**UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS**

Maria Mercedes Gamboa-Medina

**DETECÇÃO DE VAZAMENTOS E ALTERAÇÕES EM
REDES DE DISTRIBUIÇÃO DE ÁGUA PARA
ABASTECIMENTO, DURANTE A OPERAÇÃO, USANDO
SINAIS DE PRESSÃO**

São Carlos

2017

Maria Mercedes Gamboa-Medina

**DETECÇÃO DE VAZAMENTOS E ALTERAÇÕES EM
REDES DE DISTRIBUIÇÃO DE ÁGUA PARA
ABASTECIMENTO, DURANTE A OPERAÇÃO, USANDO
SINAIS DE PRESSÃO**

Tese apresentada à Escola de Engenharia de São Carlos da Universidade de São Paulo, para obtenção do título de Doutor em Ciências - Programa de Pós-Graduação em Engenharia Hidráulica e Saneamento.

Área de concentração: Hidráulica e Saneamento

Orientadora: Profa. Tit. Luisa Fernanda Ribeiro Reis

VERSÃO CORRIGIDA

São Carlos

2017

Este trabalho está licenciado com uma Licença Creative Commons Atribuição-Compartilha Igual 4.0 Internacional. É permitido copiar e compartilhar o trabalho, sempre que a atribuição seja respeitada. Todo uso ou adaptação está permitido, e deve ser compartilhado sob o mesmo tipo de licença. Documento da licença pode ser consultado em <https://creativecommons.org/licenses/by-sa/4.0/legalcode>

Este documento foi escrito com $\text{\LaTeX}$ e usando a classe USPSC

G192d	<p>Gamboa-Medina, Maria Mercedes</p> <p>DETECÇÃO DE VAZAMENTOS E ALTERAÇÕES EM REDES DE DISTRIBUIÇÃO DE ÁGUA PARA ABASTECIMENTO, DURANTE A OPERAÇÃO, USANDO SINAIS DE PRESSÃO / Maria Mercedes Gamboa-Medina ; orientadora Luisa Fernanda Ribeiro Reis. – São Carlos, 2017.</p> <p>130 p. : il. (algumas color.) ; 30 cm.</p> <p>Tese (Doutorado) - Programa de Pós-Graduação e Área de Concentração em Engenharia Hidráulica e Saneamento – Escola de Engenharia de São Carlos, Universidade de São Paulo, 2017.</p> <p>1. Redes abastecimento. 2. Vazamentos. 3. Sinais. 4. Detecção falhas. I. Reis, Luisa Fernanda, orient. II. Título.</p>
-------	---

FOLHA DE JULGAMENTO

Candidata: Engenheira **MARIA MERCEDES GAMBOA-MEDINA.**

Título da tese: "Detecção de vazamento e alterações em redes de distribuição de água para abastecimento, durante a operação, usando sinais de pressão".

Data da defesa: 16/08/2017.

Comissão Julgadora:

Resultado:

Profa. Titular **Luisa Fernanda Ribeiro Reis**
(Orientadora)
(Escola de Engenharia de São Carlos/EESC)

Aprovado

Prof. Associado **Ivan Nunes da Silva**
(Escola de Engenharia de São Carlos/EESC)

APROVADO

Prof. Titular **André Carlos Ponce de Leon Ferreira de Carvalho**
(Instituto de Ciências Matemáticas e de Computação/ICMC)

Aprovado

Prof. Dr. **Edevar Luvizotto Júnior**
(Universidade Estadual de Campinas/UNICAMP)

Aprovado

Profa. Dra. **Helena Margarida Machado da Silva Ramos**
(Instituto Superior Técnico de Lisboa)

Aprovado

Coordenadora do Programa de Pós-Graduação em Engenharia Hidráulica e Saneamento:

Profa. Associada **Maria Bernadete A. Varesche Silva**

Presidente da Comissão de Pós-Graduação:

Prof. Associado **Luis Fernando Costa Alberto**

Ao leitor que, levando algo destas páginas para si, completa-lhes o sentido.

AGRADECIMENTOS

A realização do doutorado, como de toda andança pela vida, somente foi possível pelas muitas pessoas que tenho encontrado no caminho, às quais tenho muito a agradecer. Gostaria de manifestar aqui minha gratidão especificamente:

- à professora Fernanda Reis, pela dedicada orientação, por todos os ensinamentos, e especialmente pela confiança depositada e a imensa paciência que teve para mim durante esses anos;
- ao professor Rodrigo Guido pela generosa colaboração com sugestões para este trabalho e por ter auxiliado em seu desenvolvimento desde o mestrado;
- ao professor Rodrigo de Mello por ter contribuído a este trabalho e à minha formação com recomendações e discussões sobre a pesquisa;
- aos engenheiros Fernando Lourencetti, José Roberto Meciano e Rogério do Prado, do DAAE-Araraquara, pela sua colaboração que foi fundamental para o estudo de caso e para me aproximar um pouco da realidade do gerenciamento de sistemas;
- ao Asdrubal Lozada pela colaboração na implementação dos métodos, e principalmente pela ajuda, suporte e parceria fundamentais durante a elaboração desta tese;
- à engenheira Maria Alejandra Aparicio pelo auxílio com as labores de aquisição de dados durante seu estagio;
- ao Luiz e Edson, funcionários do DAAE, pelo acompanhamento às tarefas de campo, que eles tornaram menos difíceis e mais agradáveis;
- aos colegas e amigos do Laboratório de Simulação Numérica (LabSiN) de tantas gerações que já tive o prazer de conhecer e que lembro com carinho e gratidão;
- aos professores, funcionários e estudantes do SHS-EESC pelo tempo compartilhado;
- ao Javier Mahecha, a quem tenho muito a agradecer, dentre isso a companhia e apoio durante a primeira parte do doutorado;
- ao Departamento Autônomo de Água e Esgoto da cidade de Araraquara por permitir a realização do estudo de caso e oferecer o suporte para sua realização;
- ao CNPq pela bolsa de doutorado que recebi.

"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."

John Tukey (1915-2000)

RESUMO

GAMBOA-MEDINA, M. **DETECÇÃO DE VAZAMENTOS E ALTERAÇÕES EM REDES DE DISTRIBUIÇÃO DE ÁGUA PARA ABASTECIMENTO, DURANTE A OPERAÇÃO, USANDO SINAIS DE PRESSÃO.** 2017. 130p. Tese (Doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2017.

O controle ativo de vazamentos é fundamental para o gerenciamento dos recursos hídricos, e particularmente o problema de sua detecção precisa de alternativas de solução. Nesta pesquisa foram desenvolvidos três métodos para detecção da ocorrência de vazamentos ou outras alterações em redes de distribuição de água para abastecimento, durante operação, com base na análise dos sinais de pressão adquiridos por monitoramento permanente na rede. Os métodos foram desenvolvidos com informações de um estudo de caso real, e fundamentam-se em cadeias de Markov, em controle estatístico de processos, e em comparação com padrões, respectivamente. Seu desempenho global foi quantificado com a área abaixo da curva ROC, obtendo valores médios de 0,67, 0,65 e 0,71, respectivamente. Foi encontrado que a ocorrência de vazamentos produz mudanças nas pressões medidas na rede, mas tais mudanças podem ser próximas às que acontecem pela operação normal, de forma que os métodos propostos constituem ferramentas de suporte ao gerenciamento úteis, sem atingir a detecção e diferenciação da totalidade dos vazamentos e alterações.

Palavras-chave: Vazamentos. Redes de abastecimento de água. Sinais de pressão. Séries temporais. Aprendizado de máquina.

ABSTRACT

GAMBOA-MEDINA, M. LEAK AND ANOMALY DETECTION ON WATER SUPPLY NETWORKS, WHILE OPERATING, USING PRESSURE SIGNALS.. 2017. 130p. Tese (Doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2017.

Active leak control is fundamental on water resources management, and particularly the detection problem needs alternative solutions. On this research three methods for leak or anomaly detection on operating supply systems were developed, based on analysis of pressure signals acquired by network monitoring. The proposed methods used real data from a case study, and are based on Markov chains, statistical process control, and pattern comparison, respectively. Their overall performances on case study were measured using the area under ROC curve, reaching averages 0,67, 0,65 and 0,71, respectively. It was found that a leak onset result in disturbances on measured pressure, but sometimes those disturbances are close to those of normal operation, therefore proposed methods are useful decision support tools, not reaching detection and distinction for all the leaks or anomalies.

Keywords: Leak. Burst. Water Supply Network. Pressure signals. Time series. Machine learning.

LISTA DE ILUSTRAÇÕES

Figura 1 – Vazão na saída do reservatório, informação adquirida pelo sistema de telemetria.	36
Figura 2 – Nível nos reservatórios, informação adquirida pelo sistema de telemetria. . .	37
Figura 3 – Visão geral das séries de pressão obtidas durante todo o período de monitoramento	38
Figura 4 – Esquema do modelo em EPANET do setor sob estudo	39
Figura 5 – Ganho relativo acumulado nas funções objetivo (Δf_{mean}) pela inclusão de um sensor adicional, de 2 a 21.	49
Figura 6 – Locais escolhidos pela otimização multiobjetivo para diferente número de sensores.	52
Figura 7 – Esquema da matriz de confusão.	62
Figura 8 – Curva ROC para três tipos de classificadores. Esquema.	63
Figura 9 – Exemplo do efeito da escala sobre a inspeção para classificação manual das series. Trecho da serie do sensor 1	79
Figura 10 – Visão geral dos sinais adquiridos, e a media diária	80
Figura 11 – Visão geral dos sinais corrigidos pelo translado da média.	83
Figura 12 – Correlograma das séries de pressão originais, completas, para as séries dos nove sensores.	85
Figura 13 – Correlograma das séries de pressão corrigidas (média transladada), considerando somente a classe "normal", para as séries dos nove sensores.	86
Figura 14 – Análise espectral dos sinais de pressão originais, para as series dos nove sensores.	87
Figura 15 – Análise espectral dos sinais de pressão corrigidos (médias transladadas), desconsiderando os da classe 3 "visual", para as series dos nove sensores. . .	88
Figura 16 – Valor médio de pressão para os momentos uma semana, segundo os dias completos da classe "normal" nas séries corrigidas, para os nove sensores. .	91
Figura 17 – Variância dos valores de pressão e resultado do teste de normalidade, para os momentos de uma semana, segundo os dias completos da classe "normal" nas séries corrigidas, para os nove sensores.	92
Figura 18 – Taxa de recorrência (RR) para diferentes raios de corte (distância máxima para avaliar a recorrência), para os nove sensores.	93
Figura 19 – Taxa de recorrência (RR) para diferentes raios de corte (distância máxima para avaliar a recorrência), para os nove sensores, em escala logarítmica. . .	94
Figura 20 – Gráfico de recorrência (RP) e série utilizada (trecho "normal" extraído das series corrigidas) para os sensores 1, 2, 5 e 8. Apresenta-se o trecho central da série e a região central do RP, de tamanho 7000 (9,7 dias) e 7000x7000, respectivamente.	95

Figura 21 – Curvas ROC resultado da detecção usando cadeias de Markov, para as nove séries corrigidas completas. Estados: pressão com $\Delta t = 2min$ e $\Delta e = 0,50mca$. Memória $k_{mem} = 5$. Considerando todas as classes (a. normal-todos) ou somente os experimentos (b. normal-simulado)	98
Figura 22 – AUROC resultante da detecção para as séries completas, com estados definidos pela pressão adquirida a cada $\Delta t = 2min$ ou pela pressão média a cada $\Delta t = 10min$. Valor de média e desvio padrão calculado entre os resultados de AUROC dos nove sensores, considerando todas as classes (a. normal-todos) ou somente os experimentos (b. normal-simulado), várias discretizações dos estados (Δe), e valores de memória (k_{mem}).	99
Figura 23 – AUROC resultante da detecção para as séries completas corrigidas, com estados definidos pela diferença entre pressões cada $\Delta t = 2min$ ou cada $\Delta t = 10min$. Valor de média $\pm$ desvio padrão calculado entre os resultados de AUROC dos nove sensores, considerando todas as classes (normal-todos) ou somente os experimentos (norma-simu), várias discretizações dos estados (Δe), e valores de memória (k_{mem}).	100
Figura 24 – Trecho da série e as probabilidades condicionais calculadas, e detalhe do trecho com alteração. Extraído da série corrigida do sensor 2, estados: pressões cada $\delta t = 2min$ discretizadas em $\delta e = 0,1mca$. Memória $k_{mem} = 1$	101
Figura 25 – Curvas ROC resultado da detecção usando MVMEP, para as nove séries completas. $\lambda = 0,5$, $\beta_c = 3$, e $\beta_l = 3$. a)todas as classes. b)somente os experimentos	102
Figura 26 – AUROC resultante da detecção para as séries completas aplicando controle estatístico. Valor de média e desvio padrão calculado entre os resultados de AUROC de cada um dos nove sensores, considerando todas as classes (a. normal-todos) ou somente os experimentos (b. normal-simulado), vários limiares de controle de atualização β_c e β_l , e fator de ponderação λ	103
Figura 27 – Trecho da série, média e limites do intervalo de aceite para atualização tanto de curto quanto de longo prazo. Extraído da série original do sensor 2, com $\lambda = 0,2$, $\beta_c = 3$, e $\beta_l = 2$	105
Figura 28 – Curvas ROC resultado da detecção usando comparação com o padrão, para as nove séries originais completas. Janela $30valores$, avanço $10valores$, transformação <i>ntng</i> e comparação <i>eucl</i> . a)todas as classes supervisionadas (normal-todos). b)somente os experimentos (normal-simulado)	108
Figura 29 – AUROC resultante da detecção para as séries completas originais aplicando comparação com padrão. Valor de média e desvio padrão calculado entre os resultados de AUROC de cada um dos nove sensores, considerando todas as classes supervisionadas (normal-todos), janela de $720 = 1dia$ e passo $240 = 8horas$, todas as funções de transformação e as de comparação. . . .	109

Figura 30 – AUROC resultante da detecção para as séries completas originais aplicando comparação com padrão. Valor de média e desvio padrão calculado entre os resultados de AUROC de cada um dos nove sensores, considerando somente os experimentos (normal-simulado), janela de $720 = 1dia$ e passo $240 = 8horas$, todas as funções de transformação e as de comparação.	110
Figura 31 – AUROC resultante da detecção para as séries completas originais aplicando comparação com padrão. Média e desvio padrão dos resultados dos nove sensores, janela de $180 = 6horas$, avanço $60 = 2horas$, avaliação com classes supervisionadas (normal-todos) e (normal-simulado), todas as funções de comparação, e as seis funções de transformação que atingiram melhor desempenho.	111
Figura 32 – AUROC resultante da detecção para as séries completas originais aplicando comparação com padrão. Média e desvio padrão dos resultados dos nove sensores, janela de $30 = 1hora$, avanço $10 = 20min$, avaliação com classes supervisionadas (normal-todos) e (normal-simulado), todas as funções de comparação, e as seis funções de transformação que atingiram melhor desempenho.	112
Figura 33 – AUROC resultante da detecção para as séries completas originais aplicando comparação com padrão. Média e desvio padrão dos resultados dos nove sensores, janela de $720 = 1dia$, avanço $60 = 2horas$, avaliação com classes supervisionadas (normal-todos) e (normal-simulado), todas as funções de comparação, e as seis funções de transformação que atingiram melhor desempenho.	114
Figura 34 – AUROC resultante da detecção para as séries completas originais aplicando comparação com padrão. Média e desvio padrão dos resultados dos nove sensores, janela de $180 = 6horas$, avanço $30 = 1hora$, avaliação com classes supervisionadas (normal-todos) e (normal-simulado), todas as funções de comparação, e as seis funções de transformação que atingiram melhor desempenho.	115
Figura 35 – AUROC resultante da detecção para as séries completas originais aplicando comparação com padrão. Média e desvio padrão dos resultados dos nove sensores, janela de $30 = 1hora$, avanço $5 = 10min$, avaliação com classes supervisionadas (normal-todos) e (normal-simulado), todas as funções de comparação, e as seis funções de transformação que atingiram melhor desempenho.	116
Figura 36 – Trecho da série, padrão, e distâncias calculadas. Extraído da série original do sensor 2, tamanho de janela 30, avanço 10, alternativa de transformação <i>ntng</i> , e função de comparação <i>eucl</i>	119

LISTA DE TABELAS

Tabela 1	– Características dos experimentos de simulação de vazamentos	39
Tabela 2	– Razão de vazamento e vazão total, V/Q , segundo o valor do coeficiente c . .	43
Tabela 3	– Valores das funções objetivo e seus ganhos relativos para número de sensores incremental (M).	48
Tabela 4	– Média e desvio das séries de pressão adquiridas	90
Tabela 5	– Lista das funções de transformação e de comparação utilizadas, alternativamente, no método	107

SUMÁRIO

1	INTRODUÇÃO	23
1.1	Objetivos	25
1.2	Hipótese	25
1.3	Organização do documento	26
2	ESTADO DA ARTE EM DETECÇÃO DE VAZAMENTOS	27
2.1	Métodos baseados em modelos hidráulicos	28
2.2	Métodos baseados em análise de sinais	30
2.2.1	Análise no transitório hidráulico	30
2.2.2	Análise durante operação do sistema	31
3	APRESENTAÇÃO DO ESTUDO DE CASO	35
3.1	Características gerais do setor	35
3.2	Informação existente e adquirida	35
3.2.1	Aquisição rotineira pelo sistema de telemetria	36
3.2.2	Condição do sistema	37
3.2.3	Informação de pressão adquirida	37
3.3	Modelo hidráulico de simulação da rede	40
4	PROJETO DE MONITORAMENTO	41
4.1	Método proposto	41
4.1.1	Representação dos vazamentos no modelo	41
4.1.2	Análise de correlação	43
4.1.3	Análise de sensibilidade	44
4.1.4	Otimização	45
4.1.5	Número de sensores	46
4.2	Aplicação ao estudo de caso	47
4.2.1	Número de sensores	48
4.2.2	Localização dos sensores	49
4.2.3	Análises adicionais	50
4.3	Comentários finais	50
5	MÉTODOS DE ANÁLISE	53
5.1	Métodos para descrição das séries	53
5.1.1	Estacionariedade	53
5.1.2	Autocorrelação	55
5.1.3	Análise espectral	56

5.1.4	Análise de recorrência cruzada	57
5.2	Métodos para a detecção de vazamentos e alterações	59
5.2.1	Avaliação de desempenho da detecção	60
5.2.2	Processo estocástico e cadeias de Markov	64
5.2.3	Controle estatístico	66
5.2.4	Comparação com padrão	68
5.2.4.1	Funções de transformação	69
5.2.4.2	Funções de comparação	73
5.2.4.3	Inferência para detecção	76
6	ANÁLISE DAS SÉRIES	77
6.1	Descrição das series adquiridas	77
6.1.1	Classificação manual	77
6.1.2	Estacionaridade	80
6.1.3	Autocorrelação	84
6.1.4	Análise espectral	86
6.1.5	Descrição estatística	89
6.1.6	Comportamento médio	90
6.1.7	Análise de recorrência	92
6.2	Detecção com cadeias de Markov	96
6.2.1	Resultados comentados	97
6.2.2	Discussão	99
6.3	Detecção com controle estatístico	102
6.3.1	Resultados comentados	102
6.3.2	Discussão	104
6.4	Detecção pela comparação com padrões	106
6.4.1	Resultados comentados	108
6.4.2	Discussão	115
7	CONCLUSÕES	121
	REFERÊNCIAS	125

1 INTRODUÇÃO

O aumento do estresse hídrico e do interesse por suas consequências vem incrementando a necessidade de melhoras no gerenciamento de recursos para atendimento das demandas de abastecimento urbano de água, de forma que as companhias responsáveis são chamadas a aperfeiçoar seus procedimentos. Um aspecto importante na implementação de tais melhoras está na forma como é realizado o controle de vazamentos. Sistemas de distribuição de água para abastecimento tem pelos vazamentos condições de perdas intoleravelmente altas em comparação com critérios industriais típicos. Embora a incidência desses vazamentos não possa ser quantificada com certeza, no Brasil esta é reconhecidamente elevada (SNSA, 2017), e suas consequências vão além da evidente perda econômica, já que incluem perdas de energia, desestabilização dos solos, riscos à saúde pública, exploração pré-matura de mananciais, dentre outras.

A busca pela prevenção ou redução dos vazamentos com estratégias como o gerenciamento otimizado de pressões não elimina a necessidade de contar com um sistema de controle dos vazamentos, que inclua as tarefas de detecção, localização e reparo. Enquanto a eficiência nos reparos depende principalmente da capacidade logística da companhia responsável e portanto pode ser melhorada com ações administrativas, as tarefas de detecção e de localização incluem desafios técnicos ainda em aberto.

Sob a ótica da sustentabilidade, a prática antiga de esperar por manifestações evidentes de um vazamento (controle passivo) é insuficiente. Assim, é cada vez mais urgente o controle de perdas ativo, realizado permanentemente, com abrangência de toda a rede, e de maneira condizente com a realidade dos órgãos responsáveis pelo abastecimento. Essa realidade inclui recursos limitados para a compra de equipamentos, ausência de modelos hidráulicos completos, atualizados e calibrados, informação incompleta sobre o estado do sistema e fragilidade das tubulações existentes, dentre outras limitações.

A academia vem empenhando esforços para propôr metodologias de controle de vazamentos, e embora o problema seja conhecido desde um século atrás e melhoras importantes tenham sido atingidas, atualmente essas propostas não atingem o objetivo do controle ativo com a confiabilidade e adaptabilidade necessárias para ser efetivamente implementada pelos órgãos responsáveis pelo abastecimento. Muitas das técnicas de detecção de vazamentos propostas no âmbito da engenharia hidráulica têm sido desenvolvidas e testadas somente sobre sistemas teóricos, enquanto outras foram avaliadas para sistemas controlados de laboratório ou ainda em testes de campo, mas a aplicação prática a sistemas reais tem sido conseguida em poucos casos.

A falta de sucesso das técnicas existentes, longe de desmotivar o estudo na área, estimula a pesquisa e o desenvolvimento de soluções que aproveitem os conhecimentos adquiridos e visem a superar as limitações e os problemas conhecidos. Ao analisar as razões para que as

técnicas existentes não tenham sido aplicadas amplamente encontra-se que há algumas limitações que poderiam ser superadas seguindo uma das abordagens de mais recente desenvolvimento. Essa abordagem está baseada na análise dos sinais de pressão e/ou vazão no sistema adquiridos permanentemente, durante operação da rede, como discutido no Capítulo 2.

Outra dentre as diversas motivações para a seleção da abordagem desta pesquisa está no crescimento da implementação, por parte das companhias responsáveis pelo abastecimento de água, de sistemas de aquisição de dados, supervisão e controle (SCADA, da sigla em inglês). Este tipo de sistemas pode ter diferentes graus de complexidade, e entre seus objetivos pode incluir a detecção de vazamentos. Enquanto a adoção desses sistemas aumenta, não é difícil encontrar casos em que a única automação está na aquisição de dados, mas a supervisão e controle continuam a depender unicamente de um operador que, baseado na sua expertise, analisa as medições e promove manobras de acionamento e desligamento de válvulas, basicamente. Usualmente a base para tais análises é a representação gráfica das variáveis monitoradas, em função do tempo, e relatórios com dados agregados. Sem desconhecer que boas decisões de controle e operação podem ser tomadas com o suporte dessas informações, há um desperdício de recursos, de informação e da oportunidade de melhorar os processos, em particular o de detecção de vazamentos ou alterações na rede.

Com isso, a pesquisa de doutorado apresentada neste documento está dirigida a explorar soluções para detecção de vazamentos ou outras alterações no sistema. Busca-se propor metodologias que atendam os critérios do controle ativo, sejam potencialmente aplicáveis a sistemas de abastecimento reais, e aproveitem as informações adquiridas. A abordagem utilizada é, portanto, de análise de sinais de pressão adquiridos em pontos do sistema, aplicando técnicas originárias do processamento de sinais e o aprendizado de máquina.

Uma das características principais do trabalho apresentado nesta tese é tratar com dados reais do problema encarado. Tal característica, que incrementa complexidade e, consequentemente, dificuldade para prever os resultados da pesquisa, se comparado com o uso de sistemas simplificados (dados sintéticos, modelo laboratorial, protótipos, etc.), foi adotada em consonância com o objetivo geral do trabalho. Resultados publicados, aliados à ampla experiência da orientadora desta pesquisa com o estudo de redes de abastecimento, informam sobre a grande dificuldade em incorporar todas as variações e complexidades de um sistema de distribuição em um modelo (seja ele matemático ou físico, parcial ou completo) e a consequente discrepância entre a resposta no estudo e a do sistema real.

Assim, sem desconhecer as possibilidades e vantagens de pesquisas com maiores simplificações, foi decidido aqui explorar um caso real. Essa decisão significou, simplificada, interpretar informação existente não integrada e não atualizada, e integrá-la com informação adquirida após projetar e executar o plano de monitoramento e com informação complementar, e integrar também o conhecimento transmitido pelos expertos. A informação assim integrada constitui tudo o disponível para estudar um setor com mais de 6000 ligações individuais, tubula-

ções novas operando em conjunto com outras de duas décadas, diferenças de cotas da ordem de 50 m, incidência de vazamentos, ocasionais falhas no abastecimento, sensores defeituosos e necessidade de reparos ou intervenções frequentes.

1.1 Objetivos

O objetivo geral da pesquisa de doutorado é desenvolver um método para detecção de vazamentos em tempo real para redes de distribuição de água para abastecimento, aplicável a redes urbanas existentes, usando os sinais de pressão em pontos do sistema a serem definidos, registrados mediante monitoramento permanente da rede. Para tanto, serão aplicadas ferramentas de processamento de sinais e de aprendizado de máquina não supervisionado, visando detecção de novidades nas séries, de forma a atingir a detecção de vazamentos na escala de um setor de rede. Os objetivos específicos são:

- Desenvolvimento de metodologias de aplicabilidade geral para definição do projeto de monitoramento e aplicação ao setor analisado como caso de estudo, definindo o número e localização de sensores.
- Construção de uma base de dados de sinais de pressão para o setor eleito para estudo, através do monitoramento, contendo períodos longos representativos das condições da rede durante operação normal e também da situação de início de vazamentos através de experimentos de simulação de vazamentos.
- Pré-processamento dos sinais de pressão adquiridos, e construção da informação supervisionada para avaliação.
- Aplicação das análises descritivas que permitam identificar a coerência dos sinais com os fenômenos hidráulicos que os originam, bem como caracterizá-los segundo várias abordagens potencialmente úteis para seu processamento.
- Desenvolvimento de métodos de detecção de vazamentos ou alterações baseados em mudanças no sinal que não correspondam com o comportamento normal.
- Avaliação dos métodos de detecção propostos, segundo a informação supervisionada construída.

1.2 Hipótese

A abordagem definida para esta pesquisa, visando o desenvolvimento de um método para detecção de vazamentos num setor de rede de distribuição de água para abastecimento usando os sinais de pressão adquiridos, baseia-se na hipótese seguinte: Um vazamento gera mudanças no estado de pressões da rede que são diferentes daquelas devidas ao consumo e operação, e que são medíveis com sensores de pressão. Portanto, é possível detectar em tempo quase real a

ocorrência de um vazamento através da análise dos sinais de pressão adquiridos no mesmo setor da rede.

1.3 Organização do documento

A estrutura deste documento visa facilitar a leitura, que pode ser feita para todos os capítulos em sequência ou somente para os de interesse. No Capítulo 1, que com esta seção termina, é apresentada a introdução, os objetivos e a hipótese central da pesquisa. Um panorama organizado da pesquisa recente sobre detecção de vazamentos, incluindo referências a trabalhos no tema e análise das vantagens e limitações de cada linha de pesquisa identificada, é apresentado no Capítulo 2. As diferentes etapas da pesquisa foram desenvolvidas sobre o estudo de caso que é apresentado no Capítulo 3. O projeto de monitoramento para otimização do número e localização dos sensores de pressão é apresentado no Capítulo 4, que inclui o desenvolvimento e explicação do método proposto, tanto quanto seus resultados quando aplicado ao estudo de caso. No Capítulo 5 são apresentados todos os métodos que são utilizados para analisar os sinais, tanto na etapa de descrição quanto nos três métodos de detecção propostos. Os resultados da aplicação desses métodos para as séries de pressão completas são apresentados e discutidos no Capítulo 6. No Capítulo 7, por fim, são apresentadas as conclusões e recomendações.

2 ESTADO DA ARTE EM DETECÇÃO DE VAZAMENTOS

Em resposta ao desafio da detecção e localização de vazamentos em redes de distribuição de água para abastecimento diversas abordagens são conhecidas, já um século atrás Babbitt, Amsbary e Gwinn (1920) reportavam as iniciativas de solução da época, cujas ideias fundamentais têm bastante semelhança com as das técnicas modernas, embora tanto o interesse pelo problema quanto a sofisticação das soluções tenha aumentado nos últimos tempos. Apesar da grande quantidade de trabalhos a respeito, não existe uma técnica consagrada, e para diferentes partes ou condições do problema as diferentes propostas apresentam vantagens e desvantagens.

Dentre as diferentes abordagens podem ser consideradas, por exemplo, a identificação de manifestações visíveis de vazamento (umidade, depressão no asfalto, etc), o uso de aparelhos de escuta que variam desde a haste de escuta até geofones, a injeção de gás, a fotografia no infravermelho e outras técnicas que estão baseadas na procura de indicadores da existência de um vazamento nas suas proximidades (COLOMBO; LEE; KARNEY, 2009). Todas as alternativas anteriores, aqui chamadas simplesmente de técnicas de inspeção, ficam limitadas à procura por vazamentos que estejam ativos durante a observação em locais próximos desta. A limitada abrangência espacial e temporal resulta na impossibilidade de fornecer informação sobre o estado de todo o sistema constantemente. Pelo anterior, e pela capacidade logística necessária para providências eficientes, as inspeções têm utilidade para localização de vazamentos mas não para detecção, e não constituem o foco deste trabalho.

Por outro lado, analisando-se a influência do vazamento no estado do sistema hidráulico onde se inserem, surge um grande e variado conjunto de técnicas para a detecção indireta dos vazamentos, utilizando as variáveis de estado (vazão e pressão) do sistema hidráulico e seu comportamento. Segundo classificação proposta em um trabalho anterior (GAMBOA-MEDINA, 2013) as técnicas podem ser agrupadas em duas classes principais: modelagem hidráulica (MH) e análise dos sinais (AS), sendo que o segundo grupo pode ainda considerar os sinais adquiridos durante um transitório hidráulico ou durante a operação normal do sistema.

Essa classificação é adotada por corresponder melhor aos aspectos considerados centrais na análise de métodos nesta pesquisa, mesmo que não seja a única possível. Por exemplo, Puust et al. (2010) agrupam como "técnicas de detecção (*awariness*)" métodos que nesta tese se referem como MH (principalmente) ou AS, e como "técnicas de localização" alguns de inspeção. Por outro lado, as duas classes (MH e AS) correspondem a métodos "baseados em *software*" segundo Li et al. (2015), enquanto os métodos de inspeção são "baseados em *hardware*" para os mesmos autores.

Os métodos para detecção de vazamentos e alterações desenvolvidos nesta pesquisa podem ser incluídos no grupo de técnicas baseadas na análise de sinais (AS) adquiridos durante

operação normal do sistema. A decisão de seguir essa abordagem está fundamentada no fato de que, apesar dos grandes avanços nas técnicas propostas, avalia-se que as técnicas baseadas em modelos (MH) enfrentam limitações para a implementação em sistemas de abastecimento reais que poderiam ser superadas se utilizada a segunda abordagem (AS), argumento sustentado na revisão do estado da arte.

Neste capítulo a revisão do estado da arte, segundo a literatura científica específica é brevemente apresentada, sem a pretensão de completar um estudo exaustivo das publicações sobre o tema, mas sim apresentar um panorama das diferentes linhas de pesquisa que têm sido desenvolvidas e oferecer referências de trabalhos nessas linhas.

2.1 Métodos baseados em modelos hidráulicos

O modelo hidráulico de uma rede de distribuição de água para abastecimento é uma representação simplificada dos elementos que a compõem (tubulações, reservatórios, bombas, etc, sua topologia e características), integrada às equações que definem o funcionamento do sistema hidráulico (conservação de massa nos nós e de energia em circuitos), de forma a possibilitar o cálculo das variáveis de estado (pressão nos nós e vazão nos trechos). Dentre as características que definem o modelo, ou entradas, estão as demandas nos nós, e os vazamentos devem também ser considerados como demandas adicionais e dependentes da pressão.

Comumente, algumas das características de entrada do modelo são desconhecidas, dentre elas a rugosidade das tubulações e a existência ou grandeza de vazamentos. Por outro lado, as variáveis de estado que podem ser consideradas saídas do modelo (pressão e vazão), podem ser obtidas por medição em alguns dos nós e trechos da rede. Assim, é possível dar tratamento ao problema inverso, em que são procurados os valores das entradas tais que a diferença entre as saídas calculadas via modelo e as obtidas por medição seja mínima, como proposto por Pudar e Liggett (1992). Essa ideia geral têm sido aplicada para a detecção de vazamentos por numerosos autores (LI et al., 2015). Dentre as propostas, a abordagem mais comum é o uso de modelos hidráulicos no regime permanente, utilizando para a otimização Algoritmos Genéticos, como apresentado por Soares (2003), Colombo (2007), Wu, Sage e Turtle (2010), dentre outros.

Um exemplo da aplicação da análise inverso (ou resolução do problema inverso) com modelos hidráulicos em regime permanente para a detecção e localização de vazamentos sem o uso de algoritmos genéticos é o desenvolvido por Meseguer et al. (2014). Outras propostas também baseadas em modelos no regime permanente têm sido apresentadas por Andersen e Powell (2000), Poulakis, Valougeorgis e Papadimitriou (2003), Sanz et al. (2016), Yoon et al. (2012). As propostas mencionadas foram aplicadas a redes hipotéticas, com exceção das de Colombo (2007), Wu, Sage e Turtle (2010) e Meseguer et al. (2014) que apresentam resultados para setores de redes reais.

A utilização de modelos em regime transitório apresenta desafios adicionais dado que

a resolução do modelo para essas condições ainda é objeto de pesquisa e não há uma única metodologia de cálculo estabelecida. Por outro lado, têm como vantagem a maior quantidade de informações que podem ser analisadas durante a ocorrência do fenômeno transitório. Mesmo assim, a detecção de vazamentos com esse tipo de modelos, utilizando análise inverso em regime de escoamento transitório (LIGGETT; CHEN, 1994) é considerada em conjunto com otimização via algoritmos genéticos (AG) por Vitkovsky, Simpson e Lambert (2000), Saldarriaga, Fuentes e Galvis (2006), Covas e Ramos (2010), e com AG e também método de Levenberg-Marquardt por Kapelan, Savic e Walters (2003) e por Soares, Covas e Reis (2011). A aplicação deste tipo de metodologia é, segundo reportado, limitada na maioria dos casos a redes teóricas (por exemplo, uma mesma rede hipotética é usada em três dos artigos antes mencionados) ou experimentos em laboratório, e a única referência a testes em sistemas reais é feita por Saldarriaga, Fuentes e Galvis (2006). Segundo Puust et al. (2010), a importância da precisão do modelo hidráulico constitui uma limitação para a aplicação efetiva do método de transitório inverso a redes reais, e os seus resultados não podem ser apresentados sem uma quantificação da incerteza envolvida.

Dentre as vantagens principais dos métodos baseados em modelos hidráulicos (MH), em regime permanente ou transitório, estão a possibilidade de obter informação detalhada sobre o fenômeno e suas consequências em toda a extensão da rede, assim como permitir a análise igualmente detalhada de cenários criados, o que significa que se trata de uma ferramenta muito útil para o planejamento e a tomada de decisões. Dentre as limitações que independem das particularidades de cada método pode-se apontar a necessidade de uma série de condições difíceis de completar para uma rede de abastecimento real: alto detalhamento na modelação e exatidão na simulação hidráulica, incluindo exatidão nos parâmetros e condições de contorno.

Mesmo que não pareçam condições demasiado complexas, a realidade de muitos dos sistemas urbanos de distribuição de água para abastecimento faz com que sejam muito difíceis de cumprir. Por exemplo, a informação sobre os elementos efetivamente instalados, diâmetros, cotas e condição de funcionamento dos registros é o mínimo necessário para que o modelo consiga representar a realidade, mas é usual que durante a construção e manutenção sejam feitas mudanças no projeto original ou na condição de operação que não são registradas e, portanto, não podem ser incorporadas no modelo. As deficiências na medição do consumo individual dos usuários e sua incorporação no modelo são outras das dificuldades mais comuns. Assim, na maioria de modelos de redes reais há incertezas que não podem ser resolvidas pelo analista, e que podem comprometer a detecção de vazamentos mesmo se o modelo é considerado aceitável para outros propósitos.

Adicionalmente, a rede de abastecimento é um sistema em constante transformação, onde mudanças tanto dos elementos físicos da rede quanto do comportamento de usuários ou critérios de operação, geram alterações. Portanto, a calibração do modelo hidráulico, para que este seja exato até alto nível de detalhe, deve ser refeita regularmente, o que não costuma ser fácil para as companhias que gerenciam as rede.

Não deve ser desconsiderada a relevância de que todos os sistemas de abastecimento tenham o melhor modelo hidráulico possível, nem o potencial e a importância dos desenvolvimentos usando modelos hidráulicos para diversos objetivos do projeto e gerenciamento das redes. Porém, julga-se que a dependência do modelo é um empecilho para a aplicação dos métodos de detecção de vazamento, e que métodos que superem essas limitações podem se aproximar mais da aplicação em sistemas reais.

2.2 Métodos baseados em análise de sinais

As técnicas de detecção de vazamentos baseadas na análise de sinais (AS), sem deixar de reconhecer os princípios hidráulicos que regem o comportamento das redes de distribuição de água, não dependem da sua modelagem, mas do estudo dos valores medidos para as variáveis do sistema (pressão e/ou vazão) ao longo do tempo. Esse estudo dos sinais na procura de indicadores de vazamento comumente implica construir algum tipo de padrão ou modelo dos dados, para comparar essa construção com os sinais adquiridos, e a partir da comparação é analisada a existência ou ocorrência de vazamentos ou outras alterações do sistema. Com isso, todas as técnicas que seguem esta abordagem evitam a utilização dos modelos hidráulicos e as limitações relativas a eles.

Identificam-se ainda dois subgrupos destes métodos: aqueles baseados na análise de sinais do transitório hidráulico e os baseados na análise de sinais durante operação normal da rede. Mesmo baseados na mesma ideia inicial, os tipos de métodos têm grandes diferenças na sua fundamentação e, principalmente, na aplicação. Os dois vão ser brevemente comentados, mencionando em cada caso algumas das propostas desenvolvidas segundo a literatura da área.

2.2.1 Análise no transitório hidráulico

A ideia básica que resume os métodos de detecção de vazamentos baseados na análise de sinais adquiridos durante um transitório hidráulico, é que a propagação das ondas de pressão próprias do transitório são afetadas pela presença de um vazamento, mesmo a distâncias consideráveis deste. Assim, durante um transitório um único ponto de monitoramento permitiria captar um sinal de pressão em função do tempo, e uma vez que, durante a propagação, a onda encontra uma seção com vazamento pode acontecer amortecimento e/ou reflexão, que alteram o sinal medido. A onda de pressão analisada pode ser gerada por uma manobra controlada no sistema, como fechamento ou abertura rápida de uma válvula, mas também o próprio surgimento de um vazamento gera um transitório hidráulico e ondas que se deslocam pelo sistema. No primeiro caso, a busca é por um vazamento que já existe quando o transitório começa, enquanto no segundo a detecção corresponde ao início do vazamento.

Além da natureza do transitório analisado, as diferenças entre os métodos estão no tipo de análise realizado nos sinais. Transitórios controlados gerados para teste, e analisados com a transformada *wavelet* são propostos por Ferrante e Brunone (2003b), Ferrante, Brunone e

Meniconi (2007), Ferrante, Brunone e Meniconi (2009), Meniconi et al. (2011), Srirangarajan et al. (2012). A resposta em frequência quando gerado fluxo oscilatório é analisada por Mpesha, Gassman e Chaudhry (2001), Covas, Ramos e Almeida (2005), enquanto a análise harmônica é proposta por Wang et al. (2002), Ferrante e Brunone (2003a). Dentre as análises do transitório gerado pelo início do vazamento ou ruptura podem ser citados Misiunas et al. (2005).

Segundo Colombo, Lee e Karney (2009), evidencia-se falta de aplicações em campo e verificações das técnicas de detecção baseadas na análise do transitório hidráulico, mesmo que não sejam totalmente inexistentes. Entretanto, Ghazali et al. (2012) afirma que estas técnicas têm sido satisfatoriamente testadas em condições de laboratório mas não estão estabelecidas para uso com dados reais. Essas afirmações são verificadas com as técnicas mencionadas, porquanto são apresentadas somente para sistemas hipotéticos ou, em poucos casos, de laboratório. Estas técnicas consideram unicamente sistemas abertos, na maioria dos casos consistindo de uma única tubulação e ocasionalmente sistemas com uma bifurcação. Uma exceção dentre as técnicas baseadas em processamento de sinais durante regime transitório é o estudo de Ghazali et al. (2012) que compara várias técnicas de análise de frequência instantânea utilizando dados de transitórios controlados numa tubulação de uma rede real.

Como visto, as técnicas são desenvolvidas para condições muito controladas, de forma que a alteração na propagação da onda devida ao vazamento não seja coberta por outras alterações. Essa situação, porém, dista muito dos sistemas de distribuição reais, para os quais se espera poder implementar uma solução. Além dos efeitos da complexidade da rede, há outras limitações na aplicação deste tipo de técnicas a sistemas reais, comentadas a seguir.

Inicialmente, as técnicas baseadas na análise sob transitórios conhecidos e controlados não oferecem a possibilidade de monitoramento contínuo, mas da realização de experimentos. Esses experimentos, adicionalmente, podem constituir um risco para a rede pelas pressões extremas geradas, principalmente se está fragilizada por motivos como idade das tubulações. Por outro lado, a resolução temporal necessária para análise do transitório hidráulico corresponde com frequências da ordem de kHz (10^3 valores por segundo), mas uma aquisição permanente e em diversos pontos da rede, além de transmissão e processamento das informações, é inviável com essa frequência. Uma limitação adicional, menos crítica, é a impossibilidade de utilizar medidas de vazão, pois o tempo de resposta dos sensores é incompatível com as frequências de amostragem necessárias para acompanhar o transitório (PUUST et al., 2010).

2.2.2 Análise durante operação do sistema

Pesquisas recentes visam à detecção de vazamentos pela análise dos sinais adquiridos durante a operação normal, na procura de métodos aplicáveis a sistemas reais. A ideia é aproveitar as informações de monitoramento permanente e de baixo custo, sem impor manobras específicas sobre a rede (como para a geração de transitórios, que podem ainda prejudicar estruturalmente o sistema), e sem depender do conhecimento detalhado necessário para a modelagem hidráulica de

elevada precisão. Os sinais utilizados são os de vazão e/ou de pressão em pontos da rede, e as técnicas de análise variam muito, incluindo análises estatísticas e de probabilidades, processamento digital de sinais e inteligência artificial.

Diferentemente de outras abordagens, o objetivo da análise é identificar indicadores do início de um vazamento, ou de sua persistência, mas não da sua presença desde antes do início da análise. Assim, estes métodos em geral constroem uma representação, baseada nos dados adquiridos, do comportamento base do sistema, e com ele comparam todos os novos dados que ingressam para análise, em busca de uma alteração.

Vários trabalhos que analisam os sinais de vazões de entrada em um setor de rede de abastecimento podem ser citados. A vazão no período de baixa demanda, analisada como variável aleatória, é a base da proposta de Buchberger e Nadimpalli (2004), na qual análises estatísticas são aplicadas para avaliar se a distribuição de probabilidade é estacionária ou se novos valores não correspondem à distribuição esperada e são, portanto, indícios de vazamento. O método é validado para setores residenciais pequenos, onde podem ser descartados consumos noturnos importantes e variáveis.

Usando ferramentas mais complexas, Mounce, Boxall e Machell (2010) propuseram um método para detecção em tempo real, segundo o qual valores históricos de vazão são usados para treinamento de uma rede neural artificial (RNA) e, com aplicação de lógica difusa, são avaliadas as discrepâncias entre as vazões observadas e as previstas pela rede neural para a detecção.

O método apresentado por Aksela, Aksela e Vahala (2009) utiliza vazões obtidas em situações com e sem vazamentos, e está baseado nos mapas auto-organizados (SOM, pela sigla em inglês), que correspondem a um tipo de RNA treinada usando aprendizado não supervisionado para o agrupamento de um conjunto de entradas. Depois do processo de treinamento, os “mapas” criados permitem identificar se uma sequência de vazões avaliadas corresponde ou não com um vazamento.

A análise de componentes principais (PCA, do inglês), técnica muito comum para análise linear de séries temporais, é aplicada por Palau, Arregui e Carlos (2012) para séries de vazões num setor de rede. Segundo o método, as informações são utilizadas na construção de um modelo estatístico que representa o comportamento da rede para, posteriormente, fazer uso de técnicas de controle estatístico para a identificação de anomalias.

Os consumos horários na escala de uma residência ou de um setor da rede constituem as séries analisadas por Christodoulou, Kourti e Agathokleous (2016), que aplicam inicialmente controle estatístico para definir as regiões do sinal (trechos) que podem conter alterações. Nos trechos selecionados, é calculada a transformada *wavelet* para, no domínio transformado, identificar as descontinuidades ou alterações que permitam a detecção.

Também é a vazão a variável utilizada por Malik, Ghosh e Rogers (2015) para a construção de um modelo como processo gaussiano. O modelo é construído com base na periodicidade

da série, supondo que o comportamento de um momento da semana pode ser descrito pelos valores históricos no mesmo momento, tanto quanto à tendência quanto à variação. A comparação da predição feita pelo modelo com os valores adquiridos é analisada para concluir se há indícios de ocorrência de vazamento, e a sua grandeza estimada.

Para a incorporação da análise de sinais de pressão juntamente com os de vazão para a detecção de vazamentos, foi proposto o uso de filtros Kalman (YE; PH; FENNER, 2011). O uso dos filtros pode ser interpretado, simplificada, como um método estatístico que utiliza segmentos conhecidos de séries temporais para produzir estimativas do estado do sistema, no caso as pressões e vazões, para depois avaliar as diferenças entre valores estimados e medidos, e interpretá-los para detecção.

Usando também o monitoramento de pressões e vazões na rede durante operação, Romano, Kapelan e Savic (2010) desenvolveram um método baseado em inferência bayesiana. Segundo ele, valores de pressão ou vazão observados são utilizados para criar um modelo de previsão de curto prazo, com o qual os valores para os seguintes períodos são estimados. Essas estimativas por sua vez são comparadas com os valores adquiridos para identificar discrepâncias, e depois essas informações são entregues ao sistema de inferência bayesiana para a geração de alarmes.

A mesma equipe de pesquisa introduziu melhorias na proposta (ROMANO; KAPELAN; SAVIC, 2014a; ROMANO; KAPELAN; SAVIC, 2012; ROMANO; KAPELAN; SAVIC, 2014b), desenvolvendo um sistema modular que analisa as discrepâncias usando RNA, juntamente com avaliações estatísticas de desvios e de tendência, para produzir as entradas ao sistema de inferência e geração de alarmes.

Uma técnica integrando análise de sinais de vazão adquiridos na entrada da rede com comparação entre valores de pressão modelados e observados foi proposta por Misiunas et al. (2006). Segundo a técnica proposta, as variações na vazão de entrada são avaliadas analisando os gráficos de soma cumulativa (CUSUM, pelo nome em inglês) visando a detecção do vazamento, e posteriormente um modelo hidráulico é usado para simular possíveis locais do vazamento e comparar os resultados de pressão com os valores reais adquiridos, para aproximar a localização do vazamento. O método é testado unicamente para uma rede hipotética.

A aplicação a sistemas de abastecimento reais é um objetivo importante para os métodos baseados na análise de sinais durante operação, e esse foco é evidenciado no tipo de informações utilizadas para projeto dos métodos comentados: séries de vazão ou de pressão em alguns pontos de um setor de rede de distribuição, e adquiridos com períodos de amostragem desde alguns minutos até uma hora, são suficientes para executar os métodos. Porém, em muitos dos trabalhos citados foram utilizados valores sintéticos, majoritariamente gerados a partir de suposições estacionárias adicionadas com algum ruído estocástico. Além disso, os vazamentos ou outras alterações utilizadas como exemplo são, em não poucos casos, de tal ordem de grandeza que os sinais sofrem modificações evidentes por inspeção na representação gráfica dos sinais.

Apesar do claro foco dos métodos na aplicação e do detalhamento de alguns deles, são escassos os reportes de aplicação a sistemas reais. O único caso de implementação documentado encontrado nesta revisão é o de Romano, Kapelan e Savic (2012).

Uma característica fundamental para aplicabilidade dos métodos é sua adequação ao processamento em tempo (quase) real, mas pouca discussão sobre essa capacidade é feita nos trabalhos citados. Também não há em nenhum deles avaliações ou comentários sobre os requisitos computacionais, outro aspecto importante para que os métodos sejam aplicáveis.

Apesar do curto tempo desde sua aparição, esta abordagem de análise de sinais adquiridos por monitoramento durante operação do sistema para a detecção de vazamentos tem recebido crescente interesse, e nos últimos anos vários grupos de pesquisa no mundo passaram a elaborar novas propostas, sendo que as possibilidades parecem se expandir com cada trabalho. A aplicação de técnicas de aprendizado de máquina, dentre outras áreas, tem apresentado resultados positivos e pode-se supor que aprofundar nessa linha traga progresso. Os resultados reportados somente sobre dados sintéticos, porém, devem ser avaliados com cautela, pois juntamente com a não aplicação a sistemas reais indicam que ainda há aspectos a serem melhorados nos métodos.

3 APRESENTAÇÃO DO ESTUDO DE CASO

Em colaboração com o Departamento Autônomo de Água e Esgoto de cidade de Araraquara (DAAE), no âmbito de um convenio acadêmico na área de Hidráulica e Saneamento celebrado com a Universidade de São Paulo, foi desenvolvido um estudo de caso sobre um setor de rede de distribuição de água de abastecimento, monitorado durante doze meses, dessa cidade. Neste estudo os métodos e análises que serão descritos no Capítulo 5, foram empregados.

3.1 Características gerais do setor

O município de Araraquara está localizado na região central do estado de São Paulo, a 43 km da cidade de São Carlos, sede da universidade onde foi desenvolvidas a pesquisa. Tem população de aproximadamente 212 mil habitantes, e é abastecida por um sistema que explora mananciais superficiais e subterrâneos. Em colaboração com a equipe técnica do DAAE foi selecionado para o estudo um setor isolado sob suspeita de vazamentos, denominado Martinez e localizado região central do município.

Uma área de aproximadamente 2,75 km² é abastecida pelo setor, com mais de 6000 ligações, das quais aproximadamente 90% são residenciais. A água de um único poço é impulsionada por uma bomba vertical tipo turbina através de uma tubulação de 250 mm por aproximadamente 850 m, até os reservatórios. O setor é abastecido por dois reservatórios apoiados interligados, com capacidade combinada de 2800 m³, cota de fundo 679,75 m e cota do nível d'água máximo de 683,95 m. As diferenças de cota entre nós do setor superam os 50 m.

A rede de distribuição conta com aproximadamente 57 km de tubulação, principalmente de PVC, de diâmetros desde 50 mm (72%) até 250 mm (7%), com alguns trechos de diâmetros maiores. Os trechos mais antigos da rede, com mais de 20 anos, são de ferro fundido, e em recentes substituições foram comprovadas incrustações que reduzem a área útil das tubulações. Vazamentos são reparados pelo DAAE no setor mais de uma vez por semana, em media. Os altos valores de demanda mínima noturna medidos, inesperados num setor principalmente residencial, são um indício da existência de vazamentos, mas essa hipótese não tem sido esclarecida com os resultados das campanhas de inspeção para detecção e reparo de vazamentos realizadas antes do período de estudo. Sobre esse particular, tem que ser levado em conta o problema da medição de vazão descrito na próxima seção.

3.2 Informação existente e adquirida

Informação sobre a topologia da rede, diâmetros e materiais das tubulações, disponibilizada para a pesquisa pelo DAAE-Araraquara, tinha poucas atualizações em relação aquela processada para o mesmo setor em trabalhos anteriores ((ABE, 2014) e (ODAN, 2013)). A

informação sobre consumos mensais para as ligações que integram o setor foi também recebida nas fases iniciais da pesquisa. As informações do sistema durante o período de monitoramento fornecidas pelo DAAE são os dados da telemetria e os reportes de reparos, que complementam as pressões adquiridas em pontos internos da rede segundo o plano de monitoramento desenvolvido.

No presente documento, os termos "sinais" e "séries temporais" (ou simplesmente "séries") serão usados como sinônimos, sempre com referência às sequências de dados ou segmentos delas, seja dos valores diretamente adquiridos, seja após algum processamento.

3.2.1 Aquisição rotineira pelo sistema de telemetria

O sistema de controle, supervisão e aquisição de dados do DAAE-Araraquara conta com telemetria do nível nos reservatórios, e da vazão de distribuição (saída dos reservatórios). Os valores adquiridos por este sistema, com taxa de amostragem de dois minutos, foram utilizados na construção do modelo de simulação utilizado para o projeto de monitoramento.

Durante a aquisição de dados de vazão de saída dos reservatórios, variável de interesse nos estudos de detecção de alterações, percebeu-se que o sensor instalado tinha erros de medição, como evidenciado na Figura 1, em que é apresentada uma visão geral do sinal de vazão. O equipamento foi trocado (dia 244*), levando inicialmente a mudança da vazão média de 51 para 46 L/s, e do desvio padrão de 6,39 para 23 L/s. Após essa data ocorreram ainda múltiplas alterações na equação de calibração do equipamento, refletidas em mudanças de média e desvio padrão dos valores adquiridos. Como resultado do anterior, não há uma continuidade nas medições, nem uma referência confiável para aplicar processos de reconstrução da serie, o que impossibilita o emprego da vazão nos métodos de detecção. Assim, sua aplicação nesta pesquisa foi limitada à construção do modelo hidráulico. Quanto aos valores de nível nos reservatórios, as séries são

Figura 1 – Vazão na saída do reservatório, informação adquirida pelo sistema de telemetria.

durante a maior parte do tempo estacionárias, mantendo a média em 2,7 m e o desvio padrão

* Em todo o documento a escala de tempo está em dias, com origem no primeiro dia da aquisição.

em $\pm 0,4$ m, com uma alteração principal entre os dias 234 e 243 (média 2,1 e desvio padrão $\pm 0,2$) que pode estar relacionada a mudanças na operação para a manutenção dos reservatórios. Uma visão geral da série é apresentada na Figura 2. Algumas irregularidades de menor duração coincidem com situações especiais reportadas ou de origem não conhecida.

Figura 2 – Nível nos reservatórios, informação adquirida pelo sistema de telemetria.

3.2.2 Condição do sistema

Além dos dados adquiridos pelos sensores, informações sobre atividades na rede foram fornecidas pelo DAAE atendendo às solicitações da equipe da EESC-USP. Assim, contou-se com as listagens dos reparos realizados na rede em atenção a reportes de vazamento dos usuários ou dos próprios funcionários do DAAE. Adicionalmente, situações especiais (problemas na bomba, manutenção dos reservatórios, dentre outros) durante o período de monitoramento foram consideradas.

Mesmo com essas informações adicionais, não há certeza sobre o número total de vazamentos, reparos, intervenções ou mudanças na rede que aconteceram no período. Os registros de reparos são elaborados como parte dos processos internos do DAAE, e portanto os dados contidos nestes são importantes, mas insuficientes para uma descrição detalhada dos eventos.

3.2.3 Informação de pressão adquirida

Com apoio do DAAE-Araraquara e segundo os resultados do projeto de monitoramento implantado (Capítulo 4), foram construídos no setor de estudo nove poços de inspeção da rede de abastecimento, e em cada um foram feitas as adaptações para a instalação de sensores de pressão. Foram utilizados registradores de dados com transdutores de pressão integrados, fabricados pela TECHNOLOG, do modelo METROLOG-P, com faixa de medição até 100 mca (com exceção do sensor 9, que é de 200 m), sendo 8 do patrimônio da USP e um do DAAE. Valores instantâneos de pressão em cada ponto monitorado da rede foram adquiridos a cada 10

segundos, com resolução de 0,1 m (com exceção do sensor 9, com resolução de 1 m), salvando um único valor médio a cada 2 minutos, de forma que a memória do registrador se esgotava a cada 9 dias, aproximadamente.

Assim, a base de dados obtida corresponde a nove séries de valores de pressão em metros com resolução temporal de 1 valor a cada 2 minutos, adquiridos durante 357 dias. Assim, o número total de elementos em cada série é de até 257040, porém, o total de informações disponíveis foi afetado por problemas que se apresentaram durante o monitoramento. Representados pela presença de valores atípicos (*outliers*).

Na base de dados estão organizadas as informações para o período de monitoramento, com uma referência temporal em dias iniciando o primeiro dia da aquisição. Os sinais adquiridos completos têm a aparência geral apresentada na Figura 3, onde não se pretende mostrar nenhum detalhe mas somente o conjunto completo. Cada uma das nove linhas na figura corresponde a um dos sensores e as diferenças entre suas médias representam as diferenças de nível topográfico dos pontos de instalação. Uma descrição completa das series adquiridas é feita no capítulo 6.

Figura 3 – Visão geral das séries de pressão obtidas durante todo o período de monitoramento

Adicionalmente aos dados adquiridos durante a operação normal da rede, um total de doze simulações de vazamentos mediante abertura controlada de válvulas de descarga e hidrantes na rede foram realizadas. Sua localização pode ser conferida no esquema do modelo em **EPANET**[†] (Figura 4). Na tabela Tabela 1 apresenta-se um resumo das características destas

[†] EPANET é um software para modelar redes de abastecimento de água, de domínio público, desenvolvido originalmente na USEPA (Agência de proteção ambiental dos Estados Unidos) por Rossman (2000), composto por uma biblioteca de funções em linguagem C e uma interface gráfica ligada. A última versão suportada pelos desenvolvedores originais é a 2.0.12 de 2008. Numerosos usuários profissionais e pesquisadores têm desenvolvido complementos ou extensões ao programa, mas é destacada a iniciativa da comunidade informalmente agrupada sob o nome *Open Water Analytics* por atualizar e expandir o código. Como resultado, foi lançada em julho de 2016 a versão 2.1 da biblioteca, que corrige erros e incorpora melhoras (OPEN WATER ANALYTICS, 2016). A biblioteca

simulações.

Figura 4 – Esquema do modelo em EPANET do setor sob estudo

Tabela 1 – Características dos experimentos de simulação de vazamentos

Experimento	Data(dias)	Vazão (L/s)	Local
1	98,4944	5,3	d6
2	119,400	5,0	d6
3	126,4146	10,4	d16
4	148,3222	5,0	d11
5	235,1736	4,3	d6
6	237,1785	3,6	d8
7	239,1708	5,4	d11
7	242,2014	11,5	d16
9	258,3840	9,0	d16
10	300,9167	28	h5
11	343,9167	13	h3
12	354,6465	21	h8

pode ser integrada a desenvolvimentos em diferentes linguagens de programação, e existem módulos específicos para facilitar tal integração; nesta pesquisa foi utilizada com Matlab e com Fortran.

3.3 Modelo hidráulico de simulação da rede

Dentre as motivações para pesquisar a detecção de novidades sobre dos sinais adquiridos pelo monitoramento do sistema está evitar as limitações dos métodos baseados em modelos hidráulicos Capítulo 2. Assim, os métodos de detecção propostos independem das respostas do modelo do setor em estudo. Porém, a simulação hidráulica via modelo foi utilizada para a definição do projeto de monitoramento.

O setor Martinez foi considerado nos estudos de Odan (2013) e Abe (2014), e o modelo hidráulico do setor desenvolvido e já calibrado pelos dois autores foi tomado, com sua autorização, como base para a modelação no presente trabalho. A topologia da rede de distribuição e as rugosidades calibradas foram mantidas, enquanto os consumos, condições de contorno e outros parâmetros da simulação foram atualizados. Um esquema do modelo hidráulico, constituído por 879 nós e 1059 trechos de tubulação, e construído no software EPANET. Os nove pontos de instalação dos sensores de pressão e os sete dos experimentos de simulação de vazamentos estão marcados na mesma figura.

No projeto de monitoramento é utilizado o modelo hidráulico de simulação em período estendido para 24 horas, com intervalo de 1 hora. Na construção de modelo hidráulico utilizado na etapa inicial da pesquisa, para o projeto de monitoramento, foram utilizados os valores obtidos pela telemetria nos meses anteriores ao monitoramento para construir uma curva do nível médio dos reservatórios para o dia útil típico. A curva de níveis foi utilizada com condição de contorno na simulação perante a insuficiente informação para simular o abastecimento desde o poço. Também dos dados de telemetria foi obtida a curva de consumo diário médio para os dias úteis da semana, que constitui o padrão diário de oscilação do consumo para todos os nós. O valor base de consumo em cada nó foi atualizado a partir dos consumos mensais medidos (hidrometrados) dos usuários do setor. Esse modelo base foi iterativamente alterado pela inclusão em cada um dos nós de vazamentos como consumos dirigidos pela pressão (*emitters*), conforme apresentado no Capítulo 4.

4 PROJETO DE MONITORAMENTO

Segundo Bush e Uber (1998) a utilidade e confiabilidade dos dados de pressão para conhecer o estado da rede de distribuição de água e, particularmente, para detectar vazamentos e outras alterações, depende em grande parte dos pontos onde esses dados são medidos. Assim, a decisão sobre o monitoramento para o estudo de caso, isto é, sobre número e localização dos sensores de pressão a serem instalados, esteve baseada em uma análise completa, apresentada integralmente neste capítulo. Para uma explicação resumida do método para projeto de monitoramento pode ser consultado o artigo em que a proposta foi inicialmente publicada (GAMBOA-MEDINA; REIS, 2017).

4.1 Método proposto

O método para o projeto de monitoramento desenvolvido busca pelos locais ótimos para instalação de sensores de pressão e seu número, visando à detecção de vazamentos. Segundo o método aqui proposto, a busca é orientada por quatro critérios: i) maximização da sensibilidade* total, ii) maximização da consistência da sensibilidade, iii) minimização da informação redundante e iv) minimização do número de sensores. Os três primeiros critérios são avaliados de maneira simultânea como um problema de otimização multiobjetivo, enquanto o quarto é avaliado independentemente.

Para a busca do conjunto de nós para localização de sensores, o critério de sensibilidade é analisado com base nas respostas do modelo de simulação à incorporação de vazamentos (Equação 4.5). A consistência da sensibilidade perante todos os potenciais casos de vazamento considerados é avaliada através da função de entropia (Equação 4.7). A redundância, por sua vez, é avaliada pela correlação entre as pressões nos nós de possível localização dos sensores segundo as simulações (Equação 4.2). As informações necessárias para quantificar esses três critérios para qualquer conjunto de nós estão contidas nas matrizes de sensibilidade e de correlação, ambas construídas com os resultados do modelo numérico de simulação. Após a obtenção das matrizes, a otimização é feita utilizando o algoritmo genético multiobjetivo NSGA-II para essas três funções objetivo. As análises necessárias para todas as etapas mencionadas, e a referente à quantidade de sensores, são detalhadas nas seções a seguir.

4.1.1 Representação dos vazamentos no modelo

Como apresentado na seção 3.3, a simulação de vazamentos é feita pela incorporação de uma demanda adicional dependente da pressão, que segue a equação 4.1, em cada um dos nós

* A afetação na pressão de um nó devido à presença de um vazamento é quantificada como sua sensibilidade. Ver seção 4.1.3

da rede (um de cada vez).

$$V = cP^\alpha \quad (4.1)$$

Em estudos experimentais sobre vazamentos, Greyvenstein e Zyl (2007) observaram que o valor do expoente α variava segundo o tipo de material e as características da fissuras na tubulação. Para tubulações de PVC obtiveram valores para o expoente no intervalo entre 0,40 e 1,85. Na presente pesquisa foi utilizado o expoente $\alpha = 1,18$, valor amplamente aplicado em modelos hidráulicos.

O coeficiente de vazamento c depende, além do tipo de material da tubulação, de outras condições do sistema. Segundo Germanopoulos (1985) este coeficiente depende da rede de distribuição (material, idade, entre outros), e no seu estudo refere-se a perdas da ordem de 10% a 15% da vazão distribuída para um valor de $c = 1 \times 10^{-7}$. Salienta-se também que esse coeficiente não é adimensional, e que o valor referido foi usado na equação 4.1 juntamente com o comprimento da tubulação em metros, a pressão em metros coluna de água e a vazão em megalitros por dia. Ao aplicar o mesmo coeficiente para vazão em litros por segundo, e considerando um comprimento médio das tubulações de 100 m, o valor de coeficiente resultante seria próximo a 3×10^{-4} . Porém, valores diferentes têm sido utilizados por outros autores.

Devido a dependência do coeficiente das características individuais de cada sistema, foi feito estudo da influência dos valores para a rede analisada no estudo de caso. Inicialmente foram extraídos das simulações com o modelo os valores de vazão de saída dos nós para todos os cenários de vazamentos. Todas as simulações foram realizadas para valores de $c = 0,005, 0,01, 0,02, 0,05$ e $0,10$, e para cada valor foram construídas duas matrizes: a matriz de vazão de vazamento e a matriz de vazão total, ambas de tamanho 24×879 (24 horas no dia, 879 nós do modelo).

Na matriz de vazão de vazamento (V) nos nós a cada hora do dia, é considerada só a parcela de consumo no nó devida ao vazamento. Entretanto, a matriz de vazão total de saída (Q) que corresponde ao somatório do consumo em todos os nós acrescentado pelo vazamento para cada cenário de vazamento a cada hora do dia. Com a média diária de cada nó para todas as vazões, e sua média entre todos os nós, pode-se calcular a razão entre a vazão de vazamento e a vazão total, V/Q , cujos resultados são os apresentados na Tabela 2.

No estudo de caso apresentado por Misiunas et al. (2005), a magnitude dos vazamentos analisados foi de 6% até 25% da vazão distribuída. A razão V/Q deve ter valores maiores a 4,8% para que eventos de vazamentos (hipotéticos) sejam detectados segundo reportado por Romano, Kapelan e Savic (2014b). Para a pesquisa descrita neste documento, foi utilizado o valor do coeficiente $c = 0,05$, que produz vazamentos que correspondem em média a 4,3% da vazão distribuída.

Tabela 2 – Razão de vazamento e vazão total, V/Q, segundo o valor do coeficiente c

c	V (L/s)	Q (L/s)	V/Q
0,1	4,683	65,14	7,19%
0,05	2,741	63,19	4,34%
0,02	1,227	61,68	1,99%
0,01	0,637	61,09	1,04%
0,005	0,324	60,78	0,53%
0,0 (sem)	0,000	65,14	0

4.1.2 Análise de correlação

Devido às características físicas da rede de distribuição de água, alguns nós apresentam comportamentos similares com relação as pressões e vazamentos. Assim, se um nó apresenta um valor alto de sensibilidade (ver rodapé e seção 1.4.3), aqueles nós similares também apresentarão um valor alto. Dita similitude entre nós pode gerar uma condição de redundância, que pode ser evitada mediante avaliação da correlação entre as séries de pressão nos possíveis nós de localização de sensores.

Sucessivas simulações do modelo são realizadas para produzir os vetores de pressão e vazão para o cenário sem vazamentos, e para produzir as matrizes de pressão e vazão para os cenários com vazamento em um nó. Estas são realizadas para cada hora do dia. O tamanho dos vetores e o número de linhas das matrizes correspondem aos nós onde poderiam ser instalados os sensores, enquanto o número de colunas nas matrizes corresponde aos nós onde um vazamento é avaliado. Neste estudo, todos os nós da rede pertencem aos dois tipos, instalação de sensores e ocorrência de vazamentos.

A matriz de pressões na presença de vazamentos, PL , com elementos $PL_{x,k}$ e média $\overline{PL}_x$, é utilizada para calcular os elementos da matriz de correlação, $CH_{a,b}^h$, segundo a Equação 4.2, que avalia a correlação entre as pressões para cada dupla de nós (a, b), com $a, b \in 1, 2, \dots, N$. O símbolo h faz referencia à hora do dia, e enquanto a simulação é repetida para as 24 horas do dia os valores de $CH_{a,b}^h$ são acumulados para obter a matriz de correlação total C . A matriz resultante é utilizada na terceira função objetivo da otimização.

$$CH_{a,b}^h = \frac{\sum_{l=1}^N (PL_{a,l} - \overline{PL}_a) (PL_{b,l} - \overline{PL}_b)}{\sqrt{\sum_{l=1}^N (PL_{a,l} - \overline{PL}_a)^2} \sqrt{\sum_{l=1}^N (PL_{b,l} - \overline{PL}_b)^2}} ; C_{a,b} = \sum_{h=1}^{24} CH_{a,b}^h \quad (4.2)$$

4.1.3 Análise de sensibilidade

Um nó da rede tem maior sensibilidade quanto maior for a variação da sua pressão como consequência de um vazamento, e é, portanto, mais interessante para o monitoramento. Se, ao contrário, a pressão no nó permanece quase inalterada face à ocorrência de um vazamento, o nó tem baixa sensibilidade e não será útil fazer medições nesse ponto visando detecção. Para quantificar essa sensibilidade têm sido utilizadas várias expressões, a maioria delas focadas no problema de calibração do modelo de rede. Isto é, analisam a sensibilidade na pressão ou na vazão perante mudanças na rugosidade das tubulações. Pela maior proximidade com o objeto da pesquisa são analisadas as propostas de Bush e Uber (1998), e especificamente as expressões para sensibilidade aos vazamentos de Farley, Mounce e Boxall (2010) e Misiunas et al. (2005).

Farley, Mounce e Boxall (2010) define a sensibilidade como o quadrado da variação da pressão num ponto devido a um vazamento, dividida pela pressão em ausência de vazamentos. Essa definição é apresentada na equação Equação 4.3, onde $S_{n,l}$ é a sensibilidade do nó n a um vazamento no nó l , $PL_{n,l}$ a pressão no nó n quando simulado um vazamento no nó l , e PN_n a pressão no nó n na ausência de vazamentos.

$$S_{n,l} = \frac{(PL_{n,l} - PN_n)^2}{PN_n} \quad (4.3)$$

Nas análises preliminares foi utilizada uma rede hipotética simples de 79 nós, dois reservatórios, duas bombas e um tanque, e uma versão preliminar do processo de otimização. Nos resultados para varias repetições da otimização percebe-se que diferentes respostas (conjuntos de nós para localização dos sensores) atingem o valor ótimo das funções objetivo, ainda que sem nós em comum. Inspeccionando somente a definição de sensibilidade proposta, outra característica que pode ser inconveniente é que tamanhos diferentes de vazamento conduzem a valores de sensibilidade totalmente diferentes e potencialmente difíceis de utilizar. Nos testes, por exemplo, a ordem de grandeza dos valores de sensibilidade resulta baixa (10^{-3} para $c = 0.1$, e 10^{-4} para $c = 0.01$), e com variações muito grandes (percentualmente) quando considerados diferentes locais de vazamento. As considerações anteriores podem comprometer a acurácia dos cálculos realizados a partir desses valores e, em geral, diminuem a confiabilidade dos resultados.

Uma outra definição de sensibilidade tinha sido anteriormente proposta por Misiunas et al. (2005) como a razão entre a diferença de pressão causada pelo vazamento e a vazão do mesmo vazamento. Ao incorporar as duas grandezas físicas relacionadas com o evento essa proposta tem resultados mais consistentes quando avaliados diferentes vazamentos. A expressão é adaptada na equação Equação 4.4. Além dos símbolos utilizados também na equação Equação 4.3, $QL_{l,l}$ representa a vazão no nó l quando ocorre nele um vazamento, e QN_l a vazão no nó na ausência de vazamentos.

$$S_{n,l} = \left| \frac{PN_n - PL_{n,l}}{QL_{l,l} - QN_l} \right| \quad (4.4)$$

Ao aplicar essa definição na otimização dos locais para instalação dos sensores, nenhuma resposta atinge os máximos teóricos das funções objetivo, mas elas são mais consistentes, com

variações menores entre as respostas obtidas em diversas execuções. Adicionalmente, a ordem de grandeza da sensibilidade (10^0) não afeta a acurácia dos cálculos.

Para o estudo de caso, foi decidido utilizar a definição de Misiunas et al. (2005), especificamente com a Equação 4.5 onde para cada hora h do dia, são calculados os valores de sensibilidade $SH_{n,l}$ de cada um dos nós n a cada um dos vazamentos l , para, posteriormente, acumular ao longo das 24 horas e obter os elementos $S_{n,l}$ da matriz de sensibilidade que será utilizada para as funções objetivo 1 e 2 da otimização.

$$SH_{n,l}^h = \left| \frac{PN_n - PL_{n,l}}{QL_{l,l} - QN_l} \right|^h ; S_{n,l} = \sum_{h=1}^{24} SH_{n,l}^h \quad (4.5)$$

4.1.4 Otimização

A localização ótima de M sensores em N possíveis nós pode requerer tempo exponencial, e soluções por busca exaustiva ou ainda por métodos de busca de tempo polinomial não são praticáveis para redes de tamanho intermediário ou maiores, sendo assim justificado o uso de meta heurísticas como os algoritmos genéticos (AG). Para o projeto de monitoramento foi implementada a otimização usando AG, onde as soluções candidatas são conjuntos de M nós $X = x1, x2 \dots xM : xi \in \{1, \dots N\}$. São consideradas três funções objetivo conflitantes: i) maximização da sensibilidade, ii) maximização da entropia, e iii) minimização da correlação. As funções são avaliadas a partir das matrizes de sensibilidade (S) e correlação (C) obtidas previamente. O primeiro objetivo pode ser definido como a maior sensibilidade total para todos os possíveis vazamentos, e é calculada para uma solução X como $f1_X$ usando a Equação 4.6

$$f1_X = \sum_{l=1}^N \max (S_{xi,l}) \quad \forall \quad xi \in X \quad (4.6)$$

O segundo objetivo visa ter valores de respostas do sistema dispersos, que evitem soluções com alta ponderação para um único vazamento, tal caso poderia ser bem avaliado segundo a $f1$, mas não teria utilidade real. Uma das formas de construção desta função objetivo é através da denominada entropia de Shannon, que representa uma medida da quantidade de informação, ou da incerteza dessa informação ao considerar sua distribuição de probabilidade (SHANNON, 1948).

Neste estudo, uma função derivada da entropia de Shannon é utilizada como um indicador da consistência da sensibilidade, isto é, da distribuição da sensibilidade para todos os possíveis vazamentos, pois quanto mais uniformemente esteja distribuída a sensibilidade, maior será o valor de entropia. A Equação 4.7 é utilizada para avaliar, para a solução X , a segunda função objetivo, $f2_X$.

$$f2_X = \sum_{l=1}^N p_l \log (p_l) ; p_l = \frac{\max (S_{xi,l})}{\sum_{l=1}^N \max (S_{xi,l})} = \frac{\max (S_{xi,l})}{f1_X} \quad \forall \quad xi \in X \quad (4.7)$$

A terceira função objetivo visa reduzir a redundância, tendo em vista que ao considerar somente as duas primeiras funções objetivo seria possível acabar adquirindo informação semelhante com vários sensores, o que não seria conveniente. Com base na matriz de correlação total, C , a terceira função objetivo $f3_X$ é calculada segundo a Equação 4.8,

$$f3_X = \frac{\sum_{i=1}^M \sum_{j=1}^{i-1} C_{xi,xj}}{\binom{M}{2}} \quad \forall \quad xi, xj \in X \quad (4.8)$$

onde M representa o número de nós para instalação de sensores e X uma solução ao problema de otimização.

O processo de otimização utiliza um algoritmo genético elitista denominado o *Nondominated Sorting Genetic Algorithm II* (NSGA-II) (DEB et al., 2002). Quando a busca é finalizada, seja por ter atingido o número máximo de gerações ou a mínima alteração entre valores da função objetivo em gerações sucessivas, o resultado é uma frente de Pareto contendo várias soluções não dominadas, sendo cada uma um conjunto de nós para localização dos sensores. Como resultado da aleatoriedade incorporada no AG, os resultados podem mudar para diferentes execuções, e portanto para incrementar a confiabilidade dos resultados foram feitas cinco repetições.

Sendo que todas as soluções na frente Pareto são igualmente eficientes, a mais próxima do origem dos eixos é selecionada. Essa proximidade é avaliada como a distância euclidiana, no espaço de objetivos, dos valores resultantes de função objetivo padronizados $F1$, $F2$ e $F3$, obtidos com as equações Equação 4.9. As novas funções são as diferenças entre os valores obtidos e os valores ideais para cada função objetivo original, e podem tomar valores entre 0 e 1, sendo melhor o desempenho de fi quanto mais próximo esteja Fi de 0 ($i \in \{1, 2, 3\}$).

$$Fi_X = \begin{cases} \frac{fi_{\text{máx}} - fi_X}{fi_{\text{máx}} - fi_{\text{min}}}, & \text{for } i \in \{1, 2\} \\ \frac{fi_X - fi_{\text{min}}}{fi_{\text{máx}} - fi_{\text{min}}}, & \text{for } i = 3 \end{cases} \quad (4.9)$$

4.1.5 Número de sensores

Como a instalação de grande número de sensores é limitada na prática (técnica, administrativa e financeiramente), o seu aumento deve estar justificado, em termos de qualidade da solução. Quando poucos sensores estão sendo utilizados, os valores das funções objetivo serão melhorados pela adição de um novo sensor. Porém, quando o número aumenta, a informação adquirida tende a se tornar redundante, e a melhora obtida com um novo sensor passa a ser menos importante. Assim, a decisão sobre a quantidade "ótima" de sensores é baseada na análise desses ganhos relativos.

Lembrando que a primeira função objetivo, $f1$, está baseada na máxima sensibilidade de um nó, seus valores considerando diferente número de sensores são comparáveis. A função

objetivo f_2 , calculada com a entropia das sensibilidades máximas, também pode ser comparada independente do número de sensores. No cálculo da função objetivo f_3 , o resultado é padronizado pelo número de duplas de sensores possíveis, para que possa ser comparada. Dessa maneira, o desempenho segundo as três funções objetivo conflitantes pode ser avaliado para diferentes números de sensores na busca pelo ótimo.

O processo de otimização dos locais de instalação de sensores descrito na seção anterior é executado (e repetido cinco vezes) para diferente número de sensores, produzindo as respectivas frentes Pareto. Com o objetivo de avaliar os resultados comparativamente, cada frente é representada pelo intervalo médio[†] das três funções objetivo considerando todas as soluções Pareto-ótimas, e calculada sua média para as cinco repetições. Com o anterior, cada número de sensores (M) fica relacionado com três valores de intervalo médio $\overline{f1_M}, \overline{f2_M}, \overline{f3_M}$.

O ganho relativo pela inclusão de cada sensor adicional, segundo cada função objetivo i , é calculado como Δf_{i_M} na equação Equação 4.10. Para obter um único valor para comparação, é calculada a média do ganho relativo para as três funções objetivo (Δf_{mean}). Mudanças na tendência dos ganhos relativos médios são então analisados para definir o melhor número de sensores.

$$f_{i_M} = \begin{cases} \frac{\overline{f_{i_M}} - \overline{f_{i_{M-1}}}}{\overline{f_{i_{M-1}}}}, & \text{for } i \in \{1, 2\} \\ \frac{\overline{f_{i_{M-1}}} - \overline{f_{i_M}}}{\overline{f_{i_{M-1}}}}, & \text{for } i = 3 \end{cases} ; \Delta f_{mean_M} = \frac{\sum_{i=1}^3 \Delta f_{i_M}}{3} \quad (4.10)$$

4.2 Aplicação ao estudo de caso

O método para projeto de monitoramento proposto foi implementado em Matlab[‡], integrando inicialmente a simulação hidráulica através da biblioteca EPANET2 (ver capítulo Capítulo 3) para construção das matrizes de sensibilidade e correlação. Posteriormente, a otimização da localização de sensores é realizada utilizando o *Matlab Global Optimization Toolbox*. Para o algoritmo NSGA-II foi considerada a criação de soluções inteiras aleatórias com distribuição de probabilidade uniforme, seleção por torneio, função aritmética inteira para cruzamento, mutação por função inteira usando distribuição Gaussiana, padronização das funções *fitness* segundo o intervalo, até 500 gerações, população inicial de 200 soluções e probabilidades de mutação e de cruzamento de 0.8 e 0.1, respectivamente.

[†] Intervalo médio = (máximo + mínimo)/2

[‡] MATLAB, abreviatura de *MATrix LABoratory*, é uma ferramenta de *software* matemático que oferece um entorno de desenvolvimento integrado com uma linguagem de programação própria, a linguagem intérprete M.

4.2.1 Número de sensores

A otimização foi executada para 19 tamanhos de conjuntos de sensores, M , (com $M \in [2, 20]$), com cinco repetições. Para cada conjunto de sensores os valores de intervalo médio das funções objetivo das cinco frentes de Pareto foram avaliados, e suas médias são apresentadas na Tabela 3.

Tabela 3 – Valores das funções objetivo e seus ganhos relativos para número de sensores incremental (M).

M	$f1$	$f2$	$f3$	$\Delta f1$	$\Delta f2$	$\Delta f3$	Δf_{mean}	Acumul.
2	9307.6	5.919	17.862					
3	11334.0	5.892	19.390	21.77%	-0.47%	-8.55%	4.25%	4.25%
4	12497.6	6.054	19.859	10.27%	2.74%	-2.42%	3.53%	7.78%
5	13022.8	6.055	19.569	4.20%	0.02%	1.46%	1.90%	9.68%
6	13641.4	6.055	19.466	4.75%	0.00%	0.52%	1.76%	11.44%
7	14388.3	6.031	19.675	5.47%	-0.40%	-1.07%	1.33%	12.77%
8	14906.8	6.030	19.573	3.60%	-0.01%	0.52%	1.37%	14.14%
9	15425.2	6.063	19.616	3.48%	0.54%	-0.22%	1.27%	15.40%
10	15644.6	5.989	19.691	1.42%	-1.21%	-0.38%	-0.06%	15.35%
11	15882.3	6.053	19.654	1.52%	1.06%	0.19%	0.92%	16.27%
12	15998.8	6.058	19.762	0.73%	0.09%	-0.55%	0.09%	16.36%
13	16262.9	5.987	19.587	1.65%	-1.16%	0.88%	0.46%	16.82%
14	16356.3	6.078	19.753	0.57%	1.51%	-0.84%	0.41%	17.23%
15	16537.5	6.066	19.590	1.11%	-0.20%	0.82%	0.58%	17.81%
16	17113.3	6.067	19.728	3.48%	0.01%	-0.70%	0.93%	18.74%
17	17170.4	6.069	19.742	0.33%	0.04%	-0.07%	0.10%	18.84%
18	17598.8	6.064	19.674	2.50%	-0.09%	0.34%	0.92%	19.76%
19	17563.1	6.060	19.681	-0.20%	-0.06%	-0.04%	-0.10%	19.66%
20	18115.9	6.077	20.029	3.15%	0.28%	-1.76%	0.55%	20.21%

Nota: $f1$, $f2$ e $f3$ são as funções objetivos relativas a sensibilidade, entropia e correlação, respectivamente.

Segundo os resultados, a função objetivo baseada na sensibilidade, $f1$, melhorou (aumentou) com cada sensor adicional para quase todos os sensores considerados. Entretanto, a função $f2$, relativa à consistência da sensibilidade, aumentou com a maioria dos incrementos, porém essas melhoras são pequenas, e até negativas, para alguns dos sensores, sem um padrão claro. A função relativa à correlação, por sua vez, em geral piora (aumenta) quando sensores adicionais são incorporados, mesmo que para alguns valores as mudanças sejam pequenas ou até negativas.

Ao analisar os resultados encontra-se que cada função objetivo orientaria uma decisão diferente sobre o melhor número de sensores se fosse considerada isoladamente. É condizente com sua definição que a sensibilidade fique melhor com o aumento dos sensores, mas que as informações se tornem cada vez mais correlatas, reforçando que as funções objetivo são realmente conflitantes.

A média do ganho relativo, Δf_{mean} , e seu acumulado, na Tabela 3 e Figura 5, confirmam que, quando poucos sensores estão em consideração, um sensor adicional pode trazer melhoras importantes, mas elas se tornam menores enquanto o número de sensores aumenta, e podem até dar um resultado pior. Especificamente, a incorporação de um terceiro sensor mostra grandes melhorias (4,25%), enquanto a entrada de um quarto ou nono geram melhoras de 1,27 e 3,53%. O décimo sensor, pelo contrario, traz diminuição do valor.

Aplicando o critério de permitir o incremento de sensores somente se trouxerem benefício ao resultado (incremento no ganho global), a escolha é usar nove sensores para o estudo de caso. Se bem não é possível negar que o resultado podia ser melhorado ao utilizar mais do que 10 sensores, o critério antes mencionado junta-se à limitada disponibilidade de equipamentos para a realização da pesquisa, pelo qual não sendo possível a exploração da informação de uma quantidade muito grande de sensores, decidiu-se seguir o critério e instalar somente nove.

Figura 5 – Ganho relativo acumulado nas funções objetivo (Δf_{mean}) pela inclusão de um sensor adicional, de 2 a 21.

4.2.2 Localização dos sensores

A localização dos nove sensores é escolhida dos resultados da otimização, procurando nas frentes Pareto com os critérios de transformação das funções ($F1$, $F2$, $F3$) e cálculo de distância apresentados na seção anterior. Os locais onde os nove sensores foram instalados são os apresentados na Figura 4, mas é preciso considerar que enquanto o método considera a localização dos sensores nos nós da rede, por razões práticas não é possível que a instalação seja feita exatamente nesses pontos, às vezes nem na vizinhança mais próxima.

Salienta-se que os nós para instalação escolhidos estão, em sua maioria, em seções extremas (mais a jusante) da rede. Tal resultado parece consistente com o fato de que, em geral, mudanças promovidas por um vazamento em qualquer nó são mais notórias a jusante. Além disso, os nós selecionados estão bem distribuídos considerando a topologia do sistema, sugerindo que a incorporação das funções de entropia e correlação levaram a uma resposta bem esparsa.

4.2.3 Análises adicionais

Análises adicionais foram desenvolvidos para testar aspectos específicos do método proposto, e se bem não é considerado interessante apresentá-los por extenso, a seguir são resumidas as conclusões obtidas.

Foram comparadas as localizações obtidas para diferente número de sensores $M \in (2, 10)$, mostradas na Figura 6, lembrando que pela incerteza no local exato de instalação do sensor próximo ao nó selecionado, nós primeiros vizinhos (respeitando a topologia da rede) são considerados como uma mesma localização. Um dos resultados interessantes é que uma das localizações escolhida quando considerados somente dois sensores ($M = 2$), é selecionada também em todos os outros casos, exceto um. A tendência de manter os pontos previamente selecionados quando o número de sensores aumenta é verificada para vários locais, e em vários casos um local selecionado para $M < i$ mas não para $M = i$ pode ser selecionado de novo para $M > i$.

O efeito da otimização multiobjetivo foi analisado executando a otimização para nove sensores com modificações para considerar as três, dois ou uma das funções objetivo previamente definidas. Assim, sete procedimentos alternativos foram avaliados, usando para o AG os conjuntos de funções objetivo: $\{f1, f2, f3\}$, $\{f1, f2\}$, $\{f1, f3\}$, $\{f2, f3\}$, $\{f1\}$, $\{f2\}$ e $\{f3\}$. Comparando as soluções para os últimos três casos, otimizando em cada um somente uma função objetivo, nenhum local é escolhido nos três e somente dois locais aparecem em dois dos casos, o que ratifica que os três objetivos são conflitantes. É notório também que dois locais específicos aparecem em todos os quatro casos em que a função $f1$ é considerada, mas em nenhum dos outros casos avaliados. De fato, esses locais apresentam altos valores de sensibilidade segundo a matriz S . A respeito das soluções nos casos com duas funções objetivo, os três conjuntos são diferentes entre si, mas têm maior semelhança com a resposta obtida com as três funções objetivo do que as avaliações com uma única função.

4.3 Comentários finais

O método para projeto de monitoramento visando a detecção de vazamentos foi também um passo necessário para a realização da pesquisa, e além disso é uma das contribuições do trabalho. Para a tomada de decisão sobre o projeto de monitoramento foi utilizada a melhor informação disponível do sistema, particularmente as séries de nível e vazão e a modelagem hidráulica. Porém, durante o desenvolvimento da pesquisa foi encontrado que algumas dessas

informações não correspondiam à realidade da rede de abastecimento. Além das medidas de vazão de saída do reservatório não serem confiáveis devido aos problemas com o sensor que na época operava, como comentado no Capítulo 3, também foram identificadas mudanças na topologia da rede que não estavam refletidas no modelo. As pressões medidas não estão no mesmo patamar que as simuladas para o modelo para alguns sensores, como analisado no capítulo 6, e uma das explicações para tal resultado é a alteração nas tubulações mais antigas pelas incrustações. Não é possível, portanto, afirmar com certeza que o número de pontos e localização escolhidos tenham sido os melhores possíveis para a pesquisa, mas sim que sua escolha foi feita com critérios claros e com base na melhor informação disponível sobre o sistema existente.

Figura 6 – Locais escolhidos pela otimização multiobjetivo para diferente número de sensores.

5 MÉTODOS DE ANÁLISE

Tanto para a descrição das séries de pressão adquiridas quanto para o desenvolvimento de métodos para detecção de vazamentos e alterações na rede, são aplicados conceitos e métodos empregados nas áreas de processamento de sinais, análise de séries temporais e aprendizado de máquina, dentre outras. Esses conceitos e métodos são, em geral, muito difundidos nas áreas de conhecimento nas quais foram desenvolvidos mas não são utilizados para análise de sistemas de abastecimento, razão pela qual uma apresentação clara e focada na sua aplicação é importante.

Assim, neste capítulo são apresentados todos os métodos aplicados às séries de pressão adquiridas, incluindo um breve resumo da sua fundamentação teórica e os aspectos mais importantes aplicados na pesquisa, além das referências onde o leitor pode encontrar maiores informações. A estrutura deste capítulo é construída na sequência em que os métodos foram utilizados e os resultados apresentados no Capítulo 6, começando com os métodos para a descrição das séries e posteriormente os relativos à detecção de alterações. Na primeira seção são descritas as análises de estacionariedade, autocorrelação, espectral e de recorrência das séries de pressão adquiridas. Na segunda, os fundamentos dos métodos propostos baseados em processos estocásticos, controle estatístico de processos e comparação com padrões, além da avaliação do desempenho da detecção.

5.1 Métodos para descrição das séries

Uma das características mais importantes da pesquisa apresentada neste documento é a utilização de informação monitorada em pontos estrategicamente determinados de um sistema real e complexo, com características sobrepostas que podem corresponder ou não às suposições e simplificações que são usualmente utilizadas para muitas aplicações em redes de abastecimento de água. Assim, uma descrição detalhada das séries adquiridas é necessária para conhecer as reais características e sua importância no comportamento dos sinais. Adicionalmente, o conhecimento completo das séries de pressão adquiridas dentro do estudo de caso é necessário para avaliar a validade e as limitações de diferentes abordagens do problema de detecção de alterações, assim como de diferentes tratamentos numéricos dos dados. Se tal análise for desconsiderada, existe o risco de aplicar conceitos incompatíveis ou ignorar vantagens potenciais. São resumidos nesta seção alguns dos conceitos utilizados para as descrições dos sinais que são apresentadas no Capítulo 6, em particular aqueles cuja aplicação não é rotineira na área.

5.1.1 Estacionariedade

Uma característica fundamental a ser avaliada nos sinais é a estacionariedade. A constância do comportamento de um sinal ou série temporal, particularmente o equilíbrio estatístico ao redor de uma média e uma dispersão constante, fazem tal série estacionária. Mais especifica-

mente, se uma série tem média e variância constantes ela tem estacionariedade no sentido amplo, também chamada estacionariedade fraca ou de segunda ordem. Se adicionalmente a estrutura da autocovariância é constante, isto é, se seu valor depende unicamente do intervalo de tempo entre dois momentos considerados, a série será estritamente estacionária (YAFFEE; MCGEE, 2000).

A estacionariedade pode ser avaliada em diferentes escalas (curto ou longo prazo), segundo a série analisada. Tal diferenciação é fundamental visto que sinais com comportamento periódico podem ser considerados não estacionários se analisados numa escala menor à dos períodos presentes, mas ao considerar escalas superiores sua estacionariedade pode ser verificada.

Há principalmente quatro comportamentos característicos que fazem com que uma série seja não estacionária (YAFFEE; MCGEE, 2000): o primeiro, de "passeio aleatório"*, aparece quando em ocorrências sucessivas a variável comporta-se como uma flutuação completamente aleatória, de forma que a média pode ser mantida mas a variância pode crescer indefinidamente ao longo do tempo. O segundo é o deslocamento (*drift*) que refere uma variação aleatória ao redor de uma média diferente de zero, de forma que a média deixa de ser estacionária. O terceiro comportamento indicado pelo autor é o de tendência, relacionado com mudanças da média da série em função do tempo, que podem ser lineares ou não lineares. Se essas mudanças são sistemáticas (e, portanto, passíveis de ser preditas pela informação histórica) a tendência é determinística, e em caso contrário a tendência se considera estocástica. Essa tendência pode ser de longo prazo, com as mudanças acontecendo mais lentamente, ou de curto prazo, com mudanças mais rápidas e posterior estabilização. O quarto comportamento não estacionário é a mudança de variância, que pode ser resultado de efeitos de acumulação ou da influência de outras variáveis sobre a série.

Em alguns casos é possível retirar o comportamento não estacionário das séries e dessa forma transformá-las em estacionárias, desde que exista uma tendência, seja de curto ou longo prazo, que possa ser identificada e ter seu efeito retirado da série. Essa retirada é feita, por exemplo, ao substituir pelas suas primeiras diferenças uma série cuja não-estacionariedade é do tipo de caminhada aleatória, ou transformando os dados com a função logaritmo para estabilizar a variância.

Uma forma simples mas fundamental para avaliar a estacionariedade é a inspeção visual. Em um gráfico da série sob análise, com domínio compatível com a escala de interesse, é possível identificar se a média e a variância são constantes ou, pelo contrário, algum dos comportamentos antes descritos aparece na série. Em adição à inspeção do gráfico, os valores de média e variância em trechos da série (compatíveis com a escala analisada) podem ser calculados e comparados com os respectivos valores dos trechos vizinhos. Os resultados desse dois tipos de análise para as séries de pressão na rede adquiridas são apresentados no Capítulo 6.

* De forma simplificada, o termo "passeio aleatório" refere-se ao caminho ou sequência resultante de uma sucessão de passos onde cada um é aleatório.

5.1.2 Autocorrelação

A autocorrelação é uma medida da semelhança de uma série com versões dela mesma deslocadas no tempo ou, equivalentemente, da similaridade dos valores da série em função do intervalo de tempo que transcorre entre eles. É utilizada para a detecção de comportamento não aleatório, em particular de padrões de repetição, tais como a presença de componentes periódicos na série.

O cálculo da autocorrelação da série x com um atraso k é equivalente ao do coeficiente de correlação produto-momento, ou coeficiente de correlação de Pearson, mas sem utilizar uma segunda série e sim duas vezes a série x : uma na versão original e outra com atraso k , como apresentado na Equação 5.1. Pode-se demonstrar que essa definição é equivalente a uma padronização da função de covariância entre a série original e a deslocada. Um valor de autocorrelação é calculado para toda a série, e seu resultado é o mesmo independentemente do ponto de início ou referência na série, sempre que o atraso entre cada ponto da série original e o correspondente na série deslocada seja o mesmo k (YAFFEE; MCGEE, 2000).

$$acorr(x, k) = \frac{\sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2} \quad (5.1)$$

sendo $acorr(x, k)$ a autocorrelação da série x para um atraso k , n o comprimento (tamanho) da série e k o atraso que poderia receber valores inteiros de 0 até n . Porém, a autocorrelação com atrasos equivalentes a porcentagens muito grandes do tamanho total do sinal tende a perder significado, tanto físico quanto no cálculo, razão pela qual que é recomendado analisar atrasos $k \leq 25\% \cdot n$ (MONTGOMERY, 2009).

O valor de $acorr$ está no intervalo $[-1, 1]$, sendo que 1 corresponde a correlação perfeita, isto é, valores idênticos repetem-se com período fixo. $acorr = 0$ indica não correlação, e $acorr = -1$ correlação negativa para esse atraso. Os valores extremos, porém, não são atingíveis por séries de tamanho finito dado que o deslocamento pelo atraso reduz para $n - k$ o tamanho das séries comparáveis, e portanto o valor do somatório no numerador da Equação 5.1 e o da autocorrelação diminuem.

Com o cálculo das autocorrelações com todos os valores de k de interesse pode ser produzido o autocorrelograma, um gráfico cujo eixo horizontal corresponde aos atrasos (k) e o eixo vertical aos valores de autocorrelação ($acorr$). Ao aumentar o valor de atraso espera-se que o da autocorrelação diminua pela redução das séries comparáveis, como antes explicado, e a intensidade dessa diminuição é dependente do tamanho total da série e do atraso. Com o correlograma podem ser avaliadas qualitativamente características da série como a estacionariedade e a existência e frequência de componentes periódicos.

5.1.3 Análise espectral

Uma das características fundamentais dos sinais analisados é a sua periodicidade, que tem origem no comportamento do consumo. Comumente as atividades dos usuários estão enquadradas numa rotina diária mais ou menos estável, repetindo nos mesmos horários aproximadamente o consumo de água para banho, preparo de alimentos, etc. Há usualmente particularidades nos diferentes dias da semana, como mudança de horários no final de semana, e consumo maior de água aos sábados. A operação de bombeamento do poço para os reservatórios costuma obedecer também critérios aproximadamente diários e orientados pelo consumo e pelas tarifas de energia. Essas observações, porém, são insuficientes para ter certeza da existência e valores de diferentes componentes periódicos, para o qual é de utilidade a análise espectral.

A partir dos conceitos base das séries de Fourier, os sinais podem ser representados como uma soma de funções senos e cossenos com diversas frequências e amplitudes, ou componentes. Com isso, um sinal no domínio do tempo, isto é, definido pela amplitude de uma variável (pressão, no nosso caso) em função do tempo, pode ser representado no domínio da frequência, como uma grandeza (que não tem nenhuma correspondência com a variável medida) em função da frequência (ALLEN; MILLS, 2003; PROAKIS; MANOLAKIS, 2007). Por exemplo, se o sinal no domínio do tempo coincide com uma única senoide, no domínio da frequência será representado com um único pico localizado na frequência daquela senoide. Já a maioria dos sinais obtidos de medições reais apresentam superposições de efeitos e ruído, e sua representação no domínio da frequência contém mais elementos e dá destaque a componentes periódicos que poderiam nem ser identificáveis na representação original. As ideias antes apresentadas têm aplicação tanto para sinais contínuos quanto para discretos, embora as definições matemáticas sejam diferentes para cada caso. Os sinais considerados nesta pesquisa são discretos e quantizados, e são as definições aplicáveis a eles as comentadas no restante da seção.

A ferramenta para fazer a transformação de $x(t)$, um sinal no domínio do tempo, e produzir o correspondente $X(\omega)$ no domínio da frequência é a *transformada de Fourier discreta* (PROAKIS; MANOLAKIS, 2007):

$$X(\omega) = \sum_{t=-\infty}^{\infty} x(t) e^{-i\omega t} \quad (5.2)$$

sendo $i^2 = -1$, ω a frequência medida em radianos por intervalo de amostragem ou em ciclos por intervalo de amostragem ($\frac{rad}{|tempo(t)-tempo(t-1)|}$), e $X(\omega)$ um sinal discreto de números complexos. O cálculo da transformada, na prática, não é feito aplicando diretamente essa equação mas um algoritmo muito mais eficiente e equivalente, o de *transformada rápida de Fourier* (*fft* pela sigla em inglês). A partir dessa transformada é possível definir para o sinal $x(t)$ sua densidade espectral de energia $S(\omega)$, também chamada densidade espectral de potência, densidade espectral ou simplesmente espectro, usando a equação seguinte (PROAKIS; MANOLAKIS, 2007).

$$S(\omega) = |X(\omega)|^2 \quad (5.3)$$

A densidade espectral de energia representa a distribuição da energia da série $x(t)$ como função da frequência. Assim, um valor alto de $S(\omega_i)$ pode ser interpretado como indicador de que no sinal $x(t)$ há uma componente importante da frequência ω_i , isto é, que o sinal tem (mesmo que superposto a outros comportamentos) comportamento periódico com período $\frac{1}{\omega_i}$. A densidade espectral é representada num gráfico com o eixo horizontal correspondendo às frequências ω_i e o vertical aos valores de S , e desse gráfico podem ser extraídas informações importantes sobre o sinal, especialmente sobre sua periodicidade.

Para a análise do gráfico de densidade espectral é necessária a correta interpretação das frequências obtidas, cujas unidades dependem do intervalo de amostragem. Com os algoritmos e implementações utilizadas para cálculo da transformada (função *fft* do Matlab) e da densidade espectral, a frequência é expressa em unidades de ciclos por intervalo de amostragem, com valores entre $-0,5$ e $0,5$, sendo que os valores de densidade das frequências negativas são idênticos aos das positivas. Assim, considerando que os sinais de pressão foram adquiridos com taxa de amostragem de um valor a cada 2 minutos, o intervalo de frequências de interesse é de:

$$0 \text{ a } 0,5 \frac{\text{ciclos}}{\text{intervalo de amostragem}} = 0 \text{ a } 0,5 \frac{\text{ciclos}}{2 \text{ min}} =$$

$$0 \text{ a } 0,25 \frac{\text{ciclos}}{\text{min}} = 0 \text{ a } 0,25 \text{ min}^{-1}$$

Esses intervalos correspondem ao estabelecido pelo teorema da amostragem de Nyquist–Shannon (ALLEN; MILLS, 2003), pois a frequência máxima que pode ser analisada com a taxa de amostragem de $1 \text{ valor}/2 \text{ min}$ é de $0,25 \text{ min}^{-1}$. Isto é, podem ser analisados componentes que ocorrem até uma vez a cada quatro minutos (ou 15 vezes cada hora), e nenhuma com períodos menores. As frequências de interesse, porém, estão muito abaixo desse limite.

5.1.4 Análise de recorrência cruzada

Seguindo outra abordagem que pode ser utilizada para analisar os sinais ou séries de pressões adquiridas, os conjuntos de medições de pressão em pontos da rede de abastecimento conformam sistemas dinâmicos, nos quais a variável tem diferentes valores ou estados ao longo do tempo. O espaço constituído pelos possíveis estados de todas as variáveis que constituem um sistema é denominado espaço fase (WEBBER; MARWAN, 2015), de forma que as transições entre estados do sistema podem ser vistas como deslocamentos no espaço fase. A sucessão temporal de estados no espaço fase é denominada uma trajetória, e a evolução dessa trajetória ao longo do tempo expõe a dinâmica do sistema. Como nomenclatura, quando uma trajetória chega a um ponto no espaço fase, isto é, quando as variáveis estão em um estado determinado, esse estado está sendo "visitado".

A descrição do sistema dinâmico, e em particular a construção de um modelo que permita prever os estados futuros das variáveis com base no conhecimento dos estados anteriores é um objetivo muito amplo e para o qual múltiplas abordagens podem ser utilizadas, de acordo com a

caracterização do sistema (estacionário ou não, determinístico ou estocástico, linear ou não linear, etc). Alguns dos métodos modernos para estudo de sistemas dinâmicos não lineares (aplicáveis a sistemas lineares como casos particulares) estão baseados na análise do espaço fase do sistema ou suas reconstruções, dentre eles a análise de recorrência.

Quando em uma trajetória o mesmo estado (ou estados arbitrariamente próximos) é visitado em diferentes momentos, seja produto de periodicidade ou de ciclos não regulares, existe recorrência. Para visualizar essa recorrência, a quantidade de vezes que cada estado no espaço fase é recorrente (é repetidamente visitado) é representada num gráfico de recorrência ou *RP* (pela sigla em inglês). Independentemente da dimensão do espaço fase, o gráfico de recorrência é construído com base numa matriz quadrada de duas dimensões, e seus dois eixos correspondem ao tempo, na mesma referência e escala em que os valores da série são conhecidos. Seja xm a serie da variável m -dimensional que descreve o sistema; n o comprimento da série; xm_t o elemento no tempo t , isto é, o estado do sistema no tempo t , com $t = 1, 2, \dots, n$. A recorrência do estado do tempo i para o tempo j , $RP_{i,j}$ é calculada com a expressão seguinte (WEBBER; MARWAN, 2015):

$$RP_{i,j}^\varepsilon = \Theta(\varepsilon - ||xm_i - xm_j||) \quad i, j = 1, 2, \dots, n \quad (5.4)$$

Na Equação 5.4, $||\cdot||$ é uma métrica que neste trabalho é adotada como distância euclidiana, ε é o raio de corte ou distância máxima para considerar dois estados como diferentes ou não, e Θ é a função de Heaviside ou função degrau:

$$\Theta(y) = \begin{cases} 0 & \text{se } y < 0 \\ 1 & \text{se } y \geq 0 \end{cases} \quad \forall y \in \mathbb{R} \quad (5.5)$$

Assim, o gráfico definido pela matriz *RP*, ou gráfico de recorrência, permite a identificação de aspectos gerais do sistema dinâmico, como a homogeneidade ou a presença de periodicidade e suas características, além de identificar irregularidades ou mudanças de comportamento. Analisando as estruturas presentes no gráfico em diferentes escalas é possível extrair informações sobre o sistema que muitas vezes não são evidentes em outras representações, principalmente se estão relacionadas à não linearidade. Para complementar a análise da recorrência das series tem sido proposto o cálculo de indicadores que permitam quantificar as principais propriedades representadas nos gráficos de recorrência (WEBBER; ZBILUT, 1994; ZBILUT; THOMASSON; WEBBER, 2002; MARWAN et al., 2002), e dentre eles o utilizado para a aplicação do método à descrição das series de pressão adquiridas é a taxa de recorrência, *RR* (pela sigla em inglês), definida como (WEBBER; MARWAN, 2015):

$$RR = \frac{1}{n^2 - n} \sum_{i \neq j=1}^n RP_{i,j}^\varepsilon \quad (5.6)$$

Para aplicação do método anteriormente descrito é necessária a definição de três parâmetros importantes: a dimensão para reconstrução m , o atraso (ou passo de reconstrução) d , e o raio de corte ε .

A reconstrução do espaço fase de um sistema dinâmico que inicialmente está definido por uma série temporal unidimensional é melhorada através do aumento da dimensionalidade (ZBILUT; WEBBER, 1992). Para tanto, a série original x é reconstruída ou imersa em um espaço de dimensionalidade m , onde uma das dimensões corresponde à referencia temporal original e as seguintes dimensões a momentos posteriores da série. Assim, se originalmente para cada tempo t o valor da série é $x_t = \{x_t\}$, após aumentar a dimensionalidade:

$$xm_t = \{x_t, x_{t+d}, x_{t+2d}, \dots, x_{t+(m-1)d}\} \quad (5.7)$$

Para completar essa reconstrução e obter a série m -dimensional devem ser definidas a dimensão de imersão m e o valor de atraso d . A escolha destes parâmetros depende das séries a ser analisadas, e não há uma única definição de valores ideais, mas recomendações para sua estimativa. Para o número de dimensões é utilizada a técnica de Falsos Vizinhos Próximos (*FNN* pela sigla em inglês)(KENNEL; BROWN; ABARBANEL, 1992). O método depende de vários parâmetros, principalmente da distância máxima para a vizinhança e do valor de atraso escolhido, além de características como o comprimento e nível de ruído das séries e do nível de não-linearidade e não-estacionaridade existentes. Por outro lado, a decisão do passo de reconstrução d usando os resultados da autocorrelação da série original, em particular segundo o primeiro mínimo local que exista no correlograma, é comumente utilizada e está justificada em séries estacionárias (KANTZ; SCHREIBER, 2003).

A definição do raio de corte é fundamental para a utilidade dos resultados da análise de recorrência, como evidenciado pelo seu efeito sobre taxa de recorrência RR . Valores de ε muito grandes fazem com que todos os estados sejam considerados próximos e portanto qualquer estado seja recorrente, produzindo $RR \lesssim 1$. Pequenos ε , pelo contrário, tendem a acusar recorrência somente para estados idênticos, fazendo $RR \gtrsim 0$. Porém, a definição do raio de corte ε depende da série considerada, e não foi encontrada uma única metodologia de uso amplo, mas critérios recomendados para fazer a escolha.

Zbilut, Thomasson e Webber (2002), após salientar que o foco da análise de recorrência é encontrá-la para uma região localmente definida e linear do espaço fase, recomendam avaliar essa linearidade para diferentes valores de raio. Essa avaliação pode ser feita analisando os indicadores calculados do RP , em particular a taxa de recorrência RR , e escolhendo o valor que, sem cair em nenhum dos extremos mencionados, represente uma região linear.

5.2 Métodos para a detecção de vazamentos e alterações

A aplicabilidade da metodologia para a detecção de alterações nos sinais de pressão depende da sua adequação ao tratamento de fluxo de dados. Isto é, capacidade de incrementalidade,

de processamento computacional frequente, com recursos limitados de armazenamento e memória, e capacidade de detectar e adaptar o modelo de decisão a mudanças do conceito[†] (FACELI et al., 2011). Dentre as estratégias para possibilitar que essas características sejam atingidas, nos métodos propostos são usadas janelas temporais deslizantes (GAMA; GABER, 2007), e modelo de tratamento da base de dados com operações de inserção de novas informações, recuperação de pequenas porções de informação existente, atualização, e esquecimento das informações antigas.

Em geral, os métodos de detecção propostos recebem as novas informações a cada dois minutos ou tempos maiores, e as avaliam segundo os critérios aprendidos para determinar se existe ou não uma irregularidade (vazamento ou outra alteração). Depois de ter completado e arquivado a avaliação, o aprendizado é atualizado, e as informações mais antigas não são mais necessárias. Sob este esquema de funcionamento, a diferença fundamental entre os métodos propostos está na forma em que o comportamento "normal" é aprendido ou representado e em como é avaliada a adequação dos novos dados a essa representação.

Salienta-se que na abordagem adotada para esta pesquisa, entende-se que no problema a resolver não se conta com exemplos das diferentes situações que devem ser detectadas, de forma que abordagens de classificação supervisionada não são aplicáveis. Exemplos supervisionados das anormalidades são utilizados na pesquisa somente para avaliação de desempenho, mas não para treinamento e execução dos métodos. Há nos métodos um conhecimento prévio que é aproveitado, mas na aplicação prática pode ser extraído em períodos curtos antes do início do funcionamento dos métodos, sem incluir necessariamente alterações, e atualizado regularmente.

Para as séries de pressão, e para as condições de funcionamento do sistema que as origina, é possível criar descrições ou modelos em diferentes escalas, desde minutos até várias semanas. Assim, os diferentes métodos propostos podem realizar análises em tais escalas. Também, o uso de diferentes tipos de informação é explorado, com dois métodos baseados exclusivamente no conhecimento que pode ser extraído diretamente das séries de dados, e o terceiro incorporando conhecimento adicional sobre o sistema físico.

Os dois primeiros métodos são baseados em cadeias de Markov e em controle estatístico de processos, respetivamente, enquanto o terceiro incorpora o conhecimento da periodicidade das séries e de um padrão de comportamento médio esperado e baseia-se na comparação com esse padrão. Nas sub-seções seguintes, os fundamentos dos três métodos são apresentados, antecidos pela explicação da forma de avaliação de desempenho implementada.

5.2.1 Avaliação de desempenho da detecção

Segundo o esquema geral proposto para a detecção de vazamentos ou outras alterações no sistema, para cada nova janela de informação, todos os métodos realizam uma comparação

[†] Mudança de conceito ou *Concept drift*, em aprendizado de máquina, refere-se à situação em que a distribuição que está sendo predita ou, em outras palavras, o modelo que está sendo predito, muda no tempo

entre o esperado segundo o aprendizado (não supervisionado) e o observado, e dessa comparação juntamente com um limiar α é respondido se os novos dados indicam uma alteração ou não. A escolha desse limiar depende do equilíbrio buscado entre sensibilidade e especificidade (definidos nesta mesma subseção), e para avaliação foram considerados, em cada caso, 100 diferentes valores de α , escolhidos de forma a maximizar a dispersão dos resultados abrangendo todos os resultados da comparação.

A resposta produzida pelo método somente pode estar certa ou errada se existir uma resposta "verdadeira". Pela complexidade e variabilidade da rede de abastecimento de água monitorada, para a base de dados adquiridos não existe certeza completa sobre a situação "verdadeira" do sistema em cada um dos instantes durante a aquisição de dados. Essa informação, porém, é necessária para fazer a avaliação dos métodos, e portanto é utilizado o melhor conhecimento possível obtido pela classificação manual das informações.

Para avaliação dos métodos, é considerada verdadeira a classe atribuída pela classificação manual para cada instante da série. Essa classificação supervisionada, feita para cinco classes (Capítulo 3), é simplificada com uma resposta binária, segundo duas alternativas:

- normal-todos $\rightarrow$ Neg : Normal vs. Pos : Reportado, ou Visual, ou Simulado, ou Extremo
- normal-simulado $\rightarrow$ Neg : Normal vs. Pos : Simulado

O motivo para utilizar tal simplificação é que a resposta dos métodos é binária, isto é, para cada valor da série é concluída uma de duas categorias:

- Negativo : Nenhuma alteração detectada vs. Positivo : Alteração detectada

Os métodos de detecção, pela sua capacidade de gerar essas respostas, podem ser considerados métodos de classificação. Segundo essa notação, após a aplicação aos dados para definição dos parâmetros (treinamento) é produzido um classificador.

Cada método a ser avaliado é executado para toda a série sob análise, realizando para cada valor da série a comparação entre as respostas obtidas no método e as correspondentes respostas verdadeiras. Cada resultado da comparação é agregado e o número de acertos e erros de diferente tipo constituem os elementos da *matriz de confusão*. A matriz de confusão é uma representação usual dos resultados de uma classificação, especialmente binária, em que são contabilizados os resultados da comparação das categorias obtidas (colunas) e as correspondentes categorias esperadas (filas), como apresentado na Figura 7.

A soma dos quatro elementos da matriz é igual a N , o tamanho da série avaliada considerando somente os valores passíveis de comparação, que são aqueles que têm atribuída uma das classes analisadas e para os que o método obteve conclusão. Pelo anterior, o valor de N é sempre menor quando consideradas as classes normal-simulado e maior para normal-todos.

Figura 7 – Esquema da matriz de confusão.

	Nenhuma alteração	Alteração
normal	VN	FP
todos (ou simulado)	FN	VP

A matriz de confusão de um classificador ideal terá $FN = FP = 0$, e $VN + VP = N$. Com os resultados da matriz de confusão podem ser calculados diversos indicadores de desempenho, dentre os quais destacam os seguintes (FACELI et al., 2011).

- Taxa de acerto. Indicador da acurácia geral que quantifica o porcentagem de valores corretamente classificados

$$\frac{VN + VP}{N} \quad (5.8)$$

- Taxa de erro. Percentagem de valores erroneamente classificados.

$$\frac{FN + FP}{N} = 1 - \text{acerto} \quad (5.9)$$

- Sensibilidade. Também denominada taxa de verdadeiros positivos ou *recall*. Mede a razão dos valores realmente positivos que foram detectados como positivos.

$$\frac{VP}{VP + FN} \quad (5.10)$$

- Especificidade, ou taxa de verdadeiros negativos. Quantifica a razão dos valores realmente negativos que foram detectados como negativos.

$$\frac{VN}{VN + FP} \quad (5.11)$$

- Taxa de falsos positivos. Refere-se aos valores que, sendo realmente negativos, foram classificados como positivos.

$$\frac{FP}{VN + FP} = 1 - \text{especificidade} \quad (5.12)$$

Quiçá o indicador mais comumente utilizado, ainda sem outras análises da matriz de confusão, é o acerto ou acurácia, que parece ser intuitivo tendo como objetivo se aproximar do acerto = 1 ou 100%. É preciso, porém, notar que o resultado de acerto pode conduzir a interpretações errôneas quando nos sinais de análise as classes estão mal distribuídas, isto é, quando há muito mais valores de uma das classes do que das outras. No estudo de caso, considerando para avaliação todas as classes supervisionadas (normal-todos), há 75% de valores da classe normal que deviam ter resposta negativa: *Nenhuma alteração detectada*) e somente 25% de todas as classes diferentes à normal que deviam ter resultado positivo: *Alteração detectada*. Se

somente os experimentos de simulação são considerados na avaliação (normal-simulado) esses percentagens mudam para 99,8% negativos e somente 0,2% positivos. Se para o estudo fosse escolhido um classificador que simplesmente responde sempre com negativo: Nenhuma alteração detectada, certamente não seria de utilidade, mas sua taxa de acerto seria muito alta, de 75,0% na avaliação normal-todos ou de 99,8% em normal-simulado. Essa situação de altos valores de acerto calculados para classificadores de desempenho real ruim é chamada de paradoxo da acurácia.

Como a acurácia, a taxa de erro e os outros indicadores que têm como base o total de exemplos supõem igual importância para todos os tipos de acerto ou de erro, o que não é interessante para todas as aplicações. Esses indicadores resultam portanto pouco úteis em casos de classes altamente desbalanceadas e em particular na detecção de alterações, e é necessário utilizar outras quantidades. A sensibilidade e especificidade, por exemplo, são indicadores que continuam sendo úteis mesmo para classes tão desbalanceadas quanto as das series sob análise. Esses dois indicadores são fundamentais para outra representação do desempenho muito utilizada: a curva característica de operação do receptor, ou curva ROC pelo inglês *Receiver operating characteristic*.

A curva ROC foi inicialmente utilizada na área de detecção de sinais na década de 1960, e seu nome refere-se às características de comportamento do classificador para toda a faixa potencial de sua aplicação (KRZANOWSKI; HAND, 2009). Isso significa que essa curva também resume neste trabalho o desempenho de cada método de detecção para todos os possíveis valores de limiar, permitindo a avaliação geral e não dependente de um único valor de α . Um gráfico é utilizado para representar a curva ROC, e nele o eixo vertical corresponde à sensibilidade e o eixo horizontal à taxa de falsos positivos (1-especificidade), e cada ponto da curva corresponde a um diferente limiar α . Na Figura 8 é apresentado um esquema das curvas ROC para um classificador ideal, um aleatório e um classificador que pode ser encontrado.

Figura 8 – Curva ROC para três tipos de classificadores. Esquema.

Para poder ser utilizada e comparada mais facilmente, a curva ROC deve ser resumida por indicadores numéricos, e o mais comum deles é a área abaixo da curva, ou *AUROC* (pelas

siglas em inglês). A *AUROC* constitui um bom indicador de quão separáveis são as classes com o método aplicado, e pode ser interpretada como a sensibilidade média tomada uniformemente sobre todos os possíveis resultados de especificidade. Outra interpretação possível é que se o classificador produz um valor de comparação (a ser comparado com α) para um elemento positivo qualquer da série, e outro valor de comparação para um elemento qualquer negativo, a *AUROC* é a probabilidade do valor encontrado para o positivo ser maior ao encontrado para o negativo (KRZANOWSKI; HAND, 2009). O valor de *AUROC* para cada método de detecção avaliado é o principal indicador de desempenho global utilizado nesta pesquisa, e é calculado a partir da aproximação com trapézios das curvas construídas com 100 pontos (100 valores de α).

O classificador ideal com limiares muito restritivos não detecta nada como positivo (ponto 0, 0 na figura), mas com qualquer valor de α suficiente detecta todos os positivos (*sensibilidade* = 1) e nenhum dos negativos é detectado erroneamente (ponto 0, 1). Se o limiar for muito relaxado, a sensibilidade continua máxima mas a especificidade vai diminuindo até chegar ao ponto 1, 1 na figura. A área abaixo da curva para esse ideal seria de $AUROC=1,0$. O significado dos pontos 0, 0 e 1, 1, resultantes de valores de limiar extremamente restritivos ou relaxados, é o mesmo para todas as curvas. O classificador aleatório é aquele que atribui cada resposta sem considerar nenhuma característica do problema, mas de forma simplesmente aleatória. Se bem o formato específico da curva pode variar pela própria aleatoriedade da escolha, a tendência se aumentar o número de pontos é seguir a linha reta mostrada, e esse classificador tem $AUROC=0,5$. Qualquer classificação somente tem sentido se traz resultados melhores do que não aplicar método nenhum, de forma que o desempenho do aleatório é considerado o limite inferior para um classificador ter algum interesse, mesmo que em muitos casos possam aparecer desempenhos inferiores.

Cada um dos métodos propostos, como qualquer classificador que possa ser obtido para um problema não trivial, tem um desempenho inferior ao ideal, e os valores de limiar implicam em diferentes valores de sensibilidade e especificidade. Como antes dito, a área abaixo da curva característica de operação do receptor é o indicador de desempenho global principal para as análises, e é esperado que os métodos propostos tenham $0,50 < AUROC < 1,0$.

5.2.2 Processo estocástico e cadeias de Markov

As pressões em pontos da rede de abastecimento adquiridas durante a operação do sistema dependem de grande quantidade de fatores externos e internos, e mesmo sob condições até certo ponto idênticas o valor específico da pressão pode variar, caracterizando-se como uma variável aleatória. Igual a qualquer outro conjunto de variáveis aleatórias indexadas no tempo, a pressão num ponto da rede ao longo do tempo é um *processo estocástico* (BHAT; MILLER, 2002). Particularmente, o anterior é um processo estocástico com parâmetros e espaço de estados contínuo, pois tanto o tempo (t) quanto a pressão (p_t) são variáveis contínuas. Porém, ao considerar a série de valores de pressão medidas em intervalos de tempo (p) temos um processo estocástico de parâmetros discretos (instantes da medição) e estados discretos (pressão

segundo precisão do equipamento).

Em um instante qualquer, o valor de pressão ou valor que toma a variável aleatória p não deve ser qualquer um independentemente das pressões antecedentes, mas manter alguma relação com elas. Podemos supor facilmente que o valor no momento atual depende do valor no momento anterior, de forma que o processo pode ser considerado um *processo de Markov*, e a série de pressões adquiridas uma *cadeia de Markov*. Segundo esta caracterização, a probabilidade de uma pressão $p = i$ ocorrer k passos após a ocorrência de $p = j$ denota-se P_{ij}^k e é definida na equação seguinte (BHAT; MILLER, 2002).

$$P_{ij}^k = P(p_k = i | p_0 = j) \quad (5.13)$$

O conjunto das probabilidades condicionais assim definidas para todos os possíveis estados forma a *matriz de transição* (BENAROYA, 1989). Essa matriz de transição para k passos, P^k , pode ser calculada como a potência k da matriz de transição para passo unitário, P , que é única para todo o comprimento de uma série estacionária.

$$P^k = (P)^k = (P^1)^k \quad (5.14)$$

Para aplicação ao problema da detecção, o método proposto contempla calcular inicialmente a matriz de transição para cada série. O janelamento é realizado com avanço de $1valor = 2min$ e tamanho de janela igual ao valor de memória selecionado, k . Depois da construção da matriz de transição, para cada novo valor de entrada é analisada a cadeia de Markov, isto é, calcula-se a probabilidade condicional desse valor aparecer após o valor precedente que pode ser o imediato ou outro valor anterior, segundo a memória k . Uma probabilidade muito baixa, segundo um limiar α , é interpretada como a uma ocorrência incomum, e detectada como alteração, como indica a expressão seguinte.

$$P(p_i | p_{i-k}) \begin{cases} \geq \alpha \implies \text{negativo: Nenhuma alteração detectada} \\ < \alpha \implies \text{positivo: Alteração detectada} \end{cases} \quad (5.15)$$

Um esquema simplificado para a definição dos estados possíveis do sistema supõe o uso dos valores de pressão medidos. Isto é, a definição de um conjunto de estados discretos com o total de elementos permitido pela resolução do equipamento de aquisição e os valores mínimo e máximo. Por exemplo, para uma resolução de $0,01 mca$ o número total de estados pode estar próximo de 500 e com frequências de ocorrência baixa para cada estado, situação pouco conveniente para as análises. Com isto, outras formas discretização são necessárias, agrupando os valores de pressão em intervalos de tamanho fixo.

As matrizes de probabilidade de transição são construídas a partir das frequências de ocorrência das transições na serie, e como pretendem representar o comportamento do sistema sem alterações são consideradas unicamente as transições de pressão nos momentos da classe "normal". As matrizes de probabilidade são constantes ao longo do tempo se o processo que

descrevem for estacionário, mas as séries adquiridas no estudo de caso têm comportamento não estacionário em alguns trechos, como é apresentado no Capítulo 6. Na aplicação prática, as matrizes de transição poderiam ser recalculadas após uma mudança de comportamento; porém, para conseguir analisar a metodologia com uma quantidade maior de dados é decidido utilizar as séries corrigidas, as quais podem ser consideradas estacionárias (para todos os momentos considerados "normais"), segundo apresentado no Capítulo 6.

5.2.3 Controle estatístico

Originalmente definido para processos industriais de produção nos quais a qualidade do produto está relacionada a algumas características mensuráveis, o *controle estatístico de processos (CEP)* é um conjunto de ferramentas utilizadas para verificar que o valor das características de interesse permanece com pouca variabilidade ao redor do seu valor objetivo (MONTGOMERY, 2009). A partir dessa definição inicial, diversas aplicações e variações têm sido desenvolvidas, incluindo a detecção de novidades em sinais (MARKOU; SINGH, 2003; ROSS et al., 2012), e particularmente a detecção de rupturas em sistemas de distribuição de água (JUNG et al., 2015; ROMANO; WOODWARD; KAPELAN, 2016). Assim, podemos entender a rede de abastecimento como sendo o processo base que é monitorado através das medidas de pressão adquiridas, e esses sinais ou séries adquiridas devem ser acompanhados para detectar se um novo valor está ou não dentro do esperado.

As características acompanhadas nos sinais são, comumente, os seus descritores como a média e o desvio padrão. Estes podem ser acumulados ao longo da série ou, mais interessante para a detecção de novidades em fluxo contínuo de dados, considerar somente um trecho precedente. Considerando que as séries a serem estudadas são altamente auto-correlacionadas, um bom preditor de curto prazo aplicável (MONTGOMERY, 2009) é a *média móvel exponencialmente ponderada, MMEP*, que também foi identificada por Jung et al. (2015) como a de melhor desempenho entre dentre seis técnicas de CEP avaliadas para detecção de rupturas. A detecção está baseada, então, na comparação entre o valor predito com a MMEP e o novo valor adquirido, para cada momento da série.

A MMEP prediz o novo valor como uma média ponderada de um conjunto de valores precedentes, com uma ponderação diferenciada segundo o tempo de ocorrência de cada valor até o momento atual, de forma que o imediatamente anterior têm o maior peso, dado pelo fator λ . A MMEP pode ser calculada incrementalmente, e portanto não é necessário manter em memória os valores adquiridos precedentes ($p_{t < (ti-1)}$) para fazer a predição para ti mas somente o valor anterior p_{ti-1} e a predição anterior MC_{ti-1} , usando expressão seguinte.

$$MC_{ti} = \lambda p_{ti-1} + (1 - \lambda) MC_{ti-1} \quad (5.16)$$

No Capítulo 6 é mostrado que nas séries de interesse a dispersão dos valores é periódica, e não constante ao longo do tempo. Assim, em alguns momentos é comum ter valores mais afastados da média, enquanto em outros esse mesmo afastamento não faz parte do comportamento normal.

Considera-se importante, portanto, que a comparação entre cada novo valor adquirido p_{ti} e o respetivo valor predito MC_{ti} seja feita não em termos absolutos mas do desvio padrão local. O cálculo do desvio padrão (ou da variância, mais diretamente) local, compatível com o da média, pode ser realizado também como móvel e exponencialmente ponderado, complementando a técnica de *média e variância móveis exponencialmente ponderadas*, *MVMEP*. O cálculo, também incremental, da variância estimada para o tempo ti (VC_{ti}) utiliza os valores para o momento anterior de pressão (p_{ti-1}), média (MC_{ti-1}), e variância (VC_{ti-1}), com a equação seguinte (FINCH, 2009).

$$\begin{aligned} VC_{ti} &= (1 - \lambda)VC_{ti-1} + \lambda(p_{ti} - MC_{ti})(p_{ti} - MC_{ti-1}) \\ &= (1 - \lambda) (VC_{ti-1} + \lambda(p_{ti} - MC_{ti-1})^2) \end{aligned} \quad (5.17)$$

A média e variância (e dela o desvio padrão) calculados pela MVMEP são representativos do curto prazo, daí a letra C na nomenclatura utilizada. Com esses valores, a detecção de vazamentos ou outras novidades é feita pela comparação de cada novo valor adquirido com a predição, em termos do desvio padrão, segundo um limiar α , como na Equação 5.18.

$$\frac{|p_{ti} - MC_{ti}|}{\sqrt{VC_{ti}}} \begin{cases} < \alpha \implies \text{negativo: Nenhuma alteração detectada} \\ \geq \alpha \implies \text{positivo: Alteração detectada} \end{cases} \quad (5.18)$$

O janelamento está definido por tamanho e avanço de $1valor = 2min$, salientando que no início e após cada pausa na série (por valor faltante ou por intervenção direta do operador) é definido um trecho de aprendizado sem que a avaliação seja feita, cujo comprimento é inversamente proporcional ao λ utilizado.

A definição do método de detecção anterior não está completa, porém, se não for considerado que, após a detecção, novos valores continuam a ser adquiridos e novas avaliações devem ser feitas, pelo menos durante algum período de tempo até que o operador decida realizar uma intervenção. Mesmo que não discutida na maioria de trabalhos na literatura, a decisão sobre atuação após ocorrência de alterações é fundamental para o sucesso do método. Uma primeira alternativa pode ser não dar nenhum tratamento diferencial e incorporar todos os dados na atualização da MVMEP. Porém, valores irregulares vão levar os descritores a faixas fora do comportamento normal, comprometendo sua utilidade. Outra alternativa é desconsiderar todos os valores identificados como alteração segundo a Equação 5.18, mas essa pode não ser a opção conveniente porque não incorpora nenhuma análise adicional e por ser dependente da escolha do limiar α : um operador pode escolher α que gere algum nível de falsos positivos se prefere decidir uma intervenção somente quando os alarmes repetem-se, ou pode preferir α maior para garantir capacidade de reagir a cada detecção. Assim, e visando a robustez do método, são propostos dois critérios complementares de atualização dos indicadores que utilizam avaliações diferentes à usada para detecção.

A decisão sobre a incorporação ou não do mais recente valor adquirido (e avaliado) na atualização dos indicadores é realizada segundo a estabilidade no curto e no longo prazo.

O primeiro visa excluir valores extremos (*outliers* em escala curta) que podem distorcer os indicadores, definindo para tanto uma máxima distância baseada nos indicadores de MVMEP, de forma semelhante ao método de detecção, mas com um limiar alto e pre-definido β_c . O segundo permite a incorporação da análise em escala diferente para melhorar o desempenho do método, pois utiliza uma escala suficientemente longa para avaliar a estacionariedade da série, melhorando dessa forma a detecção de alterações progressivas.

Para avaliação no longo prazo são definidos os descritores de média (ML) e variância (VL) (e desvio padrão) calculados sobre toda a série desde o início ou re-início após um dado faltante. Embora corresponde com o cálculo exato dos descritores para a série completa desde re-início até o momento atual, o cálculo proposto não requer manter na memória todos os valores (inviável na prática) mas somente três variáveis de acumulação (n , $\sum p$ e $\sum p^2$), aplicando as seguintes equações (GAMA; GABER, 2007).

$$ML_{ti} = \frac{p_{ti} + (ti - 1)ML_{ti-1}}{n_i} \quad (5.19)$$

$$VC_{ti} = \frac{\sum_{j=1}^i p_{tj}^2 - \left(\sum_{j=1}^i p_{tj}\right)^2 / n_i}{n_i - 1} \quad (5.20)$$

sendo n_i o número de elementos que têm sido incorporados na atualização, desde o início ou re-início até o momento atual ti . Cada novo valor é comparado com a média de longo prazo, segundo o desvio padrão de longo prazo, e se supera um limiar β_l não é aceito para atualização. Assim, cada novo valor de aquisição, após passar pela análise para detecção é avaliado segundo os critérios de curto e longo prazo, e somente se cumpre os dois é utilizado para atualizar os indicadores. Caso o novo valor descumpra um dos critérios, os valores dos quatro indicadores (MC , VC , ML , VL) para o momento atual ti são mantidos idênticos aos do momento anterior $ti - 1$, como resumido na Equação 5.21. Deve ser considerado também que o trecho inicial após início ou re-início é somente de aprendizado, e portanto todos os valores são usados para atualização sem realizar avaliação.

$$\frac{|p_{ti} - MC_{ti}|}{\sqrt{VC_{ti}}} < \beta_c \quad \wedge \quad \frac{|p_{ti} - ML_{ti}|}{\sqrt{VL_{ti}}} < \beta_l \quad \begin{cases} V \Rightarrow \text{Fazer atualização} \\ F \Rightarrow \text{Não atualizar} \end{cases} \quad (5.21)$$

5.2.4 Comparação com padrão

Enquanto nos dois métodos apresentados nas seções anteriores o aprendizado do comportamento "normal" das séries é realizado somente sobre os mesmos dados, é também possível incorporar na análise o conhecimento que existe sobre o sistema físico e as características que origina. Pelo conhecimento do funcionamento dos sistemas de abastecimento de água na região, e principalmente pelas conclusões obtidas da descrição das séries, apresentadas no Capítulo 6, sabe-se que as séries têm comportamento periódico, com dois ciclos principais bem definidos:

diário e semanal. Essa informação adicional é útil para a detecção, pois constitui conhecimento bem definido do comportamento que pode ser utilizado para modelar o comportamento normal.

Aplicando o anterior, o método de detecção de vazamentos ou outras alterações proposto nesta subseção utiliza o comportamento médio e normal do sistema, representado na série de pressões semanal média (ver Capítulo 6) relativa a cada um dos nove sensores, e compara com ele os mais recentes dados adquiridos. Para aproveitar essa comparação, a escala de aplicação deste método deve ser maior à dos anteriores, com janelamento na ordem de horas. A cada nova janela a comparação é feita com o trecho do padrão do mesmo tamanho da janela e relativo ao mesmo dia da semana e horários. Essa comparação produz um valor de semelhança que deve ser confrontado com um limiar α para classificar os dados como indicativos de alteração ou não.

A partir dessa definição geral muitas análises podem ser feitas, e deve ser decidido o quê comparar e como compará-lo, isto é, qual a representação mais útil dos dados adquiridos ou função de transformação (*trans*), e qual a medida de semelhança com o padrão ou função de comparação (*comp*). Assim, para uma janela p_j de tamanho w , constituída pelo trecho $\{p_{tj-w} \dots p_{tj}\}$ da série, o trecho do padrão correspondente é *padrao_j*, e a detecção é resumida na expressão seguinte,

$$comp(trans(p_j), trans(padrao_j)) \begin{cases} < \alpha \implies \text{negativo: Nenhuma alteração detectada} \\ \geq \alpha \implies \text{positivo: Alteração detectada} \end{cases} \quad (5.22)$$

Em relação à detecção, a definição anterior não é completa pois deve ser especificado como uma decisão definitiva é tomada a respeito de cada valor da série recebido, considerando a sobreposição de janelas. Este aspecto é apresentado em uma próxima subseção, após as apresentações relativas às funções de transformação e de comparação analisadas.

5.2.4.1 Funções de transformação

É comum nas aplicações de análise de sinais que a comparação direta dos sinais (séries) adquiridos seja insuficiente para a obtenção de resultados satisfatórios, seja pela interferência que causam informações contidas no sinal original que são desinteressantes para a análise, pela sua alta dimensionalidade, pela dependência entre seus valores, ou pela incompatibilidade entre a distribuição das informações (no espaço original, pressão-tempo) e o tipo de separação buscada pelos métodos de detecção. Para superar esses problemas é comum a transformação dos sinais originais do espaço pressão-tempo para espaços construídos, de forma que fiquem representados por conjunto de valores (usualmente de tamanho menor ao original) que ofereçam melhores resultados nas análises posteriores. Tal estratégia é denominada extração de características (ALLEN; MILLS, 2003), e no método de detecção aqui apresentado corresponde à aplicação das funções de transformação.

Dentro do método proposto, a transformação ou extração de características é aplicada a cada nova janela de dados adquiridos, e também ao trecho do padrão que corresponde ao

tamanho, dia e horário da janela sob análise. Após a transformação independente das duas séries, dois novos vetores passam a representá-las na etapa seguinte, a de comparação.

São muitas as características que podem ser extraídas e a sua escolha, que depende do objetivo da análise, pode ser feita semi-automaticamente através de metodologias sofisticadas (como *feature learning*) ou explicitamente pelo analista (GAMBOA-MEDINA; REIS; GUIDO, 2014). Seguindo a segunda abordagem de escolha das características e considerando que o objeto da pesquisa é diferenciar os trechos de séries que contém vazamentos ou outras alterações daqueles normais, são analisados 12 tipos de transformação ou extração (ou não) de características, descritas a seguir.

a) Não transformar - *Nothing*

Esta alternativa refere-se a não aplicar nenhuma transformação ou extração de características, e avançar à etapa de comparação com as séries de pressão adquiridas e o padrão também em termos da pressão, em mca.

É importante considerar esta alternativa como base para a avaliação das restantes, pois todas visam que o desempenho melhore graças à sua aplicação. Adicionalmente, não se deve rejeitar a possibilidade de que os melhores resultados sejam, de fato, atingidos quando nenhuma transformação é feita.

b) Decomposição *wavelet* - *wavedecom_all*, *wavedecom_detail*, *wavedecom_aproximation*

Uma técnica muito versátil para análise de sinais é a transformada de *wavelet*, que gera uma representação dos sinais a respeito de uma família de sinais ou funções elementares (funções *wavelet*) localizadas tanto em tempo quanto em frequência. Essa característica de localização é importante ao permitir a análise simultânea das características segundo o tempo e das frequências presentes no sinal, uma forma de análise tempo-frequência ou, mais precisamente, tempo-escala (MALLAT, 1999).

A decomposição *wavelet* discreta, por sua vez, permite a representação do sinal em diferentes escalas, relacionadas com as frequências presentes nele, e sem perder a estrutura temporal. Ao executar a decomposição são produzidas inicialmente duas componentes: a de altas frequências ou detalhes, e a de baixas frequências ou aproximação, e posteriormente essa componente de baixa frequência é de novo decomposta em duas partes, num processo que continua até o nível de decomposição (escala) de interesse (ADDISON, 2002). As componentes podem ser analisadas como a sequência de coeficientes que resultam do cálculo da decomposição, ou como a reconstrução do sinal obtida a partir deles.

Cada componente é uma sequência temporal de valores relativos à mesma duração que o sinal original, mas com menor número de elementos, de forma que quanto maior o nível de decomposição pior a resolução temporal das componentes, e no limite a componente de aproximação terá somente dois elementos de forma que não será possível continuar a decomposição. Por outro lado, o significado de "baixa" ou

"alta" frequência é relativo às frequências no sinal original e ao nível de decomposição, sendo que em cada separação de componentes os detalhes correspondem à metade superior das frequências originais e a aproximação à metade inferior. Assim, um maior nível de decomposição equivale ao estudo em escala menor, e permite melhor detalhamento de frequências que não são as mais altas do sinal. O nível de decomposição deve ser escolhido buscando um equilíbrio entre resolução de frequências e temporal.

Análises preliminares permitiram decidir a utilização da função de decomposição ou "wavelet mãe" Daubechies-2 e nível de decomposição quatro (4). Como resultado dessa decomposição, obtém-se para cada trecho de série x quatro vetores, sendo três das componentes de detalhe e um da componente de aproximação. A partir desses resultados são definidas três representações alternativas, cada uma procurando explorar o significado da decomposição:

- *wavedecomp_all*. Matriz de três colunas com as componentes de aproximação do nível 4, e detalhes dos níveis 2 e 1, respetivamente.
- *wavedecomp_detail* Matriz de duas colunas com as componentes de detalhe dos níveis 1 e 2.
- *wavedecomp_aproximation* Vetor com a componente de aproximação nível 4.

c) Características baseadas na energia - *energy_resolutions*, *energy_proportional*

Uma das características mais simples, comumente utilizadas e úteis para análise que podem ser extraídas do sinal é a energia (PROAKIS; MANOLAKIS, 2007; ALLEN; MILLS, 2003), definida para o trecho de série x de comprimento n como:

$$E(x) = \sum_{i=1}^n x_i^2 \quad (5.23)$$

Com base na energia vários vetores de características da série podem ser extraídos, e foram escolhidos para a pesquisa duas das propostas de Guido (2015), ambas relativas ao estudo da distribuição da energia ao longo da série, isto é, a concentração de energia em função do tempo.

- *energy_resolutions*. Este algoritmo, chamado de algoritmo a2 por Guido (2015), requer dividir cada sinal (ou trecho) em duas janelas retangulares sem sobreposição, e calcular a energia em cada janela, produzindo assim um subvetor de características. Posteriormente o mesmo sinal é dividido em cinco janelas e obtido outro subvetor, e o processo repetido usando um número primo de janelas até o máximo definido pelo analista (necessariamente menor que o tamanho do sinal). O vetor de características do sinal será, finalmente, a concatenação dos subvetores obtidos com os diferentes janelamentos.

- *energy_proportional*. O algoritmo a3, como denominado pelo mesmo autor, consiste em calcular a energia total do sinal e, posteriormente, avaliar a porcentagem de elementos do sinal (desde seu início) necessários para atingir uma porcentagem $c = 100/k$ da energia total. Depois, avalia-se a porcentagem de elementos para atingir $2c$ da energia total, depois $3c$ e continua até $kc = 100\%$. O vetor de características do sinal é formado pelas k porcentagens de elementos que correspondem com as porcentagens definidas da energia do sinal.

d) Características baseadas no número de cruzamentos por zero - *zcr_resolutions*, *zcr_proportional*

Uma característica que pode ser extraída do sinal e que traz informações relacionadas com o comportamento espectral (GUIDO, 2016b), é o número de cruzamentos por zero (*zcr* pela sigla em inglês). O *zcr* de uma serie ou trecho x está definido como o número de vezes em que um elemento positivo é seguido por um negativo, ou um elemento negativo é seguido por um positivo, na série transladada $x' = x - \bar{x}$.

$$zcr(x) = \sum_i^n \Theta(-1 \cdot x'_i \cdot x'_{i-1}) \quad (5.24)$$

onde Θ é a função de Heaviside, definida na Equação 5.5. Dois algoritmos análogos aos apresentados no item anterior são também aplicados a partir do *zcr*.

- *zcr_resolutions* Com o algoritmo b2 (segundo denominação de Guido (2016b)) é obtida a concatenação de valores de *zcr* para o sinal janelado em diferentes tamanhos.
- *zcr_proportional* O algoritmo b3 (na denominação do mesmo autor) permite obter um vetor com as porcentagens do tamanho que correspondem com porcentagens fixas do *zcr* total do sinal.

e) Características baseadas no comprimento - *discretelength_pl*, *discretelength_zcr*

A transformada discreta de comprimento (*Discrete length transform - DLT*) (GUIDO, 2016a) cria uma representação do sinal baseada no "trabalho" discretizado, interpolando entre elementos sucessivos. Numa outra interpretação, a DLT é basicamente o comprimento normalizado até cada um dos elementos do sinal, considerando sua sequência.

O resultado da DLT é um vetor com os valores de comprimento normalizado acumulado até cada elemento do sinal. Tal novo vetor, ou sinal transformado, tem o mesmo tamanho que o sinal original, e é a base para extração de duas características propostas por Guido (2016a).

- *discretelength_pl* A característica de comprimento proporcional é expressa em um vetor contendo os valores de comprimento normalizado acumulado até

porcentagens fixas do tamanho do sinal. Sua definição é muito semelhante à de *energy_proportional* e *zcr_proportional*, mas aplicada ao sinal transformado.

- *discretelength_zcr* A evolução do número de cruzamentos por zero, *zcr*, ao longo do sinal transformado, é avaliada para esta característica. O vetor resultante corresponde ao *zcr* acumulado até frações definidas do sinal.

5.2.4.2 Funções de comparação

A mais recente janela da série adquirida e transformada deve ser comparada com o padrão (também transformado) para medir a sua "semelhança", e há diversas opções para essa quantificação, que incluem medidas de distância e de similaridade. Todas elas comparam dois trechos de séries transformadas x e y (ou z), um extraído da série adquirida e outro do padrão, ambos de tamanho n , e que podem ser entendidos como dois pontos n -dimensionais. São distâncias entre os trechos todas as funções $d(x, y)$ que cumpram as três primeiras das condições a seguir. Se adicionalmente a quarta condição é cumprida, $d(x, y)$ é uma métrica (DEZA; DEZA, 2016).

- Não negatividade: $d(x, y) \geq 0$
- Simetria: $d(x, y) = d(y, x)$
- Reflexividade: $d(x, x) = 0$
- Desigualdade triangular: $d(x, y) \leq d(x, z) + d(z, y)$

As similaridades são funções $s(x, y)$ que cumprem as duas primeiras características das distâncias e também (DEZA; DEZA, 2016):

- $s(x, y) \leq s(x, x)$

Para a Equação 5.22, a função de comparação (*comp*) é considerada uma função de distância, pois quanto maior semelhança ou "proximidade" há entre os trechos de sinal menor será o valor resultante. Se, pelo contrário, a função de comparação é uma função de similaridade, o sentido da comparação com α seria contrario, pois maior proximidade está relacionada com valores maiores de similaridade. Como correção, para aplicar a mesma Equação 5.22 o resultado da comparação quando usadas funções de similaridade é multiplicado por -1 . A seguir são apresentadas as cinco funções de comparação analisadas.

a) Distância euclidiana - *euclidean*

Considerando o trecho da série transformado como o vetor $x = x_1, x_2, \dots, x_n$ e o trecho do padrão relativo o vetor $y = y_1, y_2, \dots, y_n$, a distância euclidiana é definida

por (DEZA; DEZA, 2016)

$$\begin{aligned} compare(x, y) &= eucl(x, y) \\ &= \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} \end{aligned} \quad (5.25)$$

Uma característica importante desta distância é que a comparação é feita exclusivamente ponto a ponto entre os valores da série transformada, e o resultado é um único valor que representa as comparações em todo o espaço n-dimensional. Se o espaço dos vetores for homogêneo, como no caso das sequência de valores de pressão sem transformar, a distância terá as mesmas unidades de todas as componentes. Se, pelo contrário, o espaço criado pela transformação das séries incorpora diferenças de ordem de grandeza entre os elementos de cada vetor, a comparação de elementos de grandeza maior terá uma influencia mais importante que a de elementos de menor grandeza.

b) Correlação cruzada - *correlation*

A similaridade de correlação é conhecida como coeficiente de correlação produto-momento, coeficiente de correlação de Pearson ou simplesmente ρ de Pearson (DEZA; DEZA, 2016). Com ela quantifica-se a semelhança do comportamento (ou "formato") do trecho com o padrão, mas não seus valores absolutos, e inicialmente ambos vetores x e y são deslocados subtraindo de cada um sua média, para obter x' e y' com média zero.

O resultado da correlação é um valor de -1 a 1 que quantifica a similaridade das sequências de valores x' e y' , sendo que 0 indica nula correlação ou não dependência linear entre elas, 1 é a máxima similaridade, com correlação ou dependência completa, e -1 é correlação negativa, quando todo incremento de x' implica diminuição de y' .

Levando em conta que a sincronia entre o trecho sob análise e o padrão pode não ser perfeita, é avaliada a correlação de x' com versões de y' deslocadas $\pm j$, $y'_{j,i}$, e j de 0 até 5 valores (se não for aplicada previamente nenhuma transformação, esse deslocamento máximo corresponde a 10 minutos). Como resposta final é escolhido o maior valor de correlação obtido dentre todos os deslocamentos avaliados. O cálculo é feito segundo a seguinte expressão

$$\begin{aligned} compare(x, y) &= -1 \cdot corr(x, y) \\ &= \underset{j \in \{0:5\}}{\text{maximum}} \frac{\sum_{i=0}^n x'_i y'_{j,i}}{\sqrt{\sum_{i=0}^n (x'_i)^2} \sqrt{\sum_{i=0}^n (y'_{j,i})^2}} \end{aligned} \quad (5.26)$$

c) Correlação cruzada da envoltória - *envcorrelation*

Em várias aplicações a análise de sinais não é realizada diretamente nos dados originais, mas sim usando uma envoltória. No caso do cálculo da similaridade de correlação essa estratégia também pode ser útil, considerando que os sinais podem

ter oscilações de alta frequência que gerem indesejadas alterações nos resultados, além da possibilidade de que parte importante da informação esteja sendo carregada pela envoltória.

Uma técnica poderosa que permite, dentre outras aplicações, obter a envoltória de um sinal, é a transformada de Hilbert. A transformada de Hilbert de um sinal (em função do tempo) $x(t)$, representada por $X_H(t)$, é calculada como (GHAZALI et al., 2012)

$$X_H(t) = \frac{PV}{\pi} \int_{-\infty}^{\infty} \frac{x(\tau)}{t - \tau} d\tau \quad (5.27)$$

sendo que PV representa o *valor principal de Cauchy*. Com o sinal original ($x(t)$) e transformado ($X_H(t)$) é possível definir um sinal analítico complexo $X_A(t)$, e com ele a envoltória $|X_A(t)|$, conforme as seguintes equações (ALLEN; MILLS, 2003).

$$\begin{aligned} X_A(t) &= x(t) + jX_H(t) \\ |X_A(t)| &= \sqrt{x(t)^2 + X_H(t)^2} \end{aligned} \quad (5.28)$$

Após obter a envoltória, tanto para o trecho do sinal a ser comparado quanto para o trecho do padrão, é realizado o cálculo da similaridade de correlação, segundo apresentado no item anterior.

$$compare(x, y) = -1 \cdot cor_env(x, y) = -1 \cdot corr(|X_A|, |Y_A|) \quad (5.29)$$

d) Distância dos centroides - *centroids*

Outra das formas de completar a comparação entre os trechos de séries é o tratamento deles como dois conjuntos x e y de n dados cada um, sobre os quais podem-se aplicar as medições de semelhança básicas da análise de agrupamentos (FACELI et al., 2011). A distância entre os centroides é uma métrica para quantificar a separação entre os conjuntos de dados, e no caso das séries equivale à distância entre as médias, calculada como segue (DEZA; DEZA, 2016)

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^n x_i}{n} & \bar{y} &= \frac{\sum_{i=1}^n y_i}{n} \\ compare(x, y) &= cent(x, y) = eucl(\bar{x}, \bar{y}) \end{aligned} \quad (5.30)$$

e) Distância de Hausdorff - *hausdorff*

A semelhança de dois conjuntos de dados é quantificada pela distância de Hausdorff, de forma que a comparação não é feita somente entre os pares de valores correspondentes ao mesmo elemento i do trecho, mas com todos os valores de cada conjunto x e y . Para cada elemento de x é feita a busca do mais próximo ponto em y . A distância é calculada então como a maior das distâncias obtidas após encontrar todos os pontos mais próximos, de acordo com a equação seguinte (DEZA; DEZA, 2016).

$$\begin{aligned} hau_1(x, y) &= \max_{a \in x} \min_{b \in y} eucl(a, b) \\ hau_1(y, x) &= \max_{b \in y} \min_{a \in x} eucl(b, a) \\ compare(x, y) &= haus(x, y) = \max(hau_1(x, y), hau_1(y, x)) \end{aligned} \quad (5.31)$$

5.2.4.3 Inferência para detecção

Devido ao sistema de janelamento com sobreposição (avanço < tamanho de janela), o valor correspondente a um momento ti é parte de várias janelas, aparecendo na parte inicial de uma janela e posteriormente nos trechos médios e final das seguintes. A quantidade de janelas em que cada momento da série é considerado (*visit*) pode ser calculado com a equação seguinte, válida para trechos diferentes aos primeiros ou últimos antes e depois de um re-início.

$$visit = \left\lceil \frac{\text{tamanho de janela}}{\text{avanço}} \right\rceil \quad (5.32)$$

No método proposto, para cada janela é realizada a transformação e comparação com o padrão, obtendo um resultado de acordo com a Equação 5.22, e essa resposta única aplica para todos os elementos da janela. Em consequência, após serem incorporados dados posteriores suficientes para que todos os elementos na mais recente janela correspondam com $t > ti$, o valor relativo a ti estará associado a várias (*visit*) respostas sobre detecção.

Essas respostas devem ser unificadas e, para tanto, é aplicada uma abordagem derivada do aprendizado por agrupamento (*ensemble learning*) (RUSSELL; NORVIG; DAVIS, 2010), combinando as respostas por votação com maioria simples. Assim, se a maioria das respostas para o momento ti são positivas para alteração, então o valor adquirido para o momento ti é detectado como parte de uma alteração, se essa maioria não é atingida a conclusão sobre o valor é que nenhuma alteração é detectada.

No caso dos momentos da série próximos do início, de interrupções (valores faltantes), ou de re-inícios, não serão completadas as *visit* respostas, porém a votação também é realizada com as respostas disponíveis, obtendo a conclusão por maioria simples.

6 ANÁLISE DAS SÉRIES

Com a base de dados completa adquirida segundo apresentado no Capítulo 3, e em particular os sinais ou séries de pressão medidos nos nove pontos escolhidos do sistema, são realizadas todas as análises. Os métodos descritos no Capítulo 5 são aplicados a cada uma das séries, e os resultados obtidos são apresentados e discutidos neste capítulo.

Durante o desenvolvimento da pesquisa foi comprovado que o conhecimento detalhado das séries analisadas é fundamental para orientar as escolhas e entender os resultados relativos ao problema de detecção. Portanto, a primeira seção está dedicada à descrição das séries e ao seu pré-processamento. As três secções posteriores apresentam a aplicação dos três métodos propostos para detecção de vazamentos e outras anomalias nas séries.

6.1 Descrição das series adquiridas

As séries de pressão adquiridas em nove pontos seleccionados no setor da rede de abastecimento são nesta seção descritas segundo diferentes abordagens. Completando o inicialmente descrito no capítulo 3, a classificação manual é a primeira a ser apresentada, por ser a etapa na qual está fundamentada a avaliação de desempenho de todos os métodos de detecção, e em geral todas as análises.

A descrição das séries segue, como as propostas de detecção, enfoque de múltiplas escalas, visando conhecer tanto comportamentos gerais quanto detalhados do problema. São analisadas nesta seção as séries segundo sua estacionaridade e as transformações necessárias para garanti-la, e seguidamente a autocorrelação e a análise espectral. Posteriormente está a descrição estatística e, particularmente, a obtenção do comportamento médio, que será aplicado nos métodos de detecção. A análise de recorrência, apresentada por último, oferece um ponto de vista diferente e importantes informações sobre as séries e sobre o processo que as gera.

6.1.1 Classificação manual

Após a eliminação de erros, foi completada a classificação manual dos dados nas quatro classes definidas segundo as informações coletadas, apresentadas no capítulo 3. As atribuições de classes foram inicialmente realizadas à medida que novos dados eram colocados na base de dados, e após a conclusão do monitoramento todas foram revistas e complementadas quando necessário. A identificação e definição das classes que é utilizada em todo este documento é a seguinte:

- Classe 1 ou "simulado": Dados adquiridos durante os experimentos de simulação de vazamentos ou rupturas, identificados a partir dos horários de realização dos experimentos;

- Classe 2 ou "reporte": Dados adquiridos durante o período entre a notificação de vazamento e seu reparo, segundo os registros do DAAE;
- Classe 3 ou "visual": Dados que apresentam descontinuidades visualmente identificáveis, reconhecidos por inspeção visual dos gráficos de cada uma das séries;
- Classe 5 ou "extremo": Dados maiores ou menores a 99,8% dos valores, e que não foram identificados nas classes anteriores;
- Classe 0 ou "normal": Dados considerados normais, por não se enquadrarem em nenhuma das classes anteriores.

Para o período de monitoramento completo, 46 ocorrências de vazamento foram reportadas ao DAAE e reparadas, na grande maioria dos casos, no mesmo dia. É claro que o reporte somente pode ter acontecido algum intervalo de tempo depois do início do vazamento, mas é impossível estabelecer sequer uma ordem de grandeza para tal defasagem temporal. A hora de finalização da ocorrência é informada pelos profissionais que fazem o reparo, e também não coincide necessariamente com a de retorno ao funcionamento normal da rede. Na impossibilidade de uma melhor estimativa, foram rotulados com a classe 2 os valores adquiridos entre as horas de início e fim descritas nos reportes.

A classe 3 foi atribuída segundo uma cuidadosa inspeção dos gráficos das séries de valores. Foram marcados numerosos trechos com alterações de pressão (aumentos ou decréscimos) que desaparecem após algum tempo curto (minutos a horas), e também trechos de variação mais ou menos rápida que levam a uma mudança no valor médio da série que não retornam ao valor anterior após um período de vários dias. Nesses últimos casos, somente ao trecho da mudança foi atribuída a classe 3, mas não aos posteriores à estabilização da pressão média no novo patamar.

A inspeção foi repetida diversas vezes mudando a escala de observação dos dados e características de visualização, dado que foi comprovado que esses aspectos trazem diferenças na avaliação pelo observador. Como exemplo de tal situação, apresenta-se na Figura 9 um trecho da série 1, visualizado em duas escalas diferentes. Na escala maior (de 263 a 270 dias) pode ser observada uma tendência no aumento da pressão, enquanto na escala menor (de 263 a 264 dias), que é próxima à utilizada durante um monitoramento típico, nenhuma alteração poderia ser acusada.

Salienta-se, por tanto, quão fácil é ter modificada a percepção de um trecho como irregular ou não por pequenas variações na visualização dos dados, mesmo após um tempo de repetição de observações considerado suficiente para um reconhecimento visual das séries. Essa inspeção é importante, tendo em vista que em muitas companhias operadoras das redes de distribuição esse tipo de análises é a única realizada.

Em etapas intermediárias da pesquisa, ao comparar as informações encontrou-se alguma correspondência entre os reportes de reparos (classe 2) e grandes quedas na pressão (classe 3),

Figura 9 – Exemplo do efeito da escala sobre a inspeção para classificação manual das series.
Trecho da serie do sensor 1

e uma hipótese levantada foi que aquelas grandes diminuições fossem causadas pelo início de vazamentos. Porém, tal hipótese foi descartada por duas razões principais: início da queda de pressão posterior ao horário do reporte de reparo, e não ocorrência dessas alterações durante os experimentos de simulação de vazamento. A explicação que consideramos provável é, então, que para os reparos ou outros trabalhos na rede os registros são operados, suspendendo o fluxo em alguns trechos e causando alterações na pressão muito importantes. É preciso notar, porém, que há algumas alterações não justificáveis sob esse ponto de vista, e também reportes de reparo não relacionados com alterações visíveis nos sinais.

A classe 5, dos valores extremos, não foi atribuída considerando a série completa, mas trechos estacionários dela. Assim, uma mesma pressão pode significar um dado extremo num trecho da série mas não em outro, nos casos em que acontece uma mudança na média, que serão analisados na subseção seguinte.

Os experimentos de simulação de vazamentos ou rupturas pela abertura de registros de descarga e hidrantes foram descritos no Capítulo 3, e aos valores adquiridos durante estes é atribuída a classe 1. Percebeu-se que esses trechos têm, em relação aos trechos vizinhos, variações menores às esperadas, quase sempre da mesma ordem de grandeza da variância da série. Devido à idade e tipo de materiais das redes, pode-se supor a presença de vazamentos, uma possível razão para o comportamento descrito. Segundo essa hipótese, enquanto a vazão de saída incrementa-se pelo experimento há uma queda de pressão. Essa queda, entretanto, faz com que a vazão dos

vazamentos preexistentes diminua, provocando um efeito contrário e reduzindo portanto os efeitos do experimento. Não esteve no alcance desta pesquisa a comprovação dessa hipótese, mas isso não invalida as técnicas analisadas para a identificação de alterações, mas faz com que as análises aqui apresentadas sejam mais fieis aos desafios implícitos no desenvolvimento e aplicação de técnicas a sistemas reais.

6.1.2 Estacionaridade

Para os objetivos desta análise é adotada a definição de estacionaridade fraca (Capítulo 5), e uma primeira avaliação é feita a partir dos gráficos dos sinais adquiridos. Na Figura 10 são apresentados os nove sinais, com as cores correspondendo aos rótulos segundo a classificação manual. Cada linha vermelha nos gráficos representa a pressão média diária calculada considerando só os valores classificados como normais. O aspecto diferente do sensor 9 deve-se à menor resolução. Mesmo que a escala não permita identificar detalhes do sinal, uma visão geral é útil, dado que a estacionaridade que será avaliada é a correspondente ao total do sinal adquirido e que representa o comportamento tido como normal. Esse comportamento que se espera estacionário é descrito pelos momentos calculados sobre trechos de duração de pelo menos um dia.

Figura 10 – Visão geral dos sinais adquiridos, e a media diária

Foi combinada uma inspeção visual para identificação de mudanças na média e na dispersão, com o cálculo desses valores para cada uma das series com segmentos fixos sem superposição (ALLEN; MILLS, 2003), repetindo para diferentes tamanhos de janela entre o comprimento total da série e um dia. Irregularidades em escalas menores são muito mais frequentes, e foram analisadas no processo de classificação manual, mas não são de interesse para a análise nesta seção.

É observável que vários dos sinais apresentam mudanças importantes no seu valor médio. Com maior frequência incrementos rápidos de pressão como no dia 47 no sensor 2 e dia 135 no sensor 6, além dos associados a outras alterações: dia 231 nos sensores 2 a 4, e dias 127 e 204 no sensor 5. O único decréscimo de pressão rápido considerável (nesta escala) é no sensor 1 no dia 257, seguido por uma recuperação lenta até voltar ao normal no dia 268. Um decréscimo gradual na pressão aparece para os sensores 2, 3 e 4 desde o dia 175, aproximadamente, e continua até ter uma recuperação rápida no dia 231 retornando aos patamares normais de pressão. Um comportamento irregular, iniciando com um aumento rápido na pressão seguido por diminuição gradual, é observado para o sensor 5 desde o dia 127, e termina com a rápida recuperação no dia 204.

Quanto à dispersão dos valores, não é possível identificar a partir dos gráficos mudanças consideráveis ao longo do tempo em todos os sinais, nem relação direta com os deslocamentos da média. A amplitude de tal dispersão é muito variável entre os diferentes sensores, mas não guarda relação aparente com a grandeza média das pressões. Por exemplo, o maior desvio padrão é o do sensor 8 ($\pm 4, 2m$) cujo valor médio é o segundo menor dentre os sensores ($34, 4m$), enquanto o menor desvio padrão ($\pm 1, 3m$) é do sensor 4 e sua média ($58, 5m$) é a terceira menor. Do cálculo do desvio por janelas temos que ao longo de cada sinal a dispersão não é estritamente igual, mas apresenta oscilações não bem correlacionadas com as mudanças na media. Adicionalmente, considerando uma escala mais curta, em todos os sinais as oscilações de alta frequência mais amplas acontecem quando o valor médio local é mais baixo dentro do ciclo diário, que corresponde ao período de máximo consumo, indicando que a variação na amplitude estaria relacionada ao consumo pelos usuários.

As apreciações anteriores evidenciam a existência de comportamento não estacionário na maioria das séries. A interpretação dos comportamentos não estacionários é discutida segundo o significado físico dos valores e o conhecimento do problema, antes de propor o tratamento adequado das informações.

Uma hipótese a considerar para explicar as mudanças no valor médio da pressão na rede é que são consequência da variação de nível dos reservatórios. Esse nível poderia chegar a oscilar no máximo em $4, 2m$ (sua altura útil), mas nem em casos excepcionais é menor a $1, 0m$ nem maior a $3, 9m$ segundo as medições. A amplitude da oscilação diária do nível é de $1, 4m$ em média, atingindo $2, 0m$ somente seis dias de todo o período de monitoramento, num comportamento representativo da operação conservadora. Adicionalmente, é esperado que a

influência dessas mudanças de nível sobre a pressão seja igual para todos os pontos da rede. Porém, as alterações no valor da pressão média identificadas nas séries são maiores que as mudanças de nível reportadas, e adicionalmente não se apresentam simultaneamente em todos os sensores, nem simultâneas a mudanças visíveis no sinal de nível. Assim, julga-se que o comportamento não estacionário não é gerado por mudanças nos reservatórios. Mudanças a montante dos reservatórios, como na operação da bomba, também não podem ser consideradas causas prováveis, pois seu efeito estaria diretamente relacionado com os níveis.

O surgimento de vazamentos é um fator associado a alterações na pressão, entretanto, de maneira a diminuir e não incrementar a pressão, não explicando tais observações. A diminuição de pressão rápida seguida de recuperação gradual (observada para o sensor 1) também não é compatível com os efeitos de um vazamento. Quanto à diminuição gradual de pressão observada para os sensores 2, 3 e 4, em princípio poderia corresponder a um vazamento por defeito na tubulação que piora com o passar do tempo aumentando a vazão perdida e diminuindo cada vez mais a pressão, até ser reparado após 55 dias e retornar rapidamente ao normal. Essa hipótese ganha maior importância ao verificar que coincidentemente com o retorno das pressões normais, no fim do dia 230, há reporte de um reparo de vazamento, mais especificamente troca de uma luva numa tubulação de grande diâmetro. Adicionalmente, os sensores mais próximos ao local indicado no reporte são os sensores 4, 2 e 3. Para este caso a hipótese do vazamento não intencional e crescente é aceita, e assim identificada para as análises posteriores.

A diminuição gradual na pressão no registro do sensor 5 começa após um aumento rápido de pressão, que não poderia ser explicado por um vazamento. Porém, pode ser imaginado um cenário composto, em que alguma manobra gera aumento da pressão na rede e com esse aumento foi causado um vazamento pequeno mas que cresce com o tempo, até ser reparado 77 dias depois. Diferentemente do caso analisado no parágrafo anterior, nos reportes de reparos não há atividades coincidentes com a data e horário do aumento de pressão original nem do retorno. Portanto, não é possível supor que esse tenha sido o cenário real.

Outro fenômeno que pode trazer alterações na pressão média são alterações locais no funcionamento hidráulico da rede, em particular a operação de registros ou válvulas. Segundo os reportes obtidos sobre reparos, somente metade das alterações identificadas (nesta escala) coincidem aproximadamente com o intervalo de alguma ocorrência registrada, mas não é possível ter certeza que não tenham sido realizadas outras manobras. O aumento rápido do dia 135 visível na série do sensor 6 coincide com um reparo de vazamento reportado, mas não é claro o motivo dessa alteração. Se neste último caso há uma situação semelhante à encontrada para os dias 175 a 231 (vazamento persistente por longo período até ser reparado), a velocidade da diminuição de pressão devia ser muito pequena e o vazamento ter começado antes do período de monitoramento. Não há informação suficiente para confirmar ou negar essa hipótese. Para os aumentos rápidos de pressão identificados nos registros dos sensores 2 e 3 do dia 47 não foi encontrada uma explicação plausível, além de uma operação sobre um registro ou válvula não reportado, ou a

finalização de um vazamento anterior ao início do monitoramento com um reparo não reportado.

A primeira forma de incorporação das considerações anteriores em todas as análises posteriores é garantir que na classificação manual de todos os trechos com mudanças graduais da pressão seja atribuída a classe 3 "visual", em particular ao trecho entre os dias 175 a 231. Adicionalmente, perante a incerteza sobre a origem da maioria das alterações nos sinais que fazem com que não sejam estacionários, e considerando que algumas das análises que podem ser úteis para melhor caracterizá-los carecem de sentido se aplicadas a dados não estacionários, decidiu-se pela construção de series alteradas (corrigidas) para serem usadas em etapas específicas da pesquisa.

Figura 11 – Visão geral dos sinais corrigidos pelo translado da média.

Para os sinais que apresentam deslocamento na média, passando de um valor médio localmente constante (aproximadamente) a outro maior ou menor mas também localmente constante (aproximadamente), os deslocamentos são artificialmente corrigidos. A todos os dados dos trechos com diferente média (identificados pelo procedimento de cálculo dos momentos em janelas) é somada a diferença entre sua média local e a média local do trecho mais longo, que passa a ser a nova média de toda a série. Os trechos com tendência, cuja média não é constante, são desconsiderados nos cálculos, e somente transladados junto aos trechos vizinhos. Por não ter mudanças bem definidas na dispersão dos dados, nenhuma correção é feita nesse sentido.

Em todas análises que usem as series corrigidas essa escolha será enunciada, e onde isso não for explícito as séries originais serão as utilizadas. Na Figura 11 são apresentados os sinais corrigidos.

6.1.3 Autocorrelação

A autocorrelação é calculada, utilizando a Equação 5.1, para cada um dos nove sinais adquiridos durante todo o período de monitoramento (representando 357 dias, com tamanho de 257040 valores cada). As análises apresentadas nesta seção são válidas para atrasos de 25% do tamanho do sinal (recomendados na literatura), sendo que pelo menos até ele mantém-se o comportamento mostrado na Figura 12 e Figura 13, aproximadamente.

Embora seja esperada a diminuição do valor de autocorrelação com o aumento do atraso, o grande tamanho dos sinais analisados faz com que essa diminuição seja muito fraca, e que para atrasos pequenos em comparação com o tamanho total, como os apresentados nas figuras, a diminuição não seja visível. Para todos os correlogramas é conferida uma diminuição rápida da autocorrelação desde o início do gráfico até o primeiro mínimo local (*atraso* $\approx 0,5$ dia) e o posterior aumento até o primeiros máximo local (*atraso* $\approx 1,0$ dia), e esse trecho inicial não está incluído nos comentários seguintes.

Inicialmente, a autocorrelação é calculada para os sinais completos originais, e para eles os correlogramas obtidos, apresentados na Figura 12, evidenciam a presença de comportamento não estacionário para vários dos sensores. Para o sensor 2, e em menor grau para os sensores 1, 3, 4 e 5, a autocorrelação têm uma tendência decrescente, indicadora de comportamento não estacionário (YAFFEE; MCGEE, 2000). Enquanto o valor médio de autocorrelação decai desde aproximadamente 0,6 para atraso de 1 dia até 0 para atrasos de 40 ou 50 dias, as oscilações na autocorrelação mantém-se aproximadamente constantes e próximas de 0,2, 0,3, 0,4 e 0,3 para os sensores 2, 3, 4 e 5, respectivamente.

O sensor 6 mostra uma fraca tendência de decréscimo da autocorrelação média, em geral oscilando entre $-0,1$ e $0,5$ para todos os atrasos maiores a 1 dia. Os sensores 7 a 9 mostram autocorrelação recorrente para os atrasos desde 1 dia, com valores entre $-0,3$ e $0,5$ para o sensor 7, e $-0,5$ e $0,8$ para os 8 e 9.

Uma característica comum em todas as séries, mais clara em algumas delas, é a periodicidade dos valores de autocorrelação, com máximos locais bem definidos aparecendo a cada 720 (aproximadamente) valores de atraso, correspondendo a 1 dia. Mínimos locais aparecem também com recorrência de 1 dia, e aproximadamente centrados entre os atrasos dos máximos locais.

Sabe-se, da secção anterior, que os sinais originais têm incorporados trechos de comportamento não estacionário, incluindo um vazamento de tamanho crescente atuando por um tempo prolongado, além de várias outras alterações. É esperado que essas alterações tenham efeito sobre a autocorrelação, e também que, se as correções aplicadas foram suficientes para

Figura 12 – Correlograma das séries de pressão originais, completas, para as séries dos nove sensores.

restaurar a estacionaridade, a autocorrelação das séries corrigidas o evidencie.

É necessário considerar que as situações com alteração na rede necessariamente se afastam em algum grau do comportamento normal, que é o que se estima estacionário. Dessa forma, o cálculo das autocorrelações é repetido para as séries corrigidas, e considerando somente os valores de pressão de momentos que receberam o rótulo "normal" na classificação manual. Valores não normais são desconsiderados sem alterar a sequência dos dados, evitando descaracterizar a dependência temporal entre os valores válidos. Os resultados são apresentados na Figura 13.

Os correlogramas produzidos a partir dos sinais corrigidos não têm as alterações ou tendências observadas nos relativos aos sinais originais, mantendo para os 9 sensores um comportamento estável no tempo. A principal característica dos correlogramas obtidos, além da sua estabilidade para os atrasos considerados, é a recorrência claramente definida.

Além da recorrência dos máximos e mínimos com período de 720 (1 dia) já descrita, são observadas oscilações num período maior. Uma envoltória dos correlogramas, especialmente dos máximos, evidencia uma oscilação regular com período de 7 dias. Assim, a existência de periodicidade tanto diária quanto semanal nos sinais adquiridos fica comprovada. Tal comportamento é compatível com o comportamento do consumo esperado, e comprovado também pela análise espectral.

Figura 13 – Correlograma das séries de pressão corrigidas (média transladada), considerando somente a classe "normal", para as séries dos nove sensores.

6.1.4 Análise espectral

Como verificação do comportamento periódico dos sinais adquiridos foi realizada a análise espectral, usando a *Transformada de Fourier discreta* para identificar as frequências mais importantes presentes em cada sinal.

O algoritmo de cálculo de *Transformada Rápida de Fourier* (*fft*, em inglês) foi aplicado aos sinais adquiridos. Como no cálculo da *fft* não podem ser analisados corretamente dados faltantes, de cada série são escolhidos os trechos sem descontinuidade e o maior deles é utilizado na transformação. A máxima frequência detectável, em razão da taxa de amostragem, é de $0,25min^{-1}$ ou $360dia^{-1}$ ou $2520semana^{-1}$. A frequência mínima depende também do comprimento do trecho de sinal utilizado para a transformação, e para os sinais originais é de aproximadamente $0,13semana^{-1}$, ou $0,02dia^{-1}$, com exceção do sensor 9 cuja frequência mínima é $0,08semana^{-1}$.

Com os resultados da *fft* é calculada a densidade espectral de energia, e ao analisar o gráfico dessa densidade ou potência em função da frequência constata-se que os maiores valores estão concentrados nas frequências baixas. Na Figura 14 são apresentadas os gráficos resultantes para cada um dos sensores, analisando os trechos escolhidos dos sinais, e apresentando somente as frequências baixas. O eixo horizontal foi re-escalado para unidades de frequência convenientes,

de semana^{-1} .

Figura 14 – Análise espectral dos sinais de pressão originais, para as series dos nove sensores.

Vários dos picos apresentados no gráfico coincidem em todos ou na maioria dos sensores, indicando que correspondem a frequências originadas no processo fundamental do funcionamento da rede, e não a efeitos locais. O maior dos picos corresponde, para todos os sinais, à frequência de 7semana^{-1} , isto é, que acontece sete vezes na semana, ou a cada dia. A existência do ciclo diário, aqui confirmada, coincide com o esperado pelo consumo e o identificado pela autocorrelação. Outro pico a ser levado em consideração é o de frequência 1semana^{-1} ($0,14\text{dia}^{-1}$), que aparece para todos os sinais exceto o 6. Esse pico bem definido confirma a existência de periodicidade semanal, observada também na autocorrelação.

São também notáveis, mesmo que de grandeza menor, os picos correspondentes a valores inteiros de 2 a 6semana^{-1} , particularmente claros no sinal 9, e que indicam periodicidade intermédia entre a diária e a semanal. Para frequências maiores à do período diário há também componentes importantes, para a frequência 14semana^{-1} , e em menor medida para 21semana^{-1} , os picos indicam ciclos que se repetem duas (e três) vezes a cada dia. Essa informação de ciclos de maior frequência será útil em análises posteriores. Picos de grandeza importante também aparecem para vários sensores com frequências 6 e 8semana^{-1} , e de 13 e 15semana^{-1} , aparentemente alterações ao redor dos picos principais.

Para todos os sensores, as densidades para frequências maiores às apresentadas na figura

são de ordem de grandeza inferior e, em geral, pode-se supor que sua presença no sinal é devida às oscilações do consumo e manobras na rede. Já as mudanças entre um e outro sensor nas frequências mais baixas podem estar relacionadas com os trechos não estacionários. No intuito de tirar o efeito da não estacionaridade local, a análise espectral é repetida para os sinais corrigidos.

As restrições sobre os dados a ser considerados fazem com que os trechos contínuos sejam de comprimento consideravelmente menor que nos sinais originais, para todos os sensores. Enquanto a frequência máxima continua sendo 2520semana^{-1} , o tamanho dos trechos de sinal transformados modifica as frequências mínimas detectáveis, que passam a estar entre $0,4$ e $0,6\text{semana}^{-1}$. Frequências baixas, mesmo superiores a esse mínimo teórico, têm suas densidades diminuídas pelas poucas ocorrências contidas no sinal, de forma que tendem a desaparecer. Os resultados das análises a partir dos sinais corrigidos, retirando os trechos com grandes alterações (classe 3, "visual"), são apresentados na Figura 15.

Figura 15 – Análise espectral dos sinais de pressão corrigidos (médias transladadas), desconsiderando os da classe 3 "visual", para as series dos nove sensores.

O efeito do tamanho menor dos sinais transformados é claramente observado nos baixos valores de densidade espectral encontrados para frequências uma vez que, para alguns sinais, é apagado o pico correspondente ao período semanal. À frequência de 7semana^{-1} continua correspondendo o pico maior, confirmando a importância do período diário, enquanto o segundo pico mais alto é para 14semana^{-1} . Na maioria dos sinais também há um pico notório para a frequência 3semana^{-1} , ocorrências que existem três vezes na semana. O efeito observado de

desconsiderar os trechos não estacionários está na diminuição relativa dos picos para frequências próximas mas diferentes das principais, antes descritas. Nas frequências mais baixas esse efeito está combinado com o do tamanho das séries, dificultando a análise.

Para comprovar o efeito do comprimento do trecho nas frequências baixas, foi repetida a análise espectral com os sinais originais, mas limitando o máximo comprimento de trecho àqueles utilizados na análise dos sinais corrigidos. Salienta-se que não se trata, necessariamente, dos mesmos trechos do sinal, mas de trechos com igual comprimento. Os resultados para os sinais originais com trechos curtos assim definidos também apresentam uma grande diminuição na densidade espectral de potência para as baixas frequências, diminuindo ou eliminando o pico para período semanal (frequência 1) em oito dos nove sinais. Esse comportamento sugere que grande parte das diferenças observadas entre as Figura 14 e Figura 15 para frequências menores que $3semana^{-1}$ são devidas aos tamanhos de trecho mais do que à existência da não estacionaridade. Com isso não se pretende concluir que a não estacionaridade não tenha efeitos, mas que estes não podem ser analisados adequadamente pelas limitações metodológicas desta avaliação.

6.1.5 Descrição estatística

Como parte da descrição das séries de pressão adquiridas, seus valores podem ser analisados como sendo variáveis aleatórias para calcular os descritores estatísticos de interesse, em particular os primeiros dois momentos estatísticos. A tendência central e a dispersão de cada uma das séries são calculadas pela média e o desvio padrão (escolhido ao invés da variância, pela facilidade na interpretação ao ter as mesmas unidades da variável), respectivamente, e os resultados estão na Tabela 4.

A descrição pode ser feita sobre os sinais originais completos (coluna 2 da tabela), e nesse caso os resultados estão influenciados pela existência de não-estacionaridade, descrita em subseções anteriores, como também por todas as alterações que se apresentaram durante o monitoramento. Se desconsiderados os dados adquiridos em instantes identificados com alguma alteração na classificação manual, e utilizados unicamente os da classe "normal", as médias têm um aumento de até 2% enquanto o desvio apresenta diminuição de desde 3% até 42% para os diferentes sensores (coluna 3 da tabela). O aumento da média de todos os sensores ao considerar somente dados normais é indicativo de que a maior parte das alterações geram diminuição (mais do que aumento) da pressão, além de incrementar o desvio, como esperado.

Os resultados ao utilizar as séries corrigidas, com trechos transladados para atingir o mesmo valor de média, são apresentados na coluna 4 da Tabela 4. Para os sensores 7, 8 e 9, não há mudança pois não é aplicada correção neles. O efeito das correções (para sensores 1 a 6) é observado no valor da média obtida, que aumenta até 6%, e também nos desvios calculados que diminuem entre 4% e 27% em relação aos obtidos para as séries originais completas. Se nas séries corrigidas consideramos somente os valores classificados como "normal", os resultados

(apresentados na coluna 5 da tabela) mostram para cada um dos sensores a maior média e o menor desvio, dentre as quatro versões de séries consideradas.

Tabela 4 – Media e desvio das séries de pressão adquiridas

sensor	original todos	original normal	corrigido todos	corrigido normal
1	72,42±1,66	72,65±0,96	72,58±1,23	72,59±0,95
2	52,45±2,43	53,07±1,75	53,74±1,81	53,79±1,32
3	64,54±3,24	65,54±2,17	65,31±2,63	65,47±2,08
4	58,47±1,25	58,74±1,08	58,17±1,14	58,37±0,99
5	43,07±3,40	43,93±2,95	46,02±2,45	46,06±1,96
6	40,83±1,39	40,93±1,30	40,97±1,34	41,08±1,23
7	44,66±2,34	44,84±2,19	44,66±2,34	44,84±2,19
8	34,36±4,24	34,75±4,10	34,36±4,24	34,75±4,10
9	21,77±3,39	22,18±3,18	21,77±3,39	22,18±3,18

As análises para detecção são realizadas, em geral, para os sinais originais, descritos pelos valores indicados nas colunas 2 e 3 da Tabela 4. Quando utilizados os sinais corrigidos, cujos descritores estão nas colunas 4 e 5 da tabela, essa escolha será informada.

6.1.6 Comportamento médio

Como esperado pela tendência de consumo de água, e provado pela autocorrelação e análise espectral, as séries de pressão apresentam periodicidade diária e semanal, principalmente. Assim, para algumas das análises posteriores a semana é considerada a duração de um ciclo completo. Pelo anterior, é de interesse analisar os dados adquiridos segundo o instante da semana que representam, e em particular conhecer qual o funcionamento normal do sistema para este ciclo.

Para descrever o comportamento normal são analisadas os valores das séries corrigidas que receberam a classe "normal", e seus valores são interpretados como ocorrências das variáveis aleatórias definidas segundo o instante da semana correspondente, descritas pela sua média e variância. Foi percebido, porém, que a incerteza na classificação manual, em particular quanto ao momento exato de início ou fim de uma alteração de curta duração, faria com que fossem incorporados valores que não eram realmente representativos do comportamento normal. Como solução a essa limitação foi decidido inicialmente selecionar para todas as séries os dias completos em que todos seus valores foram atribuídos à classe "normal", e somente com estes construir os grupos para cada instante da semana.

Os valores de média e de variância ao longo da semana para os nove sensores são apresentados na Figura 16 e na Figura 17, respectivamente, onde o eixo horizontal corresponde ao momento da semana em dias, a começar na hora zero do domingo. Nas séries construídas pelas médias são notáveis o ciclo diário e o comportamento diferenciado no final de semana.

Essas séries semanais de médias obtidas serão utilizadas para a detecção de alterações segundo a abordagem de comparação, como parte do conhecimento da rede incorporado. Ao analisar a dispersão dos dados, representada pela sua variância, constata-se que a grandeza da dispersão acompanha o ciclo diário, sendo maior nos períodos de maior consumo, associados à ocorrência de menores pressões. Na Figura 17 é apresentado, adicionalmente, o resultado do teste de normalidade.

Figura 16 – Valor médio de pressão para os momentos uma semana, segundo os dias completos da classe "normal" nas séries corrigidas, para os nove sensores.

Ao analisar as variáveis de pressão em cada momento da semana, uma hipótese a ser considerada é a de que cada conjunto siga uma distribuição de probabilidade normal. Para estabelecer a validade dessa hipótese foi utilizado o teste de Kolmogorov-Smirnov (com a correção de Lilliefors), testando a hipótese nula de que os dados tem origem em uma distribuição normal, com um nível de significância de 5%. Os resultados do teste, em termos de rejeição ou não rejeição da hipótese nula, são apresentados na Figura 17. Na figura, a faixa na parte superior dos gráficos da variância mostra o resultado do teste, usando uma linha vertical colorida nos momentos em que a hipótese nula é rejeitada, e mantendo em branco os momentos em que essa hipótese não é rejeitada. Consequentemente, os trechos coloridos da faixa correspondem a intervalos de tempo para os que é rejeitada a hipótese de que os dados correspondam com uma distribuição normal, enquanto os trechos brancos de cada faixa indicam que para esses intervalos não é rejeitada a hipótese de normalidade.

De forma menos definida do que nas variâncias, também os resultados do teste apresentam uma correlação com as horas do dia, sendo mais evidente para o sensor 9 (ver figura). Em geral, há mais grupos aderindo a uma distribuição normal nos períodos de menor consumo, no início de cada dia, enquanto nos horários de queda de pressão (como o primeiro pico de consumo diário) para a maioria de grupos rejeita-se a hipótese de normalidade. Os resultados anteriores são importantes ao considerar as possíveis estratégias para detecção de alterações na série, sendo

Figura 17 – Variância dos valores de pressão e resultado do teste de normalidade, para os momentos de uma semana, segundo os dias completos da classe "normal" nas séries corrigidas, para os nove sensores.

que metodologias baseadas na hipótese de normalidade devem ser desconsideradas. É importante salientar, porém, que ao limitar a análise aos valores dos sinais corrigidos para dias da classe "normal", o número disponível de ocorrências para cada momento da semana é de somente 21 a 32, o que pode diminuir a confiabilidade do resultado do teste.

6.1.7 Análise de recorrência

A descrição dos sinais de pressão adquiridos é complementada com a análise de recorrência, que permite encontrar características do sistema dinâmico de origem, e para tanto são aplicados os gráficos de recorrência (*Recurrence plots, RP*) às séries corrigidas. De forma similar à análise espectral, não podem ser processadas séries com valores faltantes, então é preciso escolher trechos de informações contínuas. Adicionalmente, e tendo como objetivo a descrição do comportamento normal, são desconsiderados os valores correspondentes à classe "visual" e à classe "simulados". Assim, todas as análises a seguir foram realizadas com trechos de série extraídas das séries corrigidas, com tamanho mínimo 17360, que representam 24 dias consecutivos sem dados faltantes nem alterações identificadas visualmente. Por ter uma resolução de ordem de grandeza menor à dos demais sensores, os resultados com o sensor 9 não são numérica ou graficamente comparáveis, e não são aqui apresentados.

Aplicando os métodos apresentados no Capítulo 5 o número de dimensões foi definido com o método de falsos vizinhos próximos. Para as diferentes séries os resultados, sem ser idênticos, indicam que a mínima dimensão que pode ser utilizada é $m = 2$. O passo de reconstrução d é definido segundo os resultados da autocorrelação, segundo o método proposto. Na análise de autocorrelação (Figura 13) vemos que o primeiro mínimo local corresponde a um atraso de $0,5$ dias, para a maioria dos sensores, enquanto para os sensores 1 e 3 há um mínimo local perto de $0,3$ dias e um máximo local não muito alto em $0,5$ dias. Apesar dessas particularidades nos registros dos sensores 1 e 3, em todos é possível ver que para atrasos na vizinhança ampla de $0,5$ dias a autocorrelação está próxima do seu valor mínimo, o que coincide com o procurado pelo método para seleção do atraso. A particularidade existente para este período também foi notada ao analisar os picos na densidade espectral de energia (Figura 15) e encontrar que dentre os mais importantes presentes em todos os sinais, o de mais alta frequência (período mais curto) é o correspondente a 14semana^{-1} . Dessa maneira é escolhido para a reconstrução das séries em duas dimensões o passo de $d = 0,5 \text{ dia} = 360 \text{ valores}$.

Outro parâmetro, fundamental na construção do gráfico de recorrência, é o raio de corte ε , que define a máxima distância (euclidiana) para considerar um estado como recorrente, como apresentado na Equação 5.4. Seguindo a metodologia descrita no Capítulo 5 o cálculo da recorrência com cada série é repetido para 100 raios diferentes, desde $\varepsilon \approx 0$ até a maior amplitude encontrada na série. Em cada caso é calculada a taxa de recorrência RR e seu valor como função do raio de corte é apresentado na Figura 18, onde cada uma das nove linhas corresponde à série de um sensor. Para melhor avaliação, os mesmos dados são representados em escala logarítmica na Figura 19.

Figura 18 – Taxa de recorrência (RR) para diferentes raios de corte (distância máxima para avaliar a recorrência), para os nove sensores.

Como esperado, um raio pequeno leva a estimar uma recorrência muito baixa, medindo basicamente o ruído. Isto entende-se porque o raio pequeno implica que somente valores muito

Figura 19 – Taxa de recorrência (RR) para diferentes raios de corte (distância máxima para avaliar a recorrência), para os nove sensores, em escala logarítmica.

próximos são considerados coincidentes e, portanto, recorrentes. Assim, resultados que produzem valores de recorrência menores a $1,0\%$, são desconsiderados. Por outro lado, com um valor muito grande de raio, tende-se a avaliar qualquer par de estados como coincidentes, de forma que a recorrência estimada é muito alta sem que seja verdadeira. Nos gráficos esse comportamento evidencia-se na baixa declividade das curvas para RR próximo de $1,0$, e para evitá-lo os valores de raio que geram taxas de recorrência maiores a 80% são desconsiderados. Adicionalmente, o tamanho do radio que envolve os estados (valores de pressão) da série está diretamente relacionado com sua dispersão, de forma que o desvio das séries é também levado em conta. Assim, é decidido utilizar para cada uma das series um raio igual ao valor do desvio da serie corrigida para a classe "normal" (última coluna da Tabela 4), que corresponde a $RR = 20\%$, aproximadamente.

Uma vez escolhidos os parâmetros é construído o gráfico de recorrência (RP) para cada um dos sensores, dentre os quais são apresentados na Figura 20 os dos sensores 1, 2, 5 e 8, visto que os restantes são muito similares a algum deles. O cálculo dos RP foi feito com séries de $17000 \text{ valores} \approx 23,6 \text{ dias}$, e para visualização é extraída a porção central de $7000 \text{ valores} \approx 9,7 \text{ dias}$. Assim, os eixos tanto vertical quanto horizontal do RP correspondem ao tempo ao longo da série escolhida, na mesma escala e intervalo que o eixo horizontal no gráfico da série, localizada na parte superior de cada RP. O eixo vertical da série corresponde aos valores de pressão (mca), lembrando que os trechos escolhidos são classificados como "normal" e foram extraídos das séries corrigidas. Salienta-se que nesta seção o objetivo é a descrição das séries representativas do comportamento normal, de forma que é calculada a auto-recorrência e portanto o RP é simétrico. Cada um dos pontos coloridos no RP indicam a existência de recorrência. Se para o ponto (i, j) há recorrência, no instante i o estado (valor de pressão) da série é o mesmo que o estado no tempo j ou próximo a ele, com distância menor que o raio escolhido. Assim, conjuntos de pontos coloridos consecutivos são resultado das trajetórias no

espaço-fase, e os padrões presentes nos gráficos revelam também características da evolução temporal dessas trajetórias (WEBBER; MARWAN, 2015). Para inferir características da série os gráficos são interpretadas segundo seus atributos em escala maior ou aspecto (*typology*), e em escala menor ou textura, seguindo a notação e orientações de Webber e Marwan (2015).

Figura 20 – Gráfico de recorrência (RP) e série utilizada (trecho "normal" extraído das series corrigidas) para os sensores 1, 2, 5 e 8. Apresenta-se o trecho central da série e a região central do RP, de tamanho 7000 (9,7 dias) e 7000x7000, respectivamente.

O aspecto dos RP para todos os sensores é periódico, com estruturas diagonais definidas criando uma impressão de xadrez, o que é representativo de sistemas oscilatórios. A periodicidade, já encontrada por outras técnicas no sinal, é notória nos RP, mas a aparência do gráfico não coincide perfeitamente com a teórica para esse tipo de sistemas, revelando a existência de ruído e de outros tipos de estrutura conjugados. A frequência do comportamento periódico identificado

pode ser avaliada segundo a distância entre uma estrutura recorrente e a seguinte, no caso medindo a distancia vertical entre sucessivas diagonais quase-contínuas. Foi encontrado que o período representado pela estrutura principal de diagonais é, aproximadamente, de 720 valores, isto é, o ciclo diário.

Ao avaliar estruturas menores nos gráficos são notórias as linhas diagonais, indicando que uma trajetória visita a mesma região do espaço-fase em diferentes momentos. Para o sensor 8, as diagonais são muito mais continuas, sendo interrompidas por faixas brancas verticais. Já no sensor 1 essa continuidade é menos bem definida, e as linhas diagonais podem sumir em trechos curtos em todo o gráfico. Esse comprimento das diagonais está relacionado com o grau de determinismo, sendo que uma série completamente determinística teria linhas diagonais tão longas quanto permitisse o tamanho do RP, enquanto sistemas puramente estocásticos gerariam pontos isolados ou diagonais curtas. Os resultados para todos os sensores permitem concluir que há uma componente determinística importante nas series, sem ser a única componente. O comprimento de diagonais é utilizado para o cálculo de alguns dos indicadores na análise de quantificação da recorrência.

Podem também ser vistas nos RP linhas verticais (e as horizontais relativas), que correspondem aos tempos nos quais os estados mudam menos, ou menos rápido, que para o restante do gráfico, revelando descontinuidades ou comportamento não estacionário do sistema. Comparando com a figura das séries, a localização dessas faixas verticais (e horizontais) coincide com comportamentos menos comuns da pressão, como valores locais muito altos ou baixos (serie 1) ou trechos da ordem de dias com comportamento diferenciado, como no final de semana. Assim, mesmo que os trechos selecionados não apresentem alterações que os coloquem em classes diferentes à "normal", as descontinuidades são muito mais notórias nos RP, e mostram que há muitas mais alterações além das identificadas na classificação manual, "poluindo" as séries. As regiões sem linhas verticais, com continuidade nas linhas diagonais, presentes para todas as series, são representativas de comportamento estacionário do sistema, confirmando que existe tal tipo de comportamento nas séries corrigidas, mas não em todos os intervalos.

6.2 Detecção com cadeias de Markov

Seguindo a abordagem da interpretação da serie de dados como ocorrências de um processo estocástico que cumprem a propriedade de Markov, apresenta-se nesta subsecção a construção, a partir dos dados de pressão adquiridos, das cadeias de Markov e seu uso para detecção de irregularidades ou alterações nas séries.

Na análise são utilizadas as séries completas, de até 257040 valores para cada um dos nove sensores, e corrigidas pela translação da média, como apresentado na secção anterior. Para a definição de estados, é testada discretização da pressão em intervalos de $\Delta e = 0.1, 0.2, 0.3$ e $0.5 mca$. Quanto à resolução temporal, avaliam-se os dados adquiridos a cada $\Delta t = 2 min$, e a media dos dados acumulados em $\Delta t = 10 min$. Para cada caso, as matrizes de transição são

construídas a partir dos dados classificados como "normal". São testados três valores de memória para o cálculo das probabilidades condicionais: $k_{mem} = 1, 2, \text{ ou } 5$ valores, equivalentes a 2, 4, ou 10 minutos.

Na execução do método, para cada novo valor da série é calculada a probabilidade condicional associada, e se for maior a um limiar α o valor é avaliado como *negativo: nenhuma alteração detectada*, enquanto se for menor o resultado é *positivo: alteração detectada* (Equação 5.15). Essa resposta é comparada com a obtida pela classificação supervisionada, considerando todos os tipos de alterações agrupadas (normal-todos) ou somente os experimentos de simulação realizados (normal-simulado). Note-se que no agrupamento das classes supervisionadas não está inclusa a classe 2, "reporte"; essa exclusão obedece a que a representação como cadeia de Markov analisa poucos antecessores de um valor, permitindo detectar o início súbito de uma alteração. Como os reportes iniciam, necessariamente, após o início da alteração, não é possível que com este método sejam detectados os valores rotulados com essa classe 2 "reporte", pelo qual é desconsiderada. Da comparação para cada serie completa é construída a matriz de confusão, e repetindo para 100 valores de α são obtidas as curvas ROC. A área abaixo dessas curvas, *AUROC*, é o indicador de desempenho global utilizado. Esse método foi mais detalhadamente apresentado no Capítulo 5.

6.2.1 Resultados comentados

As curvas ROC resultantes para as séries corrigidas dos nove sensores, com estados definidos pela pressão a cada $\Delta t = 2 \text{ min}$ discretizada cada $\Delta e = 0,50 \text{ mca}$ e memória $k_{mem} = 5$, avaliando com a supervisão de todas as classes (normal-todos) ou somente dos experimentos de simulação (normal-simulado), são apresentadas na Figura 21. A partir dos pontos da curva é calculada a área abaixo da curva (*AUROC*), que é o resultado de desempenho global obtido independentemente para cada um dos nove sensores. A média e desvio padrão dos resultados de *AUROC* para os sensores são reportados na Figura 22.

Já na Figura 21 é apreciável o pobre desempenho do método na avaliação quando consideradas somente as classes "normal" e "simulado" (b), sendo apenas superior ao aleatório ou até pior. Considerando todas as classes (a), o desempenho é razoavelmente melhor, mas com diferenças grandes entre os sensores.

Na decima segunda barra na Figura 22 (a e b) está o resumo dos resultados apresentados na Figura 21, e corresponde ao melhor desempenho médio dentre as condições avaliadas, com *AUROC* média de 0,672 para a avaliação de todas as classes e *AUROC* média 0,528 ao considerar somente os simulados. Os resultados de *AUROC* quando consideradas na avaliação somente as classes "normal" e "simulado" são, em todos os casos, piores que os obtidos considerando todas as classes, ficando muito próximos do resultado do aleatório, ou ainda abaixo deste. O resultado tende a melhorar ao aumentar a memória de 2 min para 4 min e desta para 10 min , indicando que a probabilidade condicional com valores muito próximos é um pior indicador

Figura 21 – Curvas ROC resultado da detecção usando cadeias de Markov, para as nove séries corrigidas completas. Estados: pressão com $\Delta t = 2min$ e $\Delta e = 0,50mca$. Memória $k_{mem} = 5$. Considerando todas as classes (a. normal-todos) ou somente os experimentos (b. normal-simulado)

de alterações. Quanto ao efeito de aumentar a discretização dos estados, em geral os melhores resultados de *AUROC* médio são obtidos, na ordem, com $\Delta e = 5, 2, e 1$, mas os resultados para $\Delta e = 3$ fogem dessa tendencia. Em qualquer caso, é preciso salientar que as diferenças nas médias antes comentadas são em muitos casos inferiores ao valor de desvio padrão calculado, de forma que a aparente superioridade ou inferioridade de um conjunto de condições pode não significar um desempenho realmente diferente.

Quando os estados são definidos como pressão média em 10 minutos ($\Delta t = 10$) cada estado é construído com a média de cinco valores adquiridos, e portanto a importância do conteúdo de altas frequências (ruído) deve ser menor. Porém, seus resultados de desempenho global não são significativamente diferentes aos obtidos com $\Delta t = 2min$, mas na maioria dos casos resultam inferiores. Todos os comentários anteriores referem-se à *AUROC* média, calculada entre os resultados dos nove sensores. Deve ser salientado, porém, que o desvio padrão é proporcionalmente maior para resultados mais altos, chegando até 17,3%, enquanto os resultados de *AUROC* mais baixos têm desvio padrão menor, desde 8,0% (primeira linha, quarta coluna).

Análises adicionais sobre os resultados e discussões mais gerais são desenvolvidas na subseção seguinte, mas antes é apresentada mais uma variação do método estudado. Uma definição alternativa para os estados do sistema (e, portanto, para o processo) é a que considera as primeiras diferenças, ou volatilidade. Com essa nova forma de definir os estados, tanto para $2min$ quanto para $10min$, são repetidas todas as análises antes descritas, produzindo os resultados apresentados na Figura 23.

Os estados definidos pela primeira diferença geram resultados de *AUROC* inferiores (mesmo que na mesma ordem de grandeza) aos antes comentados. Considerando na avaliação

Figura 22 – AUROC resultante da detecção para as séries completas, com estados definidos pela **pressão** adquirida a cada $\Delta t = 2min$ ou pela pressão média a cada $\Delta t = 10min$. Valor de média e desvio padrão calculado entre os resultados de AUROC dos nove sensores, considerando todas as classes (a. normal-todos) ou somente os experimentos (b. normal-simulado), várias discretizações dos estados (Δe), e valores de memória (k_{mem}).

somente as classes supervisionadas "normal" e "simulado", o desempenho médio é próximo ao do aleatório, com valores superiores e inferiores, e o máximo desvio padrão é de 0,044. Ao considerar todas as classes, o melhor resultado médio é 0,517, e o pior 0,503, enquanto os desvios são de, no máximo, 0,012. Os resultados para as diferentes condições de método estudadas (Δt , k_{mem} , Δe) são muito mais homogêneos com esta definição de estados: se consideradas todas as classes a variação máxima é de 2,8, e de 13,8 se considerados somente os simulados. Assim, não é possível estabelecer tendências claras da influência das condições no desempenho médio global.

6.2.2 Discussão

Há vários aspectos importantes para entender o baixo desempenho do método de detecção de alterações baseado em cadeias de Markov. Um dos principais é a discrepância entre a forma de avaliar o desempenho e o funcionamento do método. A tomada de decisão do método sobre a condição regular ou irregular do mais recente valor analisado depende da probabilidade condicional desse valor dado um valor anterior, e portanto a principal situação em que uma alteração seria detectada é aquela em que o valor anterior era parte do comportamento normal e

Figura 23 – AUROC resultante da detecção para as séries completas corrigidas, com estados definidos pela **diferença** entre pressões cada $\Delta t = 2min$ ou cada $\Delta t = 10min$. Valor de média $\pm$ desvio padrão calculado entre os resultados de AUROC dos nove sensores, considerando todas as classes (normal-todos) ou somente os experimentos (norma-simu), várias discretizações dos estados (Δe), e valores de memória (k_{mem}).

o novo foge desse comportamento. Essa habilidade de detecção é aplicável ao início abrupto de uma alteração. Se a alteração se mantém no tempo, e especialmente se os valores alterados continuam a estar numa faixa de valores que ocorrem na serie, o método não tem capacidade de reconhecer que os novos dados ainda fazem parte da alteração. Por outro lado, a classificação supervisionada rotula com classe diferente de "normal" não somente o momento em que a alteração acontece, mas todo o trecho em que ela se mantém.

Um exemplo ilustrando a discrepância descrita é mostrado na Figura 24, onde apresenta-se uma alteração que na classificação manual foi identificada no sinal (figura superior) desde o tempo 128,36 até o tempo 129,04 (cor verde). Na figura inferior mostram-se os valores de probabilidade condicional calculados para cada valor no mesmo trecho, usando como estados a pressão a cada $\delta t = 2min$ discretizada cada $\delta e = 0,1mca$ e memória $k_{mem} = 1$. À direita, detalhe das mesmas figuras.

A probabilidade condicional resultante é 0 para os primeiros momentos da alteração (tempo 128,39 até 128,44), e depois as probabilidades aumentam atingindo valores que, estando longe do máximo, também estão presentes em trechos normais da série (ao redor do tempo 127,5 ou de 128,5). Neste caso o método, segundo o limiar escolhido, ou não detecta como alterados a maior parte dos momentos do trecho irregular, ou detecta também trechos que não são alterados.

Assim, mesmo que nesse exemplo o método consegue detectar a alteração quando esta começa, permitindo por exemplo a emissão de alarme rápido (pensando na aplicação prática), na avaliação global realizada são contabilizadas muitas mais falhas do que acertos para este exemplo.

Figura 24 – Trecho da série e as probabilidades condicionais calculadas, e detalhe do trecho com alteração. Extraído da série corrigida do sensor 2, estados: pressões cada $\delta t = 2min$ discretizadas em $\delta e = 0,1mca$. Memória $k_{mem} = 1$.

A Figura 24 permite exemplificar também o segundo dos aspectos importantes sobre o baixo desempenho do método, e é a periodicidade das probabilidades condicionais obtidas. As matrizes de transição são calculadas usando a frequência de aparição nos trechos com rótulo "normal" das séries, e são consideradas estacionárias, mas essa consideração deve ser analisada.

O comportamento periódico da série e, em particular, dos seus desvios, faz com que as transições entre estados mais afastados e entre estados mais próximos sejam comuns (ou incomuns, complementarmente) em trechos diferentes da série. Em particular, transições curtas entre os estados de altas pressões são típicas unicamente das primeiras horas de cada dia, mas cada uma têm no total maior frequência de aparição que cada uma das transições típicas do restante do dia. Assim, mesmo durante trechos sem nenhum tipo de alteração, a probabilidade calculada para as primeiras horas do dia é muito maior que para o restante. O cálculo das matrizes de transição estacionárias com todos os dados e aplicação da cadeias de Markov sem considerações para tratar a alta periodicidade das séries pode ser apontado como um dos motivos do baixo desempenho global do método.

6.3 Detecção com controle estatístico

Dentro da abordagem de detecção baseada somente nos dados adquiridos, foi aplicado o controle estatístico, em particular o cálculo de média e variância móvel exponencialmente ponderada (MVMEP), integrado com o seguimento de descritores de longo prazo, como apresentado no Capítulo 5.

O método foi aplicado às series completas (não corrigidas), de até 257040 valores para cada sensor. Cada novo valor analisado é comparado com a média e variância na iteração anterior da MVMEP com $\lambda = 0,1, 0,5$ ou $0,7$, e é classificado como *negativo: nenhuma alteração detectada* ou *positivo: alteração detectada* usando o limiar α . Posteriormente, os valores de média e desvio padrão são atualizados, sempre que o novo valor esteja dentro das faixas de controle tanto de curto prazo quanto de longo prazo segundo os limites $\beta_c = 3$ ou 4 e $\beta_l = 1,5, 2$ ou 3 .

Para avaliar o desempenho do método o resultado obtido em cada instante (ponto na série) é comparado com o definido na classificação manual (das classes normal-todos ou normal-simulado, alternativamente) e calculada a matriz de confusão. O processo repetido com 100 valores de limiar α permite construir a curva ROC, e dela a área abaixo da curva (AUROC) que é o indicador de desempenho utilizado em todas as análises.

6.3.1 Resultados comentados

Um exemplo de curvas ROC obtidas para todos os sensores é apresentado na Figura 25, onde foi utilizado $\lambda = 0,5$, $\beta_c = 3$, e $\beta_l = 3$. Para analisar a influência dos parâmetros λ , β_c e β_l , com cada combinação de parâmetros foi calculada a AUROC de cada uma das nove series, e sua média e desvio padrão são apresentados na Figura 26.

Figura 25 – Curvas ROC resultado da detecção usando MVMEP, para as nove séries completas. $\lambda = 0,5$, $\beta_c = 3$, e $\beta_l = 3$. a) todas as classes. b) somente os experimentos

Figura 26 – AUROC resultante da detecção para as séries completas aplicando controle estatístico. Valor de média e desvio padrão calculado entre os resultados de AUROC de cada um dos nove sensores, considerando todas as classes (a. normal-todos) ou somente os experimentos (b. normal-simulado), vários limiares de controle de atualização β_c e β_l , e fator de ponderação λ .

Nos resultados mostrados na Figura 25a, quando a avaliação é feita somente considerando as classes supervisionadas "normal" e "simulado", o desempenho com todos os sensores é pobre, semelhante ao aleatório. Considerando todas as classes (b), a dispersão das respostas para os diferentes sensores é muito importante, por exemplo para o sensor 1 $AUROC = 0,75$, que poderia ser considerado um bom desempenho, enquanto para o sensor 2 $AUROC = 0,46$, desempenho ainda pior que o aleatório. O resumo desses resultados corresponde com a terceira barra no segundo bloco no quadro superior esquerdo da Figura 26 e, com $AUROC = 0,650 \pm 0,105$, tem tanto o maior valor médio quanto o maior valor inferior da faixa de desvio. Na figura observa-se que a alta dispersão comentada não aparece somente para a combinação de parâmetros utilizada, mas é ainda maior em alguns outros casos, limitando as conclusões a respeito da influencia dos parâmetros.

Os valores de limiares para atualização e de fator de ponderação apresentados correspondem aos de melhor desempenho segundo análises prévias. Ao considerar todas as classes (normal-todos), a combinação do mais rigoroso limiar de longo prazo para atualização, $\beta_l = 1,5$, junto com fatores de ponderação que dão mais importância às informações mais recentes e mais rapidamente deixam de considerar os valores antigos, $\lambda = 0,5$ e $0,7$, resulta numa maior variabilidade e instabilidade dos resultados. Outras tendências no desempenho segundo os parâmetros

estão comprometidas pela grande variabilidade dos resultados, não sendo possível estabelecer uma influência clara.

Nos casos onde somente as classes normal-simulado são levadas em conta, o desempenho médio é melhor para os menores valores de λ , isto é, considerando valores menos recentes no cálculo para cada novo instante da média e desvio padrão ponderados. Dessa forma, parece que as variações produzidas pelos experimentos são mais identificáveis se considerados mais valores anteriores do que se utilizados exclusivamente os mais recentes. Porém, esse desempenho melhor mencionado é muito baixo, e a tendência deixa de ser importante quando verificada com resultados menores que o do aleatório.

6.3.2 Discussão

Os resultados globais antes apresentados, que indicam um desempenho baixo do método, não conseguem refletir aspectos importantes do seu funcionamento, e deve ser analisada principalmente a discrepância entre a estratégia de detecção e a forma de avaliação.

O controle estatístico, em geral, busca a identificação de mudanças ou alterações assim que começam, razão pela qual a detecção é feita segundo a predição (média e desvio padrão) construída somente com os valores mais recentes (MVMEP). Após ocorrência de um evento irregular no sistema, a alteração por ele provocada pode ser identificada no método pelo afastamento do esperado. Se o evento irregular continua, os descritores são recalculados e a predição passa a corresponder à nova situação, após algum intervalo curto de tempo. Para tentar evitar esse comportamento são definidos os limiares para atualização β_c e β_t , mas não é possível que todo tipo de alteração seja excluída sem excluir também comportamentos normais. Um exemplo é apresentado na Figura 27, com um trecho da série original do sensor 2 que inclui uma alteração do tempo 128,36 até o tempo 129,04 (mesmo exemplo usado na Figura 24).

No exemplo da figura, aproximadamente no tempo 127,09 os valores da série apresentam uma grande dispersão, incomum nos trechos que correspondem ao início de cada dia, ou de mínimo consumo. A incorporação de um valor extremo gera um aumento importante do desvio padrão de curto prazo (calculado com MVMEP), que faria com que os valores seguintes tenham menos chance de ser detectados como irregulares, mesmo se afastando da média. A aplicação do limiar de curto prazo $\beta_c = 3$ para atualização impede que valores extremos sejam incorporados ao cálculo do desvio padrão, mantendo-o numa faixa mais razoável e permitindo assim a detecção de alteração. No exemplo, porém, a oscilação inusualmente ampla das pressões no trecho comentado não foi identificado durante a classificação manual.

Também na Figura 27 evidencia-se que no início da alteração identificada na classificação manual o valor de pressão tem um decréscimo rápido, mas progressivo, de forma que cada novo valor está dentro do limiar β_c e é, portanto, aceito para atualizar os descritores. Com tal atualização dos descritores de curto prazo (calculados com MVMEP) a média diminui e o desvio aumenta, o que incrementa as chances dos seguintes valores também serem considerados normais

Figura 27 – Trecho da série, média e limites do intervalo de aceite para atualização tanto de curto quanto de longo prazo. Extraído da série original do sensor 2, com $\lambda = 0, 2$, $\beta_c = 3$, e $\beta_l = 2$.

pelo método. Essa média e desvio padrão de curto prazo continuariam acompanhando os dados irregulares, sem detectá-los como tal, caso não fosse considerada a avaliação dos descritores de longo prazo. Como a média e desvio padrão de longo prazo são menos influenciadas por cada novo valor, a partir do tempo 128,39 o limiar de duas vezes o desvio de longo prazo é ultrapassado, e todos os descritores (de curto e longo prazo) param de ser atualizados. Isso faz com que a média de curto prazo deixe de acompanhar os dados irregulares, permitindo a sua detecção. Quando os valores não são inferiores ao limiar, são incorporados na atualização, de forma que ao fim da alteração os descritores de curto prazo de novo estão acompanhando os dados adquiridos.

Para esse exemplo o resultado é bom, pois pouco tempo depois do início das alterações elas podem ser detectadas. Porém, não todos os momentos próximos geram a resposta (*nenhuma alteração detectada* ou *alteração detectada*) compatível com o rótulo atribuído na classificação manual. O resultado da avaliação, desde a matriz de confusão até a *AUROC*, são para este caso estritos demais, julgando numerosos erros ainda num caso de sucesso.

O bom resultado descrito no exemplo não se repete para todas as alterações ao longo das séries. Alterações grandes acontecendo logo no início (ou re-início, após trechos sem dados) podem fazer com que os descritores de curto e, principalmente, de longo prazo fiquem fixos em valores alterados e não consigam voltar a acompanhar a série. Mudanças rápidas do valor médio da pressão são acompanhadas com bastante atraso pelos descritores de longo prazo, ou

ainda podem continuar sem ser acompanhadas, como no caso anterior. Variações muito suaves na pressão, como no início da alteração (vazamento) identificada, ao redor do tempo 175, podem não ser identificadas fora dos limiares nem de curto nem de longo prazo; quando a variação continua com igual ou maior taxa, alguns dos valores começam a ser invalidados e deixar de ser acompanhados, permitindo a detecção como irregulares. Em um caso diferente, dentro do comportamento normal do sistema podem acontecer situações de mudanças rápidas e se um valor erroneamente é desconsiderado para a atualização, é possível que os próximos sejam também desconsiderados, de forma que trechos longos, ainda que normais e com pouca variação entre eles, podem resultar detectados como irregulares.

A ocorrência e gravidade dos casos anteriormente mencionados e demais situações de não detecção ou de detecção errônea depende fortemente da seleção dos limiares para atualização e do fator de ponderação, além do limiar para detecção. Mas, além da verdadeira ocorrência de erros e acertos, a avaliação por comparação ponto a ponto com a classe supervisionada incorpora também alguma distorção ao resultado. Também é preciso salientar que, na operação real, uma situação alterada como as descritas não deveriam se estender no tempo sem ser advertida pelo operador, que deve então tomar uma decisão como recalcular o início dos descritores, o que não é considerado na avaliação global.

Outro aspecto muito importante a ser levado em conta é a integração dos resultados dos diferentes sensores. Nos resultados apresentados é notável que, utilizados os mesmos parâmetros, algumas séries tem um desempenho global (*AUROC*) muito bom enquanto outras muito ruim. Ao mudar os parâmetros, porém, são outras as séries com bom e mal desempenho global. Assim, todas as séries podem atingir um valor de desempenho global alto ($AUROC > 0,8$) com alguma combinação de parâmetros, mas nenhuma combinação permite que todas as séries, simultaneamente, o consigam. Ao analisar casos específicos, vê-se que é muito frequente que, com algum parâmetro, as séries apresentem algum dos casos problemáticos antes mencionados e continuem por um tempo longo sem correção, comprometendo o indicador de desempenho. Assim, são evidenciados dois problemas que limitam a generalização dos resultados obtidos: o primeiro, a avaliação global utilizada, que pode ser mais negativa que a situação em aplicação real onde erros podem ser atendidos oportunamente, e o segundo, a agregação dos resultados globais em uma única média e desvio padrão entre os resultados de todos os sensores, calculados simultaneamente com os mesmos parâmetros. Em ambos os casos, o efeito é a diminuição do valor de desempenho global, e entende-se que com critérios adaptados para tal fim poderiam ser obtidos melhores valores.

6.4 Detecção pela comparação com padrões

Diferentemente dos dois métodos de detecção antes mencionados, com a detecção pela comparação com padrões procura-se ir além do que pode ser conhecido diretamente dos dados recebidos (matrizes de transição, descritores estatísticos) e incorporar adicionalmente o

conhecimento previamente adquirido sobre o sistema. O conhecimento do sistema é representado pelo padrão semanal, cujo cálculo é apresentado na primeira secção deste capítulo, e com ele é comparada cada nova informação adquirida.

Para a incorporação e análise de novas informações são propostas e avaliadas configurações alternativas, que incluem janelas de tamanho relativo a *1hora* (30 valores de pressão), *6horas* (180 valores) ou *1dia* (720 valores). As janelas são atualizadas com intervalos ou avanço de 10 ou *20minutos* (5 ou 10 valores) para o primeiro tamanho, avanço de 1 ou *2horas* (30 ou 60 valores) para o segundo tamanho e de 2 ou *8horas* (60 ou 240 valores de pressão) para a janela maior. A cada janela é aplicada uma das 12 transformações analisadas, e posteriormente a janela transformada é comparada com a correspondente janela do padrão (também transformada) usando, para tanto, uma das 5 funções de comparação propostas. O resumo das funções é apresentado na Tabela 5.

Tabela 5 – Lista das funções de transformação e de comparação utilizadas, alternativamente, no método

Transformação		
ntng	nothing	Mantém a serie de entrada, sem transformação
wave	wavedecomp_all	Coeficientes da decomposição <i>wavelet</i> discreta. Função <i>wavelet</i> db2, nível 4
wav_d	wavedecomp_detail	Coeficientes de detalhe dos níveis 1 e 2 da decomposição <i>wavelet</i> discreta. Função <i>wavelet</i> db2.
wav_a	wavedecomp_aprox.	Coeficientes de aproximação do nível 4 da decomposição <i>wavelet</i> discreta. Função <i>wavelet</i> db2.
ene_2	energy_resolutions	Distribuição da energia na serie, com várias resoluções.
ene_3	energy_proportional	Distribuição da energia ao longo da série, segundo incrementos proporcionais.
zcr_2	zcr_resolutions	Distribuição do número de cruzamentos por zero na série, com várias resoluções.
zcr_3	zcr_proportional	Distribuição do número de cruzamentos por zero ao longo da série, segundo incrementos proporcionais.
ent_1	entropy_distribution	Distribuição da entropia ao longo da série.
ent_2	entropy_resolutions	Distribuição da entropia na série, com varias resoluções.
dlt_pl	discretelength_pl	Comprimento proporcional calculado da transformada discreta de comprimento.
dlt_zcr	discretelength_zcr	Número de cruzamentos por zero calculado da transformada discreta de comprimento.
Comparação		
eucl	euclidean	Distância euclidiana.
corr	correlation	Correlação cruzada máxima, com atrasos restritos.
cor_env	envcorrelation	Correlação cruzada das envoltórias de Hilbert.
cen	centroids	Distância entre os centroides (médias).
haus	hausdorff	Distância de Hausdorff.

6.4.1 Resultados comentados

Para avaliação do método, a detecção é feita para as séries completas originais (sem correção) e as relativas séries semanais de comportamento médio, usando uma função de transformação e uma de comparação em cada caso. A resposta obtida (*nenhuma alteração detectada* ou *alteração detectada*) para cada ponto (por votação), considerando um limiar α , é comparada com o rótulo atribuído na classificação manual e construída a matriz de confusão. Com os resultados produzidos para 100 valores diferentes de α são construídas as curvas ROC, como as apresentadas na Figura 28, e delas a AUROC para cada série, que é o indicador de desempenho global utilizado.

Figura 28 – Curvas ROC resultado da detecção usando comparação com o padrão, para as nove séries originais completas. Janela 30valores, avanço 10valores, transformação *ntng* e comparação *eucl*. a) todas as classes supervisionadas (normal-todos). b) somente os experimentos (normal-simulado)

Combinando as 12 funções de transformação, 5 funções de comparação, e 6 configurações de janelas, são avaliadas 360 configurações do método. Adicionalmente, a etapa de avaliação inclui duas alternativas: considerar todas as alterações presentes segundo a classificação manual (normal-todos) ou somente as relativas aos experimentos de simulação de vazamentos (normal-simulado). Para apresentação da análise desse grande volume de resultados, os resultados são agrupados segundo a configuração das janelas, e são comentados inicialmente os apresentados na Figura 29 e na Figura 30. Esses comentários serão tidos como gerais e para os resultados com outras configurações de janela, nas figuras seguintes (31 a 35), somente serão comentados os aspectos que difiram dos já tratados.

O caso apresentado na Figura 29 corresponde à adoção de janela de tamanho $720 = 1dia$ e avanço de $240 = 8horas$ ou um terço do tamanho, equivalente a sobreposição de 66,6%. Talvez o resultado mais interessante quando analisadas as funções de transformação é o alto desempenho obtido ao não realizar nenhuma transformação (*ntng*, pois é a segunda alternativa de transformação com o máximo desempenho atingido. Pelo princípio de parcimônia, tende-se

Figura 29 – AUROC resultante da detecção para as séries completas originais aplicando comparação com padrão. Valor de média e desvio padrão calculado entre os resultados de AUROC de cada um dos nove sensores, considerando todas as classes supervisionadas (**normal-todos**), janela de 720 = 1dia e passo 240 = 8horas, todas as funções de transformação e as de comparação.

a preferir a metodologia mais simples perante as mais complexas de desempenho equivalente, porém as combinações de melhor desempenho não podem ser negligenciadas, e essa conclusão não é aqui aceita sem antes completar as análises para todos os casos.

Ainda segundo apresentado na Figura 29, os melhores desempenhos estão concentrados nas funções baseadas na transformada *wavelet* (*wave*, *wave_d*, *wave_a*) e na alternativa sem transformação (*ntng*), enquanto a extração de características baseada em *zcr*, *entropia* e *dlt* (metade inferior na figura) produzem resultados próximos do aleatório. Os desvios padrão apresentam relação com as funções de transformação utilizadas, sendo maiores para o grupo de funções que tem os melhores desempenhos, e pequenos para as características de pior desempenho, com exceção das funções de transformação *wave* e *wav_d* combinadas com a função de comparação *corr*, que têm alto desempenho e pequeno desvio padrão. Resultados semelhantes aparecem também para outras configurações de janela, quando consideradas todas as classes supervisionadas (*normal-todos*).

Em relação às funções de comparação, a função de correlação (*corr*) apresenta o desempenho mais alto em combinação com as funções de transformação *wave* e *wave_d*, mas não com as restantes. As distâncias euclidiana (*eucl*), dos centroides (*cen*) e de Hausdorff (*haus*)

Figura 30 – AUROC resultante da detecção para as séries completas originais aplicando comparação com padrão. Valor de média e desvio padrão calculado entre os resultados de AUROC de cada um dos nove sensores, considerando somente os experimentos (**normal-simulado**), janela de $720 = 1\text{dia}$ e passo $240 = 8\text{horas}$, todas as funções de transformação e as de comparação.

apresentam os seguintes melhores desempenhos, semelhantes entre si, e menos variáveis para diferentes funções de transformação. Esse comportamento é mais claro quando combinado com as primeiras funções de transformação, casos nos que se obtêm os melhores resultados globais, e também pode ser verificado para outras configurações de janelas. Todas as funções de comparação em combinação com as de transformação da sexta à decima segunda, apresentaram desempenho próximo do aleatório. Não há nenhuma tendência que possa ser extraída dos resultados a respeito do desvio padrão, segundo a função de comparação.

Em resumo, para este tamanho e sobreposição de janelas (720 e 240, respetivamente) o melhor desempenho (tanto médio quanto do valor inferior da faixa de um desvio padrão) foi obtido com a combinação de transformação a coeficientes de detalhe da decomposição *wavelet* com comparação por correlação, *wav_d – corr*, atingindo $AUROC = 0,865 \pm 0,017$. O segundo melhor resultado ($AUROC = 0,842 \pm 0,041$) é obtido com a combinação *wave – corr*. As seis funções de transformação mais interessantes coincidem com as seis primeiras apresentadas.

As observações anteriores não se aplicam completamente aos resultados obtidos ao considerar somente as classes "normal" e "simulado" da supervisão, como pode ser visto na Figura 30. Neste caso os resultados são, em média, piores que quando consideradas todas as

classes supervisionadas, com muitas combinações apresentando desempenho médio inferior ao aleatório ($AUROC < 0,5$), e duas maiores a $AUROC = 0,70$. Os dois melhores resultados aparecem para as mesmas combinações do caso normal-todos, $wave - corr$ e $wav_d - corr$, com $AUROC = 0,742 \pm 0,04$ e $AUROC = 0,715 \pm 0,05$. Os seguintes valores mais altos, porém, são diferentes, sem que possa ser estabelecida uma tendência geral para cada tipo de função, considerando que combinações que geram desempenhos piores que o do aleatório não podem ser consideradas válidas e, portanto, perdem significado as diferenças entre elas.

Como o pobre desempenho para muitas das funções de transformação mantém-se para as outras configurações de janela, na maioria dos casos na mesma ordem de grandeza ($AUROC \leq 0,5$), seus resultados são omitidos nas figuras seguintes. Assim, nas figuras 31, 32 e seguintes são apresentados os resultados somente para as seis funções de transformação que atingem melhor desempenho considerando todas as classes supervisionadas (normal-todos) na parte superior das figuras, e na parte inferior de cada figura as seis funções de melhor desempenho considerando somente as classes normal-simulado.

Figura 31 – AUROC resultante da detecção para as séries completas originais aplicando comparação com padrão. Média e desvio padrão dos resultados dos nove sensores, janela de $180 = 6horas$, avanço $60 = 2horas$, avaliação com classes supervisionadas (**normal-todos**) e (**normal-simulado**), todas as funções de comparação, e as seis funções de transformação que atingiram melhor desempenho.

Os resultados com janelas de $180 = 6horas$ e avanço de $33,3\%$ ($60 = 2horas$) são apresentados na Figura 31 considerando todas as classes supervisionadas (normal-todos) ou

somente os experimentos de simulação (normal-simulado). No primeiro caso o comportamento geral é bastante parecido com o descrito para a primeira configuração de janelas, com exceção da função de comparação *corr*, que diminui seu desempenho nas combinações em que tinha tido os melhores resultados. O melhor desempenho no caso é $AUROC = 0,74 \pm 0,12$ e é atingido por sete combinações: *wav_a – haus*, *ntng – cen*, *wave – cen*, *wav_a – cen*, *ntng – eucl*, *wav_d – eucl* e *wav_a – eucl*. A extração de características baseadas na energia teve um desempenho menor mas ainda útil, diferentemente das demais características que tiveram resultados todos próximos do aleatório (média geral de $AUROC = 0,49$ e desvio padrão médio de 0,03).

Para avaliação com normal-simulado, os resultados para a janela de 180 são melhores que os relativos às janelas de 720, mas continuam menores que os resultados para normal-todos. 44 das 60 combinações avaliadas têm desempenho abaixo de $AUROC \leq 0,55$, e 5 atingem $AUROC \geq 0,60$ mas dessas últimas as combinações de funções com melhor desempenho não coincidem nem com as identificadas com a janela maior, nem com nenhuma das de melhor desempenho para normal-todos. Usando como função de transformação dlt_{pl} e comparação *cen* obtém-se o melhor desempenho, de $AUROC = 0,68 \pm 0,06$.

Figura 32 – AUROC resultante da detecção para as séries completas originais aplicando comparação com padrão. Média e desvio padrão dos resultados dos nove sensores, janela de $30 = 1\text{ hora}$, avanço $10 = 20\text{ min}$, avaliação com classes supervisionadas (**normal-todos**) e (**normal-simulado**), todas as funções de comparação, e as seis funções de transformação que atingiram melhor desempenho.

Os resultados para o menor tamanho de janela considerado, de $30\text{ valores} = 1\text{ hora}$

e avanço de $10\text{valores} = 20\text{min}$, são apresentados na Figura 32. Na avaliação com normal-todos, a superioridade das funções de comparação *euc*, *cen* e *haus* em combinação com as funções de transformação *ntng*, *wave* e *wav_a* é verificada, de forma muito semelhante ao obtido com os tamanhos de janela maiores. Entretanto, a combinação de *wave* e *wav_d* com *corr*, que tinham produzido os melhores resultados para a janela maior e resultado intermediário para a janela de 180, têm aqui um desempenho baixo. Adicionalmente, os resultados para funções de transformação baseadas na *dlt*, mesmo que próximos de aleatório, resultam um pouco melhores que os baseados na energia, ocupando seu lugar na figura. Em particular, o melhor desempenho foi de $AUROC = 0,72 \pm 0,12$, atingido pelas combinações *wave - eucl* e *wav_a - eucl*. O desempenho próximo do aleatório e com baixo desvio padrão para todas as funções de transformação não apresentadas (quinta, sexta, e da nona à décimo-segunda na Tabela 5) é conferido neste caso, com resultado médio geral para essas seis funções de $AUROC = 0,43 \pm 0,02$.

Para esta configuração de janelas (tamanho 30, avanço 10 valores) os resultados considerando na avaliação as classes normal-simulado são, na maioria, menores ao do aleatório, e nenhum atinge $AUROC = 0,60$. Também não é fácil identificar superioridade de alguma função, e as de melhor desempenho não coincidem nem com as da avaliação normal-todos nem com as de outras configurações de janela. As combinações *ent_2 - eucl*, *ent_2 - cen* e *ent_2 - haus* têm o melhor desempenho, com $AUROC = 0,58 \pm 0,06$.

Outra das alterações realizadas no método para sua avaliação é o avanço das janelas, e com isso sua sobreposição. Ao reduzir o avanço de $1/3$ do tamanho para $1/6$ ou $1/12$, aumenta a quantidade de respostas calculadas para cada valor e que são posteriormente integradas para definir a sua classe prevista, enquanto o tempo para a obtenção da primeira resposta diminui. O custo computacional, porém, aumenta, e as diferenças no desempenho são analisadas a seguir.

Ao comparar a figura dos resultados obtidos com janelas de tamanho $720\text{valores} = 1\text{dia}$ e avanço $60\text{valores} = 2\text{horas}$ (Figura 33) com a de avanço de 240 antes analisada (Figura 29), a primeira impressão é que são idênticas, principalmente se consideradas as classes normal-todos e as combinações de funções dos melhores desempenhos. Pode-se conferir, porém, que existem diferenças nesses resultados, mas essas diferenças são pequenas e não fazem com que os aspectos anteriormente analisados mudem, nem mude o máximo desempenho e as combinações que permitem obtê-lo. Se a comparação é feita com os resultados da avaliação com as classes supervisionadas normal-simulado as diferenças devidas à diminuição do avanço das janelas são mais evidentes, mas sem mudar o comportamento geral comentado para o avanço maior. O melhor desempenho é obtido pela combinação *wave - corr*, igual que com o avanço maior, mas com desempenho um pouco mais alto, de $AUROC = 0,76 \pm 0,04$.

Janelas com tamanho $180\text{valores} = 6\text{horas}$ e avanço $30\text{valores} = 1\text{hora}$ produzem os resultados apresentados na Figura 34. Mesmo que mudanças tanto de média quanto de desvio padrão sejam perceptíveis na figura, também o comportamento geral dos resultados é o mesmo

Figura 33 – AUROC resultante da detecção para as séries completas originais aplicando comparação com padrão. Média e desvio padrão dos resultados dos nove sensores, janela de $720 = 1\text{dia}$, avanço $60 = 2\text{horas}$, avaliação com classes supervisionadas (**normal-todos**) e (**normal-simulado**), todas as funções de comparação, e as seis funções de transformação que atingiram melhor desempenho.

do obtido para avanço maior, principalmente ao considerar a avaliação com classes normal-todas. Neste caso, as funções de transformação que atingem os dois melhores desempenhos não mudam, e o máximo é de $AUROC = 0,75 \pm 0,11$, obtido pela combinação $wav_a - haus$ e muito próximo das outras combinações identificadas também com o avanço de 60. Para a avaliação com as classes normal-simulado, há diferenças mais notórias, principalmente de aumento de médias e desvios padrão, mas somente muda uma das seis funções de transformação com maior valor de $AUROC$ máximo. Igual ao resultado com avanço maior, o melhor desempenho é obtido com a combinação $dlt_pl - cen$, e esse valor aumenta para $0,69 \pm 0,04$.

Na Figura 35 são apresentados os resultados obtidos ao diminuir nas janelas de tamanho $30\text{valores} = 1\text{hora}$ o avanço de 10valores para $5\text{valores} = 10\text{minutos}$. Mais uma vez, o desempenho relativo das diferentes funções muda pouco como resultado do avanço menor, mas é suficiente para mudar uma das seis funções com melhor desempenho máximo quando a avaliação usa normal-todos. Nesse caso, cinco combinações ($wav_a - eucl$, $wave - eucl$, $ntng - eucl$, $wave - cen$ e $wave_a - cen$) atingem o melhor desempenho, de $AUROC = 0,72 \pm 0,12$. Quando a avaliação considera somente normal-simulado, também uma das seis funções de transformação com maiores valores de desempenho máximo muda, e o mais alto é de $AUROC$

Figura 34 – AUROC resultante da detecção para as séries completas originais aplicando comparação com padrão. Média e desvio padrão dos resultados dos nove sensores, janela de 180 = 6horas, avanço 30 = 1hora, avaliação com classes supervisionadas (**normal-todos**) e (**normal-simulado**), todas as funções de comparação, e as seis funções de transformação que atingiram melhor desempenho.

= $0,61 \pm 0,06$ atingido com a combinação *ent_2* – *eucl*, a mesma de melhor desempenho no avanço maior.

6.4.2 Discussão

São muitas as diferentes configurações que podem ser utilizadas para o método de detecção por comparação com padrões, e dentre as analisadas é necessário buscar algumas conclusões, salientando que não é possível propor uma única configuração "certa", nem sequer uma ótima. Isso é devido às variações importantes do desempenho entre sensores e entre diferentes tipos de alteração, além de incertezas introduzidas na avaliação e limitações no uso de um único indicador de desempenho global, como discutido mais adiante. Com essa ressalva, utilizaremos os resultados apresentados anteriormente e baseados na *AUROC* global, juntamente com considerações práticas, para tentar concluir sobre as configurações mais interessantes com o método.

Os resultados quando consideradas todas as classes supervisionadas na avaliação (*normal-todos*) são bastante consistentes perante as mudanças em elementos da configuração, diferentemente dos resultados considerando *normal-simulado*, nos quais cada mudança de configuração

Figura 35 – AUROC resultante da detecção para as séries completas originais aplicando comparação com padrão. Média e desvio padrão dos resultados dos nove sensores, janela de $30 = 1\text{hora}$, avanço $5 = 10\text{min}$, avaliação com classes supervisionadas (**normal-todos**) e (**normal-simulado**), todas as funções de comparação, e as seis funções de transformação que atingiram melhor desempenho.

tem consequências variadas nos resultados. Esse fato reforça a hipótese de que as alterações nas pressões do sistema obtidas com a execução de experimentos de simulação de início de vazamento não são detectáveis por todos os sensores, e sua detecção é mais difícil que a de outras irregularidades ou alterações nas séries. Portanto, para a definição das configurações mais interessantes dar-se-á mais atenção aos resultados considerando a avaliação normal-todos.

Quanto às funções de comparação, os resultados com as janelas intermediária e menor mostram superioridade das distâncias euclidiana (*eucl*), dos centroides (*cen*) e de Hausdorff (*haus*), enquanto com a janela maior a correlação (*corr*) tem um desempenho melhor que as outras. Assim, métricas globais para cada janela parecem mais adequadas que comparações do comportamento quando o período considerado é mais curto, e para períodos suficientemente longos o comportamento tem melhor resultado. Isso pode ser entendido ao lembrar que o comportamento cíclico mais claro é o diário, de forma que ao considerar essa escala maior há um formato claro da série que pode acusar alterações mais facilmente, enquanto para janelas menores esse comportamento não é tão bem definido, e começa a ter mais influência o ruído. Já o desempenho das distâncias não muda tão radicalmente com os diferentes tamanhos de janela, ao não estar focado nesse comportamento, o que pode ser considerado uma vantagem.

Intuitivamente, cada uma dessas quatro funções de comparação é melhor para detecção de diferentes situações: alterações que resultam em diminuição ou aumento fixo e continuado na média da série seriam bem identificadas pela distância dos centroides, por exemplo. Atrasos relativos na série ou outras diferenças de tendência podem ser detectadas melhor com a distância euclidiana, enquanto alterações locais como quedas súbitas tem maior incidência na distância de Hausdorff. A correlação será mais efetiva para identificar mudanças no comportamento relativo da série, independentemente dos seus valores absolutos. Embora exemplos tenham sido identificados para as afirmações anteriores, a análise mais detalhada necessária para se concluir algo a respeito escapa à metodologia de análise definida. O desempenho global, eixo da avaliação adotada, é mais dependente da função de transformação do que da escolha entre as três distâncias comentadas (*eucl*, *cen*, *haus*), enquanto o tamanho de janela tem também grande influencia no desempenho da quarta função (*corr*).

Na seleção das funções de transformação mais interessantes há também um comportamento diferenciado das combinações de transformações *wave* e *wav_d* com a comparação *corr*, e outro claro quando consideradas as alternativas restantes. A julgar pelos resultados de desempenho global na avaliação normal-todos, as características baseadas no número de cruzamentos por zero, entropia, transformada de comprimento e energia são, nessa sequência, as de pior desempenho. Tem resultados bons a transformada discreta *wavelet*, especialmente para tamanhos de janela maiores, como é de se esperar. Em particular o uso da componente de detalhe ou alta frequência é mais interessante para janelas maiores e comparação *corr*, enquanto a de aproximação ou de baixa frequência tem desempenho mais estável e bom com as funções de comparação de distância. Desempenho global semelhante ou ainda mais alto ao atingido usando transformação baseada em *wavelet* é obtido quando utilizada a opção *ntng*, que significa não realizar nenhuma transformação e fazer as comparações entre a janela com dados de pressão adquiridos e a janela do padrão correspondente, também com os valores de pressão em mca.

Adicionalmente, se nenhuma transformação é feita não é necessário aplicar limitações aos outros elementos da configuração por possíveis incompatibilidades, como tamanho da janela mínimo para aplicar a transformada ou inconsistência da avaliação de distância dos centroides com características de número de cruzamentos por zero (que requer subtrair a média da série). À conclusão anterior junta-se a diminuição do custo computacional associado à opção de não realizar transformação, e também o princípio de parcimônia, tudo levando a concluir que a alternativa de não transformação das séries (opção *ntng*) é indicada, concorrendo somente com as combinações *wav_d* – *corr* e *wave* – *corr* com janela de tamanho 720.

Para a configuração de janela, além do desempenho global, é importante a análise da aplicabilidade do método. O tamanho da janela e o avanço definem o tempo mínimo de detecção, visto que após uma alteração começar passa um tempo (máximo) equivalente ao avanço até o primeiro valor anômalo adquirido entrar na análise. Somente após o tempo equivalente ao tamanho de janela o momento inicial da anomalia terá sido analisado o total de vezes e então a

decisão sobre sua classe estimada é concluída. Para aplicação do método executando quase em tempo real, é desejável diminuir o intervalo até a primeira análise (avanço) e, principalmente, o intervalo até a conclusão da mesma (tamanho de janela). Adicionalmente, com tamanhos de janela e avanços grandes, a mesma resposta é aplicada a um número maior de pontos, trazendo menor resolução aos resultados e dificuldade em detectar mudanças de curta duração, que não são infrequentes, segundo a informação adquirida. A diminuição no tamanho da janela, porém, implica em maior custo computacional (mais frequente recálculo do método) e em redução do desempenho global avaliado no estudo de caso, mas é necessário lembrar que essa diminuição é percentualmente pequena e que um pior indicador de desempenho global não é necessariamente devido a piores resultados em detectar cada uma das alterações presentes.

Segundo os resultados obtidos, a diminuição no avanço, ou aumento na sobreposição, não parece ter efeito significativo no desempenho do método. Para entender tal resultado, deve-se considerar que para cada um dos tamanhos de janela analisados, todas as janelas utilizadas com o maior avanço existem também no janelamento com avanço menor. Assim, os avanços menores trazem as mesmas informações do avanço maior, acrescida de informações intermediárias entre elas. A julgar pelos resultados do indicador de desempenho, essas informações intermediárias não trazem nenhuma diferença para a detecção, além da de incrementar o número de análises para cada valor da série. Segundo o anterior, o incremento em complexidade e esforço de computo ligado ao uso dos menores avanços analisados, de $1/6$ ou $1/12$ do tamanho, não é compensado por nenhuma melhora significativa no desempenho, de forma que não é recomendado.

A definição do tamanho de janela, associado ao avanço (calculado como $1/3$ do tamanho), dependerá do equilíbrio procurado entre tempo de detecção, resolução da detecção e consumo de recursos computacionais na atualização. Considerando a velocidade de cômputo atingida nas rotinas de análise, a importância de reduzir o tempo de detecção e a frequência de aparição de alterações de curta duração, aliado à incerteza associada ao indicador de desempenho global, considera-se conveniente o uso de janelas de $1hora = 30valores$ com avanço de $20minutos = 10valores$.

Como conclusão, uma recomendação de configuração para o método proposto seria janelas de 30 e avanço 10, sem realizar transformação e comparando com norma euclidiana, distância dos centroides ou distância de Hausdorff. Essa conclusão está sujeita às ressalvas anteriores, algumas das quais discutidas mais amplamente a seguir.

As considerações sobre a avaliação discutidas nas subsecções referentes aos outros dois métodos desenvolvidos são parcialmente aplicáveis ao método de detecção por comparação com padrões. Ficam fora da discussão, por não estarem relacionadas ao método, as dificuldades e incertezas envolvidas na classificação manual e a consequente incerteza incorporada na avaliação, mas um aspecto que tem relação é a influência da forma em que a decisão é tomada, no indicador de desempenho. Imaginando o início de uma alteração rápida e de grandeza detectável, a janela que contém esses novos dados irregulares contém também dados anteriores, normais. A

comparação é então feita e seu resultado faz parte da avaliação para todos os elementos da janela, incluindo os anteriores ao início da alteração. O uso de sobreposição e o tamanho da janela estão relacionados com esse efeito de expansão para atrás (ou para a frente) da detecção no caso de grandes alterações rápidas, mas não há como garantir que não aconteça. Já para mudanças continuadas no tempo, como a que produz um vazamento, o efeito da avaliação por janelas para possibilitar conclusões sobre cada valor adquirido pode trazer modificações tanto positivas quanto negativas para a detecção, a depender do caso. Para exemplificar, é utilizado o mesmo trecho analisado na discussão dos outros métodos, extraído da série original e que inclui uma alteração do tempo 128,36 até o tempo 129,04.

Figura 36 – Trecho da série, padrão, e distâncias calculadas. Extraído da série original do sensor 2, tamanho de janela 30, avanço 10, alternativa de transformação *ntng*, e função de comparação *eucl*.

Na Figura 36 chama a atenção um aparente deslocamento entre o padrão e os dados adquiridos que, apresentando o mesmo comportamento relativo, têm diferentes valores para as médias. Mesmo que para o trecho em consideração não tinham sido feitas correções de média, há ao longo da série outras mudanças de média que geram essas diferenças. Esse comportamento não é igual para todos os sensores pois há diferenças de desempenho como mostradas na Figura 28 para a mesma configuração de método usada no exemplo. Se o mesmo trecho for analisado para a série 1 o padrão está muito mais próximo dos dados adquiridos. A respeito do trecho com alteração, vê-se um importante aumento na distância próximo do tempo 128,4, que permite a detecção rápida do início da alteração. Quando a pressão aumenta, porém, essa distância passa a ser menor e próxima das distâncias de trechos sem alterações, gerando erros na detecção.

7 CONCLUSÕES

Neste documento é apresentada a pesquisa de doutorado cujo objetivo foi desenvolver um método para a detecção de vazamentos em redes de distribuição de água para abastecimento baseado na análise de sinais de pressão adquiridos por monitoramento na rede, durante sua operação. As propostas são desenvolvidas sobre um estudo de caso de um setor de rede de abastecimento real monitorado por aproximadamente um ano. O setor de estudo, com mais de 6000 ligações, trechos de tubulação de diferentes materiais e de idades distintas (alguns implementados há décadas), diferenças de cota de mais de 50 m e alta incidência de vazamentos, trouxe para a pesquisa muitas das dificuldades e complexidades que estão implícitas no trabalho com redes de abastecimento comuns. Se por um lado a complexidade do sistema impediu o sucesso ou até o desenvolvimento de várias iniciativas de análise (não apresentadas neste documento), por outro lado fez com que a pesquisa ficasse mais próxima do desafio da aplicação, assumido quando da decisão de trabalhar com uma rede real.

Nos planos originais de aquisição de informação os registros de vazão de entrada ao setor, de aquisição rotineira, seriam utilizados. Porém, foi descoberto que o medidor de vazão instalado tinha erros e problemas de calibração que impediam a utilização dos valores adquiridos. Essa situação ilustra a incerteza que pode existir na informação coletada nos sistemas de distribuição de água, e alerta sobre as informações de má qualidade cuja utilização pode comprometer os resultados de qualquer análise. Assim, as descrições e análises foram realizadas exclusivamente para as séries de pressão, adquiridas a cada dois minutos em pontos escolhidos segundo o método proposto.

Perante a necessidade de decidir, criteriosamente, o número e a localização de sensores para o estudo de caso, foi identificada a lacuna de métodos para projeto de monitoramento visando, especificamente, a detecção de vazamentos. Um método para definição desse projeto de monitoramento ótimo, ou pelo menos bem elaborado a partir de critérios que se mostraram robustos, foi desenvolvido. No método combinam-se análises de sensibilidade e de correlação realizadas com suporte do modelo de simulação hidráulica, com a otimização multiobjetivo usando a tecnologia dos Algoritmos Genéticos. Quatro critérios são levados em conta: maximização da sensibilidade, maximização da distribuição da sensibilidade (medida com a entropia), minimização da correlação e minimização do número de sensores. Os três primeiros critérios revelaram-se conflitantes no processo de otimização, e o quarto é avaliado posteriormente analisando-se o ganho residual representado pela adição dos sensores. A aplicação do método ao estudo de caso teve como resultado a seleção de nove pontos para instalação de sensores, localizados preferencialmente nas extremidades da rede, e mais concentrados nas regiões de pressão estática elevada.

Para conseguir avaliar os métodos de detecção foi necessária a classificação manual das

séries de pressão adquiridas. O exercício permite concluir que os resultados da inspeção visual são sensíveis não somente ao critério do analista mas também às condições da visualização, tais como escala e tipo de gráfico construído. Salienta-se a importância desta conclusão ao lembrar que a inspeção visual é o único aproveitamento da informação com que contam muitas companhias responsáveis pelo abastecimento urbano de água.

Um resultado específico interessante foi a identificação de um vazamento que, começando com tamanho pequeno, manteve-se ativo e aumentando por aproximadamente 56 dias, até ser reportado e reparado. Este caso é um bom exemplo da importância de se contar com um sistema de controle de vazamentos ativo que permita a detecção mais rápida, pois o volume de água perdido e a afetação durante os quase dois meses em que o problema passou despercebido certamente resultam em um alto custo oculto para o sistema. Salienta-se, porém, que o longo tempo de atividade do vazamento está relacionado com a dificuldade de sua detecção sem ferramentas adequadas, e não a falhas de procedimento da companhia ou a erros dos seus funcionários, que completaram o reparo poucas horas depois da recepção do reporte.

Outro aspecto importante levantado pelo exemplo desse vazamento encontrado é o horizonte considerado para a detecção. Dentre as premissas da pesquisa esteve a necessidade de que detecção do início do vazamento fosse rápida, em poucos passos de tempo; entretanto, para o exemplo tanto o método de detecção baseado em análise estatística quanto o de comparação com padrões, conseguiram detectar o problema dias ou horas após seu início, respectivamente. Pode-se concluir que o critério de detecção rápida não é de aplicação absoluta, e que há casos em que a detecção, mesmo que não consiga ser quase-instantânea, continua sendo útil e evitando a persistência de vazamentos por longos períodos.

Da descrição das séries obtidas são extraídas várias características que, segundo estimamos, não devem ser únicas do setor de rede estudado mas podem estar presentes em outros sistemas semelhantes. Foi identificado comportamento não estacionário em trechos das séries, entre eles os relacionados ao vazamento de longa duração antes comentado ou a outras ocorrências identificadas, mas para outros as causas continuaram desconhecidas. A periodicidade diária e semanal dos valores de pressão foi comprovada e, adicionalmente, a existência de componentes de notável amplitude em períodos menores a um dia. A presença de ruído, e os valores de dispersão seguindo um comportamento quase-periódico com amplitudes máximas grandes foram também evidenciados.

A análise de recorrência realizada é, segundo as publicações revisadas, a primeira aplicação dessa técnica a sinais de pressão de redes de abastecimento de água. Os gráficos de recorrência obtidos confirmam alguns elementos como a existência de comportamento periódico diário, mas também permitem conclusões sobre outros aspectos, como a presença de componentes determinísticas nas séries, conjugadas com componentes estocásticas. Também foram encontradas evidências de descontinuidades e de comportamento não estacionário não perceptíveis com outras análises. De esses gráficos pode-se concluir também que existem outras

alterações nas séries além das identificadas na classificação manual, e que os comportamentos periódico e estacionário supostos existem mas não se cumprem para toda a série. As características distinguíveis nos gráficos relacionadas com mudanças ou irregularidades sugerem que a análise de recorrência poderia ser também aplicada à detecção, possibilidade que deverá constituir alvo de estudos futuros.

Quanto aos métodos de detecção de vazamentos ou outras alterações através da análise dos sinais de pressão, três métodos diferentes foram propostos, o primeiro baseado nas cadeias de Markov, o segundo no controle estatístico de processos, e o terceiro na comparação com padrões. O desempenho de cada um dos três métodos foi quantificado com o valor médio de área abaixo da curva ROC, AUROC médio, calculado para todo o comprimento das séries e para 100 valores diferentes de limiar de detecção α . Uma característica fundamental da forma de avaliação escolhida é o uso de um único indicador que agrega múltiplos resultados, o que é uma vantagem ao permitir comparações que não poderiam ser feitas com resultados individuais, mas que em ocasiões apaga diferenças grandes entre respostas ou mistura comportamentos contrários, casos nos que a média não é um bom estimador dos resultados.

A detecção com cadeias de Markov produziu resultados de desempenho baixos, com o melhor valor atingido de $AUROC = 0,67$. Encontrou-se que o método pode ser apropriado para detecção de início abrupto de um vazamento, mas deixa de sê-lo se a nova situação se mantém no tempo, tendendo a estabelecer uma nova média. Além disso, o comportamento cíclico da variância contradiz a suposição de estacionaridade da variável aleatória.

Com o método proposto de detecção proposto com base no controle estatístico, o valor médio de desempenho global de apenas $AUROC = 0,65 \pm 0,10$ foi produzido. O método acompanha simultaneamente, tanto os descritores estatísticos do comportamento de curto prazo (minutos), quanto os do longo prazo (dias), numa combinação que depende dos parâmetros escolhidos. Encontrou-se que a escolha dos parâmetros convenientes é diferente para cada série, razão pela qual ao forçar a integração em um único indicador o valor de desvio resulta alto. Além disso, nesse caso a execução para a série completa tem um efeito importante devido a que os valores dos descritores podem cair numa situação de aprendizado errado que se estende pelo restante da série, o que não aconteceria no cenário de aplicação real em que haveria alguma reação do tomador de decisão após um tempo ocorrendo de valores falhos.

A proposta da detecção pela comparação com padrões permitiu avaliar doze alternativas de representação dos trechos de sinal e cinco funções de comparação, além de seis configurações de janelamento. Sobre o janelamento, notou-se que a diminuição do avanço das janelas e o consequente aumento na redundância das análises, não implica em mudanças do desempenho. Mesmo que algumas combinações de funções atinjam valores de indicador de desempenho (AUROC) mais altos com janelas grandes (diárias), a perda de resolução da resposta não justifica a sua adoção, e é recomendado o uso de janela equivalente a duração de 1 hora.

Ainda sobre o terceiro método de detecção, um resultado interessante é que a representa-

ção dos sinais com algumas das funções avaliadas, em particular com a transformada *Wavelet*, permite atingir valores de desempenho global altos, mas não melhores aos obtidos pela comparação direta, sem realizar nenhuma transformação. A conclusão de que nenhuma das funções de transformação avaliadas favorece a detecção não poderia ser obtida antes da comparação dos resultados, mas com os resultados do estudo de caso a configuração mais recomendável foi não realizar nenhuma transformação e fazer comparação com uma função de distância (euclidiana, centroides ou Hausdorff), além de usar janelas de tamanho equivalente a 1 hora e sobreposição equivalente a 20 minutos.

Por fim, pode se concluir que os métodos propostos para detecção de vazamentos não são mutuamente excludentes, mas agem sobre diferentes escalas temporais, e tem melhor resposta para diferentes tipos de irregularidades. Dessa forma é possível imaginar seu funcionamento acoplado. A implementação desse acoplamento, porém, foge do escopo deste documento e fica sugerido como trabalho futuro.

A respeito da hipótese da pesquisa, foi encontrado que a ocorrência de vazamentos efetivamente gera mudanças nas pressões medidas na rede. Essa mudança, porém, tem em alguns casos a mesma ordem de grandeza que a variação dos valores de pressão em condição sem alteração. Dessa forma, é difícil estabelecer as diferenças entre os efeitos do vazamento e os das condições próprias do funcionamento da rede. Esse problema de diferenciação para casos gerais não pode-se considerar resolvido com os resultados desta pesquisa, mas a combinação de métodos de detecção em diferentes escalas e que respondam preferencialmente a diferentes tipos de alteração mostra-se uma estratégia útil para buscar superar essas dificuldades em trabalhos futuros.

REFERÊNCIAS

- ABE, N. **Proposição de Calibração de Redes de Sistemas de Abastecimento de Água Através de Método Híbrido**. 2014. Tese (Doutorado) — Universidade de São Paulo, 2014.
- ADDISON, P. S. **The Illustrated Wavelet Transform Handbook**. Bristol, UK: IOP Publishing Ltd, 2002. ISBN 0750306920.
- AKSELA, K.; AKSELA, M.; VAHALA, R. Leakage detection in a real distribution network using a SOM. **Urban Water Journal**, v. 6, n. 4, p. 279–289, oct 2009. ISSN 1573-062X.
- ALLEN, R. L.; MILLS, D. W. **Signal Analysis**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2003. ISBN 0471234419.
- ANDERSEN, J. H.; POWELL, R. S. Implicit state-estimation technique for water network monitoring. **Urban Water**, v. 2, n. 2, p. 123–130, jun 2000. ISSN 14620758.
- BABBITT, H. E.; AMSBARY, F.; GWINN, D. The detection of leaks in underground pipes. **American Water Works Association**, v. 7, n. 4, p. 589–595, 1920.
- BENAROYA, H. Markov chain transition probabilities and experimental data. **Applied Mathematics and Computation**, v. 29, n. 2, p. 107–121, 1989. ISSN 00963003.
- BHAT, U. N.; MILLER, G. K. G. K. **Elements of applied stochastic processes**. 3. ed. [S.l.]: Wiley-Interscience, 2002. 461 p. ISBN 0471414425.
- BUCHBERGER, S. G.; NADIMPALLI, G. Leak Estimation in Water Distribution Systems by Statistical Analysis of Flow Readings. v. 130, n. 4, p. 321–329, 2004.
- BUSH, C. A.; UBER, J. G. Sampling Design Methods for Water Distribution Model Calibration. **Journal of Water Resources Planning and Management**, v. 124, n. 6, p. 334–344, nov 1998. ISSN 0733-9496.
- CHRISTODOULOU, S. E.; KOURTI, E.; AGATHOKLEOUS, A. Waterloss Detection in Water Distribution Networks using Wavelet Change-Point Detection. **Water Resources Management**, v. 31, p. 1–16, dec 2016. ISSN 0920-4741.
- COLOMBO, A. F.; LEE, P.; KARNEY, B. W. A selective literature review of transient-based leak detection methods. **Journal of Hydro-environment Research**, v. 2, n. 4, p. 212–227, apr 2009. ISSN 15706443.
- COLOMBO, F. **Calibração de modelos hidráulicos de redes de abastecimento de água de sistemas reais admitindo vazamentos**. 2007. Dissertação (Mestrado) — Universidade de São Paulo, São Carlos, 2007.
- COVAS, D.; RAMOS, H. Case Studies of Leak Detection and Location in Water Pipe Systems by Inverse Transient Analysis. **Journal of Water Resources Planning and Management**, v. 136, n. 2, p. 248–257, 2010.
- COVAS, D.; RAMOS, H.; ALMEIDA, A. B. D. Standing Wave Difference Method for Leak Detection in Pipeline Systems. **Journal of Hydraulic Engineering**, v. 131, n. 12, p. 1106–1116, 2005.

- DEB, K. et al. A fast and elitist multiobjective genetic algorithm: NSGA-II. **IEEE Transactions on Evolutionary Computation**, v. 6, n. 2, p. 182–197, apr 2002. ISSN 1089778X.
- DEZA, M. M.; DEZA, E. **Encyclopedia of distances**. 4. ed. [S.l.]: Springer-Verlag Berlin Heidelberg, 2016. XXII, 756 p. ISBN 978-3-662-52843-3.
- FACELI, K. et al. **Inteligência artificial. Uma abordagem de aprendizado de máquina**. Rio de Janeiro: LTC, 2011. ISBN 978-85-216-1880-5.
- FARLEY, B.; MOUNCE, S. R.; BOXALL, J. B. Field testing of an optimal sensor placement methodology for event detection in an urban water distribution network. **Urban Water Journal**, v. 7, n. 6, p. 345–356, dec 2010. ISSN 1573-062X.
- FERRANTE, M.; BRUNONE, B. Pipe system diagnosis and leak detection by unsteady-state tests. 1. Harmonic analysis. **Advances in Water Resources**, v. 26, n. 1, p. 95–105, jan 2003. ISSN 03091708.
- _____. Pipe system diagnosis and leak detection by unsteady-state tests. 2. Wavelet analysis. **Advances in Water Resources**, v. 26, n. 1, p. 107–116, jan 2003. ISSN 03091708.
- FERRANTE, M.; BRUNONE, B.; MENICONI, S. Wavelets for the Analysis of Transient Pressure Signals for Leak Detection. **Journal of hydraulic engineering**, v. 133, n. 11, p. 1274–1282, 2007.
- _____. Leak detection in branched pipe systems coupling wavelet analysis and a Lagrangian model. **Journal of Water Supply: Research and Technology—AQUA**, v. 58, n. 2, p. 95, mar 2009. ISSN 0003-7214.
- FINCH, T. Incremental calculation of weighted mean and variance. 2009.
- GAMA, J.; GABER, M. M. **Learning from Data Streams: Processing Techniques in Sensor Networks**. [S.l.: s.n.], 2007. 244 p. ISBN 9783540736783.
- GAMBOA-MEDINA, M. M. **Deteção de vazamentos em redes sob pressão baseada na análise dos sinais de pressão e vazão com um sistema de reconhecimento de padrões**. 2013. 123 p. Dissertação (Mestrado) — Universidade de São Paulo, São Carlos, 2013.
- GAMBOA-MEDINA, M. M.; REIS, L. F. R. Sampling Design for Leak Detection in Water Distribution Networks. **Procedia Engineering**, v. 186, p. 460–469, 2017. ISSN 18777058.
- GAMBOA-MEDINA, M. M.; REIS, L. F. R.; GUIDO, R. C. Feature extraction in pressure signals for leak detection in water networks. **Procedia Engineering**, v. 70, p. 688–697, 2014.
- GERMANOPOULOS, G. A technical note on the inclusion of pressure dependent demand and leakage terms in water supply network models. **Civil Engineering Systems**, v. 2, n. 3, p. 171–179, sep 1985. ISSN 0263-0257.
- GHAZALI, M. et al. Comparative study of instantaneous frequency based methods for leak detection in pipeline networks. **Mechanical Systems and Signal Processing**, v. 29, p. 187–200, may 2012. ISSN 08883270.
- GREYVENSTEIN, B.; ZYL, J. E. V. An experimental investigation into pressure - leakage relationship of some failed water pipes. **Journal of Water Supply: Research and Technology—AQUA**, v. 56, n. 2, p. 117–124, 2007.

GUIDO, R. C. A tutorial on signal energy and its applications. **Neurocomputing**, v. 179, p. 264–282, 2015. ISSN 18728286.

_____. Introducing the Discrete Length Transform (DLT) and its Applications for Speech Processing. 2016.

_____. ZCR-aided neurocomputing: A study with applications. **Knowledge-Based Systems**, v. 105, p. 248–269, aug 2016. ISSN 09507051.

JUNG, D. et al. Improving the rapidity of responses to pipe burst in water distribution systems: a comparison of statistical process control methods. **Journal of Hydroinformatics**, v. 17, n. 2, p. 307, 2015. ISSN 1464-7141.

KANTZ, H.; SCHREIBER, T. **Nonlinear Time Series Analysis**. Cambridge: Cambridge University Press, 2003. ISBN 9780511755798.

KAPELAN, Z.; SAVIC, D.; WALTERS, G. A hybrid inverse transient model for leakage detection and roughness calibration in pipe networks. **JOURNAL OF HYDRAULIC RESEARCH**, v. 41, n. 5, p. 481–492, 2003. ISSN 0022-1686.

KENNEL, M. B.; BROWN, R.; ABARBANEL, H. D. I. Determining embedding dimension for phase-space reconstruction using a geometrical construction. **Physical Review A**, v. 45, n. 6, p. 3403–3411, mar 1992. ISSN 1050-2947.

KRZANOWSKI, W. J.; HAND, D. J. **ROC Curves for Continuous Data**. FL,USA: Taylor&Francis Group, 2009. 226 p. ISBN 978-1-4398-0021-8.

LI, R. et al. A Review of Methods for Burst/Leakage Detection and Location in Water Distribution Systems. **Water Science & Technology: Water Supply**, v. 15, n. 3, p. 429–441, 2015. ISSN 1606-9749.

LIGGETT, J. A.; CHEN, L.-C. Inverse transient analysis in pipe networks. **Journal of hydraulic engineering**, v. 120, n. 8, p. 934–955, 1994.

MALIK, O.; GHOSH, S.; ROGERS, A. A Noise Scaled Semi-Parametric Gaussian Process Model for Real Time Water Network Leak Detection in the Presence of Heteroscedasticity. In: **Computational Sustainability: Papers from the 2015 AAAI Workshop**. Austin, TX, USA: [s.n.], 2015. p. 84–90.

MALLAT, S. **A Wavelet Tour of Signal Processing**. 2. ed. London: Academic Press, 1999. ISBN 9780124666061.

MARKOU, M.; SINGH, S. Novelty detection: a review—part 1: statistical approaches. **Signal Processing**, v. 83, n. 12, p. 2481–2497, dec 2003. ISSN 01651684.

MARWAN, N. et al. Recurrence-plot-based measures of complexity and their application to heart-rate-variability data. **Physical Review E - Statistical, Nonlinear, and Soft Matter Physics**, v. 66, n. 2, p. 1–8, 2002. ISSN 15393755.

MENICONI, S. et al. Potential of transient tests to diagnose real supply pipe systems : What can be done with a single extemporaneous test. **Journal of Water Resources Planning and Management**, v. 137, n. 2, p. 238–241, 2011.

- MESEGUER, J. et al. A decision support system for on-line leakage localization. **Environmental Modelling & Software**, v. 60, p. 331–345, oct 2014. ISSN 13648152.
- MISIUNAS, D. et al. Pipeline Break Detection Using Pressure Transient Monitoring. **Journal of Water Resources Planning and Management**, American Society of Civil Engineers, v. 131, n. 4, p. 316–325, jul 2005. ISSN 0733-9496.
- _____. Failure monitoring in water distribution networks. **Water Science & Technology**, IWA Publishing, v. 53, n. 4-5, p. 503–511, 2006.
- MONTGOMERY, D. **Introduction to statistical quality control**. 6. ed. USA: John Wiley & Sons, Inc., 2009. 754 p. ISSN 14337851. ISBN 9780470169926.
- MOUNCE, S. R.; BOXALL, J. B.; MACHELL, J. Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows. **Journal of Water Resources Planning and Management**, v. 136, n. 3, p. 309–318, 2010.
- MPESHA, W.; GASSMAN, S. L.; CHAUDHRY, M. H. Leak detection in pipes by frequency response method. **Journal of hydraulic engineering**, v. 127, n. 2, 2001.
- ODAN, F. K. **Estudo de confiabilidade aplicado à otimização da operação em tempo real de redes de abastecimento de água**. 2013. 210 p. Tese (Doutorado) — Universidade de São Paulo, 2013.
- OPEN WATER ANALYTICS. **Epanet 2.1**. 2016. Disponível em: <http://wateranalytics.org/EPANET/release_2_1.html>.
- PALAU, C. V.; ARREGUI, F. J.; CARLOS, M. Burst Detection in Water Networks Using Principal Component Analysis. **Journal of Water Resoucer planning and management**, v. 138, n. 1, p. 47–54, 2012.
- POULAKIS, Z.; VALOUGEORGIS, D.; PAPADIMITRIOU, C. Leakage detection in water pipe networks using a Bayesian probabilistic framework. **Probabilistic Engineering Mechanics**, v. 18, n. 4, p. 315–327, oct 2003. ISSN 02668920.
- PROAKIS, J. G.; MANOLAKIS, D. G. **Digital Signal Processing**. 4. ed. UpperSaddleRiver: Pearson Education, 2007.
- PUDAR, R. S.; LIGGETT, J. A. Leaks in Pipe Networks. **Journal of Hydraulic Engineering**, v. 118, n. 7, p. 1031–1046, jul 1992. ISSN 0733-9429.
- PUUST, R. et al. A review of methods for leakage management in pipe networks. **Urban Water Journal**, v. 7, n. 1, p. 25–45, 2010. ISSN 1573-062X.
- ROMANO, M.; KAPELAN, Z.; SAVIC, D. **Anomaly Detection Based in Bayesian Inference**. [S.l.]: Patent number: PCT/GB2010/000961, 2010.
- _____. Testing of the system for detection of pipe bursts and other events in a uk water distribution system. **14th Water Distribution System Analysis Conference**, p. 1109–1126, 2012.
- _____. Automated Detection of pipe burst and other events in water distribution systems. **Journal of Water Resources Planning and Management**, v. 140, n. 4, 2014.

- ROMANO, M.; KAPELAN, Z.; SAVIC, D. A. Evolutionary Algorithm and Expectation Maximization Strategies for Improved Detection of Pipe Bursts and Other Events in Water Distribution Systems. **Journal of water resources planning and management**, n. May, p. 572–584, 2014.
- ROMANO, M.; WOODWARD, K.; KAPELAN, Z. Statistical process control based system for approximate location of pipe bursts and leaks in water distribution systems. In: **XVIII International Conference on Water Distribution Systems, WDSA2016 Statistical**. Cartagena, COL: [s.n.], 2016. ISBN 0000000000.
- ROSS, G. J. et al. Exponentially weighted moving average charts for detecting concept drift. **Pattern Recognition Letters**, v. 33, n. 2, p. 191–198, 2012. ISSN 01678655.
- ROSSMAN, L. A. **EPANET 2 User manual**. Cincinnati, OH, USA: U.S. Environmental protection agency, 2000.
- RUSSELL, S. J. S. J.; NORVIG, P.; DAVIS, E. **Artificial intelligence : a modern approach**. [S.l.]: Pearson Education, 2010. 1132 p. ISBN 0136042597.
- SALDARRIAGA, J. G.; FUENTES, D. A. A.; GALVIS, L. F. C. Implementation of the Hydraulic Transient and Steady Oscillatory Flow with Genetic Algorithms for Leakage Detection in Real Water Distribution Networks. In: **8th Annual Water Distribution Systems Analysis Symposium**. Cincinnati, Ohio, USA: [s.n.], 2006. ISBN 978-0-7844-0941-1.
- SANZ, G. et al. Leak Detection and Localization through Demand Components Calibration. **Journal of Water Resources Planning and Management**, American Society of Civil Engineers, v. 142, n. 2, 2016. ISSN 0733-9496.
- SHANNON, C. A Mathematical Theory of Communication. **The bell system technical journal**, v. 27, n. 3, p. 379–423, 1948.
- SNSA. **Diagnóstico dos serviços de água e esgotos - 2015**. [S.l.], 2017. 212 p.
- SOARES, A. **Calibração de modelos de redes de distribuição de água para abastecimento considerando vazamentos e demandas dirigidas pela pressão**. 2003. Dissertação (Mestrado) — Universidade de São Paulo, São Carlos, 2003.
- SOARES, A. K.; COVAS, D. I. C.; REIS, L. F. R. Leak detection by inverse transient analysis in an experimental PVC pipe system. **Journal of Hydroinformatics**, v. 13, n. 2, p. 153, apr 2011. ISSN 1464-7141.
- SRIRANGARAJAN, S. et al. Wavelet-based Burst Event Detection and Localization in Water Distribution Systems. **Journal of Signal Processing Systems**, v. 72, n. 1, p. 1–16, 2012.
- VITKOVSKY, J. P.; SIMPSON, A. R.; LAMBERT, M. F. Leak detection and calibration using transient and genetic algorithms. **Journal of Water Resources Planning and Management**, v. 126, n. 4, p. 262–265, 2000.
- WANG, X.-j. et al. Leak Detection in Pipelines using Damping of Fluid Transients. **Journal of hydraulic engineering**, v. 128, n. 7, p. 697–711, 2002.
- WEBBER, C. L.; MARWAN, N. **Recurrence Quantification Analysis**. [S.l.: s.n.], 2015. 421 p. ISBN 978-3-319-07155-8.

WEBBER, C. L.; ZBILUT, J. P. Dynamical assessment of physiological systems and states using recurrence plot strategies. **Journal of Applied Physiology**, v. 76, n. 2, 1994.

WU, Z. Y.; SAGE, P.; TURTLE, D. Pressure-Dependent Leak Detection Model and Its Application to a District Water System. **Journal of Water Resources Planning and Management**, v. 136, n. 1, p. 116–128, 2010.

YAFFEE, R.; MCGEE, M. **Introduction to Time Series Analysis and Forecasting**. [S.l.]: Academic Press, 2000. ISBN 9780080478708.

YE, G.; PH, D.; FENNER, R. A. Kalman Filtering of Hydraulic Measurements for Burst Detection in Water Distribution Systems. **Journal of pipeline systems engineering and practice**, v. 2, n. 1, p. 14–22, 2011.

YOON, J. S. et al. The optimal leakage detection model for water distribution systems considering uncertainties in roughness coefficient. **14th Water Distribution Systems Analysis Conference 2012, WDSA 2012**, v 2, p. 1316 – 1323, 2012.

ZBILUT, J. P.; THOMASSON, N.; WEBBER, C. L. Recurrence quantification analysis as a tool for nonlinear exploration of nonstationary cardiac signals. **Medical Engineering & Physics**, v. 24, n. 1, p. 53–60, jan 2002. ISSN 13504533.

ZBILUT, J. P.; WEBBER, C. L. Embeddings and delays as derived from quantification of recurrence plots. **Physics Letters A**, v. 171, n. 3-4, p. 199–203, 1992. ISSN 03759601.