a qualidade de dados. Por fim, sugere-se incluir novas dimensões na monitoração da QD, assim como a credibilidade dos dados. Nas séries financeiras,

essa credibilidade poderia ser medida a partir do conhecimento da QD do banco de dados completo obtido do provedor.

Referências Bibliográficas

[1] DANTAS, A. C. H. *Sistema de Monitoração da Qualidade de Dados*. Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2006.

[2] LIMA, L. F. R., MAÇADA, A. C. G., VARGAS, L. M. "Research Into Information Quality: A Study of the State-of-Art in IQ and its Consolidation". In: *Proceedings of the International Conference on Information Quality – ICIQ*, 2006.

[3] EPPLER, M. J., WITTIG, D. "Conceptualizing Information Quality: A review of Information Quality Frameworks from the Last Ten Years". In: *Proceedings of the International Conference on Information Quality – ICIQ*, 2000.

[4] MILEK, J., REIGROTZKI, M., BOSCH, H., et al. "Monitoring and Data Quality Control of Financial Databases from a Process Control Perspective". In: *Proceedings of the International Conference on Information Quality – ICIQ*, 2001.

[5] *Oxford Dictionaries Online*. http://oxforddictionaries.com, Oxford University Press, 2011.

[6] CHURCHMAN, C. W. *The design of inquiring systems*. New York, Basic Books, 1971.

[7] FERREIRA, A. B. H. *Dicionario Aurélio da Língua Portuguesa*. São Paulo, Positivo, 2010.

[8] JURAN, J. M., GRYNA, F. M., BINGHAM, R. S. *Quality Control Handbook*. 3 ed. New York, McGraw-Hill, 1974.

[9] WANG, R. Y., STRONG, D. M. "Beyond accuracy: What data quality means to data consumers", *J. Manage. Info. Syst.*, v. 12, n. 4, pp. 5–34, 1996.

[10] KLEIN, B. D., GOODHUE, D. L., DAVIS, G. B. "Can humans detect errors in data? Impact of base rates, incentives, and goals", *MIS Quarterly*, v. 21, n. 2, pp. 169–194, 1997.

[11] GE, M., HELFERT, M. "A Review of Information Quality Research". In: *Proceedings of the International Conference on Information Quality – ICIQ*, 2007.

[12] REDMAN, T. *Data Quality For The Information Age*. Artech House Publishers, 1996.

[13] BALLOU, D. P., PAZER, H. L. "Modeling Data and Process Quality in Multi-input, Multi-output Information Systems", *Manage Science*, v. 31, n. 2, pp. 150–162, 1985.

[14] WANG, R. Y., KON, H. B., MADNICK, S. E. "Data Quality Requirements Analysis and Modeling". In: *Ninth International Conference on Data Engineering*, Viena, Austria, 1993.

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

[16] HELFERT, M. "Managing and Measuring Data Quality in Data Warehousing". In: *Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics*, 2001.

[17] BOVEE, M., SRIVASTAVA, R. P., , et al. "A conceptual framework and belief- function approach to assessing overall information quality". In: *Proceedings of the Sixth International Conference on Information Quality*, 2001.

[18] KAHN, B., STRONG, D., , et al. "Information Quality Benchmarks: Product and Service Performance", *Communications of the ACM*, pp. 184–192, 2002.

[19] WANG, R. Y., REDDY, M., , et al. "An object-oriented implementation of quality data products". In: *Proceedings of Third Workshop on Information Technology and Systems*, pp. 670–677, 1993.

[20] PIPINO, L., LEE, Y. W., WANG, R. Y. "Data Quality Assessment", *Communications of the ACM*, pp. 211–218, 2002.

[21] WANG, R. Y. "A Product Perspective on Total Data Quality Management", *Communications of the ACM*, pp. 58–63, 1998.

[22] HAMILTON, J. D. *Time Series Analysis*. New Jersey, Princeton University Press, 1994.

[23] HUGHES, E., SMITH, K. "The Quality of Temporal Information". In: *Proceedings of the International Conference on Information Quality – ICIQ*, pp. 173–178, 2000.

[24] MALETIC, J. I., , MARCUS, A. "Data Cleansing: Beyond Integrity Analysis". In: *Proceedings of the International Conference on Information Quality – ICIQ*, pp. 200–209, 2000.

[25] DANTAS, A. C. H., SEIXAS, J. M. "Neural Networks for Data Quality Monitoring of Time Series". In: *9th International Conference on Enterprise Information Systems*, pp. 411–415, Funchal, Madeira, 2007.

[26] CAPORELLO, G., MARAVAL, A. *A Tool for Quality Control of Time Series Data*. Program terror:, Bank of Spain, Spain., 2004.

[27] BUSATTO, R. "Using Time Series to Assess Data Quality in Telecommunications Data Warehouses". In: *Proceedings of the International Conference on Information Quality – ICIQ*, pp. 129–136, 2000.

[28] GERTLER, J. *Fault Detection and Diagnosis in Engineering Systems*. Marcel Dekker, 1998.

[29] LJUNG, L. *System Identification: Theory for the User*. Prentice Hall, 1998.

[30] HYVÄRINEN, A., KARHUNEN, J., OJA, E. *Independent Component Analysis*. John Wiley & Sons, 2001.

[31] REIGROTZKI, M., MILEK, J., BOSH, H., et al. *A Holistic Approach to Data Quality Management*. In: Report, Predict AG, 2001.

[32] PERNICI, B., SCANNAPIECO, M. "Data quality in web information systems", *Journal on Data Semantics I - Springer Verlag*, pp. 48–69, 2003.

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

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

[35] CAPPIELLO, C. *Data Quality and Multichannel Services*. Tese de D.Sc., 2005.

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

[37] 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.

[38] JARKE, M., JEUSFELD, M. A., QUIX, C., et al. "Architecture and Quality in Data Warehouses: An Extended Repository Approach. Information Systems", *Information Systems*, v. 24, n. 3, 1999.

[39] DALCIN, E. C. *Data Quality Concepts and Techniques Applied to Taxonomic Databases*. Tese de D.Sc., School of Biological Sciences, Faculty of Medicine, Health and Life Sciences, University of Southampton, 2004.

[40] CHAPMAN, A. D., BUSBY, J. R. *Linking plant species information to continental biodiversity inventory, climate and environmental monitoring*. London, Chapman and Hall, 1997.

[41] TAGUCHI, G. *Introduction to Off-line Quality Control*. Magaya, Japan, Central Japan Quality Control Association, 1979.

[42] NAVATHE, S., BATINI, C., CERI, S. *The Entity Relationship Approach*. New York, Wiley and Sons, 1992.

[43] TEOREY, J. D. *Database Modeling and Desing: The Entity-Relationship Approach*. San Mateo, CA, Morgan Kaufman Publisher, 1990.

[44] *Industrial data – ISO 8000*. Relatório técnico, International Organization for Standardization - ISO TC 184/SC 4 - http://www.iso.org, 2011.

[45] G.DARMOIS. "Analyse générale des liaisons stochastiques", *Rev. Inst. Internationale Statist.*, v. 21, n. 2-8, 1953.

[46] KAGAN, A. M., LINNIK, Y. V., RAO, C. R. *Characterization Problems in Mathematical Statistics*. New York, Wiley, 1973.

[47] HERAULT, J., JUTTEN, C. "Space or time adaptive signal processing by neural network models". In: *Neural networks for computing: Proceedings of the AIP Conference*, pp. 206–211, New York: American Institute of Physics, 1986.

[48] COMON, P. "Independent component analysis—a new concept?" *Signal Processing*, v. 36, pp. 287–314, 1994.

[49] COVER, T. M., THOMAS, J. A. *Elements of Information Theory*. John Wiley & Sons, 1991.

[50] PAJUNEN, P. "Blind Source Separation Using Algorithmic Information Theory", *Neurocomputing*, v. 22, pp. 35–48, 1998.

[51] PAJUNEN, P. *Extensions of Linear Independent Component Analysis: Neural and Information-Theoretic Methods*. Tese de Ph.D., Helsink Univertity of Technology, 1998.

[52] MATSUOKA, K., OHYA, M., , et al. "A Neural Net for Blind Separation of Nonstationary Signal", *Neural Networks*, v. 8, n. 3, 1995.

[53] RISSANEN, J. "Modeling by Shortest Data Description", *Automatica*, v. 14, pp. 465–471, 1978.

[54] RISSANEN, J. "A Universal Prior for Integers and Estimation by Minimum Description Length", *Annual of Statistics*, v. 11, n. 2, pp. 416–431, 1983.

[55] L.TONG, SOON, V., HUANG, Y. F., et al. "Indeterminacy and identifiability of blind identification", *IEEE Trans. CAS*, v. 38, n. 2, pp. 499–509, 1991.

[56] SZUPILUK, R., CICHOCKI, A. "Blind signal separation using second order statistics", *IEEE Trans. on Signal Processing*, pp. 485–488, 2001.

[57] GEORGIEV, P., CICHOCKI, A. "Blind source separation via symmetric eigenvalue decomposition". In: *Proceedings of Sixth International Symposium on Signal Processing and its Applications*, pp. 17–20, Kuala Lumpur, Malaysia, 2001.

[58] BELOUCHRANI, A., ABED-MERAIM, K., CARDOSO, J. F., et al. "Second-order blind separation of temporally correlated sources". In: *Proc. Int. Conf. on Digital Sig. Proc.(Cyprus)*, pp. 346–351, 2002.

[59] BELOUCHRANI, A., ABED-MERAIM, K. "Second-order blind separation of temporally correlated sources". In: *Proc. Int. Conf. on Digital Sig. Proc.(Cyprus)*, pp. 346–351, 2002.

[60] F.CARDOSO, J., SOULOUMIAC, A. "Jacobi angles for simultaneous diagonalization", *SIAM Journal of Matrix Analysis and Applications*, v. 17, n. 1, pp. 161–164, 1996.

[61] LI, X.-L., ZHANG, X.-D. "Sequential blind extraction adopting second-order statistics", *IEEE Signal Processing Letters*, v. 14, n. 1, pp. 58–61, 2007.

[62] ZIEHE, A., LASKOV, P., NOLTE, G., et al. "A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind signal separation", *J Mach. Learn. Res.*, v. 5, pp. 1777–800, 2004.

[63] R.VOLLGRAF, OBERMAYER, K. "Quadratic optimization for simultaneous matrix diagonalization", *IEEE Trans. Signal Processing*, v. 54, n. 9, pp. 3270–3278, 2006.

[64] CHOI, S., CICHOCK, A., BELOUCHRAN, A. "Blind separation of nonstationary sources in noisy mixtures", *Journal of VLSI Signal Processing*, 2002.

[65] BELOUCHRAN, A., CICHOCK, A. "Robust whitening procedure in blind source separation context", *Electronics Letters*, v. 36, n. 24, pp. 2050–2053, 2000.

[66] GHARIEB, R. R., CICHOCK, A. "Second order statistics based blind signal separation using a bank of subband filters", *Journal of Digital Signal Processing*, 2003.

[67] TICHAVSKÝ, P., KOLDOVSKÝ, Z., DORON, E., et al. "Blind signal separation by combining two ICA algorithms: HOS-based EFICA and time structure-based WASOBI". In: *Proceedings of The European Signal Processing Conference (EUSIPCO'2006)*, Florence, 2006.

[68] YEREDOR, A. "Blind Separation of Gaussian Sources via Second-Order Statistics with Asymptotically Optimal Weighting", *IEEE Signal Processing Letters*, v. 7, n. 7, pp. 197–200, 2000.

[69] CHOI, S., CICHOCKI, A. "Blind separation of nonstationary sources in noisy mixtures, Electronics Letters", *Electronics Letters*, v. 36, pp. 848–849, 2000.

[70] CARDOSO, J. F., SOULOUMIAC, A. "Blind beam-forming for non Gaussian signals", *IEE Proceedings-F*, v. 140, pp. 362–370, 1993.

[71] CICHOCKI, A., AMARI, S., SIWEK, K., et al. *ICALAB*. http://www.bsp.brain.riken.jp/icalab, 2011.

[72] GEORGIEV, P., CICHOCKI, A. "Robust blind source separation utilizing second and fourth order statistics". In: *International Conference on Artificial Neural Networks (ICANN-2002)*, Madrid, Spain, 2002.

[73] GEORGIEV, P., CICHOCK, A. "Robust Independent Component Analysis via Time-Delayed Cumulant Functions", *IEICE Trans. Fundamentals*, v. E86-A, n. 3, 2003.

[74] HYVÄRINEN, A. "A family of fixed-point algorithms for independent component analysis". In: *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, pp. 3917–3920, Munich, Germany, 1997.

[75] HYVÄRINEN, A., OJA, E. "A fast fixed-point algorithm for independent component analysis", *Neural Computation*, v. 9, n. 7, pp. 1483–1492, 1997.

[76] HYVÄRINEN, A. "Fast and robust fixed-point algorithms for independent component analysis", *IEEE Trans. on Neural Networks*, v. 10, n. 3, pp. 626–634, 1999.

[77] DING, S. "Power Iteration Algorithm for ICA Based on Diagonalizations of Non-Linearized Covariance Matrix." In: *ICICIC*, pp. 730–733, 2006.

[78] KOLDOVSKÝ, Z., TICHAVSKÝ, P., OJA, E. "Efficient Variant of Algorithm FastICA for Independent Component Analysis Attaining the Cramér-Rao Lower Bound", *IEEE Trans. on Neural Networks*, v. 17, n. 5, pp. 1265–1277, 2006.

[79] TICHAVSKÝ, P., KOLDOVSKÝ, Z., DORON, E., et al. "Blind Signal Separation by Combining Two ICA Algorithms: HOS-Based EFICA and Time Structure-Based WASOBI". In: *Proceedings of The 2006 European Signal Processing Conference (EUSIPCO'2006)*, 2006.

[80] LU, W., RAJAPAKSE, J. C. "Constrained Independent Component Analysis". In: *NIPS*, pp. 570–576, 2000.

[81] LU, W., RAJAPAKSE, J. C. "Approach and Applications of Constrained ICA", *IEEE Transactions on Neural Networks*, v. 16, n. 1, 2005.

[82] LIN, R. Q. H., ZHENG, Y.-R., L. YIN, F., et al. "A fast algorithm for one-unit ICA-R, Information Science", *Information Science*, 2007.

[83] AMARI, S., CHEN, T., CICHOCKI, A. "Nonholonomic orthogonal learning algorithms for blind source separation", *Neural Computation*, v. 12, pp. 1463–1484, 2000.

[84] AKYZAWA, T. ""Extended quasi-Newton method for the ICA". In: *Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation*, p. 521–525, 2000.

[85] CRUCES, S., CICHOCKI, A. "Combining blind source extraction with joint approximate diagonalization: Thin Algorithms for ICA". In: *Proc. of the Fourth Symposium on Independent Component Analysis and Blind Signal Separation*, pp. 463– 469, Japan, 2003.

[86] CRUCES, S., CASTEDO, L., CICHOCKI, A. "Novel Blind Source Separation Algorithms Using Cumulants". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 3152–3155, Istanbul, Turkey, 2000.

[87] AMARI, S. "Natural Gradient Learning for over- and under-complete bases in ICA", *Neural Computation*, v. 11, pp. 1875–1883, 1999.

[88] CRUCES, S., CICHOCKI, A., AMARI, S. "Criteria for the Simultaneous Blind Extraction of Arbitrary Groups of Sources". In: *3rd international conference on Independent Component Analysis and Blind Signal Separation*, San Diego, California, USA, 2001.

[89] AMARI, S. "The Minimum Entropy and Cumulant Based Contrast Functions for Blind Source Extraction", *Lecture Notes in Computer Science, Springer-Verlag, IWANN'2001*, v. II, pp. 786–79, 2001.

[90] ZANG, C., FRISWELL, M. I. "Decomposition of Time Domain Vibration Signals Using the Independent Component Anaysis Technique". In: *Proceedings of the Third International Conference*, pp. 434–445, Suécia, 2002.

[91] XU, Y., WANG, H., CUI, Z., et al. "Independent component analysis of interface fluctuation of gas/liquid two-phase flows – experimental study", *Flow Meas. Instrum*, v. 20, n. 6, pp. 220–229, 2009.

[92] GADHOK, N., KINSNER, W. "Robust independent component analysis for cognitive informatics (ICCI 2005)". In: *Fourth IEEE Conference on Cognitive Informatics*, pp. 86–92, 2005.

[93] 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, Lviv, Ukraine, 2003.

[94] CHI-JIE, LEE, T.-S., CHIU, C.-C. "Financial time series forecasting using independent component analysis and support vector regression", *Journal Decision Support Systems*, v. 47, pp. 2936–2945, 2009.

[95] CHONGHUI, G., HONGFENG, J., ZHANG, N. "Time Series Clustering Based on ICA for Stock Data Analysis". In: *4th International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM*, 2008.

[96] LU, C.-J., WU, J.-Y., LEE, T.-S. "Application of Independent Component Analysis Preprocessing and Support Vector Regression in Time Series Prediction". In: *International Joint Conference on Computational Sciences and Optimization - CSO*, 468-471, 2009.

[97] J.TOHKA, FOERDE, K., ARON, A. R., et al. "Automatic independent component labeling for artifact removal in fMRI", *NeuroImage*, v. 39, n. 1, pp. 1227–1245, 2008.

[98] M.CHEUNG, Y., XU, L. "Independent component ordering in ICA time series analysis", *Neurocomputing*, v. 41, pp. 145–152, 2001.

[99] CHEUNG, Y.-M., XU, L. "An empirical method to select dominant independent components in ICA for time series analysis". In: *International Joint Conference on Neural Networks*, p. 3883 – 3887, 1999.

[100] ESPOSITO, F., SEIFRITZ, E., FORMISANO, E., et al. "Real-time independent component analysis of fMRI time-series", *NeuroImage*, v. 20, n. 4, pp. 2209–2224, 2003.

[101] SUGIMOTO, K., KONDO, H. "Multivariate time series prediction by blind signal deconvolution". In: *ICCAS-SICE*, 2510 – 2513, 2009.

[102] WANG, F., LI, H., LI, R. "Data Mining with Independent Component Analysis". In: *The Sixth World Congress on Intelligent Control and Automation- WCICA*, v. 2, pp. 6043–6047, 2006.

[103] SCHIMERT, J. "Data-Driven Fault Detection Based on Process Monitoring using Dimension Reduction Techniques". In: *Aerospace Conference, IEEE*, pp. 1–12, 2008.

[104] DAVIES, M. E., JAMES, J. C. "Source separation using single channel ICA", *Signal Processing*, v. 87, pp. 1819–183, 2007.

[105] JAMES, J. C., WANG, S. "Blind Source separation in single-channel EEG analysis: An application to BCI". In: *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, p. 6544–6547, 2006.

[106] FERREIRA, D. D. *Análise de Distúrbios Elétricos em Sistemas de Potência*. Tese de D.Sc., Coordenação dos Programas de Pós-graduação de Engenharia da Universidade Federal do Rio de Janeiro, Programa de Engenharia Elétrica, 2011.

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

[108] 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.

[109] PHILLIPS, P. C. B. "Time series regression with a unit root", *Econometrica*, v. 55, n. 2, pp. 277–301, 1987.

[110] GOLDFELD, S. M., QUANDT, R. E. "Some Tests for Homoscedasticity", *Journal of the American Statistical Association*, v. 60, pp. 539–547, 1965.

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

[112] HAYKIN, S. *Neural Networks and Learning Machines*. Prentice Hall, 2008.

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

[114] BELOUCHRANI, A., CICHOCKI, A. "Robust Whitening Procedure in Blind Source Separation Context", *Electronics Letters*, v. 36, n. 24, pp. 2050–2053, 2001.

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

[116] CHATFIELD, C. *The Analysis of Time Series: an Introduction*. 5th ed. United Kingdon, Chapman & Hall/CRC, 1996.

[117] CICHOCKI, A., AMARI, S., SIWEK, K., et al. *ICALAB for Signal Processing Toolbox for BSS, ICA, cICA, ICA-R, SCA and MCA*. 3 ed. , 1997.

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

[119] DUDA, R. O., HART, P. E., STORK, D. G. *Pattern classification*. 2 ed. New York, John Wiley & Sons, 2000.

[120] *EUNITE - European Network on Intelligent Tech-nologies for Smart Adaptive Systems*. http://neuron.tuke.sk/competition, Acessado em junho 2010.

[121] FERREIRA, V. H. *Desenvolvimento de Modelos Neurais Autônomos para Previsão de Carga Elétrica*. Tese de M.Sc., Coordenação dos Programas de Pós-graduação de Engenharia da Universidade Federal do Rio de Janeiro, Programa de Engenharia Elétrica, 2008.

[122] *Stockwiz*. http://www.stockwiz.com, Acessado em março 2007.

[123] *Yahoo Finance*. http://finance.yahoo.com, Acessado em fevereiro 2007.

Apêndice A

Trabalhos produzidos

No decorrer diversos artigos foram produzidos e apresentados em congressos nacionais e internacionais. Além disso, são apresentados trabalhos submetidos a revistas, trabalhos a serem submetidos e trabalhos que estendem a aplicação dos conhecimentos desta tese em outras áreas.

1. ARTIGOS EM REVISA

   FAIER, J. M., SEIXAS, J. M., *Time-Series Data Quality Monitoring Using Independent Component Analysis*, Submetido a Communications of the ACM, 2011.

   Resumo: Na era da informação, bancos de dados em companhias e centros de pesquisa estão ficando cada vez maiores, o que torna a qualidade dos dados uma tarefa chave. Neste artido, uma metodologia utilizando Análise de Componentes Independentes (ICA) é desenvolvida para monitoração da qualidade de dados. A metodologia é abordada sob os conceitos da teoria de controle para séries temporais. A Análise de Componentes Independentes é parte de um pré-processamento, que produz séries mais estruturadas e aumenta a performance do sistema. O sistema de monitoração proposto é desenvolvido com séries temporais simuladas e testado com sucesso em ambientes reais, para séries de carga elétrica e séries financeiras. A inclusão de ICA como passo do pré-processamento estreitou os corredores utilizados para validação dos dados e agregou informação acurada para melhorar a qualidade dos dados.

   FAIER, J. M., SEIXAS, J. M., *Time-Series Data Quality Monitoring using Single Channel Independent Component Analysis*, A submeter.

   Resumo: Neste artigo, uma metodologia utilizando a Análise de Componentes Independentes para canais simples (SCICA) é desenvolvida para a detecção de anomalias em séries temporais. O método com SCICA é eficaz na extração de fontes de ruídos e amostras anômalas nas séries temporais, o que permite

a monitoração através de corredores construídos na fonte dos problemas. A metodologia é testada com sucesso na monitoração da qualidade de dados em séries de carga elétrica e séries financeiras.

2. CONGRESSOS INTERNACIONAIS

   FAIER, J.M., SEIXAS, J.M, *Data Quality Monitoring: Independent Component Analysis for Time-Series*, 18th European Signal Processing Conference (EUSIPCO), Aalborg/Dinamarca, 2010.

   Resumo: Neste artigo, a Análise de Componentes Independentes (ICA) é utilizada para a monitoração de séries temporais de carga elétrica. O método foi aplicado na fase de pré-processamento, o que melhorou o desempenho do sistema de qualidade. A extração das fontes revelou informação relevante e estreitou os corredores de validação utilizados para validação dos dados.

   FAIER, J.M., SEIXAS, J.M.,*Time Delayed Independent Component Analysis for Data Quality Monitoring*, 17th International Conference on Systems, Signals and Image Processing (IWSSIP), Rio de Janeiro/RJ, 2010.

   Resumo: Neste artigo, Análise de Componentes Independentes atrasada no tempo (TDICA) é utilizada para a monitoração da qualidade de dados de séries temporais do setor elétrico. A Análise de Componentes Independentes foi aplicada na fase de pré-processamento, o que aumentou o desempenho do sistema de qualidade de dados. A extração de fontes reduziu o erro de previsão, revelou informações relevantes e estreitou o comprimento dos corredores de validação.

3. CONGRESSOS NACIONAIS

   FAIER, J. M., SEIXAS, J. M., *Monitoração Neural da Qualidade de Dados em Séries Temporais Utilizando a Análise de Componentes Independentes*, Congresso Brasileiro de Automática - CBA, Juiz de Fora/MG, 2008.

   Resumo: Este artigo propõe a aplicação da análise de componentes independentes e de processamento neural para monitorar a qualidade de dados em séries temporais. Um sistema de monitoração flexível é desenvolvido para analisar a qualidade de séries temporais a serem armazenadas em grandes bancos de dados. Realizou-se um estudo de caso em séries temporais financeiras, observando-se que a extração de componentes independentes melhorou o desempenho do sistema, reduzindo o erro de previsão e estreitando os corredores de validação. Assim, a análise de componentes independentes pode ser utilizada como ferramenta adicional no pré-processamento de séries temporais, visando a qualidade de dados.

FAIER, J. M., SEIXAS, J.M., *Monitoração da Qualidade de Dados em Séries Temporais de Carga Elétrica Utilizando a Análise de Componentes Independentes*, Congresso Brasileiro de Redes Neurais - CBRN, Ouro Preto/MG, 2009.

Resumo: Este artigo propõe a aplicação de processamento neural sobre componentes independentes para monitorar a qualidade de dados em séries temporais de carga elétrica. Observou-se que a extração de componentes independentes melhorou o desempenho do sistema, estreitando os corredores de validação, agregando informações relevantes e fornecendo previsões de carga mais acuradas. Assim, a análise de componentes independentes pode ser utilizada como ferramenta adicional no pré-processamento de séries de energia, visando a qualidade de dados.

4. EXTENSÕES

- CAPÍTULO DE LIVRO

  MAIDANTCHIK, C., SEIXAS, J.M., GRAEL, F. F., TORRES, R. C., FERREIRA, F. G., GOMES, A. S. G, FAIER, J. M., LAPA e SILVA, J. R., SILVA, MELLO, F. C. Q., KRITSKI, A., SOUZA, J. B. O., *A Decision Support System Based on Artificial Neural Networks for Pulmonary Tuberculosis Diagnosis*, Efficient Decision Support Systems: Practice and Challenges From Current to Future, 2011.

  Resumo: Neste capítulo, mostra-se uma aplicação web para a coleta informações de pacientes de tuberculose (TB) através de formulários eletrônicos e aplicam-se Redes Neurais Artificiais para detectar padrões de TB. Métodos de Qualidade de Dados foram utilizados para garantir a acurácia dos dados armazenados, evitando informação imprecisa e melhorando o valor do resultado final.